Agrégation à poids exponentiels et estimation penalisée : Inégalités oracles

Duy Tung Luu, Jalal M. Fadili, Christophe Chesneau

Agrégation à poids exponentiels et estimation penalisée : Inégalités oracles

Duy Tung LUU[1], Jalal FADILI[1], Christophe CHESNEAU[2]
[1] Normandie Univ, ENSICAEN, CNRS, GREYC, France [2] Normandie Univ, UNICAEN, CNRS, LMNO, France
duy-tung.luu@ensicaen.fr,Jalal.Fadili@ensicaen.fr, Christophe.Chesneau@unicaen.fr

Résumé – Un problème classique en traitement du signal et des images vise à estimer un signal/image à partir de ses mesures linéaires sous-déterminées et bruitées. Le problème étant mal-posé, l'approche usuelle consiste à imposer des a priori sur les objets recherchés. Dans cet article, nous présentons une analyse unifiée des garanties théoriques de performance de deux familles d'estimateurs: les estimateurs par agrégation à poids exponentiels et les estimateurs pénalisés pour une classe d'a priori favorisant une certaine notion de simplicité/faible complexité. Plus précisément, nous montrons que les deux classes d'estimateurs satisfont des inégalités oracles fines en probabilité lorsque le bruit est gaussien ou sous-gaussien. Ces résultats sont ensuite appliqués à plusieurs pénalités classiques notamment les norme ℓ_1 (LASSO), ℓ_1–ℓ_2 (LASSO par groupes et sa version analyse), ℓ_∞ (anti-parcimonie), et la norme nucléaire.

Abstract – A classical problem in signal and image processing aims at recovering a signal/image from a set of underdetermined and noisy linear measurements. The problem is generally ill-posed and the usual approach consists in imposing prior knowledge on the set of sought-after objects. In this paper, we present an unified analysis of the theoretical performance guarantees of two classes of estimators: exponential weighted aggregation and penalized estimators for a general class of priors which promote objects complying with some notion of simplicity/low complexity. More precisely, we show that these two estimators satisfy sharp oracle inequalities in probability when the noise is Gaussian or subgaussian. These results are then applied to several popular penalties including the ℓ_1 (LASSO), the ℓ_1–ℓ_2 (group LASSO and its analysis version), ℓ_∞ (anti-sparsity), and the nuclear norms.

1 Introduction

1.1 Formulation du problème

Soient $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n)$ n observations identiquement distribuées de distribution marginale $\mathbb{P}$, et $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ joue le rôle de l'opérateur de dégradation dans un problème de restauration en traitement du signal/image, ou encore la matrice des covariables pour un problème de régression en statistique. Le but est d'estimer un paramètre $\boldsymbol{\theta} \in \mathbb{R}^p$ de cette distribution connaissant $\boldsymbol{y}$ et $\boldsymbol{X}$. Soit $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ une fonction de perte générale supposée convex et différentiable. Soit $\boldsymbol{\theta}_0 \in \operatorname{Argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathbb{E}\left[F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y})\right]$ un minimiseur du risque. Une instance usuelle de ce cadre est celui du modèle linéaire

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_0 + \boldsymbol{\xi}, \tag{1}$$

où $\boldsymbol{\xi}$ est centré, et où $F(\boldsymbol{u}, \boldsymbol{y}) = \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{u}\right\|_2^2$.

Notre objectif est de construire des estimateurs jouissant de garanties (non-asymptotiques) vis-à-vis du vecteur $\boldsymbol{\theta}_0$. Pour cela, nous tirons profit du fait que, même si la dimension ambiante p de $\boldsymbol{\theta}_0$ est grande, ce dernier présente une structure simple se traduisant par le fait qu'il vit dans un sous-ensemble de $\mathbb{R}^p$ qui est à la fois structuré est de faible dimension. Ceci correspond à ce qu'on appelle ici une notion de simplicité ou de faible complexité (au sens d'une faible dimension intrinsèque). On peut alors imposer ce type de structure simple par l'introduction d'a priori la favorisant. Pour ce faire, deux types d'approches seront considérées ici.

Estimateurs pénalisés Ces estimateurs nécessitent de résoudre le problème d'optimisation suivant

$$\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} \in \operatorname*{Argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ V_n(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \frac{1}{n} F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) + \lambda J(\boldsymbol{\theta}) \right\}, \tag{2}$$

où $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ est une fonction de perte générale supposée différentiable, J est une pénalité (ou une régularisation) favorisant justement la notion spécifique de faible complexité que l'on recherche, et $\lambda > 0$ est le paramètre de régularisation. Il faut noter que bien que (2) puisse se prêter parfois à une interprétation bayésienne MAP, ce n'est pas la seule possible.

Agrégation à poids exponentiels (EWA) L'EWA (signifiant *Exponential Weighted Aggregation* en Anglais) consiste à remplacer le problème de minimisation dans (2) par l'espérance associée à la mesure de probabilité liée à V_n. L'agrégat est alors défini par

$$\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} = \int_{\mathbb{R}^p} \boldsymbol{\theta}\mu_n(\boldsymbol{\theta}) d\boldsymbol{\theta}, \;\; \mu_n(\boldsymbol{\theta}) \propto \exp\left(-V_n(\boldsymbol{\theta})/\beta\right), \tag{3}$$

où $\beta > 0$ est appelé paramètre de température. Encore une fois, (3) peut être vu comme une espérance conditionnelle a posteriori, mais uniquement dans certains cas.

Inégalités oracles Les inégalités oracles quantifient la qualité d'un estimateur par rapport au meilleur estimateur théorique possible. Dans nos inégalités, la qualité d'un estimateur $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^p$ de $\boldsymbol{\theta}_0$ est quantifiée par une perte $R_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$, qui est une mesure d'erreur entre les prédictions $\boldsymbol{X}\widehat{\boldsymbol{\theta}}$ et $\boldsymbol{X}\boldsymbol{\theta}_0$. On

cherche à prouver que $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ et $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ imitent autant que possible la performance du meilleur candidat dans la famille $\Theta \subseteq \mathbb{R}^p$. Cette idée se traduit par le type d'inégalités suivant (indiqué ici pour l'EWA)

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\big) \leq C \inf_{\boldsymbol{\theta}\in\Theta} \big(R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \Delta_{n,p,\lambda,\beta}(\boldsymbol{\theta})\big),$$

où $C \geq 1$ est une constante absolue et $\Delta_{n,p,\lambda,\beta}(\boldsymbol{\theta})$ est le reste qui dépend de la performance de l'estimateur, la complexité de $\boldsymbol{\theta}$, la taille de l'échantillon n, la dimension p, et les paramètres de régularisation et de température (λ, β). Un estimateur avec de bonnes propriétés oracles correspondrait à C proche de 1 (idéalement $C = 1$, l'inégalité est dite *fine*), et $\Delta_{n,p,\lambda,\beta}(\boldsymbol{\theta})$ petit, avec une convergence rapide vers 0 lorsque $n \to +\infty$.

1.2 Contributions

Les contributions principales de ce travail sont :

- Nous établissons tout d'abord des inégalités oracles fines déterministes pour les estimateurs $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ et $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ avec une perte F générale dans un cadre unifié.
- Pour le modèle (1) avec un bruit gaussien ou sous-gaussien, et F quadratique, nous établissons des inégalités oracles fines en probabilité avec un reste optimal.
- Nous illustrons ensuite ces inégalités sur plusieurs pénalités notamment celles favorisant la parcimonie et un faible rang. Nous retrouvons ainsi comme cas particuliers des résultats de la littérature et établissons de nouveaux.

Toutes les preuves des résultats énoncés peuvent être trouvées dans la version longue [5].

1.3 Notations

Vecteurs et matrices Soient $\langle ., . \rangle$ le produit scalaire euclidien, et $\|.\|_r$, pour $r \geq 1$, la norme ℓ_r d'un vecteur avec l'adaptation habituelle pour $r = +\infty$. Soit $\mathbf{I}_d$ la matrice identité sur $\mathbb{R}^d$. P_T désigne le projecteur orthogonal sur le sous-espace vectoriel T. On dénote aussi $\boldsymbol{\theta}_T = P_T\boldsymbol{\theta}$ et $\boldsymbol{X}_T = \boldsymbol{X}P_T$. Pour $p \in \mathbb{N}$, $[p] = \{1, \ldots, p\}$. Pour $I \subset [p]$, $\boldsymbol{\theta}_I$ est le sous-vecteur de $\boldsymbol{\theta} \in \mathbb{R}^p$ dont les entrées sont restreintes aux indices dans I, et $\boldsymbol{X}_I$ la sous-matrice dont les colonnes sont celles de $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ indexées par I.

Ensembles Pour un ensemble fini $\mathcal{C}$, $|\mathcal{C}|$ désigne sa cardinalité. Pour un ensemble convexe non vide $\mathcal{C}$, son *enveloppe affine* $\mathrm{aff}(\mathcal{C})$ est le plus petit sous-espace affine le contenant. Elle est une translation de son *espace parallèle* $\mathrm{par}(\mathcal{C})$, i.e. $\mathrm{par}(\mathcal{C}) = \mathrm{aff}(\mathcal{C}) - \boldsymbol{\theta} = \mathbb{R}(\mathcal{C} - \mathcal{C})$, pour tout $\boldsymbol{\theta} \in \mathcal{C}$. Soit $\mathcal{C}$ un ensemble convexe non vide, l'ensemble $\mathcal{C}^\circ = \big\{\boldsymbol{\eta} \in \mathbb{R}^p \ : \ \langle \boldsymbol{\eta}, \boldsymbol{\theta}\rangle \leq 1, \ \forall \boldsymbol{\theta} \in \mathcal{C}\big\}$ est appelé la *polaire* de $\mathcal{C}$.

Fonctions Une fonction $f : \mathbb{R}^p \to \mathbb{R}$ est coercive, si $\lim_{\|\boldsymbol{\theta}\|_2\to+\infty} f(\boldsymbol{\theta}) = +\infty$. Une fonction f est propre si $f(\boldsymbol{\theta}) > -\infty$ pour tout $\boldsymbol{\theta}$ et ne vaut pas l'infini partout. Une fonction est sous-linéaire si elle est convexe et 1-positivement homogène. Pour une fonction f propre et convexe, $\partial f(\boldsymbol{\theta})$ est sa sous-différentielle en $\boldsymbol{\theta}$. Lorsque $f \in C^1(\mathbb{R}^p)$, sa sous-différentielle se réduit à son gradient $\nabla f(\boldsymbol{\theta})$.

Jauges et fonctions d'appui Soit $\mathcal{C} \subseteq \mathbb{R}^p$ un ensemble convexe fermé non vide contenant l'origine. La *jauge* de $\mathcal{C}$ est la fonction $\gamma_{\mathcal{C}}$ définie sur $\mathbb{R}^p$ par $\gamma_{\mathcal{C}}(\boldsymbol{\theta}) = \inf\big\{\lambda > 0 \ : \ \boldsymbol{\theta} \in \lambda\mathcal{C}\big\}$. La polaire d'une jauge $\gamma_{\mathcal{C}}$ est la fonction $\gamma_{\mathcal{C}}^\circ$ définie par $\gamma_{\mathcal{C}}^\circ(\boldsymbol{\omega}) = \inf\big\{\mu \geq 0 \ : \ \langle \boldsymbol{\theta}, \boldsymbol{\omega}\rangle \leq \mu\gamma_{\mathcal{C}}(\boldsymbol{\theta}), \forall \boldsymbol{\theta}\big\}$. La *fonction d'appui* de $\mathcal{C} \subset \mathbb{R}^p$ est $\sigma_{\mathcal{C}}(\boldsymbol{\omega}) = \sup_{\boldsymbol{\theta}\in\mathcal{C}} \langle \boldsymbol{\omega}, \boldsymbol{\theta}\rangle$. Lorsque $\mathcal{C} \subseteq \mathbb{R}^p$ est un ensemble convexe fermé contenant l'origine, on a $\gamma_{\mathcal{C}}^\circ = \gamma_{\mathcal{C}^\circ} = \sigma_{\mathcal{C}}$.

Soient deux jauges à valeurs finies et coercives $J_1 = \gamma_{\mathcal{C}_1}$ et $J_2 = \gamma_{\mathcal{C}_2}$. On définit $|||\boldsymbol{A}|||_{J_1\to J_2}$ la *borne d'opérateur* par

$$|||\boldsymbol{A}|||_{J_1\to J_2} = \sup_{\boldsymbol{\theta}\in\mathcal{C}_1} J_2(\boldsymbol{A}\boldsymbol{\theta}).$$

Lorsque J_1 et J_2 sont des normes, on retrouve une norme matricielle subordonnée. Pour alléger la notation lorsque J_i est une norme, on écrira l'indice de la norme au lieu de J_i (par exemple r pour ℓ_r, $*$ pour la norme nucléaire).

Sous-espace modèle Soit $\boldsymbol{\theta} \in \mathbb{R}^p$. On pose $e_{\boldsymbol{\theta}} = \mathrm{P}_{\mathrm{aff}(\partial J(\boldsymbol{\theta}))}(0)$, $S_{\boldsymbol{\theta}} = \mathrm{par}(\partial J(\boldsymbol{\theta}))$ et $T_{\boldsymbol{\theta}} = S_{\boldsymbol{\theta}}^\perp$. $T_{\boldsymbol{\theta}}$ est appelé le *sous-espace modèle* de $\boldsymbol{\theta}$ associé à J.

2 Hypothèses et Préliminaires

La classe des fonctions F que nous considérons obéit aux hypothèses suivantes :

($\mathbf{H}.1$) Régularité : $F(\cdot, \boldsymbol{y}) \in C^1(\mathbb{R}^n)$ et fortement convexe de module $\nu > 0$ uniformément sur $\boldsymbol{y}$.

($\mathbf{H}.2$) Moment : Pour tout $\overline{\boldsymbol{\theta}} \in \mathbb{R}^p$, $\int_{\mathbb{R}^p} \exp\left(-F(\boldsymbol{X}\boldsymbol{\theta})/\beta\right) \big|\big\langle \nabla F(\boldsymbol{X}\boldsymbol{\theta}), \boldsymbol{X}(\overline{\boldsymbol{\theta}} - \boldsymbol{\theta})\big\rangle\big| d\boldsymbol{\theta} < +\infty$.

C'est une classe assez générale en lien notamment avec l'anti log-vraisemblance de la famille exponentielle régulière. Elle couvre en particulier toute perte quadratique. Concernant la pénalité J, nous supposons que :

($\mathbf{H}.3$) $J : \mathbb{R}^p \to \mathbb{R}$ est la jauge d'un ensemble compact convexe non vide contenant l'origine comme point intérieur.

Cette hypothèse équivaut à dire que $J \overset{\text{def}}{=} \gamma_{\mathcal{C}}$ est propre, convexe, 1-positivement homogène, coercive et à valeurs finies.

Nos inégalités oracles feront intervenir, ce qui est assez naturel d'ailleurs, une mesure du *conditionnement* de $\boldsymbol{X}$ quand elle est restreinte à un sous-espace modèle T. Pour $c > 0$, on introduit alors le coefficient suivant, dit de compatibilité,

$$\Upsilon(T, c) = \inf_{\left\{\boldsymbol{\omega}\in\mathbb{R}^p \ : \ J(\boldsymbol{\omega}_S) < cJ(\boldsymbol{\omega}_T)\right\}} \frac{|||\mathrm{P}_T|||_{2\to J} \big\|\boldsymbol{X}\boldsymbol{\omega}\big\|_2}{n^{1/2}(J(\boldsymbol{\omega}_T) - J(\boldsymbol{\omega}_S)/c)}.$$

En particulier, on peut voir que $\Upsilon(T, c)$ est plus grand que la plus petite valeur singulière de $\boldsymbol{X}_T$.

3 Inégalités Oracles Fines (IOF)

3.1 IOF déterministe avec perte générale

On est maintenant en mesure d'énoncer nos inégalités oracles. Celles-ci le seront en termes de la divergence de Bregman associée à F, i.e.

$$R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \frac{F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y}) - F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) - \langle \nabla F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}), \boldsymbol{X}(\boldsymbol{\theta}_0 - \boldsymbol{\theta})\rangle}{n}.$$

Théorème 3.1. *Considérons l'estimateur $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ (resp. $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$) où F et $J \overset{\text{def}}{=} \gamma_{\mathcal{C}}$ satisfont* $(\mathbf{H}.\mathbf{1})$-$(\mathbf{H}.\mathbf{2})$ *et* $(\mathbf{H}.\mathbf{3})$ *(resp.* $(\mathbf{H}.\mathbf{1})$ *et* $(\mathbf{H}.\mathbf{3})$*). Alors, $\forall \tau > 1$ tel que $\lambda \geq \tau J^\circ\big(-\boldsymbol{X}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, y)\big)/n$, on a*

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\big) \leq \Psi(J, \lambda, \tau, \nu, \boldsymbol{\theta}) + p\beta, \tag{4}$$

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}, \boldsymbol{\theta}_0\big) \leq \Psi(J, \lambda, \tau, \nu, \boldsymbol{\theta}), \tag{5}$$

où $\Psi(J, \lambda, \tau, \nu, \boldsymbol{\theta}) =$

$$\inf_{\boldsymbol{\theta} \in \mathbb{R}^p} \left(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \frac{\lambda^2 \big(\tau J^\circ(e_{\boldsymbol{\theta}}) + 1\big)^2 \|\mathrm{P}_{T_{\boldsymbol{\theta}}}\|_{2\to J}^2}{2\tau^2 \nu \Upsilon\Big(T_{\boldsymbol{\theta}}, \frac{\tau J^\circ(e_{\boldsymbol{\theta}})+1}{\tau-1}\Big)^2} \right).$$

Ce résultats mérite quelques remarques. Tout d'abord, les inégalités oracles sont fines. Le reste dans Ψ, exprime la complexité du modèle (comme nous l'illustrerons dans les exemples par la suite). La différence entre la performance de prédiction de $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ et $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ est le terme $p\beta$ qui capture l'influence du paramètre de température dans les agrégateurs. Prenant β suffisamment petit de l'ordre $O\big((pn)^{-1}\big)$, ce terme devient $O(n^{-1})$.

3.2 IOF en probabilité avec perte quadratique

Considérons dorénavant le modèle linéaire (1) et $F(\boldsymbol{u}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{u}\|_2^2$. Dans ce cas $R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\|_2^2$. On fera intervenir l'épaisseur gaussienne d'un ensemble $\mathcal{S}$: $w(\mathcal{S}) \overset{\text{def}}{=} \mathbb{E}\,[\sigma_{\mathcal{S}}(\boldsymbol{g})]$, où $\boldsymbol{g} \sim \mathcal{N}(0, \mathbf{I}_n)$.

Théorème 3.2. *Soit le modèle* (1) *où $\boldsymbol{\xi}$ est un vecteur sous-gaussien centré de paramètre σ. Supposons que $J \overset{\text{def}}{=} \gamma_{\mathcal{C}}$ satisfait* $(\mathbf{H}.\mathbf{3})$*. Supposons que $\lambda \geq \frac{\tau\sigma c_1 \sqrt{2\log(c_2/\delta) w(\boldsymbol{X}(\mathcal{C}))}}{n}$, pour $\tau > 1$ et $0 < \delta < \min(c_2, 1)$, où c_1 et c_2 sont des constantes absolues positives. Alors* (4) *et* (5) *sont satisfaites avec une probabilité d'au moins $1 - \delta$.*

Lorsque la pénalité J est la jauge d'un polytope, le résultat est encore plus fin.

Théorème 3.3. *Soit le modèle* (1) *où $\boldsymbol{\xi}$ est un vecteur sous-gaussien centré de paramètre σ. Supposons que $J \overset{\text{def}}{=} \gamma_{\mathcal{C}}$ avec $\mathcal{C}$ un polytope de sommets $\mathcal{V}$. Supposons que $\max_{\boldsymbol{v}\in\mathcal{V}} \|\boldsymbol{X}\boldsymbol{v}\|_2 \leq \sqrt{n}$ et $\max_{\boldsymbol{v}\in\mathcal{V}} \|\boldsymbol{X}\boldsymbol{v}\|_2 \leq \sqrt{n}$, et $\lambda \geq \tau\sigma\sqrt{2\delta\log(|\mathcal{V}|)/n}$, pour $\tau > 1$ et $\delta > 1$. Alors* (4) *et* (5) *sont satisfaites avec une probabilité d'au moins $1 - |\mathcal{V}|^{1-\delta}$.*

4 Applications

Illustrons maintenant les Théorèmes 3.2 et 3.3 pour quelques pénalités utilisées dans la littérature.

LASSO La pénalité LASSO est utilisée pour favoriser la parcimonie

$$J(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1 = \textstyle\sum_{i=1}^p |\boldsymbol{\theta}_i|. \tag{6}$$

Soient $(\boldsymbol{a}_i)_{1\leq i\leq p}$ la base canonique de $\mathbb{R}^p$ et $\mathrm{supp}(\boldsymbol{\theta}) \overset{\text{def}}{=} \big\{i \in [p] \ : \ \boldsymbol{\theta}_i \neq 0\big\}$. Pour la pénalité LASSO, Théorème 3.3 se spécialise ainsi.

Corollaire 4.1. *Soit le modèle* (1) *où $\boldsymbol{\xi}$ est un vecteur sous-gaussien centré de paramètre σ, et $\boldsymbol{X}$ telle que $\max_i \|\boldsymbol{X}_i\|_2 \leq \sqrt{n}$. Supposons que $\lambda \geq \tau\sigma\sqrt{2\delta\log(2p)/n}$, pour $\tau > 1$ et $\delta > 1$. Alors,* (4) *et* (5) *sont satisfaites avec une probabilité d'au moins $1 - (2p)^{1-\delta}$ avec $\Psi(J, \lambda, \tau, \nu, \boldsymbol{\theta})$ égal à*

$$\inf_{\substack{I\subset[p] \\ \boldsymbol{\theta}:\ \mathrm{supp}(\boldsymbol{\theta})=I}} \Big(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \frac{\lambda^2(\tau+1)^2|I|}{2\tau^2\nu\Upsilon\Big(\mathrm{Span}\{\boldsymbol{a}_i\}_{i\in I}, \frac{\tau+1}{\tau-1}\Big)^2} \Big).$$

Le reste est de l'ordre $|I|\log(2p)/n$ qui est une vitesse classique dans la littérature. Nos inégalités pour $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ et $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ recouvrent respectivement celles de [1, Théorème 1] et [7, Théorème 4].

LASSO par groupes Le LASSO par groupes est utilisé pour promouvoir la parcimonie par groupes. La pénalité du LASSO par groupes de taille K est

$$J(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_{1,2} \overset{\text{def}}{=} \textstyle\sum_{l=1}^L \|\boldsymbol{\theta}_{\mathcal{G}_l}\|_2. \tag{7}$$

où $\bigcup_{l=1}^L \mathcal{G}_l = [M], \mathcal{G}_i, \mathcal{G}_j \subset [M]$, et $\mathcal{G}_i \cap \mathcal{G}_j = \emptyset$ lorsque $i \neq j$. Soit $\mathcal{B} = \{\mathcal{G}_1, \ldots, \mathcal{G}_L\}$, on definit le support de groupe par

$$\mathrm{supp}_{\mathcal{B}}(\boldsymbol{\theta}) \overset{\text{def}}{=} \big\{l \in [L] \ : \ \boldsymbol{\theta}_{\mathcal{G}_l} \neq 0\big\}.$$

On obtient ainsi le corollaire suivant.

Corollaire 4.2. *Soit le modèle* (1) *où $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$. Supposons que $\max_l \Big\|\Big\|\boldsymbol{X}_{\mathcal{G}_l}^\top \boldsymbol{X}_{\mathcal{G}_l}\Big\|\Big\|_{2\to 2} \leq n$. Supposons que $\lambda \geq \tau\sigma\Big(\sqrt{K} + \sqrt{2\delta\log(L)}\Big)/\sqrt{n}$, pour $\tau > 1$ et $\delta > 1$. Alors,* (4) *et* (5) *sont satisfaites avec une probabilité d'au moins $1 - L^{1-\delta}$ et $\Psi(J, \lambda, \tau, \nu, \boldsymbol{\theta})$ égal à*

$$\inf_{\substack{I\subset[L] \\ \boldsymbol{\theta}:\ \mathrm{supp}_{\mathcal{B}}(\boldsymbol{\theta})=I}} \Big(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \frac{\lambda^2(\tau+1)^2|I|}{2\tau^2\nu\Upsilon\Big(\mathrm{Span}\{\boldsymbol{a}_j\}_{j\in\mathcal{G}_l, l\in I}, \frac{\tau+1}{\tau-1}\Big)^2} \Big).$$

Le reste est de l'ordre $|I|\big(\sqrt{K} + \sqrt{2\log(L)}\big)^2/n$, qui est similaire à celui proposé pour $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ dans [6, 2] avec d'autres a priori de parcimonie par groupes et pour $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ dans [4, 8].

LASSO à l'analyse par groupes Le LASSO à l'analyse par groupes est utilisé pour promouvoir la parcimonie par groupes dans l'image d'un opérateur linéaire. Cet a priori est par exemple sous-jacent à la variation totale isotrope très utilisée en traitement d'image. Étant donné un opérateur linéaire $\boldsymbol{D} : \mathbb{R}^q \to \mathbb{R}^p$, la pénalité du LASSO à l'analyse par groupes de taille K s'écrit

$$J(\boldsymbol{\theta}) = \left\|\boldsymbol{D}^\top \boldsymbol{\theta}\right\|_{1,2}, \tag{8}$$

où maintenant $\mathcal{G}_l \subset [q]$, $\forall l \in [L]$. Soit $\Lambda_{\boldsymbol{\theta}} = \cup_{l \in \mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top \boldsymbol{\theta})} \mathcal{G}_l$ et $\Lambda_{\boldsymbol{\theta}}^c$ son complément. Posons

$$e_{\boldsymbol{D}^\top \boldsymbol{\theta}}^{\|\|_{1,2}} \stackrel{\text{def}}{=} P_{\mathrm{aff}\left(\partial \left\|\boldsymbol{D}^\top \boldsymbol{\theta}\right\|_{1,2}\right)}(0) = \left(H([\boldsymbol{D}^\top \boldsymbol{\theta}]_{\mathcal{G}_l})\right)_{l \in \{1,\dots,L\}},$$

où $H(\boldsymbol{a}) = \boldsymbol{a}/\left\|\boldsymbol{a}\right\|_2$ lorsque $\boldsymbol{a} \neq 0$, et $H(0) = 0$. Nous avons les inégalités oracles suivantes.

Corollaire 4.3. *Soit le modèle* (1) *où* $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. *Supposons que* $\boldsymbol{D}$ *est surjectif, et* $\max_l \left\|\left\|\boldsymbol{D}_{\mathcal{G}_l}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{D}_{\mathcal{G}_l}\right\|\right\|_{2\to 2} \leq n$. *Supposons que* $\lambda \geq \tau\sigma(\sqrt{K} + \sqrt{2\delta \log(L)})/\sqrt{n}$, *pour* $\tau > 1$ *et* $\delta > 1$. *Alors* (4) *et* (5) *sont satisfaites avec une probabilité d'au moins* $1 - L^{1-\delta}$ *et* $\Psi(J, \lambda, \tau, \nu, \boldsymbol{\theta})$ *égal à*

$$\inf_{\substack{I \subset [L] \\ \boldsymbol{\theta}:\ \mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top \boldsymbol{\theta}) = I}} \left(R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \frac{c\lambda^2 \alpha_\tau(\boldsymbol{\theta})^2 |I|}{2\tau^2 \nu \Upsilon\left(\mathrm{Ker}(\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}^c}^\top), \frac{\alpha_\tau(\boldsymbol{\theta})}{\tau - 1}\right)^2}\right),$$

où $\alpha_\tau(\boldsymbol{\theta}) = \tau \left\|\boldsymbol{D}^\top (\boldsymbol{D}\boldsymbol{D}^\top)^{-1} \mathrm{P}_{\mathrm{Ker}(\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}^c}^\top)} \boldsymbol{D} e_{\boldsymbol{D}^\top \boldsymbol{\theta}}^{\|\|_{1,2}}\right\|_{\infty,2} + 1$.

À notre connaissance, ce résultat est nouveau. Le reste est de l'ordre $\left|I\right|(\sqrt{K} + \sqrt{\log(2L)})^2/n$, qui est similaire à celui dans [2, 6] mais pour un a priori de parcimonie analyse des groupes qui est différent du nôtre. De plus, [6] supposent que $\boldsymbol{D}$ est inversible.

Anti-parcimonie Si $\boldsymbol{\theta}_0$ est a priori plat (anti-parcimonieux), une pénalité adéquate est la norme ℓ_∞,

$$J(\boldsymbol{\theta}) = \left\|\boldsymbol{\theta}\right\|_\infty = \max_{i \in [p]} \left|\boldsymbol{\theta}_i\right|. \tag{9}$$

On définit le support de saturation de $\boldsymbol{\theta}$ comme $I_{\boldsymbol{\theta}}^{\mathrm{sat}} \stackrel{\text{def}}{=} \left\{i \in [p] \ :\ \left|\boldsymbol{\theta}_i\right| = \left\|\boldsymbol{\theta}\right\|_\infty\right\}$. Le Théorème 3.3 devient alors.

Corollaire 4.4. *Soit le modèle* (1) *où* $\boldsymbol{\xi}$ *est un vecteur sous-gaussien centré de paramètre* σ, *et* $\boldsymbol{X}$ *telle que* $\max_{i,j} |\boldsymbol{X}_{i,j}| \leq 1/p$. *Supposons que* $\lambda \geq \tau\sigma\sqrt{2\delta \log(2)}\sqrt{p/n}$, *pour* $\tau > 1$ *et* $\delta > 1$. *Alors,* (4) *et* (5) *sont satisfaites avec une probabilité d'au moins* $1 - e^{(1-\delta)\log(2)p}$ *avec* $\Psi(J, \lambda, \tau, \nu, \boldsymbol{\theta}) =$

$$\inf_{\substack{I \subset [M] \\ \boldsymbol{\theta}:\ I_{\boldsymbol{\theta}}^{\mathrm{sat}} = I}} \left(R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \frac{\lambda^2(\tau+1)^2}{2\tau^2 \nu \Upsilon\left(\left\{\overline{\boldsymbol{\theta}} : \overline{\boldsymbol{\theta}}_I \in \mathbb{R}\,\mathrm{sign}(\boldsymbol{\theta}_I)\right\}, \frac{\tau+1}{\tau-1}\right)^2}\right).$$

Le reste est de l'ordre p/n. Nous ne connaissons aucun résultat de ce genre dans la littérature.

Norme nucléaire L'a priori de la norme nucléaire est l'extension spectrale de l'a priori de parcimonie aux matrices $\boldsymbol{\theta} \in \mathbb{R}^{p_1 \times p_2}$. Il pénalise les valeurs singulières de la matrice. Soit $\mathrm{rang}(\boldsymbol{\theta}) = r$ et $\boldsymbol{\lambda}(\boldsymbol{\theta}) \in (\mathbb{R}_+ \setminus \{0\})^r$, le vecteur des valeurs singulières $(\boldsymbol{\lambda}_1(\boldsymbol{\theta}), \cdots, \boldsymbol{\lambda}_r(\boldsymbol{\theta}))$ dans l'ordre décroissant. La norme nucléaire de $\boldsymbol{\theta}$ est

$$J(\boldsymbol{\theta}) = \left\|\boldsymbol{\theta}\right\|_* = \left\|\boldsymbol{\lambda}(\boldsymbol{\theta})\right\|_1. \tag{10}$$

Pour les matrices $\boldsymbol{\theta} \in \mathbb{R}^{p_1 \times p_2}$, une application $\boldsymbol{X}$ prend la forme d'un opérateur linéaire dont la ith composante est donnée par le produit scalaire de Frobenius $\boldsymbol{X}(\boldsymbol{\theta})_i = \mathrm{tr}((\boldsymbol{X}^i)^\top \boldsymbol{\theta})$, où $\boldsymbol{X}^i \in \mathbb{R}^{p_1 \times p_2}$.

Corollaire 4.5. *Soit le modèle* (1) *avec un opérateur linéaire* $\boldsymbol{X} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n$, $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ *et* $\max\left(\left\|\left\|\sum_{i=1}^n \boldsymbol{X}^i (\boldsymbol{X}^i)^\top\right\|\right\|_{2\to 2}, \left\|\left\|\sum_{i=1}^n (\boldsymbol{X}^i)^\top \boldsymbol{X}^i\right\|\right\|_{2\to 2}\right) \leq n$. *Supposons que* $\lambda \geq \tau\sigma\sqrt{2\delta \log(p_1 + p_2)/n}$, *pour* $\tau > 1$ *et* $\delta > 1$. *Alors,* (4) *et* (5) *sont satisfaites avec une probabilité d'au moins* $1 - (p_1 + p_2)^{1-\delta}$ *et* $\Psi(J, \lambda, \tau, \nu, \boldsymbol{\theta}) =$

$$\inf_{\substack{r \in [\min(p_1, p_2)] \\ \boldsymbol{\theta}:\ \mathrm{rang}(\boldsymbol{\theta}) = r}} \left(R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \frac{\lambda^2(\tau+1)^2 r}{2\tau^2 \nu \Upsilon\left(T_{\boldsymbol{\theta}}, \frac{\tau+1}{\tau-1}\right)^2}\right).$$

Le reste est de l'ordre $r \log(p_1 + p_2)/n$, et on retrouve la vitesse que dans [1, Théorème 3] pour $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ et dans [3, Théorème 2] pour $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$.

Références

[1] A. S. Dalalyan, E. Grappin, and Q. Paris. On the Exponentially Weighted Aggregate with the Laplace Prior. Technical report, arXiv :1611.08483, Nov. 2016.

[2] T. Duy Luu, J. M. Fadili, and C. Chesneau. PAC-Bayesian risk bounds for group-analysis sparse regression by exponential weighting. Technical report, hal-01367742, Sept. 2016.

[3] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5) :2302–2329, 2011.

[4] K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39(4) :2164–2204, 08 2011.

[5] T. D. Luu, J. Fadili, and C. Chesneau. Sharp oracle inequalities for low-complexity priors. Technical Report arXiv : 1702.03166, 2017.

[6] P. Rigollet and A. B. Tsybakov. Sparse estimation by exponential weighting. *Statist. Sci.*, 27(4) :558–575, 11 2012.

[7] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4) :879, 2012.

[8] S. van de Geer. Weakly decomposable regularization penalties and structured sparsity. *Scandinavian Journal of Statistics*, 41(1) :72–86, 2014.