

# Rééchantillonnage

Sylvain Arlot

sylvain.arlot@universite-paris-saclay.fr

<http://www.imo.universite-paris-saclay.fr/~arlot/>

7 septembre 2020

Ce texte est une introduction aux méthodes de rééchantillonnage en général, et notamment à leurs applications en apprentissage, par exemple pour la sélection d'estimateurs.

Une référence possible : le livre d'Efron et Tibshirani (1993).

## 1 Principe. Exemples

### 1.1 Heuristique du bootstrap (Efron, 1979)

On observe  $\xi_1, \dots, \xi_n$  i.i.d. de loi commune  $P$ , on note  $P_n$  leur mesure empirique  $n^{-1} \sum_{i=1}^n \delta_{\xi_i}$ , et on cherche à estimer la distribution d'une quantité de la forme  $F(P, P_n)$ .

**Exemple 1** On s'intéresse à une quantité inconnue  $\mu(P) \in \mathbb{R}$  (par exemple l'espérance), que l'on estime par  $\hat{f}(P_n)$  (par exemple la moyenne empirique). On pose alors  $F(P, P_n) = \hat{f}(P_n) - \mu(P)$  et l'on s'intéresse naturellement à sa loi : espérance (qui correspond au biais de l'estimateur), variance (variabilité de l'estimateur), espérance du carré (risque quadratique), quantiles (pour construire des intervalles de confiance), etc.

Dans ce but, conditionnellement aux observations  $\xi_1, \dots, \xi_n$ , on génère un rééchantillon  $\xi_1^*, \dots, \xi_n^*$  i.i.d. de loi commune  $P_n$ , et l'on définit sa mesure empirique :

$$P_n^* = P_n^W = \frac{1}{n} \sum_{i=1}^n \delta_{\xi_i^*}.$$

L'heuristique du bootstrap est que

$$\mathcal{L}(F(P_n, P_n^W) | P_n) \text{ estime } \mathcal{L}(F(P, P_n))$$

(voir la figure 1).

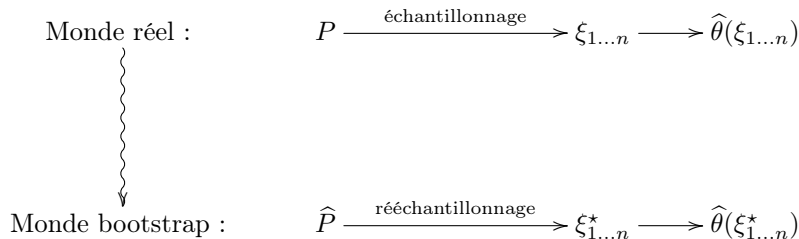


FIGURE 1 – L’heuristique de rééchantillonnage, selon Efron (1979). Schéma inspiré de la figure 1 d’un article d’Efron (2003).

## 1.2 Bootstrap à poids échangeables

On applique l’heuristique ci-dessus avec

$$P_n^W = \frac{1}{n} \sum_{i=1}^n W_{i,n} \delta_{\xi_i},$$

où  $W = (W_{i,n})_{1 \leq i \leq n}$  est un vecteur aléatoire, indépendant de  $P_n$ , tel que  $W_{i,n} \geq 0$  p.s.,  $\mathbb{E}[W_{i,n}] = 1$  pour tout  $i$ , et  $W$  est échangeable — pour toute permutation  $\tau$  de  $\{1, \dots, n\}$ ,  $W$  a la même loi que  $(W_{\tau(i),n})_{1 \leq i \leq n}$ . On note  $\mathbb{E}_W[\cdot]$  l’espérance relative à l’aléa de  $W$  uniquement. Plus précisément, l’heuristique de rééchantillonnage devient ici que  $\mathcal{L}(C_W F(P_n, P_n^W) | P_n)$  estime  $\mathcal{L}(F(P, P_n))$  où  $C_W$  est un facteur d’échelle déterministe, fonction de la loi de  $W$ .

Ceci généralise le bootstrap, car celui-ci correspond à prendre  $W$  de loi multinomiale de paramètres  $(n; 1/n, \dots, 1/n)$ . Mentionnons plusieurs autres exemples :

- le bootstrap «  $m$  out of  $n$  » — où l’on génère un rééchantillon bootstrap  $\xi_1^*, \dots, \xi_m^*$  de taille  $m \neq n$  — correspond à prendre  $(m/n)W$  de loi multinomiale de paramètres  $(m; 1/n, \dots, 1/n)$ .
- le bootstrap poissonnisé est un bootstrap «  $m$  out of  $n$  » avec  $m$  aléatoire (indépendant de l’aléa multinomial) suivant une loi de Poisson. Il correspond à prendre des poids tels que  $\lambda W_{1,n}, \dots, \lambda W_{n,n}$  sont indépendants et de même loi de Poisson de paramètre  $\lambda > 0$  (et alors,  $m$  suit une loi de Poisson de paramètre  $\lambda n$ ).
- le *sous-échantillonnage* est une autre famille importante de procédures de rééchantillonnage. Elle correspond à poser  $W_{i,n} = \kappa \mathbf{1}_{i \in I}$  pour une partie  $I$  de  $\{1, \dots, n\}$  choisie aléatoirement. Autrement dit, le rééchantillon bootstrap  $\xi_1^*, \dots, \xi_n^*$  est remplacé par un sous-échantillon aléatoire  $(\xi_i)_{i \in I}$ .

Par exemple, le hold-out aléatoire choisit  $I$  uniformément parmi les parties de taille  $m$  de  $\{1, \dots, n\}$ .

Une autre méthode, appelée « rééchantillonnage Bernoulli ( $p$ ) », est de choi-

soit  $pW_{1,n}, \dots, pW_{n,n}$  indépendants et de même loi de Bernoulli de paramètre  $p \in ]0; 1[$ . Le cas particulier  $p = 1/2$  correspond aux poids « Rademacher ».

## 2 Utilisations en statistique

Voici quelques exemples d'utilisation du rééchantillonnage pour des problèmes statistiques usuels.

### 2.1 Estimation ponctuelle

**Estimation ou correction du biais** de  $\hat{f}(P_n) \in \mathbb{R}$  comme estimateur de  $\mu(P)$  (Efron, 1979) :

$$\mathbb{E}[\hat{f}(P_n) - \mu(P)] \quad \text{estimé par} \quad C_W \mathbb{E}_W[\hat{f}(P_n^W) - \mu(P_n)] = C_W \mathbb{E}[\hat{f}(P_n^W) - \mu(P_n) \mid P_n].$$

**Estimation de la variance** de  $\hat{f}(P_n) \in \mathbb{R}$  (Efron, 1979) :  $\text{var}[\hat{f}(P_n)]$  est estimée par

$$\begin{aligned} C_W^2 \text{var}_W[\hat{f}(P_n^W)] &= C_W^2 \text{var}[\hat{f}(P_n^W) \mid P_n] \\ &= C_W^2 \mathbb{E}\left[\left(\hat{f}(P_n^W) - \mathbb{E}[\hat{f}(P_n^W) \mid P_n]\right)^2 \mid P_n\right]. \end{aligned}$$

**Estimation du risque quadratique** de  $\hat{f}(P_n) \in \mathbb{R}$  comme estimateur de  $\mu(P)$  :

$$\mathbb{E}\left[(\hat{f}(P_n) - \mu(P))^2\right] \quad \text{estimé par} \quad C_W^2 \mathbb{E}_W\left[\left(\hat{f}(P_n^W) - \mu(P_n)\right)^2 \mid P_n\right].$$

### 2.2 Étude détaillée d'un estimateur du risque quadratique

Étudions ce dernier exemple d'estimateur ponctuel de manière détaillée dans un cas particulier.

**Définition** Soient  $\xi_1, \dots, \xi_n$  des variables aléatoires i.i.d. de moyenne  $\mu$  et de variance  $\sigma^2$ . Un estimateur naturel de  $\mu$  est  $\hat{f}(P_n) = n^{-1} \sum_{i=1}^n \xi_i$  la moyenne empirique, dont le risque quadratique vaut

$$\mathbb{E}\left[(\hat{f}(P_n) - \mu(P))^2\right] = \mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^n \xi_i - \mu\right)^2\right] = \frac{\sigma^2}{n}. \quad (1)$$

L'estimateur par rééchantillonnage du risque quadratique vaut alors (en renormalisant les poids  $W_i$  pour que leur somme soit toujours égale à  $n$ ) :

$$C_W^2 \mathbb{E}_W \left[ \left( \underbrace{\frac{1}{n} \sum_{i=1}^n \frac{W_i}{\bar{W}} \xi_i}_{=\hat{f}(P_n^W)} - \underbrace{\frac{1}{n} \sum_{i=1}^n \xi_i}_{=\mu(P_n)} \right)^2 \right] \quad \text{avec} \quad \bar{W} = \frac{1}{n} \sum_{i=1}^n W_i.$$

Au vu de la formule (1), on note cet estimateur  $\widehat{\sigma_W^2}/n$ , puisque

$$\widehat{\sigma_W^2} = nC_W^2 \mathbb{E}_W \left[ \left( \frac{1}{n} \sum_{i=1}^n \frac{W_i}{\overline{W}} \xi_i - \frac{1}{n} \sum_{i=1}^n \xi_i \right)^2 \right]$$

peut être vu comme un estimateur de  $\sigma^2$ .

**Formule close** Calculons  $\widehat{\sigma_W^2}$  pour des poids  $W$  échangeables généraux :

$$\begin{aligned} \widehat{\sigma_W^2} &= nC_W^2 \mathbb{E}_W \left[ \left( \frac{1}{\sum_{k=1}^n W_k} \sum_{i=1}^n (W_i \xi_i) - \frac{1}{n} \sum_{i=1}^n \xi_i \right)^2 \right] \\ &= \frac{C_W^2}{n} \mathbb{E}_W \left[ \left( \sum_{i=1}^n \left[ \left( \frac{nW_i}{\sum_{k=1}^n W_k} - 1 \right) \xi_i \right] \right)^2 \right] \end{aligned}$$

Ainsi,

$$\begin{aligned} \widehat{\sigma_W^2} &= \frac{C_W^2}{n} \mathbb{E}_W \left[ \sum_{i,j} \left[ \left( \frac{nW_i}{\sum_{k=1}^n W_k} - 1 \right) \left( \frac{nW_j}{\sum_{k=1}^n W_k} - 1 \right) \xi_i \xi_j \right] \right] \\ &= \frac{C_W^2}{n} \sum_{i=1}^n [R_V^{(W)} \xi_i^2] + \frac{C_W^2}{n} \sum_{i \neq j} [R_C^{(W)} \xi_i \xi_j] \end{aligned}$$

en posant

$$\begin{aligned} R_V^{(W)} &:= \mathbb{E}_W \left[ \left( \frac{nW_i}{\sum_{k=1}^n W_k} - 1 \right)^2 \right] \\ \text{et } R_C^{(W)} &:= \mathbb{E}_W \left[ \left( \frac{nW_i}{\sum_{k=1}^n W_k} - 1 \right) \left( \frac{nW_j}{\sum_{k=1}^n W_k} - 1 \right) \right] \end{aligned}$$

pour des  $i \neq j$  quelconques (ces quantités ne dépendent pas de  $(i, j)$  si  $i \neq j$  car  $W$  est échangeable). Remarquons maintenant que

$$0 = \mathbb{E}_W \left[ \left( \sum_{i=1}^n \left( \frac{nW_i}{\sum_{k=1}^n W_k} - 1 \right) \right)^2 \right] = nR_V^{(W)} + n(n-1)R_C^{(W)}$$

et donc, en supposant  $n \geq 2$ ,

$$R_C^{(W)} = \frac{-1}{n-1} R_V^{(W)}.$$

Comme  $R_V^{(W)} = 0$  lorsque  $n = 1$ , on en déduit que

$$\widehat{\sigma_W^2} = \frac{C_W^2 R_V^{(W)}}{n} \mathbf{1}_{n \geq 2} \left[ \sum_{i=1}^n \xi_i^2 - \frac{1}{n-1} \sum_{i \neq j} \xi_i \xi_j \right]. \quad (2)$$

**Remarque 1** L'estimateur  $\widehat{\sigma}_W^2$  est invariant par translation des données car

$$\sum_{i=1}^n \left( \frac{W_i}{\sum_{j=1}^n W_j} - 1 \right) = 0.$$

C'est une propriété appréciable dans la mesure où la variance  $\sigma^2$  des  $\xi_i$  ne change pas quand on les translate ! Une autre conséquence est que la formule (2) est encore valable en remplaçant  $(\xi_i)_{1 \leq i \leq n}$  par  $(\xi_i - \mu)_{1 \leq i \leq n}$  ; on en déduit immédiatement le fait que

$$\mathbb{E} \left[ \widehat{\sigma}_W^2 \right] = \mathbf{1}_{n \geq 2} C_W^2 R_V^{(W)} \sigma^2. \quad (3)$$

En particulier, ceci suggère de prendre

$$C_W^2 = \frac{1}{R_V^{(W)}}$$

pour que l'estimateur par rééchantillonnage du risque quadratique soit sans biais pour tout  $n \geq 2$ .

**Comparaison avec l'estimateur classique** L'estimateur sans biais classique de la variance s'écrit, pour  $n \geq 2$ ,

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{1}{n-1} \sum_{i=1}^n \left( \xi_i - \frac{1}{n} \sum_{k=1}^n \xi_k \right)^2 = \frac{n}{n-1} \left[ \frac{1}{n} \sum_{i=1}^n \xi_i^2 - \left( \frac{1}{n} \sum_{i=1}^n \xi_i \right)^2 \right] \\ &= \frac{1}{n} \left( \sum_{i=1}^n \xi_i^2 - \frac{1}{n-1} \sum_{i \neq j} \xi_i \xi_j \right). \end{aligned} \quad (4)$$

Ainsi, pour tout  $n \geq 2$ ,

$$\widehat{\sigma}_W^2 = C_W^2 R_V^{(W)} \widehat{\sigma}^2,$$

ce qui démontre à nouveau que

$$\mathbb{E} \left[ \widehat{\sigma}_W^2 \right] = \mathbf{1}_{n \geq 2} C_W^2 R_V^{(W)} \sigma^2,$$

comme indiqué à la remarque 1.

### 2.3 Estimation par intervalle

On peut construire par rééchantillonnage un intervalle de confiance pour la moyenne  $\mu(P)$  de  $\xi_1, \dots, \xi_n \in \mathbb{R}$ . Pour cela, une première idée est de chercher  $t_\alpha \in \mathbb{R}$  tel que

$$[\mu(P_n) - t_\alpha, \mu(P_n) + t_\alpha]$$

est un intervalle confiance de probabilité de couverture  $1 - \alpha$  pour  $\mu(P)$ , où

$$\mu(P_n) = \frac{1}{n} \sum_{i=1}^n \xi_i$$

est la moyenne empirique de l'échantillon. Autrement dit, on voudrait avoir

$$\mathbb{P}\left(|\mu(P_n) - \mu(P)| \leq t_\alpha\right) \geq 1 - \alpha.$$

Le choix idéal serait  $t_\alpha^*$ , le quantile d'ordre  $(1 - \alpha)$  de  $\mathcal{L}(|\mu(P_n) - \mu(P)|)$ . On peut l'estimer par rééchantillonnage à l'aide du quantile d'ordre  $(1 - \alpha)$  de

$$\mathcal{L}\left(C_W |\mu(P_n^W) - \mu(P_n)| \mid P_n\right).$$

Une autre manière (meilleure) de construire un intervalle de confiance pour  $\mu(P)$  est de chercher  $t'_\alpha \in \mathbb{R}$  tel que

$$[\mu(P_n) - t'_\alpha \hat{\sigma}(P_n), \mu(P_n) + t'_\alpha \hat{\sigma}(P_n)]$$

est un intervalle confiance de probabilité de couverture  $1 - \alpha$  pour  $\mu(P)$ , où  $\hat{\sigma}^2(P_n)$  est un estimateur de la variance de  $\mu(P_n)$ . Par exemple, lorsque  $\mu(P_n)$  est la moyenne empirique, on peut prendre l'estimateur classique de la variance (divisé par  $n$ ) :

$$\hat{\sigma}^2(P_n) = \frac{1}{n(n-1)} \sum_{i=1}^n (\xi_i - \mu(P_n))^2.$$

Le choix idéal serait  $t'_\alpha^*$ , le quantile d'ordre  $(1 - \alpha)$  de  $\mathcal{L}(|\mu(P_n) - \mu(P)| / \hat{\sigma}(P_n))$ . Le résultat est meilleur (Hall, 1992) : on gagne au moins un ordre de grandeur dans la précision d'estimation du quantile, car la quantité dont on cherche la loi est « pivotale » (sa loi asymptotique ne dépend pas de  $P$ ), contrairement à la première méthode proposée (où elle est proportionnelle à l'écart-type  $\sigma$ ).

### 3 Avantages et limites

Le rééchantillonnage a de nombreux avantages :

- sa *généralité* : l'estimateur par rééchantillonnage peut être défini quelles que soient la fonction  $F$  et la loi  $P$ .
- son *adaptativité* : les estimateurs par rééchantillonnage s'adaptent souvent automatiquement à certaines caractéristiques inconnues du problème considéré, caractéristiques qui (si on les connaît) influent sur la vitesse d'estimation optimale. On parle d'adaptation lorsqu'un estimateur arrive à être (quasi) optimal sans la connaissance a priori de ces caractéristiques du problème.

Par exemple, on peut démontrer dans un cadre de régression (sous des hypothèses que l'on ne détaille pas) que les pénalités par rééchantillonnage (définies en section 4) s'adaptent automatiquement au fait que la variance du bruit dépend (ou pas) du point d'observation : on parle d'adaptation à l'hétéroscédasticité du bruit (Arlot, 2009).

- on dispose de *résultats théoriques asymptotiques* sur le processus empirique rééchantillonné (van der Vaart et Wellner, 1996, chapitre 3.6).
- *stabilisation* : le rééchantillonnage est souvent utilisé pour stabiliser un estimateur  $\hat{f}$  instable. Par exemple, le « bagging »<sup>1</sup> consiste à remplacer l'estimateur  $\hat{f}(D_n)$  par l'agrégation des estimateurs

$$\hat{f}(D_n^{*,1}), \dots, \hat{f}(D_n^{*,K})$$

obtenus sur  $K$  rééchantillons (bootstrap) indépendants. Selon les cas, on procède à l'agrégation de différentes manières : moyenne si  $\hat{f}$  est à valeurs dans un espace vectoriel, vote majoritaire si  $\hat{f}(D_n)$  est un classifieur, etc. Et l'on peut faire du bagging avec d'autres procédures de rééchantillonnage<sup>2</sup>

Il ne faut toutefois pas oublier les importantes limites du rééchantillonnage. Le pendant de l'universalité est en effet le risque d'utilisations abusives, dont les deux situations suivantes (où les données ne sont pas i.i.d.) sont les plus classiques.

En régression avec un plan d'expérience déterministe<sup>3</sup>, on observe

$$\forall i \in \{1, \dots, n\}, \quad \xi_i = (t_i, Y_i) \quad \text{avec} \quad Y_i = \eta(t_i) + \varepsilon_i,$$

où  $t_1, \dots, t_n$  sont déterministes et seuls les termes de bruit  $\varepsilon_1, \dots, \varepsilon_n$  sont aléatoires. On peut penser au cas où l'on observe un signal échantillonné régulièrement (c'est-à-dire, si  $\eta$  est définie sur  $[0, 1]$ ,  $t_i = i/n$ ) perturbé par un bruit additif. Alors, en général, les  $\xi_i$  ne sont pas de même loi ! Que faire ? Lorsque les  $\varepsilon_i$  sont indépendantes et de même loi, une solution est de *rééchantillonner les résidus* : on estime d'abord  $\eta$  par  $\hat{\eta}$  (avec une méthode de régression donnée), puis on calcule les *résidus* :

$$\hat{\varepsilon}_i = Y_i - \hat{\eta}(t_i), \quad i = 1, \dots, n.$$

On construit alors chaque rééchantillon en tirant  $\hat{\varepsilon}_1^*, \dots, \hat{\varepsilon}_n^*$  indépendantes et de même loi uniforme sur  $\{\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n\}$  et en posant

$$\forall i \in \{1, \dots, n\}, \quad \xi_i^* = (t_i, \hat{\eta}(t_i) + \hat{\varepsilon}_i^*).$$

De cette manière, la loi du rééchantillon  $(\xi_1^*, \dots, \xi_n^*)$  sachant l'échantillon « imite » mieux la loi de l'échantillon  $(\xi_1, \dots, \xi_n)$  que si l'on avait procédé à un rééchantillonnage classique (avec lequel on perd la structure du plan d'expérience  $t_1, \dots, t_n$  déterministe).

Un autre problème intervient lorsque les données sont stationnaires mais dépendantes. Dans le cas d'une dépendance à courte portée, il vaut mieux utiliser le bootstrap par blocs, à la place du bootstrap classique. Pour cela, on regroupe les données  $(\xi_i)_{1 \leq i \leq n}$  en  $K$  blocs

$$(\xi_i)_{1 \leq i \leq \frac{n}{K}}, \quad (\xi_i)_{\frac{n}{K}+1 \leq i \leq \frac{2n}{K}}, \quad \dots, \quad (\xi_i)_{\frac{n(K-1)}{K}+1 \leq i \leq n}$$

---

1. Le terme « bagging » est la contraction des deux mots anglais « bootstrap aggregating ».  
 2. Dans le cas du sous-échantillonnage, on parle de « subagging », contraction des mots anglais « subsampling aggregating ».  
 3. En anglais, on parle de « fixed-design ».

puis on tire  $B_1^*, \dots, B_K^*$  des variables aléatoires indépendantes et de même loi uniforme sur  $\{1, \dots, \frac{n}{K}\}$ . On définit alors le rééchantillon comme la concaténation des blocs numéros  $B_1^*, \dots, B_K^*$  (avec multiplicité) :

$$(\xi_i^*)_{1 \leq i \leq n} = \bigcup_{k=1}^K (\xi_i)_{\frac{(B_k^*-1)n}{K} + 1 \leq i \leq \frac{B_k^*n}{K}}.$$

De cette manière, si  $K$  est bien choisi, on conserve une bonne part de la structure de dépendance (à l'intérieur des blocs) tout en ayant rééchantillonné des blocs qui sont (presque) indépendants les uns des autres.

## 4 Pénalités par rééchantillonnage

Pour le problème de sélection de modèles (décrit à la section 3.9 du chapitre « fondamentaux »), une quantité clé (mais inconnue) est la *pénalité idéale* :

$$\text{pen}_{\text{id}}(m; P_n) := \mathcal{R}_P(\hat{f}_m(P_n)) - \hat{\mathcal{R}}_n(\hat{f}_m(P_n)).$$

L'heuristique de rééchantillonnage conduit à estimer son espérance par

$$C_W \mathbb{E}_W \left[ \hat{\mathcal{R}}_n(\hat{f}_m(P_n^W)) - \hat{\mathcal{R}}_n^W(\hat{f}_m(P_n^W)) \right],$$

où  $\hat{\mathcal{R}}_n^W$  désigne le risque empirique sur le rééchantillon. Une telle quantité est appelée pénalité par rééchantillonnage (Efron, 1983; Arlot, 2009). Ce type de pénalité peut être utilisé avec succès pour la sélection de modèles : des inégalités-oracle ont notamment été démontrées en régression hétéroscédastique (Arlot, 2009) et en estimation de densité (Lerasle, 2012).

Comme indiqué au chapitre « fondamentaux », un majorant classique de la pénalité idéale est la « pénalité idéale globale » :

$$\text{pen}_{\text{id,g}}(m) := \sup_{f \in \mathcal{S}_m} \{ \mathcal{R}_P(f) - \hat{\mathcal{R}}_n(f) \},$$

que l'on peut majorer comme indiqué à la section 3.7 du chapitre « fondamentaux ». Un majorant intermédiaire est notamment deux fois l'espérance de la *complexité de Rademacher* (globale) de  $\mathcal{S}_m$ , dont on rappelle la définition :

$$\text{Rad}_n(\mathcal{B}_{\mathcal{S}_m}(D_n)) = \mathbb{E} \left[ \sup_{f \in \mathcal{S}_m} \left\{ \frac{1}{n} \sum_{i=1}^n \varepsilon_i c(f(X_i), Y_i) \right\} \middle| D_n \right].$$

Or, l'estimateur par rééchantillonnage de l'espérance de la pénalité idéale globale est

$$C_W \mathbb{E} \left[ \sup_{f \in \mathcal{S}_m} \left\{ \hat{\mathcal{R}}_n(f) - \hat{\mathcal{R}}_n^W(f) \right\} \middle| D_n \right] = C_W \mathbb{E} \left[ \sup_{f \in \mathcal{S}_m} \left\{ \frac{1}{n} \sum_{i=1}^n (1 - W_i) c(f(X_i), Y_i) \right\} \middle| D_n \right].$$

Si l'on prend des poids Bernoulli (1/2), les  $1 - W_i$  sont des variables de Rademacher indépendantes, et l'on obtient exactement la complexité de Rademacher globale.



## Références

- Sylvain ARLOT : Model selection by resampling penalization. *Electronic Journal of Statistics*, 3:557–624 (electronic), 2009.
- Bradley EFRON : Bootstrap methods : another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.
- Bradley EFRON : Estimating the error rate of a prediction rule : improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983.
- Bradley EFRON : Second thoughts on the bootstrap. *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, 18(2):135–140, 2003. Silver anniversary of the bootstrap.
- Bradley EFRON et Robert J. TIBSHIRANI : *An Introduction to the Bootstrap*, volume 57 de *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York, 1993.
- Peter HALL : *The Bootstrap and Edgeworth Expansion*. Springer Series in Statistics. Springer-Verlag, New York, 1992.
- Matthieu LERASLE : Optimal model selection in density estimation. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 48(3):884–908, 2012.
- Aad W. VAN DER VAART et Jon A. WELLNER : *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.