Sélection de modèles et sélection d'estimateurs pour l'Apprentissage statistique (Cours Peccot)

Troisième cours: Rééchantillonnage et pénalisation

SYLVAIN ARLOT
CNRS – ÉQUIPE SIERRA

Table des matières

Date: 24 Janvier 2011.

1. Régressogrammes en régression hétéroscédastique

Une bonne partie des résultats de ce cours et du suivant portent sur le problème de sélection parmi des régressogrammes en régression hétéroscédastique sur un plan d'expérience aléatoire. Commençons par décrire formellement ce cadre.

1.1. **Cadre de la régression hétéroscédastique.**

On considère le cadre de la régression sur un plan d'expérience aléatoire, où l'on observe (X_1, Y_1), …, $(X_n, Y_n) \in \mathcal{X} \times \mathbb{R}$ i.i.d. de loi commune P. Si $(X, Y) \sim P$, on note

$$Y = \eta(X) + \varepsilon$$

où la fonction de régression $\eta : \mathcal{X} \mapsto \mathbb{R}$ est définie par $\eta(X) = \mathbb{E}[Y \mid X]$ P-p.s. et le bruit ε vérifie donc P-p.s.

$$\mathbb{E}[\varepsilon \mid X] = 0 \quad \text{et} \quad \mathbb{E}\left[\varepsilon^2 \,\middle|\, X\right] = \sigma^2(X) \ .$$

La fonction inconnue $\sigma : \mathcal{X} \mapsto [0, +\infty[$ mesure le niveau de bruit des observations en fonction de leur position. On parle de *régression hétéroscédastique* dans la mesure où la distribution conditionnelle de ε sachant X est autorisée à varier en fonction de X, en particulier via le niveau de bruit $\sigma(\cdot)$.

Notons ici l'intérêt fondamental de supposer le plan d'expérience aléatoire : cela nous permet de conserver un échantillon $D_n = (X_i, Y_i)_{1 \le i \le n}$ de données i.i.d., alors que si le plan d'expérience était déterministe, les résidus $\varepsilon_1, \ldots, \varepsilon_n$ ne seraient pas identiquement distribués.

Rappelons enfin que l'objectif est de construire un prédicteur $t \in \mathbb{S}$ (l'ensemble des applications mesurables $\mathcal{X} \mapsto \mathbb{R}$) dont la perte quadratique

$$P\gamma(t) = \mathbb{E}_{(X,Y)\sim P}\left[\gamma(t; (X, Y))\right] = \mathbb{E}_{(X,Y)\sim P}\left[\left(t(X) - Y\right)^2\right]$$

est minimale, avec $\gamma : \mathcal{X} \times \mathbb{R} \mapsto \mathbb{R}$ le contraste des moindres carrés défini par

$$\forall (x, y) \in \mathcal{X} \times \mathbb{R} \ , \quad \forall t \in \mathbb{S} \ , \qquad \gamma\left(t; (x, y)\right) = (t(x) - y)^2 \ .$$

La perte est minimale pour $t = s^\star$ et l'on définit alors la perte relative

$$\ell\left(s^\star, t\right) = P\gamma(t) - P\gamma(s^\star) = \mathbb{E}_{(X,Y)\sim P}\left[\left(s^\star(X) - t(X)\right)^2\right] \ .$$

1.2. **Régressogrammes.**

Pour toute partition finie m de $\mathcal{X}$, on définit l'ensemble des fonctions constantes sur chaque élément de m (ou «modèle des histogrammes associé à m») par

$$S_m := \left\{ t = \sum_{\lambda \in m} \alpha_\lambda \mathbb{1}_\lambda \text{ t.q. } \alpha \in \mathbb{R}^m \right\} \ .$$

Un estimateur des moindres carrés sur S_m (ou *régressogramme associé à la partition* m) est alors défini par

$$\widehat{s}_m = \widehat{s}_m(D_n) \in \operatorname{argmin}_{t\in S_m} \{ P_n\gamma(t) \} \quad \text{où} \quad P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}$$

est la mesure empirique associée à l'échantillon D_n.

Pour tout échantillon $D_n = (X_i, Y_i)_{1\le i\le n}$ et tout $\lambda \subset \mathcal{X}$, on pose

$$\widehat{p}_\lambda = \widehat{p}_\lambda(D_n) = \frac{1}{n} \operatorname{Card} \{ i \text{ t.q. } X_i \in \lambda \} \ .$$

Le régressogramme $\widehat{s}_m$ associé à m est défini de manière unique si et seulement si

$$\min_{\lambda\in m} \{ \widehat{p}_\lambda \} > 0 \ ,$$

et dans ce cas

$$\widehat{s}_m = \sum_{\lambda\in m} \widehat{\beta}_\lambda \mathbb{1}_\lambda \quad \text{avec} \quad \forall \lambda \in m \ , \qquad \widehat{\beta}_\lambda = \widehat{\beta}_\lambda(D_n) := \frac{1}{n\widehat{p}_\lambda} \sum_{i \text{ t.q. } X_i\in\lambda} Y_i \ . \tag{1}$$

Preuve de l'Équation (1). Pour tout $t = \sum_{\lambda\in m} t_\lambda \mathbb{1}_\lambda \in S_m$, on a

$$P_n\gamma(t) = \frac{1}{n} \sum_{\lambda\in m} \left[\sum_{X_i\in\lambda} (Y_i - t_\lambda)^2 \right] \ . \tag{2}$$

Pour tout $\lambda \in m$, si $\widehat{p}_\lambda > 0$, $t_\lambda \mapsto \sum_{X_i\in\lambda} (Y_i - t_\lambda)^2$ admet un unique minimum sur $\mathbb{R}$ en $t_\lambda = \widehat{\beta}_\lambda$, ce qui prouve (1) lorsque $\min_{\lambda\in m} \{ \widehat{p}_\lambda \} > 0$. À l'inverse, s'il existe un $\lambda \in m$ pour lequel $\widehat{p}_\lambda = 0$, le risque empirique de t ne dépend pas de t_λ, et donc $\widehat{s}_m$ n'est pas défini de manière unique sur chaque $\lambda \in m$ pour lequel $\widehat{p}_\lambda = 0$. □

De la même manière, pour tout $t = \sum_{\lambda\in m} t_\lambda \mathbb{1}_\lambda \in S_m$, on a

$$\begin{aligned} \ell(s^\star, t) &= \mathbb{E}\left[(t(X) - s^\star(X))^2 \right] \\ &= \sum_{\lambda\in m} \left(p_\lambda \mathbb{E}\left[(t_\lambda - s^\star(X))^2 \,\middle|\, X \in \lambda \right] \right) \end{aligned} \tag{3}$$

ce qui prouve que

$$s^\star_m(P) = s^\star_m \in \operatorname{argmin}_{t\in S_m} \{ P\gamma(t) \}$$

existe toujours, et est unique si et seulement si

$$\min_{\lambda\in m} p_\lambda > 0 \quad \text{où} \quad \forall \lambda \in m \ , \qquad p_\lambda := \mathbb{P}_{(X,Y)\sim P}(X \in \lambda) \ .$$

Dans ce cas, on a

$$s^\star_m = \sum_{\lambda\in m} \beta_\lambda \mathbb{1}_\lambda \quad \text{avec} \quad \forall \lambda \in m \ , \qquad \beta_\lambda := \mathbb{E}_{(X,Y)\sim P}[Y \mid X \in \lambda] \tag{4}$$

et l'erreur d'approximation du modèle S_m vaut

$$\ell\,(\,s^\star, s_m^\star\,) = \sum_{\lambda\in m}\left(p_\lambda\left(\sigma_\lambda^{(d)}\right)^2\right) \tag{5}$$

$$\text{où}\quad \left(\sigma_\lambda^{(d)}\right)^2 := \mathbb{E}\left[\,(\,\beta_\lambda - s^\star(X)\,)^2\,\middle|\; X\in\lambda\right]\ .$$

1.3. **Exemples de collections de partitions.** On peut notamment considérer les deux exemples suivants de partitions de $\mathcal{X} = [0,1[$:

(1) Collection des partitions régulières $\mathcal{M}^{(\mathrm{reg})} = \mathcal{M}^{(\mathrm{reg})}(M_n)$, avec un nombre de morceaux maximal $M_n \geq 1$:

$$\mathcal{M}^{(\mathrm{reg})} := \{\,m_D \text{ t.q. } 1\leq D\leq M_n\,\}\quad\text{avec}\quad m_D := \left(\left[\frac{k-1}{D},\frac{k}{D}\right[\right)_{1\leq k\leq D}\ .$$

(2) Collection des partitions bi-régulières avec rupture en $1/2$ $\mathcal{M}_n^{(\mathrm{reg},1/2)}$, avec un nombre de morceaux maximal $M_n \geq 2$:

$$\mathcal{M}^{(\mathrm{reg})} := \{\,m_1\,\}\cup\left\{m_{(D_1,D_2)} \text{ t.q. } 1\leq D_1,D_2\leq\frac{M_n}{2}\right\}$$

$$\text{avec}\quad m_{(D_1,D_2)} := \left(\left[\frac{k-1}{2D_1},\frac{k}{2D_1}\right[\right)_{1\leq k\leq D_1}\cup\left(\left[\frac12+\frac{k-1}{2D_2},\frac12+\frac{k}{2D_2}\right[\right)_{1\leq k\leq D_2}\ .$$

Lorsque $\mathcal{X} = [0,1[^{\,d}$, on peut généraliser la collection $\mathcal{M}^{(\mathrm{reg})}(M_n)$ comme suit :

$$\mathcal{M}^{(\mathrm{reg})} := \left\{m_T \text{ t.q. } 1\leq T\leq M_n^{1/d}\right\}\quad\text{avec}\quad m_T := \left(\prod_{i=1}^d\left[\frac{k_i-1}{T},\frac{k_i}{T}\right[\right)_{1\leq k_1,\ldots,k_d\leq T}\ .$$

1.4. **Pénalité idéale.** Nous sommes désormais en mesure de calculer la perte relative et la pénalité idéale associées au modèle S_m. En utilisant les notations de la Section 3.3 du deuxième cours, il s'agit de calculer les termes p_1 et p_2.

On suppose désormais que $\min_{\lambda\in m} p_\lambda > 0$ (sinon, il suffit de fusionner les $\lambda\in m$ tels que $p_\lambda = 0$ avec un autre élément de m, sans rien changer au régressogramme correspondant.

Formules closes. Tout d'abord, d'après (3) et (4), pour tout $t = \sum_{\lambda\in m} t_\lambda \mathbb{1}_\lambda \in S_m$,

$$\begin{aligned} P\,(\,\gamma\,(t)-\gamma\,(s_m^\star)\,) &= \sum_{\lambda\in m}\left[p_\lambda\left(\mathbb{E}\left[\,(\,t_\lambda - s^\star(X)\,)^2\,\middle|\; X\in\lambda\right] - \mathbb{E}\left[\,(\,\beta_\lambda - s^\star(X)\,)^2\,\middle|\; X\in\lambda\right]\right)\right] \\ &= \sum_{\lambda\in m}\left[p_\lambda\,(\,t_\lambda-\beta_\lambda\,)^2\right]\ . \end{aligned}$$

Ainsi, en supposant que $\min_{\lambda\in m}\{\,\widehat{p}_\lambda\,\} > 0$,

$$p_1(m) = P\,(\,\gamma\,(\,\widehat{s}_m\,)-\gamma\,(\,s_m^\star\,)\,) = \sum_{\lambda\in m}\left[p_\lambda\left(\widehat{\beta}_\lambda-\beta_\lambda\right)^2\right]\ . \tag{6}$$

De la même manière, d'après (2) et (1), pour tout $t = \sum_{\lambda\in m} t_\lambda \mathbb{1}_\lambda \in S_m$,

$$P_n\left(\gamma\left(t\right) - \gamma\left(\widehat{s}_m\right)\right) = \frac{1}{n}\sum_{\lambda\in m}\sum_{X_i\in\lambda}\left[\left(t_\lambda - Y_i\right)^2 - \left(\widehat{\beta}_\lambda - Y_i\right)^2\right]$$
$$= \frac{1}{n}\sum_{\lambda\in m}\sum_{X_i\in\lambda}\left(t_\lambda - \widehat{\beta}_\lambda\right)^2 = \sum_{\lambda\in m}\widehat{p}_\lambda\left(t_\lambda - \widehat{\beta}_\lambda\right)^2$$

si bien que

$$p_2(m) = P_n\left(\gamma\left(s_m^\star\right) - \gamma\left(\widehat{s}_m\right)\right) = \sum_{\lambda\in m}\left[\widehat{p}_\lambda\left(\widehat{\beta}_\lambda - \beta_\lambda\right)^2\right] \quad . \tag{7}$$

Notons que (7) est encore valable lorsque $\min_{\lambda\in m}\widehat{p}_\lambda = 0$ en choisissant n'importe quelle valeur pour $\widehat{\beta}_\lambda$, grâce au facteur multiplicatif $\widehat{p}_\lambda$.

Rappelons qu'on a alors calculé les termes principaux de la pénalité idéale

$$\mathrm{pen}_{\mathrm{id}}(m) = p_1(m) + p_2(m) - \delta(m) \quad \text{où} \quad \delta(m) = (P_n - P)\gamma\left(s_m^\star\right) \quad , \tag{8}$$

le terme $\delta(m)$ étant d'espérance nulle.

Espérances conditionnelles. En vue d'obtenir une formule pour l'excès de risque et l'espérance de la pénalité idéale, commençons par calculer les espérances de p_1 et p_2 conditionnellement aux positions

$$\mathcal{P}_m = \mathcal{P}_m(D_n) := \left(\mathbb{1}_{X_i\in\lambda}\right)_{1\le i\le n\,,\,\lambda\in m}$$

des X_i dans les éléments de m. Pour tout λ, sachant $\mathcal{P}_m$, si $\widehat{p}_\lambda > 0$, $\widehat{\beta}_\lambda$ est la moyenne de $n\widehat{p}_\lambda$ variables aléatoires i.i.d. de moyenne β_λ et de variance σ_λ^2, où

$$\sigma_\lambda^2 := \mathbb{E}_{(X,Y)\sim P}\left[\left(Y - \beta_\lambda\right)^2 \,\middle|\, X\in\lambda\right] = \left(\sigma_\lambda^{(d)}\right)^2 + \left(\sigma_\lambda^{(a)}\right)^2$$
$$\text{et} \quad \left(\sigma_\lambda^{(a)}\right)^2 := \mathbb{E}_{(X,Y)\sim P}\left[\left(Y - s^\star(X)\right)^2 \,\middle|\, X\in\lambda\right]$$
$$= \mathbb{E}_{(X,Y)\sim P}\left[\left(\sigma(X)\right)^2 \,\middle|\, X\in\lambda\right] \quad .$$

Par conséquent, pour tout $\lambda\in m$ tel que $\widehat{p}_\lambda > 0$,

$$\mathbb{E}\left[\left(\widehat{\beta}_\lambda - \beta_\lambda\right)^2 \,\middle|\, \mathcal{P}_m\right] = \frac{\sigma_\lambda^2}{n\widehat{p}_\lambda} \quad ,$$

et donc, en supposant $\min_{\lambda\in m}\widehat{p}_\lambda > 0$, (6) et (7) impliquent

$$\mathbb{E}\left[p_1(m) \mid \mathcal{P}_m\right] = \frac{1}{n}\sum_{\lambda\in m}\left(\frac{\sigma_\lambda^2 p_\lambda}{\widehat{p}_\lambda}\right) \tag{9}$$

$$\mathbb{E}\left[p_2(m) \mid \mathcal{P}_m\right] = \frac{1}{n}\sum_{\lambda\in m}\sigma_\lambda^2 \quad . \tag{10}$$

Espérances. Commençons par p_2 : d'après (10), on a

$$\mathbb{E}\left[p_2(m)\right] = \frac{1}{n}\sum_{\lambda\in m}\sigma_\lambda^2 \quad . \tag{11}$$

On voit avec (9) que le cas $\widehat{p}_\lambda = 0$ va poser problème pour parler d'espérance de $p_1(m)$. En effet, on ne peut jamais exclure cet événement (qui a une probabilité strictement positive sauf dans le cas trivial $p_\lambda = 1$), si bien qu'il va falloir choisir une convention pour $\widehat{\beta}_\lambda$ lorsque $\widehat{p}_\lambda = 0$. Cette convention est purement formelle (on ne l'utilisera jamais en pratique, considérant que $\widehat{\beta}_\lambda$ est indéfini lorsque $\widehat{p}_\lambda = 0$), mais indispensable pour une étude théorique rigoureuse. Notons enfin que ce problème ne se pose pas pour $p_2(m)$ dont l'espérance (11) est parfaitement définie en toute généralité. Nous ne parlerons plus dans ces notes de ce problème avant tout technique, renvoyant à [3, 4] pour donner un sens précis aux résultats énoncés dans la suite.

À l'aide de (9), on obtient l'expression suivante pour l'«espérance»[1] de $p_1(m)$:

$$\mathbb{E}\left[p_1(m)\right] = \frac{1}{n}\sum_{\lambda\in m}\left(\sigma_\lambda^2 \mathbb{E}\left[\frac{p_\lambda}{\widehat{p}_\lambda}\,\middle|\,\widehat{p}_\lambda > 0\right]\right) \quad . \tag{12}$$

Pour rendre plus claire l'expression obtenue avec (12), on note

$$\delta_{n,p_\lambda} := \mathbb{E}\left[\frac{p_\lambda}{\widehat{p}_\lambda}\,\middle|\,\widehat{p}_\lambda > 0\right] - 1 \ ,$$

qui devrait intuitivement être petite dès que np_λ est assez grand, car alors $n\widehat{p}_\lambda$ se concentre autour de son espérance np_λ. Plus précisément, comme $n\widehat{p}_\lambda$ suit une loi binômiale de paramètres (n, p_λ), le résultat suivant — extrait de [3, Lemme 3] — nous donne un contrôle non-asymptotique de δ_{n,p_λ}.

Lemme 1. *Soit $n \in \mathbb{N}\setminus\{0\}$, $p \in\ ]0,1]$, Z une variable aléatoire binômiale de paramètres (n,p). On pose $\kappa_1 = 5.1$ et $\kappa_2 = 3.2$. Alors, si $np \geq 1$,*

$$1 - e^{-np} \leq \mathbb{E}\left[Z\right]\mathbb{E}\left[Z^{-1}\,\middle|\,Z>0\right] \leq \min\left\{1 + \frac{\kappa_1}{(np)^{1/4}}\ ,\ \kappa_2\right\} \quad . \tag{13}$$

La preuve du Lemme 1 repose principalement sur l'inégalité de Bernstein [39, Proposition 2.9] appliquée à Z, voir [3].

En combinant (11), (12) et le Lemme 1, on obtient que p_1 et p_2 sont proches en espérance (clé de l'heuristique de pente, voir le deuxième cours et [6]) et l'expression suivante pour l'espérance de la pénalité idéale :

$$\begin{aligned}\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] &= \frac{1}{n}\sum_{\lambda\in m}\left[\left(2+\delta_{n,p_\lambda}\right)\left(\sigma_\lambda\right)^2\right] \\ &= \frac{1}{n}\sum_{\lambda\in m}\left[\left(2+\delta_{n,p_\lambda}\right)\left(\left(\sigma_\lambda^{(a)}\right)^2 + \left(\sigma_\lambda^{(d)}\right)^2\right)\right] \quad .\end{aligned} \tag{14}$$

De même, en combinant (5), (12) et le Lemme 1, on obtient une expression de l'excès de risque :

$$\mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m\right)\right] = \sum_{\lambda\in m} p_\lambda\left(\sigma_\lambda^{(d)}\right)^2 + \frac{1}{n}\sum_{\lambda\in m}\left[\left(1+\delta_{n,p_\lambda}\right)\sigma_\lambda^2\right] \quad . \tag{15}$$

[1] ou du moins une quantité déterministe autour de laquelle p_1 se concentre effectivement

Comparaison avec le cadre du deuxième cours. Il est intéressant de comparer l'expression (14) de l'espérance de la pénalité idéale avec l'expression $2\sigma^2 D_m/n$ obtenue au premier cours dans le cas homoscédastique avec un plan d'expérience déterministe. On peut relever deux différences principales :

- l'hétéroscédasticité (via les variations de $\sigma(\cdot)$) transforme $\sigma^2 D_m$ en $\sum_{\lambda\in m} \sigma_\lambda^{(a)\,2}$
- le caractère aléatoire du plan d'expérience ajoute les termes $\sigma_\lambda^{(d)\,2}$, qui sont le plus souvent négligeables car $\sigma_\lambda^{(d)\,2} \leq \|s^\star - s_m^\star\|_\infty^2 \to 0$ dès que $\sup_{\lambda\in m} \mathrm{Leb}(\lambda) \to 0$ si $s^\star$ est un peu régulière, tandis que $\sigma_\lambda^{(a)\,2} \geq \inf_{x\in\lambda}(\sigma(x))^2$.

2. Nécessité d'estimer la forme de la pénalité

L'objectif de cette section est de montrer que l'on ne peut se contenter d'une pénalité fonction de la dimension des modèles lorsque les données sont hétéroscédastiques.

2.1. **Illustration sur un exemple.** À titre d'illustration, considérons le cas $\mathcal{M}_n = \mathcal{M}_n^{(\mathrm{reg},1/2)}$ avec $\mathcal{X} = [0,1]$. Intuitivement, une telle collection de modèles laisse la possibilité de s'adapter à l'hétéroscédasticité, pour peu que l'on en tienne compte dans la pénalité. On peut commencer à formaliser cette intuition en calculant plus précisément l'espérance de la pénalité idéale. La quantité-clé apparaissant dans (14) est, pour $\lambda \subset [0,1]$,

$$\begin{aligned}
\sigma_\lambda^2 &= \mathbb{E}\left[(Y-\beta_\lambda)^2 \,\middle|\, X\in\lambda \right] \\
&= \mathbb{E}\left[(s^\star(X) - s_m^\star(X))^2 \,\middle|\, X\in\lambda \right] + \mathbb{E}\left[(\sigma(X))^2 \,\middle|\, X\in\lambda \right] \\
&= \frac{1}{\mathrm{Leb}(\lambda)} \int_\lambda (s^\star(x) - s_m^\star(x))^2\,dx + \frac{1}{\mathrm{Leb}(\lambda)} \int_\lambda (\sigma(X))^2\,dx \ .
\end{aligned}$$

Ainsi,

$$\begin{aligned}
\mathbb{E}\left[p_2(m)\right] &= \frac{1}{n}\sum_{\lambda\in m} \sigma_\lambda^2 \\
&= \frac{2D_{m,1}}{n} \int_0^{1/2} \left[(s^\star(X) - s_m^\star(x))^2 + (\sigma(x))^2 \right] dx \\
&\qquad + \frac{2D_{m,2}}{n} \int_{1/2}^1 \left[(s^\star(X) - s_m^\star(x))^2 + (\sigma(x))^2 \right] dx \\
&= \frac{2\sigma_a^2 D_{m,1}}{n} + \frac{2\sigma_b^2 D_{m,2}}{n} + R(m,n)
\end{aligned}$$

$$\text{où} \quad \sigma_a^2 := \int_0^{1/2} (\sigma(x))^2\,dx \qquad \sigma_b^2 := \int_{1/2}^1 (\sigma(x))^2\,dx$$

$$\text{et} \quad 0 \leq R(m,n) = \frac{2}{n}\left(D_{m,1} \int_0^{1/2} (s^\star(X) - s_m^\star(x))^2\,dx \right.$$

$$+D_{m,2}\int_{1/2}^{1}\left(s^\star(X)-s^\star_m(x)\right)^2dx\Bigg)$$
$$\leq \frac{\ell\left(s^\star,s^\star_m\right)}{\left(\ln(n)\right)^2}\ ,$$

car $D_{m,i}\leq n/(2(\ln(n))^2)$. On en déduit donc que

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]\approx\frac{4}{n}\left[\sigma_a^2D_{m,1}+\sigma_b^2D_{m,2}\right]\ ,$$

qui est proportionnelle à $D_m=D_{m,1}+D_{m,2}$ si et seulement si

$$\sigma_a^2=\int_0^{1/2}\left(\sigma(x)\right)^2dx=\sigma_b^2=\int_{1/2}^{1}\left(\sigma(x)\right)^2dx\ .$$

On va voir que lorsque cette condition est vérifiée, avec $\mathcal{M}_n=\mathcal{M}_n^{(\mathrm{reg},1/2)}$, utiliser une pénalité fonction de la dimension est sous-optimal.

2.2. **Caractérisation des pénalités fonction de la dimension.** Il est très simple de caractériser l'ensemble des modèles pouvant être sélectionnés si l'on utilise une fonction de la dimension comme pénalité. En effet, pour tout $D\in\mathcal{D}_n=\{D_m\text{ t.q. }m\in\mathcal{M}_n\}$, on pose

$$\mathcal{M}_{\mathrm{dim}}(D):=\mathrm{argmin}_{m\in\mathcal{M}_n\text{ t.q. }D_m=D}\left\{P_n\gamma\left(\widehat{s}_m\right)\right\}$$
$$\text{et}\quad\mathcal{M}_{\mathrm{dim}}:=\bigcup_{D\in\mathcal{D}_n}\mathcal{M}_{\mathrm{dim}}(D)\ .$$

Alors, le lemme ci-dessous montre que toute procédure de choix de modèles par pénalisation avec une pénalité ne dépendant que de D_m sélectionne un modèle $\widehat{m}\in\mathcal{M}_{\mathrm{dim}}$.

Lemme 2. *Pour toute fonction $F:\mathcal{M}_n\mapsto\mathbb{R}$ et tout échantillon $(X_i,Y_i)_{1\leq i\leq n}$,*

$$\mathrm{argmin}_{m\in\mathcal{M}_n}\left\{P_n\gamma\left(\widehat{s}_m\right)+F(D_m)\right\}\subset\mathcal{M}_{\mathrm{dim}}\ .$$

Preuve du Lemme 2. Soit $\widehat{m}_F\in\mathrm{argmin}_{m\in\mathcal{M}_n}\left\{P_n\gamma\left(\widehat{s}_m\right)+F(D_m)\right\}$. Alors, pour tout $m\in\mathcal{M}_n$,

$$P_n\gamma\left(\widehat{s}_{\widehat{m}_F}\right)+F(D_{\widehat{m}_F})\leq P_n\gamma\left(\widehat{s}_m\right)+F(D_m)\ . \tag{16}$$

En particulier, (16) est vraie pour tout $m\in\mathcal{M}_n$ tel que $D_m=D_{\widehat{m}_F}$, pour lesquels $F(D_{\widehat{m}_F})=F(D_m)$. Par conséquent, (16) entraîne que $\widehat{m}_F\in\mathcal{M}_{\mathrm{dim}}(D_{\widehat{m}_F})$, soit $\widehat{m}_F\in\mathcal{M}_{\mathrm{dim}}$. □

Mentionnons ici que Breiman [13] avait remarqué que seul un petit nombre de modèles — appelés «RSS-extreme submodels» — peuvent être sélectionnés par des pénalités de la forme $F(D_m)=KD_m$ avec $K\geq 0$. Là où Breiman insistait sur les avantages de cette limitation du point de vue de la complexité algorithmique, nous allons voir que c'est précisément cette limitation qui empêche d'avoir un excès de risque équivalent à celui de l'oracle lorsque le niveau de bruit varie.

2.3. **Sous-optimalité des pénalités fonction de la dimension.** D'un point de vue théorique, en supposant pour simplifier que $\mathcal{M}_n = \mathcal{M}_n^{(\mathrm{reg},1/2)}$, on peut prouver le résultat négatif suivant (extrait de [5]) pour les pénalités fonction de la dimension en présence d'hétéroscédasticité.

Théorème 1. *On suppose que* $(X_1, Y_1), \ldots, (X_n, Y_n) \in [0,1] \times \mathbb{R}$ *sont i.i.d. avec* X_i *uniforme sur* $[0,1]$ *et* $\forall i = 1, \ldots, n$, $Y_i = X_i + \varepsilon_i$ *avec* $(\varepsilon_i)_{1 \leq i \leq n}$ *indépendantes telles que* $\mathbb{E}[\varepsilon_i \mid X_i] = 0$ *et* $\mathbb{E}\left[\varepsilon_i^2 \mid X_i\right] = (\sigma(X_i))^2$ *. On suppose de plus que* $s^\star \in \mathcal{C}^2$*,*

$$\|\varepsilon_i\|_\infty \leq E < \infty \ , \qquad \min\left\{(\sigma_a)^2, (\sigma_b)^2\right\} > 0 \quad et \quad (\sigma_a)^2 \neq (\sigma_b)^2$$

$$o\grave{u} \quad (\sigma_a)^2 := \int_0^{1/2} (\sigma(x))^2 \, dx \quad et \quad (\sigma_b)^2 := \int_{1/2}^1 (\sigma(x))^2 \, dx \ .$$

Soit $\mathcal{M}_n = \mathcal{M}_n^{(\mathrm{reg},1/2)}$ *la collection définie en Section 1.3, avec une dimension maximale* $M_n = \lfloor n/(\ln(n))^2 \rfloor$ *. Alors, il existe des constantes* $K_1, \mathbb{C}_1 > 0$ *et un événement de probabilité au moins* $1 - K_1 n^{-2}$ *sur lequel, pour toute fonction* $F : \mathcal{M}_n \mapsto \mathbb{R}$ *et tout* $\widehat{m}_F \in \operatorname{argmin}_{m \in \mathcal{M}_n} \{P_n \gamma(\widehat{s}_m) + F(D_m)\}$ *,*

$$\ell\left(s^\star, \widehat{s}_{\widehat{m}_F}\right) \geq \mathbb{C}_1 \inf_{m \in \mathcal{M}_n} \{\ell(s^\star, \widehat{s}_m)\} \quad avec \quad \mathbb{C}_1 > 1 \ . \tag{17}$$

La constante $\mathbb{C}_1$ *dépend uniquement de* $(\sigma_a)^2 / (\sigma_b)^2$ *; la constante* K_1 *peut uniquement dépendre de* E, $(\sigma_a)^2$, $(\sigma_b)^2$, $\|(s^\star)'\|_\infty$ *et* $\|(s^\star)''\|_\infty$ *.*

Le Théorème 1 est prouvé dans [5]. Outre le calcul de l'espérance de $\mathrm{pen}_{\mathrm{id}}(m)$, il repose essentiellement sur un contrôle de l'erreur d'approximation $\ell(s^\star, \widehat{s}_m)$ en fonction de $(D_{m,1}, D_{m,2})$ et sur les inégalité de concentration suivantes pour les composantes de $\mathrm{pen}_{\mathrm{id}}(m)$:

Proposition 3. *Soit* $\gamma > 0$, $B_n \geq 2$. *Soit* m *une partition de* $\mathcal{X}$ *telle que* $\min_{\lambda \in m} \{np_\lambda\} \geq B_n$ *et* $p_1(m), p_2(m)$ *définies par* (6) *et* (7). *On suppose que*

(**Ab**) *Les données sont bornées :* $\|Y_i\|_\infty \leq A < \infty$.

(**An**) *Le niveau de bruit est uniformément minoré :* $\sigma(X_i) \geq \sigma_{\min} > 0$ *p.s.*

Alors, une constante absolue $L_1 > 0$ *et une constante* $L_2(A/\sigma_{\min}, \gamma)$ *existent telles qu'avec probabilité au moins* $1 - L_1 n^{-\gamma}$*,*

$$|p_1(m) - \mathbb{E}[p_1(m) \mid \mathcal{P}_m]| \leq L_2 \left[\frac{(\ln(n))^2}{\sqrt{D_m}} + D_m e^{-LB_n}\right] \mathbb{E}[p_2(m)] \tag{18}$$

$$|p_2(m) - \mathbb{E}[p_2(m) \mid \mathcal{P}_m]| \leq L_2 \frac{\ln(n)}{\sqrt{D_m}} \mathbb{E}[p_2(m)] \ . \tag{19}$$

De plus, sous la seule hypothèse (**Ab**), l'inégalité de Bernstein [39, Proposition 2.9] implique : pour tout $x \geq 0$, avec probabilité au moins $1 - 2e^{-x}$,

$$\forall \theta \in \ ]0,1] \ , \ |(P_n - P)(\gamma(s_m^\star) - \gamma(s^\star))| \leq \theta \ell(s^\star, s_m^\star) + \frac{6A^2 x}{\theta n} \ . \tag{20}$$

2.4. **Pourquoi estimer la forme de la pénalité ?** Le résultat montré par le Théorème 1 est très fort : toute pénalité de la forme $F(D_m)$, même avec

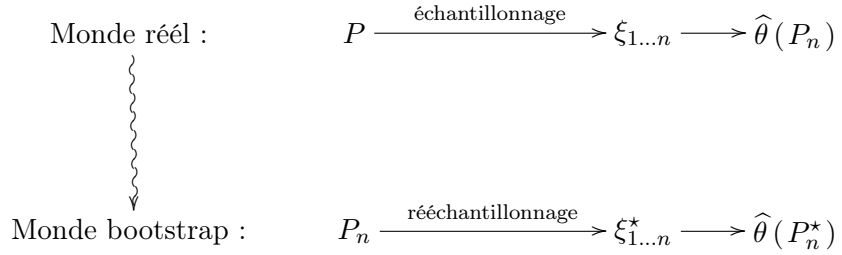

FIG. 1. L'heuristique de rééchantillonnage, selon Efron [19]. Schéma inspiré de [22, Figure 1].

une fonction F utilisant la connaissance complète de P et P_n, conduit à la perte d'un facteur multiplicatif $\mathbb{C}_1 > 1$ par rapport à l'oracle, en termes de perte relative. Si l'on veut une procédure satisfaisant une inégalité-oracle avec constante $(1 + o(1))$ dans un cadre hétéroscédastique, il est donc nécessaire d'estimer la forme de la pénalité.

Notons en revanche que si l'on se contente d'une inégalité-oracle avec constante multiplicative $\mathcal{O}(1)$, on peut utiliser une pénalité linéaire en D_m avec une constante suffisamment grande, par exemple $\mathrm{pen}(m) = 2\,\|\sigma\|^2_\infty\, D_m n^{-1}$ [5, Proposition 2]. On risque toutefois le sur-apprentissage (et une explosion de la perte relative) si l'on sous-estime $\|\sigma\|^2_\infty$ [5, Proposition 3].

3. RÉÉCHANTILLONNAGE

Rééchantillonner, c'est utiliser un jeu de données (l'*échantillon*) pour construire un ou plusieurs nouveaux échantillons (les *rééchantillons*). Derrière une description aussi simple se cachent de nombreuses méthodes statistiques :

- le *sous-échantillonnage* consiste dans la sélection (aléatoire ou non) d'une partie des données initiales.
- le *bootstrap* (non-paramétrique) consiste dans la génération aléatoire d'un n-échantillon i.i.d., suivant la loi uniforme sur l'échantillon initialement observé. Autrement dit, on effectue n tirages avec remise dans les observations.

Dans ces deux cas, on peut — entre autres — estimer le biais ou la variance d'un estimateur en comparant le(s) rééchantillon(s) à l'échantillon initial, ou les rééchantillons entre eux. Pour une introduction plus complète au rééchantillonnage, nous renvoyons par exemple à [2, Section 1.1].

3.1. **Heuristique d'Efron.**

L'heuristique qui sous-tend la plupart des utilisations du rééchantillonnage est le principe de «plug-in», décrit par Efron au sujet du bootstrap [19, 22]. Une autre description intuitive de cette heuristique se trouve dans l'introduction de [31]. Considérons la situation suivante : n observations $\xi_1, \dots, \xi_n \in \Xi$ ont été générées indépendamment avec une même loi P inconnue. On souhaite estimer un paramètre d'intérêt $\theta = t(P)$ (par exemple la moyenne, la variance, un quantile ; mais aussi des paramètres

plus complexes telles qu'une fonction de régression ou un prédicteur). Pour cela, on dispose d'un estimateur $\widehat{\theta}(P_n) = \widehat{\theta}(\xi_{1\dots n})$. La question qui se pose alors est : comment évaluer la qualité de cet estimateur ? C'est-à-dire, évaluer son biais, sa variance, construire un intervalle de confiance, ou — s'il s'agit d'un prédicteur — déterminer son erreur de prédiction.

L'heuristique proposée par Efron est de construire un «monde bootstrap», miroir du «monde réel» mais dans lequel aucune quantité n'est inconnue (voir Figure 1). La loi inconnue P est remplacée par une estimation $\widehat{P}$. Nous nous limiterons ici au cadre «non-paramétrique» pour lequel $\widehat{P}$ est la mesure empirique $P_n = n^{-1}\sum_{i=1}^n \delta_{\xi_i}$. Le processus d'échantillonnage est remplacé par celui de rééchantillonnage. Dans le cas du bootstrap, les processus d'échantillonnage et de rééchantillonnage sont identiques : $\xi_1^\star, \dots, \xi_n^\star$ sont i.i.d. de loi P_n. On note $P_n^\star = n^{-1}\sum_{i=1}^n \delta_{\xi_i^\star}$ leur mesure empirique. L'heuristique d'Efron se ramène donc à un principe de «plug-in» : on remplace P par son estimateur P_n, les autres quantités en découlant selon des processus inchangés.

La qualité d'un estimateur $\widehat{\theta}(P_n)$ — de même que de nombreuses autres quantités — s'écrit sous la forme $F(P, P_n)$. Cette quantité est inaccessible car P est inconnue, mais son équivalent dans le monde bootstrap $F(P_n, P_n^\star)$ ne dépend que des observations. On peut donc le calculer, au besoin en faisant une approximation de type Monte-Carlo (voir [31, Appendice II] et [23, Chapitre 23]). L'heuristique de rééchantillonnage se formalise donc ainsi, $\mathcal{L}(X)$ désignant la loi de la variable aléatoire X :

$$\mathcal{L}\left(F\left(P, P_n\right)\right) \approx \mathcal{L}\left(F\left(P_n, P_n^\star\right) \mid P_n\right) \quad .$$

3.2. **Rééchantillonnage à poids échangeables.** Dans l'heuristique ci-dessus, nous n'avons fait apparaître le rééchantillon que via sa mesure empirique

$$P_n^\star = \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i^\star} = \frac{1}{n}\sum_{i=1}^n W_i \delta_{\xi_i}$$

où, pour tout i, on a noté W_i le nombre d'occurrences de ξ_i dans le rééchantillon $\xi_{1\dots n}^\star$.

On peut ainsi reformuler et généraliser l'heuristique ci-dessus au «bootstrap à poids échangeables» (ou plus exactement rééchantillonnage à poids échangeables) [38, 44], en remplaçant $P_n^\star$ par

$$P_n^W := \frac{1}{n}\sum_{i=1}^n W_i \delta_{\xi_i}$$

où $W = (W_1, \dots, W_n) \in \mathbb{R}^n$ est un vecteur de poids aléatoire, échangeable[2] et indépendant de $\xi_{1\dots n}$. Notons que l'on ne suppose pas nécessairement que les W_i sont entiers, ni même positifs. Par contre, il est en général préférable d'avoir $\mathbb{E}[W_i] = 1$. Parmi les poids classiques, on trouve :

[2] $W \in \mathbb{R}^n$ est échangeable si et seulement si pour toute permutation σ de $\{1, \dots, n\}$, W a la même distribution que $\left(W_{\sigma(1)}, \dots, W_{\sigma(n)}\right)$. Voir par exemple [1].

- *Efron (m)*, $m \in \mathbb{N}\setminus\{0\}$: $mn^{-1}W$ suit une loi multinomiale de paramètres $(m; n^{-1}, \dots, n^{-1})$. Lorsque $m = n$, c'est le *bootstrap*. Lorsque $m < n$, c'est le *«m out of n» bootstrap*.
- *Random hold-out*[3] *(q)*, $q \in \{1, \dots, n-1\}$: $W_i = nq^{-1}\mathbb{1}_{i\in I}$ avec I un sous-ensemble aléatoire de $\{1, \dots, n\}$, choisi uniformément parmi les parties de cardinal q. Mise à part la contrainte sur la loi de I, il s'agit de poids de sous-échantillonnage (subsampling), également appelé «bootstrap without replacement». Lorsque $q = n-1$, on retrouve le leave-one-out.
- *Rademacher (p)*, $p \in (0,1)$: $pW_1, \dots, pW_n$ sont i.i.d., de loi Bernoulli de paramètre p. Le nom Rademacher provient du cas $p = 1/2$ où les poids sont (à translation près) des variables Rademacher i.i.d. À la normalisation de P_n^W près, il s'agit de poids de sous-échantillonnage.

Mentionnons pour finir les poids *Poisson (μ)*, $\mu > 0$: $\mu W_1, \dots, \mu W_n$ sont i.i.d. de loi Poisson de paramètre μ. Ces poids correspondent à un argument classique, dit de «Poissonisation», car on peut également les définir comme les poids Efron (M) avec $M \sim \mathcal{P}(\mu n)$. En prenant une taille de rééchantillon aléatoire, on peut ainsi rendre les poids W_i indépendants.

3.3. **Arguments théoriques asymptotiques.** Les études théoriques sur le rééchantillonnage se sont multipliées depuis les travaux d'Efron [19]. On peut citer par exemple [20, 23, 31, 46] sur le bootstrap, et [7, 25] sur le bootstrap à poids. Nous ne détaillerons ici qu'un seul résultat [50, Théorème 3.6.13] montrant que la version rééchantillonnée $\widehat{\mathbb{G}}_n^W$ du processus empirique recentré $\sqrt{n}(P_n - P)$ converge (à constante multiplicative près) vers le même processus Gaussien $\mathbb{G}$. Dans l'énoncé qui suit, on a omis des conditions de mesurabilité sur $\mathcal{F}$ pour simplifier.

Théorème 2. *Soit $\mathcal{F}$ une classe de Donsker de fonctions mesurables. Pour tout $n \in \mathbb{N}$, soit $(W_{n,1}, \dots, W_{n,n}) \in \mathbb{R}^n$ un vecteur aléatoire positif, échangeable, indépendant de $\xi_{1\dots n}$ tel que*

$$\sup_{n\in\mathbb{N}} \left\| W_{n,1} - \overline{W}_n \right\|_{2,1} < \infty \quad \textit{avec} \quad \overline{W}_n = n^{-1}\sum_{i=1}^n W_{n,i}$$

$$\textit{et} \quad \|Z\|_{2,1} := \int_0^\infty \sqrt{\mathbb{P}\left(|Z| > t\right)}dt \ ,$$

$$n^{-1/2}\mathbb{E}\left[\max_{1\le i\le n} \left|W_{n,i} - \overline{W}_n\right|\right] \xrightarrow{(p)} 0 \quad \textit{et} \quad n^{-1}\sum_{i=1}^n \left(W_{n,i} - \overline{W}_n\right)^2 \xrightarrow{(p)} c^2 > 0 \ .$$

Alors, lorsque n tend vers l'infini,

$$\sup_{h\in BL_1} \left|\mathbb{E}_W\left[h\left(\widehat{\mathbb{G}}_n^W\right)\right] - \mathbb{E}\left[h\left(c\mathbb{G}\right)\right]\right| \xrightarrow{(p)} 0$$

$$\textit{avec} \quad \widehat{\mathbb{G}}_n^W := \frac{1}{\sqrt{n}}\sum_{i=1}^n W_{n,i}\left(\delta_{\xi_i} - P_n\right) = \sqrt{n}\left(P_n^W - \overline{W}_n P_n\right) \ ,$$

[3] que l'on peut traduire par «validation aléatoire», à rapprocher de la validation croisée.

$\mathbb{G}$ *étant un processus gaussien de moyenne nulle et de fonction de covariance* $\mathrm{cov}(f,g) = P(fg) - P(f)P(g)$ *et* BL_1 *l'ensemble des fonctions 1-lipschitziennes et bornées par 1 (pour la norme* $\|\cdot\|_\infty$*).*

3.4. **Usages du rééchantillonnage.** De par sa formulation simple et très générale, le rééchantillonnage est désormais un outil statistique utilisé dans un grand nombre de domaines, voir notamment [53, 16]. À l'origine, l'objectif d'Efron se limitait à utiliser le bootstrap pour l'estimation du biais et de la variance d'un estimateur. Le bootstrap a ensuite été utilisé fructueusement pour :

- construire des intervalles de confiance [17],
- calculer des p-valeurs pour des statistiques de test [11, 10, 27],
- estimer une erreur de prédiction [51, 24, 41],
- faire de la sélection de modèles [21, 45],
- *etc.*

Notons que l'on peut utiliser d'autres types de rééchantillonnage, en particulier le sous-échantillonnage, pour construire des intervalles de confiance ou des tests [43], et pour bien d'autres applications encore.

Outre sa simplicité et sa généralité, le rééchantillonnage est souvent utilisé pour ses propriétés *stabilisatrices*. Ceci est particulièrement utile en classification, où nombre d'algorithmes sont très sensibles à une perturbation par un petit nombre de données. Une partie des observations étant absente dans chaque rééchantillon, on peut ainsi obtenir un algorithme dont la sortie varie très peu si l'on supprime un petit nombre de données. Citons ici le «bagging» (contraction de «bootstrap aggregating») [15] et les forêts aléatoires [14] entre autres algorithmes utilisant cette propriété du rééchantillonnage.

3.5. **Limites du rééchantillonnage.** La simplicité du rééchantillonnage le rend cependant facile à utiliser abusivement. Mentionnons ici quelques-unes de ses limites [37, 16] :

- l'estimation de queues de distributions plus lourdes que Gaussiennes [30].
- l'utilisation de poids échangeables correspond à l'hypothèse implicite que les données sont *échangeables*. Si ce n'est pas le cas, il faut alors faire très attention à la manière d'appliquer l'heuristique de rééchantillonnage. Par exemple, dans le cadre de la régression sur un plan d'expérience déterministe, les (X_i, Y_i) ne sont pas échangeables et l'on est amené à *rééchantillonner les résidus* lorsque les erreurs ε_i sont i.i.d. [23]. Le problème se corse en présence d'hétéroscédasticité, même s'il peut être contourné en choisissant bien les poids W [51].
- lorsque les données sont dépendantes, le bootstrap ne fonctionne pas en général [48], la manière la plus classique de résoudre le problème étant de rééchantillonner par blocs [42].

Avant d'utiliser le rééchantillonnage pour construire une pénalité, étudions de manière détaillée un estimateur par rééchantillonnage dans un cadre assez simple.

4. Un estimateur de la variance par rééchantillonnage

Soient $\xi_1, \dots, \xi_n$ des variables aléatoires i.i.d. de moyenne μ et de variance σ^2. Leur variance s'écrit alors

$$\sigma^2 = \mathbb{E}\left[n \left(\frac{1}{n} \sum_{i=1}^{n} \xi_i - \mu \right)^2 \right] = n\mathbb{E}\left[\left(\mathbb{E}_{\xi \sim P_n} \xi - \mathbb{E}_{\xi \sim P} \xi \right)^2 \right] = n\mathbb{E}\left[F(P, P_n) \right] \ , \tag{21}$$

si bien que l'on dispose de l'estimateur par rééchantillonnage

$$\widehat{\sigma^2_W} = n\mathbb{E}_W \left[F(P_n, P_n^W) \right] \ .$$

4.1. **Formule close.**

Calculons-le pour des poids W échangeables généraux :

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

car

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

Ainsi,

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

en posant

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

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

$$\begin{aligned} \widehat{\sigma_W^2} &= \frac{R_V^{(W)}}{n} \mathbb{1}_{n\geq 2} \left[\sum_{i=1}^{n} (\xi_i - \mu)^2 - \frac{1}{n-1} \sum_{i \neq j} \left((\xi_i - \mu)(\xi_j - \mu) \right) \right] \qquad (22) \\ &= \frac{R_V^{(W)}}{n} \mathbb{1}_{n\geq 2} \left[\sum_{i=1}^{n} \xi_i^2 - \frac{1}{n-1} \sum_{i \neq j} (\xi_i \xi_j) \right] \ , \end{aligned}$$

la dernière expression provenant de l'invariance de $\widehat{\sigma_W^2}$ par translation des données.

4.2. **Comparaison avec l'estimateur classique.** On suppose désormais que $n \geq 2$. Remarquons que l'estimateur sans biais classique de la variance s'écrit

$$\begin{aligned} \widehat{\sigma^2} &= \frac{1}{n-1} \sum_{i=1}^{n} \left(\xi_i - \frac{1}{n} \sum_{k=1}^{n} \xi_k \right)^2 \\ &= \frac{1}{n-1} \sum_{i=1}^{n} \left((\xi_i - \mu) - \frac{1}{n} \sum_{k=1}^{n} (\xi_k - \mu) \right)^2 \\ &= \frac{n}{n-1} \left[\frac{1}{n} \sum_{i=1}^{n} (\xi_i - \mu)^2 - \left(\frac{1}{n} \sum_{i=1}^{n} (\xi_i - \mu) \right)^2 \right] \\ &= \frac{1}{n} \left[\sum_{i=1}^{n} (\xi_i - \mu)^2 - \frac{1}{n-1} \sum_{i \neq j} \left((\xi_i - \mu)(\xi_j - \mu) \right) \right] \ . \qquad (23) \end{aligned}$$

Ainsi,

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

d'où l'on déduit notamment que

$$\mathbb{E}\left[\widehat{\sigma_W^2} \right] = R_V^{(W)} \sigma^2 \ . \qquad (24)$$

4.3. **Rééchantillonnage et structure.** Une explication de la relation (24) est que l'estimateur par rééchantillonnage de $\mathbb{E}[F(P, P_n)]$ respecte la *structure* de $F(P, P_n)$. En effet, lorsque $F(P, P_n)$ est définie par (21), elle possède

les trois propriétés suivantes :

échangeabilité : $F(P, P_n)$ ne dépend pas de l'ordre des données (25)

invariance par translation de la loi de ξ d'une constante $c \in \mathbb{R}$ (26)

homogénéité : si $\xi - \mu$ est multipliée par $\lambda > 0$, (27)
alors $F(P, P_n)$ est multipliée par λ^2 .

De plus, $F(P, P_n)$ ne dépend de P que via une quantité de la forme $\mathbb{E}_{\xi\sim P} Q(\xi)$ pour un polynôme Q, et $F(P, P_n)$ est un polynôme en $\mathbb{E}_{\xi\sim P} Q(\xi)$ et $\xi_1, \ldots, \xi_n$. Par conséquent, lorsque W est échangeable,

$$\widehat{\sigma^2_W} = \mathbb{E}_W \left[F(P_n, P_n^W) \right]$$

possède automatiquement les mêmes propriétés structurelles.

Comme remarqué dans [2, Section 8.2], ces conditions sont suffisamment fortes pour contraindre $\widehat{\sigma^2_W}$ à satisfaire (24) pour une constante $R_V^{(W)}$ dès lors que W est échangeable. Plus précisément, on a le lemme suivant, prouvé initialement dans [2, Lemme 8.2].

Lemme 4. *Soit $n \in \mathbb{N}\setminus\{0\}$ et $f : z = (z_i)_{1\le i\le n} \in \mathbb{R}^n \mapsto \mathbb{R}$ vérifiant :*

(25) *f est échangeable, c'est-à-dire, pour toute permutation σ de $\{1, \ldots, n\}$ et tout $z \in \mathbb{R}^n$, $f\left((z_{\sigma(i)})_{1\le i\le n}\right) = f(z)$.*

(26) *pour tout $c \in \mathbb{R}$ et $z \in \mathbb{R}^n$, $f\left((z_i + c)_{1\le i\le n}\right) = f(z)$.*

(27) *$f(z)$ est un polynôme en $z_1, \ldots, z_n$ et pour tout $\lambda \in \mathbb{R}$ et $z \in \mathbb{R}^n$, $f\left((\lambda z_i)_{1\le i\le n}\right) = \lambda^2 f(z)$.*

Alors, il existe $\alpha \in \mathbb{R}$ tel que pour tout $z \in \mathbb{R}^n$,

$$f(z) = \alpha \left(\sum_{i=1}^n z_i^2 - \frac{1}{n} \left(\sum_{1\le i\le n} z_i \right)^2 \right) .$$

Réciproquement, une fonction de cette forme possède les propriétés (25)–(27).

Preuve du Lemme 4. D'après (27), il existe $(a_{i,j})_{1\le i,j\le n} \in R^{n^2}$ tel que pour tout $z \in \mathbb{R}^n$,

$$f(z) = \sum_{1\le i,j\le n} a_{i,j} z_i z_j .$$

D'après (25), $a_{i,j}$ peut uniquement dépendre de $\mathbb{1}_{i=j}$, c'est-à-dire, pour tout $z \in \mathbb{R}^n$,

$$f(z) = \alpha \sum_{i=1}^n z_i^2 + 2\beta \sum_{1\le i<j\le n} z_i z_j = (\alpha - \beta) \sum_{i=1}^n z_i^2 + \beta \left(\sum_{i=1}^n z_i \right)^2$$

avec $a_{1,1} = \alpha$ et $a_{1,2} = \beta$. En utilisant maintenant (26), on a pour tout $c \in \mathbb{R}$ et $z \in \mathbb{R}^n$,

$$f(z + c) = f(z) + (\alpha + (n-1)\beta) \left(2c \sum_{i=1}^n z_i + nc^2 \right) .$$

On a donc $\alpha + (n-1)\beta = 0$, d'où le résultat. La réciproque se vérifie simplement. □

4.4. **Rééchantillonnage et concentration.** Il est intéressant de comparer la variance de

$$Z_1 = F(P, P_n) = n\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i - \mu\right)^2 = \frac{1}{n}\left(\sum_{i=1}^{n}(\xi_i - \mu)\right)^2$$

à celle de l'estimateur par rééchantillonnage de son espérance $\widehat{\sigma^2_W}$.

Le Lemme 5 ci-dessous (avec $a = 0$ et $b = 1/n$) montre que

$$\mathrm{var}(Z_1) = 2\sigma^4 + \frac{\mathbb{E}\left[(\xi_1 - \mu)^4\right] - 3\sigma^4}{n} \ . \tag{28}$$

De même, le Lemme 5 (avec $a = 1/(n-1)$ et $b = -1/(n(n-1))$) montre que

$$\mathrm{var}\left(\frac{1}{R_V^{(W)}}\widehat{\sigma^2_W}\right) = \frac{1}{n}\left(\mathbb{E}\left[(\xi_1 - \mu)^4\right] - \sigma^4\right) + \frac{2}{n(n-1)}\sigma^4 \ . \tag{29}$$

La comparaison de (28) et (29) met en lumière un phénomène de surconcentration, classique pour les estimateurs d'espérances par rééchantillonnage : la variance de $\widehat{\sigma^2_W}$ est de l'ordre de n^{-1} alors que la quantité-cible $F(P, P_n)$ a une variance de l'ordre de 1.

Pour finir, énonçons et prouvons le Lemme 5.

Lemme 5. *Soit $\xi_1, \ldots, \xi_n$ des v.a.i.i.d. centrées de variance σ^2, et l'on pose $\kappa^4 = \mathbb{E}\left[\xi_1^4\right]$. Soient $a, b \in \mathbb{R}$ et*

$$Z = a\sum_{i=1}^{n}\xi_i^2 + b\left(\sum_{i=1}^{n}\xi_i\right)^2 = (a+b)\sum_{i=1}^{n}\xi_i^2 + b\sum_{i\neq j}\xi_i\xi_j \ .$$

Alors,

$$\mathrm{var}(Z) = n\,(a+b)^2\left(\kappa^4 - \sigma^4\right) + 2n(n-1)b^2\sigma^4 \ .$$

Preuve du Lemme 5. D'une part,

$$\mathbb{E}\left[Z\right] = (a+b)n\sigma^2 \ .$$

D'autre part,

$$\begin{aligned}
\mathbb{E}\left[Z^2\right] &= a^2\mathbb{E}\left[\left(\sum_{i=1}^{n}\xi_i^2\right)^2\right] + b^2\mathbb{E}\left[\left(\sum_{i=1}^{n}\xi_i\right)^4\right] + 2ab\sum_{i,j,k}\mathbb{E}\left[\xi_i^2\xi_j\xi_k\right] \\
&= (a+b)^2\sum_{i,j}\mathbb{E}\left[\xi_i^2\xi_j^2\right] + 2b^2\sum_{i\neq j}\mathbb{E}\left[\xi_i^2\xi_j^2\right] \\
&= (a+b)^2\left(n\kappa^4 + n(n-1)\sigma^4\right) + 2b^2n(n-1)\sigma^4 \ .
\end{aligned}$$

On en déduit le résultat en retirant $(\mathbb{E}Z)^2$ à $\mathbb{E}[Z^2]$. □

$\mathcal{L}(W)$	Efr(m)	Rad(p)	Rho(q)	Loo
$R_V^{(W)}$	$\frac{n-1}{m}$	$\frac{1+\delta_{n,p}}{p}-1$	$\frac{n}{q}-1$	$\frac{1}{n-1}$

TAB. 1. Valeur de $R_V^{(W)}$, défini par (30), pour différents rééchantillonnages.

4.5. **Calcul de la constante multiplicative.** Pour finir cette section, calculons pour les exemples classiques de poids la valeur de la constante multiplicative

$$R_V^{(W)} := \mathbb{E}_W\left[\left(\frac{nW_i}{\sum_{k=1}^n W_k}-1\right)^2\right] \tag{30}$$

par laquelle il faut normaliser $\widehat{\sigma_W^2}$ pour en faire un estimateur sans biais de σ^2. La Table 1 synthétise les résultats obtenus.

Démonstration. Pour les poids Efron(m) et Random hold-out (q), on a

$$\sum_{i=1}^n W_i = n \quad \text{p.s.} \quad \text{et} \quad \mathbb{E}\left[W_i\right] = 1 \ ,$$

si bien que $R_V^{(W)} = \mathrm{var}(W_1)$. On en déduit :

– pour Efron(m) (abrégé Efr(m)) : $W_1 \sim (n/m) \times \mathcal{B}(m, 1/n)$ et donc

$$R_V^{(W)} = \mathrm{var}(W_1) = \frac{n^2}{m^2} \times m \times \frac{1}{n}\left(1-\frac{1}{n}\right) = \frac{n-1}{m} \ .$$

– pour Random hold-out (q) (abrégé Rho(q)) : $W_1 \sim (n/q) \times \mathcal{B}(q/n)$ et donc

$$R_V^{(W)} = \mathrm{var}(W_1) = \frac{n^2}{q^2} \times \frac{q}{n}\left(1-\frac{q}{n}\right) = \frac{n}{q}-1 \ .$$

En prenant $q = n-1$, on obtient la valeur de $R_V^{(W)}$ pour les poids Leave-one-out (abrégé Loo).

Pour les poids Rademacher (p) (abrégé Rad(p)), rappelons que les W_i sont i.i.d. de loi $(1/p) \times \mathcal{B}(p)$. La somme des W_i n'étant pas déterministe, nous allons commencer par raisonner conditionnellement à $\overline{W} = n^{-1}\sum_{i=1}^n W_i$. Comme W_i ne prend que deux valeurs 0 ou $1/p$, sachant $\overline{W}$, $W_i \sim (1/p) \times \mathcal{B}(p\mathbb{E}[W_i \mid \overline{W}])$. Or, les W_i étant échangeables, $\mathbb{E}[W_i \mid \overline{W}]$ ne dépend pas de i, si bien que

$$\mathbb{E}\left[W_i \mid \overline{W}\right] = \frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[W_k \mid \overline{W}\right] = \mathbb{E}\left[\overline{W} \,\middle|\, \overline{W}\right] = \overline{W} \ .$$

On en déduit que

$$\mathbb{E}\left[\left(W_i-\overline{W}\right)^2 \,\middle|\, \overline{W}\right] = \mathrm{var}\left(W_i \mid \overline{W}\right) = \frac{1}{p^2}p\overline{W}\left(1-p\overline{W}\right) = \frac{\overline{W}}{p}\left(1-p\overline{W}\right) \ ,$$

et donc

$$R_V^{(W)} = \mathbb{E}\left[\frac{\left(W_i - \overline{W}\right)^2}{\overline{W}^2}\right] = \mathbb{E}\left[\frac{1}{p\overline{W}}\left(1 - p\overline{W}\right)\right] = \mathbb{E}\left[\frac{1}{p\overline{W}}\right] - 1 \ .$$

Or, $np\overline{W} \sim \mathcal{B}(n,p)$, si bien que l'on peut utiliser le Lemme 1 :

$$\mathbb{E}\left[\frac{1}{\overline{W}}\right] = 1 + \delta_{n,p}$$

avec $\delta_{n,p} \to 0$ lorsque $np \to \infty$, si bien que

$$R_V^{(W)} = \frac{1+\delta_{n,p}}{p} - 1 \approx \frac{1}{p} - 1 \ .$$

□

5. Pénalités par rééchantillonnage

Nous avons vu (notamment avec le Lemme 2 du premier cours) qu'estimer sans biais l'espérance de la pénalité idéale

$$\mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma\left(\widehat{s}_m(P_n)\right)$$

est une bonne stratégie pour obtenir une procédure de choix de modèles par pénalisation optimale. Comme la pénalité idéale s'écrit comme une fonction de (P, P_n) (et ceci en supposant seulement que les estimateurs $\widehat{s}_m$ ne sont pas fonctions de l'ordre des données, mais seulement des fonctions de P_n), l'heuristique de rééchantillonnage (Section 3.1) nous conduit directement à l'estimateur suivant de l'espérance de $\mathrm{pen}_{\mathrm{id}}(m)$, que nous appellerons pénalité par rééchantillonnage :

$$\mathrm{pen}_W(m) := C_W \mathbb{E}_W\left[(P_n - P_n^W)\gamma\left(\widehat{s}_m(P_n^W)\right)\right] \tag{31}$$

La pénalité (31) a d'abord été proposée par Efron dans le cas du bootstrap [21], avec un premier résultat théorique sur son optimalité asymptotique dans [47]. La même pénalité avec des poids Efron(m) a été proposée par Shao [45] dans un objectif d'identification (voir Section 3.2 du premier cours), en prenant $m \ll n$. La définition (31) avec des poids échangeables généraux provient de [4].

Notons la présence d'une constante C_W devant la pénalité (31). Elle joue le même rôle que le facteur $1/R_V^{(W)}$ nécessaire pour l'estimation de la variance en Section 4. Nous proposerons une valeur pour C_W pour les poids les plus classiques en Section 6, dans le cas des régressogrammes, mais sans garantie aucune sur leur validité dans un cadre général. Il nous semble raisonnable de proposer d'estimer la forme de la pénalité seulement à l'aide de (31), et de calibrer la constante multiplicative à l'aide de l'heuristique de pente lorsque c'est possible (voir le deuxième cours).

D'autres pénalités par rééchantillonnage ont été proposées, en particulier les complexités de Rademacher (globales [33, 8] ou locales [9, 32]) en classification, les premières ayant été généralisées avec les pénalités globales par

rééchantillonnage échangeables en classification [25]. Nous ne parlerons ici (brièvement) que des complexités globales ; voir aussi [12, 39] au sujet des complexités globales ou locales en classification.

Les complexités de Rademacher (globales) reposent sur la majoration suivante de la pénalité idéale lorsque $\widehat{s}_m \in S_m$ p.s. :

$$\mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma\left(\widehat{s}_m(P_n)\right) \leq \sup_{t \in S_m} \left\{ (P - P_n)\gamma\left(t\right) \right\} =: \mathrm{pen}_{\mathrm{id,g}}(m) \ .$$

Une telle majoration peut paraître large, mais elle reste raisonnable pour obtenir des vitesses standard en classification avec un contraste borné. Son intérêt est que pour obtenir une inégalité-oracle, nous avons vu avec le Lemme 2 du premier cours qu'il est plus important d'avoir la minoration $\mathrm{pen}(m) \geq \mathrm{pen}_{\mathrm{id}}(m)$ que d'avoir une majoration de $\mathrm{pen}(m)$: au pire, le membre de droite de l'inégalité-oracle est augmenté de $\mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m)$, mais surpénaliser ne risque pas de conduire au sur-apprentissage. Ainsi, utiliser une estimation de $\mathrm{pen}_{\mathrm{id,g}}(m)$ (ou de son espérance) comme pénalité peut sembler raisonnable.

À l'aide d'un argument de symétrisation, on a obtenu la complexité de Rademacher globale

$$\mathbb{E}\left[\sup_{t \in S_m} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\varepsilon_i \gamma\left(t; \xi_i\right) \right) \right\} \middle| \ P_n \right] \quad \text{avec} \quad \varepsilon_1, \ldots, \varepsilon_n \text{ i.i.d. } \sim \mathcal{U}(\{-1, +1\}) \tag{32}$$

indépendantes de P_n. À constante multiplicative près (comprise entre un et deux), on peut montrer que (32) est proche de $\mathrm{pen}_{\mathrm{id,g}}(m)$, d'où une inégalité-oracle (fine si $\mathrm{pen}_{\mathrm{id,g}}$ n'est pas trop grande par rapport à $\mathrm{pen}_{\mathrm{id}}$). Mais si la principale justification théorique de (32) est un argument de symétrisation, il est intéressant de noter que cette pénalité est l'estimation par rééchantillonnage (avec des poids Rademacher (1/2), d'où leur nom) de $\mathrm{pen}_{\mathrm{id,g}}$. En effet,

$$\begin{aligned} &\mathbb{E}_W \left[\sup_{t \in S_m} \left\{ (P_n - P_n^W)\gamma(t) \right\} \middle| \ P_n \right] \\ &= \mathbb{E}_W \left[\sup_{t \in S_m} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left((1 - W_i)\gamma\left(t; \xi_i\right) \right) \right\} \middle| \ P_n \right] \end{aligned} \tag{33}$$

coïncide avec (32) car $1 - W_i \sim \varepsilon_i$ pour les poids Rademacher (1/2). En particulier, il est naturel de généraliser (32) à des poids généraux via (33), ce qui a été fait dans [25].

6. Garanties théoriques pour les régressogrammes

Cette section rassemble des résultats issus de [4] qui donnent des garanties théoriques pour les pénalités par rééchantillonnage échangeables (31) pour la sélection de régressogrammes en régression hétéroscédastique sur un plan d'expérience aléatoire (voir Section 1).

6.1. **Calcul de la pénalité par rééchantillonnage.** Commençons par calculer explicitement la pénalité (31) lorsque $\widehat{s}_m$ est un régressogramme associé à une partition finie m de $\mathcal{X}$. En appliquant directement la formule (31) au cadre des régressogrammes, on obtient la pénalité

$$C_W \mathbb{E}_W \left[P_n \gamma \left(\widehat{s}_m^W \right) - P_n^W \gamma \left(\widehat{s}_m^W \right) \right] \ , \tag{34}$$

$$\text{où} \quad \widehat{s}_m^W := \operatorname{argmin}_{t \in S_m} P_n^W \gamma(t) = \sum_{\lambda \in m} \widehat{\beta}_\lambda^W \mathbb{1}_\lambda, \quad \widehat{\beta}_\lambda^W := \frac{1}{n \widehat{p}_\lambda^W} \sum_{X_i \in \lambda} W_i Y_i \ ,$$

$$\widehat{p}_\lambda^W := P_n^W(X \in \lambda) = \widehat{p}_\lambda \widehat{W}_\lambda \quad \text{et} \quad \widehat{W}_\lambda := \frac{1}{n \widehat{p}_\lambda} \sum_{X_i \in \lambda} W_i \ .$$

Convention. De la même manière que l'espérance de $p_1(m)$ n'est pas bien définie à cause de l'événement $\min_{\lambda \in m} \widehat{p}_\lambda = 0$ (voir (12)), l'événement $\min_{\lambda \in m} \widehat{p}_\lambda^W = 0$ pose problème dans la définition (34) et a en général une probabilité non-nulle. Commençons par séparer la pénalité par rééchantillonnage en trois termes, à la suite de la décomposition (8) de la pénalité idéale :

$$\mathbb{E}_W \left[P_n \gamma \left(\widehat{s}_m^W \right) - P_n^W \gamma \left(\widehat{s}_m^W \right) \right] = \widehat{p}_0(m) + \widehat{p}_1(m) + \widehat{p}_2(m)$$

avec

$$\widehat{p}_0(m) := \mathbb{E}_W \left[(P_n - P_n^W) \gamma(\widehat{s}_m) \right] = \frac{1}{n} \sum_{i=1}^n \left(\mathbb{E}_W \left[1 - W_i \right] \gamma \left(\widehat{s}_m; (X_i, Y_i) \right) \right) = 0$$

car $\mathbb{E}_W[W_i] = 1$ pour tout i,

$$\widehat{p}_1(m) := \sum_{\lambda \in m} \mathbb{E}_W \left[\widehat{p}_\lambda \left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda \right)^2 \right]$$

$$\text{et} \quad \widehat{p}_2(m) := \sum_{\lambda \in m} \mathbb{E}_W \left[\widehat{p}_\lambda^W \left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda \right)^2 \right] \ .$$

Avec la convention $\widehat{p}_\lambda^W (\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda)^2 = 0$ lorsque $\widehat{p}_\lambda^W = 0$, $\widehat{p}_2(m)$ est bien définie car $\widehat{\beta}_\lambda^W$ est bien défini lorsque $\widehat{p}_\lambda^W > 0$. Il reste à définir $\widehat{p}_1(m)$. Notre proposition est de remplacer l'espérance sur W par une espérance conditionnellement à l'événement $\widehat{W}_\lambda > 0$, *séparément pour chaque* $m \in \mathcal{M}_n$ *et* $\lambda \in m$, ce qui garantit que la grande majorité des vecteurs de poids restent admissible ; voir [2, Section 8.1] pour une justification de ce choix dans un cadre plus général. En résumé, (31) est remplacée par

$$\begin{aligned} C_W \sum_{\lambda \in m} \Bigg(& \mathbb{E}_W \left[\widehat{p}_\lambda \left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda \right)^2 \,\middle|\, \widehat{W}_\lambda > 0 \right] \\ & + \mathbb{E}_W \left[\widehat{p}_\lambda^W \left(\widehat{\beta}_\lambda^W - \widehat{\beta}_\lambda \right)^2 \right] \Bigg) \ . \end{aligned} \tag{35}$$

Formule close. Nous pouvons désormais énoncer une formule close pour la pénalité par rééchantillonnage, correctement définie par (35).

Lemme 6. *Soit m une partition de $\mathcal{X}$ et $\widehat{s}_m$ le régressogramme associé. Soit $W \in [0, \infty[^{\,n}$ un vecteur aléatoire échangeable indépendant de $(X_i, Y_i)_{1 \le i \le n}$.*

Soit $\mathrm{pen}_W(m)$ *définie par* (35). *On suppose que* $\min_{\lambda\in m}\{n\widehat{p}_\lambda\}\geq 1$. *Alors,*

$$\mathrm{pen}_W(m)=\frac{C_W}{n}\sum_{\lambda\in m}\left(R_{1,W}(n,\widehat{p}_\lambda)+R_{2,W}(n,\widehat{p}_\lambda)\right)\frac{n\widehat{p}_\lambda S_{\lambda,2}-S_{\lambda,1}^2}{n\widehat{p}_\lambda(n\widehat{p}_\lambda-1)}\mathbb{1}_{n\widehat{p}_\lambda\geq 2} \tag{36}$$

$$\textit{avec}\quad S_{\lambda,1}:=\sum_{X_i\in\lambda}(Y_i-\beta_\lambda)\qquad S_{\lambda,2}:=\sum_{X_i\in\lambda}(Y_i-\beta_\lambda)^2 \tag{37}$$

$$R_{1,W}(n,\widehat{p}_\lambda):=\mathbb{E}\left[\frac{(W_1-\widehat{W}_\lambda)^2}{\widehat{W}_\lambda^2}\,\middle|\,X_1\in\lambda\ ,\ \widehat{W}_\lambda>0\right] \tag{38}$$

$$\textit{et}\quad R_{2,W}(n,\widehat{p}_\lambda):=\mathbb{E}\left[\frac{(W_1-\widehat{W}_\lambda)^2}{\widehat{W}_\lambda}\,\middle|\,X_1\in\lambda\right]\ . \tag{39}$$

Preuve du Lemme 6. La pénalité idéale s'écrit sous la forme

$$\mathrm{pen}_{\mathrm{id}}(m)=\sum_{\lambda\in m}F_\lambda\left(\mathcal{L}\left((X,Y)\,|\,X\in\lambda\right),(X_i,Y_i)_{X_i\in\lambda}\right)\ .$$

Le rééchantillonnage conservant cette structure, il n'est pas surprenant que l'on puisse décomposer de même la pénalité par rééchantillonnage :

$$\begin{aligned}
&\mathrm{pen}_W(m)=C_W\left(\widehat{p}_1(m)+\widehat{p}_2(m)\right)\\
\text{avec}\quad &\widehat{p}_1(m):=\sum_{\lambda\in m}\mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}_\lambda^W-\widehat{\beta}_\lambda\right)^2\,\middle|\,\widehat{W}_\lambda>0\right]\\
&\qquad=\frac{1}{n}\sum_{\lambda\in m}\mathbb{E}_W\left[\widehat{v}_\lambda^W\left(\widehat{W}_\lambda\right)\,\middle|\,\widehat{W}_\lambda>0\right]\\
&\widehat{p}_2(m):=\sum_{\lambda\in m}\mathbb{E}_W\left[\widehat{p}_\lambda^W\left(\widehat{\beta}_\lambda^W-\widehat{\beta}_\lambda\right)^2\right]=\frac{1}{n}\sum_{\lambda\in m}\mathbb{E}_W\left[\widehat{W}_\lambda\widehat{v}_\lambda^W\left(\widehat{W}_\lambda\right)\right]\\
&\widehat{v}_\lambda^W\left(\widehat{W}_\lambda\right)=\widehat{v}_\lambda^W\left(\widehat{W}_\lambda,(X_i,Y_i)_{X_i\in\lambda}\right):=n\widehat{p}_\lambda\mathbb{E}_W\left[\left(\widehat{\beta}_\lambda^W-\widehat{\beta}_\lambda\right)^2\,\middle|\,\widehat{W}_\lambda\right]\ .
\end{aligned}$$

Notons que le conditionnement suivant la valeur de $\widehat{W}_\lambda$ permet de faire apparaître la quantité que nous avons calculée à la Section 4. En effet, sachant $\widehat{W}_\lambda$, les poids $(W_i)_{X_i\in\lambda}$ sont échangeables et indépendants de $(X_i,Y_i)_{X_i\in\lambda}$, si bien que $\widehat{v}_\lambda^W\left(\widehat{W}_\lambda\right)$ est l'estimateur par rééchantillonnage échangeable de la variance σ_λ^2 de Y sachant $X\in\lambda$. Ceci permet non seulement de comprendre pourquoi la pénalité par rééchantillonnage estime bien l'espérance de la pénalité idéale (14) malgré l'hétéroscédasticité, mais également d'obtenir que

$$\begin{aligned}
&\widehat{v}_\lambda^W\left(\widehat{W}_\lambda\right)=\frac{R_V^{(W)}(\widehat{W}_\lambda,n\widehat{p}_\lambda)}{n\widehat{p}_\lambda-1}\left[S_{\lambda,2}-\frac{1}{n\widehat{p}_\lambda}S_{\lambda,1}^2\right]\mathbb{1}_{n\widehat{p}_\lambda\geq 2}\\
\text{avec}\quad &R_V^{(W)}(\widehat{W}_\lambda,n\widehat{p}_\lambda):=\mathbb{E}_W\left[\left(\frac{W_1}{\widehat{W}_\lambda}-1\right)^2\,\middle|\,\widehat{W}_\lambda\ ,\ X_1\in\lambda\right]\ .
\end{aligned}$$

On en déduit le résultat annoncé. □

$\mathcal{L}(W)$	Efr(m)	Rad(p)	Poi(μ)	Rho(q)	Loo
$R_{2,W}(n,\widehat{p}_\lambda)$	$\frac{n}{m}\left(1-\frac{1}{n\widehat{p}_\lambda}\right)$	$\frac{1}{p}-1$	$\frac{1}{\mu}\left(1-\frac{1}{n\widehat{p}_\lambda}\right)$	$\frac{n}{q}-1$	$\frac{1}{n-1}$
$C_{W,\infty}$	m/n	$p/(1-p)$	μ	$q/(n-q)$	$n-1$

TAB. 2. $R_{2,W}(n,\widehat{p}_\lambda)$ et $C_{W,\infty}$ pour quelques poids de rééchantillonnage.

Constantes multiplicatives pour des poids classiques. Pour finir de calculer complètement la pénalité par rééchantillonnage, il faut préciser les valeurs de $R_{1,W}(n,\widehat{p}_\lambda)$ et $R_{2,W}(n,\widehat{p}_\lambda)$. Pour les poids considérés en Section 4.5, on peut montrer que $R_{1,W}(n,\widehat{p}_\lambda) \approx R_{2,W}(n,\widehat{p}_\lambda)$ lorsque $n\widehat{p}_\lambda$ est assez grand, et l'on obtient les formules de la Table 2 pour $R_{2,W}(n,\widehat{p}_\lambda)$ [4, Lemme 17].

Espérances. Comme

$$\mathbb{E}\left[Y_i-\beta_\lambda \,\middle|\, X_i\in\lambda\right]=0 \quad \text{et} \quad \mathbb{E}\left[\left(Y_i-\beta_\lambda\right)^2 \,\middle|\, X_i\in\lambda\right]=\sigma_\lambda^2 \ ,$$

on déduit simplement de (36) que

$$\mathbb{E}\left[\mathrm{pen}_W(m) \,\middle|\, \mathcal{P}_m\right] = \frac{C_W}{n}\sum_{\lambda\in m}\left[\left(R_{1,W}(n,\widehat{p}_\lambda)+R_{2,W}(n,\widehat{p}_\lambda)\right)\sigma_\lambda^2 \mathbb{1}_{n\widehat{p}_\lambda\geq 2}\right] \ , \tag{40}$$

à comparer avec (9) et (10). En particulier, la pénalité par rééchantillonnage s'adapte bien à l'hétéroscédasticité des données dans le cas des régressogrammes.

En supposant $\min_{\lambda\in m}\{n\widehat{p}_\lambda\}$ suffisamment grand, compte-tenu de l'heuristique $R_{1,W}(n,\widehat{p}_\lambda)\approx R_{2,W}(n,\widehat{p}_\lambda)$, on en déduit que le bon choix (asymptotique) pour la constante C_W, celui qui conduit à un estimation sans biais de $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$, est de prendre

$$C_W = C_{W,\infty} := \lim_{n\widehat{p}_\lambda\to\infty}\frac{1}{R_{2,W}(n,\widehat{p}_\lambda)} \ ,$$

dont les valeurs[4] sont reportées en Table 2 pour les poids les plus classiques. Il est remarquable que les valeurs classiques des paramètres des poids (Efr(n), Rad(1/2), Poi(1), Rho($n/2$)) conduisent tous à une constante $C_{W,\infty}=1$, essentiellement car ce sont les valeurs des paramètres pour lesquelles $\mathrm{var}(W_1)\approx 1$.

Plus précisément, on peut montrer pour ces poids [4, Section 4] que

$$\mathbb{E}\left[\mathrm{pen}_W(m)\right] = \frac{C_W}{C_{W,\infty}}\frac{1}{n}\sum_{\lambda\in m}\left(2+\overline{\delta}_{n,p_\lambda}^{(\mathrm{penW})}\right)\sigma_\lambda^2$$

avec $\overline{\delta}_{n,p_\lambda}^{(\mathrm{penW})}\to 0$ lorsque $np_\lambda\to+\infty$.

[4]Précisons que la définition proposée pour $C_{W,\infty}$ n'a pas toujours du sens, dans la mesure où $R_{2,W}$ dépend parfois autant de $n\widehat{p}_\lambda$ que de n, et que l'on ne peut faire tendre $n\widehat{p}_\lambda$ vers l'infini en gardant n fixé. Pour être parfaitement rigoureux, il faut donc ne considérer $C_{W,\infty}$ définie que pour certains exemples de poids, par les valeurs données dans la Table 2.

Concentration. En utilisant la formule close (36), on peut montrer une inégalité de concentration pour la pénalité par rééchantillonnage [4, Proposition 3], dont nous énonçons ici une version simplifiée :

Proposition 7. *Soit $\gamma > 0$, $A_n \geq 2$, W un vecteur de poids échangeable indépendant de $(X_i, Y_i)_{1 \leq i \leq n}$. Soit m une partition de $\mathcal{X}$ et $\mathrm{pen}_W(m)$ définie par* (35)*. On suppose que*

(**Ab**) *Les données sont bornées : $\|Y_i\|_\infty \leq A < \infty$.*

(**An**) *Le niveau de bruit est uniformément minoré : $\sigma(X_i) \geq \sigma_{\min} > 0$ p.s.*

Alors, une constante absolue $L_1 > 0$ et une constante $L_2(A/\sigma_{\min}, \gamma) > 0$ existent telles qu'avec probabilité au moins $1 - L_1 n^{-\gamma}$,

$$|\mathrm{pen}_W(m) - \mathbb{E}[\mathrm{pen}_W(m) \mid \mathcal{P}_m]| \mathbb{1}_{\min_{\lambda \in m}\{n\widehat{p}_\lambda\} \geq A_n} \leq L_2 C_W \times \sup_{np \geq A_n} \{R_{1,W}(n,p) + R_{2,W}(n,p)\} \frac{\ln(n)}{\sqrt{A_n D_m}} \mathbb{E}[p_2(m)] \quad . \tag{41}$$

En particulier, pour les poids Efr(n), Poi(1), Rad(1/2), Rho(n/2) et Loo, une constante absolue $L_3 > 0$ existe telle que sur le même événement,

$$|\mathrm{pen}_W(m) - \mathbb{E}[\mathrm{pen}_W(m) \mid \mathcal{P}_m]| \mathbb{1}_{\min_{\lambda \in m}\{n\widehat{p}_\lambda\} \geq A_n} \leq \frac{C_W}{C_{W,\infty}} \frac{L_2 L_3 \ln(n)}{\sqrt{A_n D_m}} \mathbb{E}[p_2(m)] \quad .$$

On remarque que la borne (41) sur les déviations de pen_W est réduite d'un facteur $1/\sqrt{A_n}$ par rapport à la borne (18)–(19) sur les déviations de la pénalité idéale. Même s'il ne s'agit que de majorations, on retrouve une trace du phénomène de sur-concentration pour les estimateurs par rééchantillonnage que l'on avait mis en évidence dans un cas simple en Section 4.4.

6.2. **Inégalité-oracle.** L'événement $\min_{\lambda \in m}\{n\widehat{p}_\lambda\} \geq A_n$ ayant une grande probabilité (avec $A_n \propto \ln(n)$) dès que $\min_{\lambda \in m}\{np_\lambda\} \geq L \ln(n)$, la Proposition 7, combinée à la concentration de $\mathrm{pen}_{\mathrm{id}}$ (Proposition 3) et aux calculs d'espérance (14) et (40), conduit à une inégalité-oracle via le Lemme 2 du premier cours. Plus précisément, on obtient le théorème suivant [4, Théorème 1].

Théorème 3. *Soit $\mathcal{M}_n$ une famille de partitions de $\mathcal{X}$ et $(X_i, Y_i)_{1 \leq i \leq n}$ un échantillon i.i.d. satisfaisant les hypothèses suivantes :*

(**P1**) *Famille de complexité polynômiale* : $\mathrm{Card}(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$.

(**P2**) *Richesse minimale de $\mathcal{M}_n$: $\exists m_0 \in \mathcal{M}_n$ t.q. $D_{m_0} \in [\sqrt{n}; c_{\mathrm{rich}}\sqrt{n}]$.*

(**Ab**) *Les données sont bornées : $\|Y_i\|_\infty \leq A < \infty$.*

(**An**) *Le niveau de bruit est uniformément minoré : $\sigma(X_i) \geq \sigma_{\min} > 0$* p.s.

(**Ap**) *Décroissance polynômiale du biais : il existe $\beta_1 \geq \beta_2 > 0$ et $C_{\mathrm{b}}^+, C_{\mathrm{b}}^- > 0$ telles que*

$$\forall m \in \mathcal{M}_n \ , \qquad C_{\mathrm{b}}^- D_m^{-\beta_1} \leq \ell(s^\star, s_m^\star) \leq C_{\mathrm{b}}^+ D_m^{-\beta_2} \quad .$$

($\mathbf{Ar}_\ell^{\mathbf{X}}$) *Régularité inférieure des partitions $m \in \mathcal{M}_n$ vis-à-vis de $\mathcal{L}(X)$: il existe $c_{\mathrm{r},\ell}^X > 0$ telle que*

$$\forall m \in \mathcal{M}_n \ , \qquad D_m \min_{\lambda \in m} p_\lambda \geq c_{\mathrm{r},\ell}^X \ .$$

Soit pen_W *la pénalité définie par* (35) *avec* $C_W = C_{W,\infty}$ *(les poids étant choisis parmi Efr(n), Poi(1), Rad(1/2), Rho(n/2) et Loo).*

Alors, il existe une constante $K_1 > 0$ telle qu'avec probabilité au moins $1 - K_1 n^{-2}$, pour tout

$$\widehat{m} \in \mathrm{argmin}_{m \in \mathcal{M}_n \text{ t.q. } \min_{\lambda \in m}\{n\widehat{p}_\lambda\} \geq 3} \left\{ P_n \gamma\left(\widehat{s}_m\right) + \mathrm{pen}_W(m) \right\} \ ,$$

on a

$$\ell\left(s^\star, \widehat{s}_{\widehat{m}}\right) \leq \left(1 + (\ln(n))^{-1/5}\right) \inf_{m \in \mathcal{M}_n} \left\{ \ell\left(s^\star, \widehat{s}_m\right) \right\} \ . \tag{42}$$

De plus,

$$\mathbb{E}\left[\ell\left(s^\star, \widehat{s}_{\widehat{m}}\right)\right] \leq \left(1 + (\ln(n))^{-1/5}\right) \mathbb{E}\Big[\inf_{m \in \mathcal{M}_n} \left\{ \ell\left(s^\star, \widehat{s}_m\right) \right\} \Big] + \frac{A^2 K_1}{n^2}. \tag{43}$$

La constante K_1 peut dépendre des constantes apparaissant dans les hypothèses (**Ab**), (**An**), (**Ap**), ($\mathbf{Ar}_\ell^{\mathbf{X}}$), (**P1**) *et* (**P2**) *mais pas de n.*

Pour une discussion des hypothèses du Théorème 3, en particulier (**Ap**), et des jeux d'hypothèses alternatifs, nous renvoyons à [4, Section 3.3].

6.3. **Adaptation.** L'inégalité-oracle fournie par le Théorème 3 est l'occasion d'illustrer comment l'on peut construire un estimateur adaptatif via une procédure de sélection de modèles vérifiant une inégalité-oracle.

On suppose $\mathcal{X} = [0,1[^{\,d}$ pour simplifier, et l'on considère la famille $\mathcal{M}_n^{(\mathrm{reg})}$ des partitions régulières de $\mathcal{X}$ avec un nombre de morceaux maximal $M_n = n$ (voir Section 1.3). Dans la lignée des hypothèses du Théorème 3, on suppose également :

(**Ab**) Les données sont bornées : $\|Y_i\|_\infty \leq A < \infty$.

(**An**) Le niveau de bruit est uniformément minoré : $\sigma(X_i) \geq \sigma_{\min} > 0$ p.s.

($\mathbf{Ad}_\ell$) La densité de X est bornée inférieurement :

$$\exists c_X^{\min} > 0 \ , \qquad \forall I \subset \mathcal{X} \ , \qquad \mathbb{P}\left(X \in I\right) \geq c_X^{\min} \mathrm{Leb}(I) \ .$$

(**Ah**) La fonction de régression $s^\star = \eta$ est α-hölderienne avec $\alpha \in (0;1]$: il existe $R > 0$ tel que

$$s^\star \in \mathcal{H}(\alpha, R) \quad \text{c'est-à-dire} \quad \forall x_1, x_2 \in \mathcal{X} \ , \ |s^\star(x_1) - s^\star(x_2)| \leq R \|x_1 - x_2\|_\infty^\alpha \ .$$

On définit l'estimateur

$$\widetilde{s} := \widehat{s}_{\widehat{m}} \quad \text{avec} \quad \widehat{m} \in \mathrm{argmin}_{m \in \mathcal{M}_n^{(\mathrm{reg})} \text{ t.q. } \min_{\lambda \in m}\{n\widehat{p}_\lambda\} \geq 3} \left\{ P_n \gamma\left(\widehat{s}_m\right) + \mathrm{pen}_W(m) \right\}$$

et pen_W la pénalité définie par (35) avec $C_W = C_{W,\infty}$ (les poids étant choisis parmi Efr(n), Poi(1), Rad(1/2), Rho($n/2$) et Loo). Alors, d'après le Théorème 3, pour majorer le risque de $\widetilde{s}$, il suffit de majorer le risque de l'oracle

$$\begin{aligned}
&\inf_{m\in\mathcal{M}_n^{(\mathrm{reg})}} \{\mathbb{E}\left[\ell\left(s^\star,\widehat{s}_m\right)\right]\} \\
&= \inf_{1\le T\le n^{1/d}} \left\{ \ell\left(s^\star, s^\star_{m_T}\right) + \frac{1}{n}\sum_{\lambda\in m_T} \left(1+\delta_{n,p_\lambda}\right)\sigma_\lambda^2 \right\} \\
&\le \inf_{1\le T\le n^{1/d}(c_X^{\min})^{1/d}} \left\{ \left\|s^\star - s^\star_{m_T}\right\|_\infty^2 + \frac{D_{m_T}}{n}\left(1+\kappa_2\right)\left(\sigma_{\max}^2 + \left\|s^\star - s^\star_{m_T}\right\|_\infty^2\right) \right\}
\end{aligned}$$

Or, $D_{m_T} = T^d$ et comme $s^\star \in \mathcal{H}(\alpha, R)$, on a

$$\left\|s^\star - s^\star_{m_T}\right\|_\infty \le RT^{-\alpha}$$

et donc

$$\begin{aligned}
&\inf_{m\in\mathcal{M}_n^{(\mathrm{reg})}} \{\mathbb{E}\left[\ell\left(s^\star,\widehat{s}_m\right)\right]\} \\
&\le \inf_{1\le T\le n^{1/d}(c_X^{\min})^{1/d}} \left\{ R^2T^{-2\alpha} + \frac{T^d}{n}\left(1+\kappa_2\right)\left(\sigma_{\max}^2 + R^2T^{-2\alpha}\right) \right\} \\
&\le \left(2+\kappa_2\right) \inf_{1\le T\le n^{1/d}(c_X^{\min})^{1/d}} \left\{ R^2T^{-2\alpha} + \frac{\sigma_{\max}^2 T^d}{n} \right\} \\
&\le 2\left(2+\kappa_2\right) R^{\frac{2d}{2\alpha+d}} n^{\frac{-2\alpha}{2\alpha+d}} \sigma_{\max}^{\frac{4\alpha}{2\alpha+d}}
\end{aligned}$$

en choisissant

$$T = T_0 \approx \left(\frac{R^2 n}{\sigma_{\max}}\right)^{\frac{1}{2\alpha+d}} .$$

On en déduit que

$$\mathbb{E}\left[\ell\left(s^\star,\widetilde{s}\right)\right] \le K_2 R^{\frac{2d}{2\alpha+d}} n^{\frac{-2\alpha}{2\alpha+d}} \sigma_{\max}^{\frac{4\alpha}{2\alpha+d}} + \frac{K_3 A^2}{n^2} \tag{44}$$

pour des constantes $K_2 > 0$ numérique et $K_3(d, A, c_X^{\min}, R, \alpha, \sigma_{\min}) > 0$.

Nous renvoyons à [4, Théorème 2] pour un énoncé et une preuve précis de ce résultat, où il est également prouvé que si le niveau de bruit est régulier :

($\mathbf{A}\sigma$) σ est K_σ-Lipschitz par morceaux, avec au plus J_σ sauts,

alors, on peut supprimer l'hypothèse (**An**) et l'on a

$$\mathbb{E}\left[\ell\left(s^\star,\widetilde{s}\right)\right] \le K_2 R^{\frac{2d}{2\alpha+d}} n^{\frac{-2\alpha}{2\alpha+d}} \|\sigma\|_{L^2(\mathrm{Leb})}^{\frac{4\alpha}{2\alpha+d}} + \frac{K_4 A^2}{n^2} \tag{45}$$

$$\text{avec} \quad \|\sigma\|_{L^2(\mathrm{Leb})} := \left[\frac{1}{\mathrm{Leb}(\mathcal{X})}\int_{\mathcal{X}} \sigma^2(t)dt\right]^{1/2} > 0$$

et $K_4 = K_4(d, A, c_X^{\min}, R, \alpha, K_\sigma, J_\sigma) > 0$.

Les majorations du risque (44) et (45) montrent que l'estimateur $\widetilde{s}$ est adaptatif, au sens où :

– dans le cas homoscédastique, lorsque $s^\star \in \mathcal{H}(\alpha, R)$ avec $\alpha > 0$ (avec une définition appropriée lorsque $\alpha > 1$) et $\mathcal{X} \subset \mathbb{R}^d$, la vitesse minimax d'estimation de $s^\star$ pour $\|\cdot\|_{L^2}^2$ est

$$R^{\frac{2d}{2\alpha+d}} n^{\frac{-2\alpha}{2\alpha+d}} \sigma^{\frac{4\alpha}{2\alpha+d}} \ ,$$

à un facteur multiplicatif près ne dépendant pas de n, R et σ [49, 34, 52]. Ainsi, aucun estimateur ne peut atteindre une vitesse plus rapide que (44) (à une constante près) uniformément sur toutes les lois P telles que $s^\star \in \mathcal{H}(\alpha, R)$ avec $\alpha \in \ ]0, 1]$.

– dans le cas hétéroscédastique, la dépendance du risque minimax en $\sigma(\cdot)$ n'est connue que sous des hypothèses plus fortes. Lorsque $d = 1$ et $\alpha = 1$ et $\sigma(\cdot)$ est suffisamment régulière, la majoration (45) correspond aux minorations minimax à une constante multiplicative près indépendante de $\sigma(\cdot)$ [18, 26].

– l'estimateur $\widetilde{s}$ est construit sans utiliser la connaissance exacte de $R, \alpha, \sigma(\cdot)$, condition nécessaire pour pouvoir parler d'adaptation.

Notons que l'exposant de n n'est pas suffisant pour donner du sens à une majoration telle que (44) en pratique : en effet, à horizon n fixé, on ne peut jamais distinguer une fonction $s^\star \in \mathcal{C}^\infty$ d'une fonction très irrégulière, la valeur de la constante R pour laquelle $s^\star \in \mathcal{H}(\alpha, R)$ est tout aussi importante. La majoration (44) peut en effet être plus fine en choisissant α plus petit que le niveau de régularité de $s^\star$ si cela permet de réduire la constante R.

Insistons pour finir sur le fait que les pénalités par rééchantillonnage n'ont pas été construites dans le but spécifique de s'adapter à l'hétéroscédasticité. On peut notamment citer à ce sujet une discussion d'un article de Wu [51] par Efron :

> «The jackknife and bootstrap are general-purpose devices, not specifically adapted to take advantage of a special model like
>
> $$y = X\beta + e \ , \qquad \mathrm{var}(e) = \Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2) \ .$$
>
> Comparisons with specially adapted methods (...) are misleading if this is not made clear.»

En particulier, ici, il n'y a pas lieu de chercher à comparer l'estimateur $\widetilde{s}$ à des estimateurs *ad hoc*, construits avec cet objectif particulier [18, 26, 28, 29], et qui ont toutes les chances de se comporter un peu mieux que $\widetilde{s}$ pour l'analyse de données dont on sait qu'elles sont hétéroscédastiques. L'étude ci-dessus a essentiellement pour but d'illustrer la capacité naturelle du rééchantillonnage à s'adapter à des propriétés inconnues des données telles que l'hétéroscédasticité, y compris dans un cadre de sélection de modèles.

Il est donc assez tentant de conjecturer que les pénalités par rééchantillonnage s'adaptent effectivement à de nombreuses propriétés inconnues de la loi P des données pour de nombreux problèmes de sélection d'estimateurs. Les résultats mentionnés en Section 7 plaident également en ce sens.

6.4. **Comparaison des poids.** Nous renvoyons à [4, Sections 4 et 6.2] pour la comparaison des différents poids de rééchantillonnage.

7. Estimation de densité par moindres carrés

Les résultats présentés pour l'estimation de densité sont extraits de [40] pour le cas indépendant et de [36] pour le cas dépendant. Nous renvoyons à ces deux références pour plus d'informations, ainsi qu'à [35] dont l'introduction présente ces résultats de manière synthétique.

Références

[1] David J. Aldous. Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, volume 1117 of *Lecture Notes in Math.*, pages 1–198. Springer, Berlin, 1985.

[2] Sylvain Arlot. *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11, December 2007. `http://tel.archives-ouvertes.fr/tel-00198803/`.

[3] Sylvain Arlot. V-fold cross-validation improved : V-fold penalization, February 2008. arXiv :0802.0566v2.

[4] Sylvain Arlot. Model selection by resampling penalization. *Electron. J. Stat.*, 3 :557–624 (electronic), 2009.

[5] Sylvain Arlot. Choosing a penalty for model selection in heteroscedastic regression, June 2010. arXiv :0812.3141v2.

[6] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10 :245–279 (electronic), 2009.

[7] Philippe Barbe and Patrice Bertail. *The Weighted Bootstrap*, volume 98 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1995.

[8] Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48 :85–113, 2002.

[9] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4) :1497–1537, 2005.

[10] Rudolf Beran. The impact of the bootstrap on statistical algorithms and theory. *Statist. Sci.*, 18(2) :175–184, 2003. Silver anniversary of the bootstrap.

[11] Dennis D. Boos. Introduction to the bootstrap world. *Statist. Sci.*, 18(2) :168–174, 2003. Silver anniversary of the bootstrap.

[12] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification : a survey of some recent advances. *ESAIM Probab. Stat.*, 9 :323–375 (electronic), 2005.

[13] Leo Breiman. The little bootstrap and other methods for dimensionality selection in regression : X-fixed prediction error. *J. Amer. Statist. Assoc.*, 87(419) :738–754, 1992.

[14] Leo Breiman. Random forests. *Machine Learning*, 45 :5–32, 2001. 10.1023/A :1010933404324.

[15] Peter Bühlmann and Bin Yu. Analyzing bagging. *Ann. Statist.*, 30(4) :927–961, 2002.

[16] George Casella, editor. *Silver Anniversary of the Bootstrap*. Institute of Mathematical Statistics, Bethesda, MD, 2003. Statist. Sci. **18** (2003), no. 2.

[17] Thomas J. DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statist. Sci.*, 11(3) :189–228, 1996. With comments and a rejoinder by the authors.

[18] Sam Efromovich and Mark Pinsker. Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statist. Sinica*, 6(4) :925–942, 1996.

[19] Bradley Efron. Bootstrap methods : another look at the jackknife. *Ann. Statist.*, 7(1) :1–26, 1979.

[20] Bradley Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*, volume 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., 1982.

[21] Bradley Efron. Estimating the error rate of a prediction rule : improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382) :316–331, 1983.

[22] Bradley Efron. Second thoughts on the bootstrap. *Statist. Sci.*, 18(2) :135–140, 2003. Silver anniversary of the bootstrap.

[23] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York, 1993.

[24] Bradley Efron and Robert J. Tibshirani. Improvements on cross-validation : the .632+ bootstrap method. *J. Amer. Statist. Assoc.*, 92(438) :548–560, 1997.

[25] Magalie Fromont. Model selection by bootstrap penalization for classification. *Mach. Learn.*, 66(2–3) :165–207, 2007.

[26] Leonid Galtchouk and Sergey Pergamenshchikov. Adaptive asymptotically efficient estimation in heteroscedastic nonparametric regression via model selection, October 2008. arXiv :0810.1173v1.

[27] Yongchao Ge, Sandrine Dudoit, and Terence P. Speed. Resampling-based multiple testing for microarray data analysis. *Test*, 12(1) :1–77, 2003. With comments and a rejoinder by the authors.

[28] Xavier Gendre. Simultaneous estimation of the mean and the variance in heteroscedastic Gaussian regression. *Electron. J. Stat.*, 2 :1345–1372, 2008.

[29] Xavier Gendre. *Model selection in heteroscedastic regression*. PhD thesis, University Nice Sophia-Antipolis, June 2009. http ://tel.archives-ouvertes.fr/tel-00397608/.

[30] Peter Hall. Asymptotic properties of the bootstrap for heavy-tailed distributions. *Ann. Probab.*, 18(3) :1342–1360, 1990.

[31] Peter Hall. *The Bootstrap and Edgeworth Expansion*. Springer Series in Statistics. Springer-Verlag, New York, 1992.

[32] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6) :2593–2656, 2006.

[33] Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47 of *Progr. Probab.*, pages 443–457. Birkhäuser Boston, Boston, MA, 2000.

[34] A. P. Korostelëv and A. B. Tsybakov. *Minimax Theory of Image Reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1993.

[35] Matthieu Lerasle. *Rééchantillonnage et sélection de modèles optimale pour l'estimation de la densité de variables indépendantes ou mélangeantes*. PhD thesis, INSA de Toulouse, June 2009.

[36] Matthieu Lerasle. Optimal model selection for stationary data under various mixing conditions. arXiv :0911.1497, October 2010.

[37] Enno Mammen. *When Does Bootstrap Work ? Asymptotic Results and Simulations*, volume 77 of *Lecture Notes in Statistics*. Springer, 1992.

[38] David M. Mason and Michael A. Newton. A rank statistics approach to the consistency of a general bootstrap. *Ann. Statist.*, 20(3) :1611–1624, 1992.

[39] Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[40] Matthieu Lerasle. Optimal model selection in density estimation. arXiv :0910.1654v2, 2009.

[41] Annette M. Molinaro, Richard Simon, and Ruth M. Pfeiffer. Prediction error estimation : a comparison of resampling methods. *Bioinformatics*, 21(15) :3301–3307, 2005.

[42] Dimitris N. Politis. The impact of bootstrap methods on time series analysis. *Statist. Sci.*, 18(2) :219–230, 2003. Silver anniversary of the bootstrap.

[43] Dimitris N. Politis, Joseph P. Romano, and Michael Wolf. *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York, 1999.

[44] Jens Præstgaard and Jon A. Wellner. Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, 21(4) :2053–2086, 1993.

[45] Jun Shao. Bootstrap model selection. *J. Amer. Statist. Assoc.*, 91(434) :655–665, 1996.

[46] Jun Shao and Dong Sheng Tu. *The Jackknife and Bootstrap*. Springer Series in Statistics. Springer-Verlag, New York, 1995.

[47] Ritei Shibata. Bootstrap estimate of Kullback-Leibler information for model selection. *Statist. Sinica*, 7(2) :375–394, 1997.

[48] Kesar Singh. On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.*, 9(6) :1187–1195, 1981.

[49] Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8(6) :1348–1360, 1980.

[50] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[51] Chien-Fu Jeff Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, 14(4) :1261–1350, 1986. With discussion and a rejoinder by the author.

[52] Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5) :1564–1599, 1999.

[53] George Alastair Young. Bootstrap : more than a stab in the dark ? *Statist. Sci.*, 9(3) :382–415, 1994. With discussion and a rejoinder by the author.

Cnrs – Équipe Sierra, Laboratoire d'Informatique de l'École Normale Supérieure, (CNRS/ENS/INRIA UMR 8548), INRIA - 23 avenue d'Italie - CS 81321, 75214 PARIS Cedex 13 - France

E-mail address: sylvain.arlotRETIRERCECI@ens.fr

URL: http://www.di.ens.fr/~arlot/