Agrégation PAC-bayésienne d'estimateurs par projection

Lucie Montuelle [1] & Erwan Le Pennec [2]

[1] *Select, Inria Saclay Idf / LMO, Université Paris-Sud*
E-mail:lucie.montuelle@math.u-psud.fr
[2] *Select, Inria Saclay Idf / CMAP, École Polytechnique*
E-mail:erwan.le_pennec@inria.fr

Résumé. L'agrégation d'estimateurs à l'aide de poids exponentiels dépendant de leur risque offre de bonnes performances en moyenne. Malheureusement, il est impossible d'obtenir un aussi bon contrôle du risque de l'estimateur agrégé en probabilité. Pour contourner ce problème, nous considérons des poids exponentiels du risque pénalisé. Cette technique permet d'obtenir une inégalité oracle inexacte en probabilité. En surpénalisant, avec une prise en compte de la norme de la fonction estimée, une inégalité exacte est accessible.

Mots-clés. Agrégation à poids exponentiels, régression, inégalité oracle

Abstract. Aggregating estimators using exponential weights depending on their risk performs well in expectation, but sadly not in probability. A way to overcome this issue is considering exponential weights of a penalized risk. In this case, an oracle inequality can be obtained in probability, but is not sharp. Taking into account the estimated function's norm in the penalty offers a sharp inequality.

Keywords. exponentially weighted aggregation, regression, oracle inequality

1 Cadre de travail

Nous considérons le modèle de régression

$$Y_i = f(x_i) + W_i, \quad i = 1, \dots, n,$$

où le design $(x_i)_{1 \leq i \leq n}$ est fixe, la fonction réelle f est inconnue et les variables aléatoires de bruit W_i sont centrées, indépendantes, de loi normale de variance σ^2 connue. Notre but est d'estimer la fonction f à partir de l'observation des $(x_i, Y_i)_{1 \leq i \leq n}$. Pour cela, nous disposons d'une collection d'estimateurs par projection $\{\hat{f}_J\}_J = \{P_J Y\}_J$, où J parcourt l'ensemble des parties de $\{1, \dots, n\}$. Nous allons construire à l'aide de ce dictionnaire un nouvel estimateur, appelé estimateur agrégé, qui mimera les performances du meilleur estimateur de la collection.

L'estimateur agrégé est la moyenne des estimateurs de la collection par rapport à une mesure bien choisie. Nous nous concentrons sur l'agrégation à l'aide de poids exponentiels. Classiquement, les poids sont de la forme $\exp(-r_J/\lambda)$, où r_J désigne un estimateur du risque de $\hat{f}_J$ et λ est un paramètre à calibrer, appelé température. L'idée est de favoriser les estimateurs dont le risque est faible. Lorsque la température est grande, l'exponentielle tend vers 1, donc les poids deviennent uniformes. Aucun estimateur n'est privilégié. Au contraire, lorsque la température tend vers 0, le seul estimateur retenu lors de l'agrégation est celui qui a le plus petit risque. Il est donc crucial de bien choisir ce paramètre.

Cette procédure a montré son efficacité, notamment en offrant un risque plus faible que la sélection de modèle, puisque nous parions sur plusieurs estimateurs. Leung et Barron (2006) ont déjà étudié l'agrégation de projections et montré qu'en moyenne l'estimateur agrégé à des performances similaires à la meilleure projection de la collection, via une inégalité oracle exacte. Cependant, Dai et al (2012) ont montré la sous-optimalité des poids exponentiels en déviation, qui ne permettent pas d'obtenir en probabilité une inégalité oracle exacte. Néanmoins, en pénalisant l'estimateur du risque dans les poids, Dai et al (2013) ont obtenu borne sur le risque de l'estimateur agrégé, à condition de prendre le paramètre d'apprentissage des poids, appelé température, supérieur à 20 fois la variance du bruit. Cette inégalité est dite inexacte car la constante devant le risque du meilleur estimateur est plus grande que 1.

Nous allons montrer qu'en modifiant un peu les poids, une température plus basse peut être utilisée. Si l'on consent à prendre en compte dans une certaine mesure le rapport signal sur bruit, il est possible d'avoir une inégalité exacte ou de s'en approcher. A température et pénalité fixées, il peut-être intéressant de prendre en compte une partie de ce ratio.

2 Notations et résultat

Tout d'abord, introduisons quelques notations afin de définir le risque considéré et les poids employés. Pour toute fonction réelle g, notons

$$\|g\|_n = \left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)^2\right)^{1/2}.$$

Pour chaque sous-ensemble d'indices J de $\{1,\dots,n\}$, l'estimateur par projection $\hat{f}_J$ est défini par

$$\hat{f}_J = P_J Y.$$

La i-ième composante de $\hat{f}_J$ vaut Y_i si i est dans J et 0 sinon.

Pour évaluer les performances d'un estimateur, nous considérons son risque empirique, défini comme l'estimateur sans biais du risque de Stein

$$r_J = \|Y - \hat{f}_J\|_n^2 + \frac{2}{n}|J|\sigma^2 - \sigma^2,$$

et son risque intégré

$$\mathbb{E}\|f - \hat{f}_J\|_n^2 = \|f - P_J f\|_n^2 + \frac{\sigma^2}{n}|J| = \mathbb{E}[r_J].$$

Soit $\bar{J} \in \arg\min_J \mathbb{E}\|f - \hat{f}_J\|_n^2$, un minimiseur du risque intégré. Dans les poids d'agrégation, nous considérons un risque pénalisé. Pour tous $p \geq 1, \delta \in (0,1)$ et $\lambda > \frac{4p\sigma^2}{n(p-1)}$, définissons la pénalité

$$pen(J) = \frac{2p\sigma^2|J|}{n}\left[\frac{2\sigma^2}{\lambda n(p-1) - 4p\sigma^2} + (1-\delta)\frac{\|f\|_\infty^2}{\lambda n}\right]$$

Pour toute mesure de probabilité π sur l'ensemble $\mathcal{P}(\{1,\dots,n\})$ des parties de $\{1,\dots,n\}$, l'estimateur agrégé à poids exponentiels est défini par $f_{EWA} = \int \hat{f}_J \rho(dJ)$, avec

$$\rho(dJ) = \frac{\exp\left(-\frac{1}{\lambda}\left[r_J + pen(J)\right]\right)}{\int \exp\left(-\frac{1}{\lambda}\left[r_J + pen(J)\right]\right)\pi(dJ)}\pi(dJ).$$

Théorème 1. *Pour tout $p \geq 1, \delta \in [0,1]$ l'estimateur agrégé f_{EWA} défini ci-dessus avec $\lambda > 2\sigma^2 \frac{p}{n}\max\left(\delta, \frac{2}{p-1}\right)$, satisfait pour tout $\eta \in (0,1)$, avec probabilité au moins $1-\eta$,*

$$\|f - f_{EWA}\|_n^2 \leq \inf_\rho \left(1 + \frac{4p\sigma^2\delta}{\lambda n - 2p\sigma^2\delta}\right)\int \|f - \hat{f}_J\|_n^2 \rho(dJ) + \frac{4p\sigma^2\delta}{\lambda n - 2p\sigma^2\delta}\|f - \hat{f}_{\bar{J}}\|_n^2$$
$$+ 4p\sigma^2\left[\frac{2\lambda\sigma^2}{(\lambda n(p-1) - 4p\sigma^2)(\lambda n - 2p\sigma^2\delta)} + (1-\delta)\frac{\|f\|_\infty^2}{n(\lambda n - 2p\sigma^2\delta)}\right]\left(\int |J|\rho(dJ) + |\bar{J}|\right)$$
$$+ \frac{2n\lambda^2}{\lambda n - 2p\sigma^2\delta}\left(KL(\rho,\pi) + \ln\frac{2}{\eta}\right).$$

En particulier

$$\|f - f_{EWA}\|_n^2 \leq \left(1 + \frac{4p\sigma^2\delta}{\lambda n - 2p\sigma^2\delta}\right)\|f - \hat{f}_{\bar{J}}\|_n^2 + 4p\sigma^2\left[\frac{2\lambda\sigma^2}{(\lambda n(p-1) - 4p\sigma^2)(\lambda n - 2p\sigma^2\delta)}\right.$$
$$\left. + (1-\delta)\frac{\|f\|_\infty^2}{n(\lambda n - 2p\sigma^2\delta)}\right]|\bar{J}| + \frac{2n\lambda^2}{\lambda n - 2p\sigma^2\delta}\left(\ln\frac{1}{\pi(\bar{J})} + \ln\frac{2}{\eta}\right).$$

Si $\delta = 1$, le théorème annonce que pour toute température supérieure à $6\sigma^2/n$, il existe une pénalité indépendante de la norme de f qui permet d'obtenir une inégalité oracle inexacte (avec constante supérieure à 1). Dai (2013) obtient le même type de résultat pour des températures supérieures à $20\sigma^2/n$ en pénalisant r_J par $2\sigma^2|J|/n$.

Corollaire 1. *Pour tout $p \geq 1$, l'estimateur agrégé f_{EWA} défini ci-dessus avec $\lambda > 2\sigma^2 \frac{p}{n} \max\left(1, \frac{2}{p-1}\right)$ et $pen(J) = \frac{2p\sigma^2|J|}{n} \frac{2\sigma^2}{\lambda n(p-1)-4p\sigma^2}$, satisfait pour tout $\eta \in (0,1)$, avec probabilité au moins $1-\eta$,*

$$\|f - f_{EWA}\|_n^2 \leq \inf_{\rho} \left(1 + \frac{4p\sigma^2}{\lambda n - 2p\sigma^2}\right) \int \|f - \hat{f}_J\|_n^2 \rho(dJ) + \frac{4p\sigma^2}{\lambda n - 2p\sigma^2} \|f - \hat{f}_{\bar{J}}\|_n^2$$
$$+ \frac{8p\lambda\sigma^4}{(\lambda n(p-1) - 4p\sigma^2)(\lambda n - 2p\sigma^2)} \left(\int |J| \rho(dJ) + |\bar{J}|\right) + \frac{2n\lambda^2}{\lambda n - 2p\sigma^2} \left(KL(\rho, \pi) + \ln \frac{2}{\eta}\right).$$

Dans le cas où δ vaut 0, pour toute température supérieure à $4\sigma^2/n$, il existe une pénalité prenant en compte la norme de f, qui assure une inégalité oracle exacte (cf Catoni (2004)).

Corollaire 2. *Pour tout $p \geq 1$, l'estimateur agrégé f_{EWA} défini ci-dessus avec $\lambda > \frac{4p\sigma^2}{n(p-1)}$ et $pen(J) = \frac{2p\sigma^2|J|}{n} \left[\frac{2\sigma^2}{\lambda n(p-1)-4p\sigma^2} + \frac{\|f\|_\infty^2}{\lambda n}\right]$, satisfait pour tout $\eta \in (0,1)$, avec probabilité au moins $1-\eta$,*

$$\|f - f_{EWA}\|_n^2 \leq \inf_{\rho} \int \|f - \hat{f}_J\|_n^2 \rho(dJ) + \frac{4p\sigma^2}{n} \left[\frac{2\sigma^2}{\lambda n(p-1) - 4p\sigma^2} + \frac{\|f\|_\infty^2}{\lambda n}\right] \left(\int |J| \rho(dJ) + |\bar{J}|\right)$$
$$+ 2\lambda \left(KL(\rho, \pi) + \ln \frac{2}{\eta}\right).$$

Selon la valeur de la norme de f, il peut être plus intéressant de faire peser son influence sur le biais ou la variance dans l'inégalité oracle, afin d'obtenir la borne la plus précise possible. En pratique, une température $\lambda = 4\sigma^2/(n\theta)$, avec $\theta \in (0,1)$ et une pénalité de la forme $\gamma\sigma^2|J|$ sont fixées. Le théorème s'énonce alors comme suit:

Corollaire 3. *Pour tout $p \geq 1$ et $\delta \in [0,1]$, tels que $\theta < \min\left(\frac{2}{p}, \frac{p-1}{p}\right)$ et $\gamma \geq \frac{2p}{n} \left[\frac{2\sigma^2}{\lambda n(p-1)-4p\sigma^2} + (1-\delta)\frac{\|f\|_\infty^2}{\lambda n}\right]$, alors l'estimateur agrégé défini par λ et γ vérifie l'inégalité oracle énoncée dans le théorème, avec probabilité au moins $1-\eta$ pour tout $\eta \in (0,1)$, pour ces valeurs de p et δ.*

S'il y en a plusieurs, en choisissant la plus petite valeur de δ, l'inégalité peut être rendue la plus exacte possible. Inversement, à température et constante de l'inégalité oracle données, il existe une pénalité minimale.

Bibliographie

[1] Catoni, O., (2004), *Statistical Learning Theory and Stochastic Optimization, Ecole d'Été de Probabilités de Saint-Flour XXXI*, Springer-Verlag Berlin/Heidelberg.

[2] Dai, D., Rigollet, P. et Zhang, T, (2012), *Deviation optimal learning using greedy Q-aggregation*, The Annals of Statistics, 40,3, 1878–1905.
[3] Dai, D., Rigollet, P. , Xia L. et Zhang T.,(2013) *Aggregation of Affine Estimators*, preprint.
[4] Leung, G. et Barron, A. R., (2006) *Information theory and mixing least-squares regressions*, Institute of Electrical and Electronics Engineers. Transactions on Information Theory, 52, 8, 3396–3410.