Sharp Oracle Inequalities for Low-complexity Priors

Duy Tung Luu, Jalal M. Fadili, Christophe Chesneau

Sharp Oracle Inequalities for Low-complexity Priors

Tung Duy Luu · Jalal Fadili (corresponding author) · Christophe Chesneau

Abstract In this paper, we consider a high-dimensional statistical estimation problem in which the the number of parameters is comparable or larger than the sample size. We present a unified analysis of the performance guarantees of exponential weighted aggregation and penalized estimators with a general class of data losses and priors which encourage objects which conform to some notion of simplicity/complexity. More precisely, we show that these two estimators satisfy sharp oracle inequalities for prediction ensuring their good theoretical performances. We also highlight the differences between them. When the noise is random, we provide oracle inequalities in probability using concentration inequalities. These results are then applied to several instances including the Lasso, the group Lasso, their analysis-type counterparts, the ℓ_∞ and the nuclear norm penalties. All our estimators can be efficiently implemented using proximal splitting algorithms.

Keywords High-dimensional estimation · Exponential weighted aggregation · Penalized estimation · Oracle inequality · Low complexity models

1 Introduction

1.1 Problem statement

Our statistical context is the following. Let $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n)$ be n observations with common marginal distribution, and $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ a deterministic design matrix. The goal to estimate a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$ of the observations marginal distribution based on the data $\boldsymbol{y}$ and $\boldsymbol{X}$.

Let $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be a loss function supposed to be smooth and convex that assigns to each $\boldsymbol{\theta} \in \mathbb{R}^p$ a cost $F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y})$. Let $\boldsymbol{\theta}_0 \in \mathrm{Argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathbb{E}\left[F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y})\right]$ be any minimizer of the population risk. We regard $\boldsymbol{\theta}_0$ as the true parameter. A usual instance of this statistical setting is the standard linear regression

Tung Duy Luu, Jalal Fadili
Normandie Univ, ENSICAEN, CNRS, GREYC, 6, Bd du Maréchal Juin, 14050 Caen, France
E-mail: {duy-tung.luu,Jalal.Fadili}@ensicaen.fr

Christophe Chesneau
Normandie Univ, Université de Caen Normandie, CNRS, LMNO, BP 5186, 14032 Caen, France
E-mail: christophe.chesneau@unicaen.fr

model based on n pairs $(\boldsymbol{y}_i, \boldsymbol{X}_i)$ of response-covariate that are linked linearly $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_0 + \boldsymbol{\xi}$, and $F(\boldsymbol{u}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{u}\|_2^2$.

Our goal is to provide general oracle inequalities in prediction for two estimators of $\boldsymbol{\theta}_0$: the penalized estimator and exponential weighted aggregation. In the setting where "p larger than n (possibly much larger), the estimation problem is ill-posed since the rectangular matrix $\boldsymbol{X}$ has a kernel of dimension at least $p - n$. To circumvent this difficulty, we will exploit the prior that $\boldsymbol{\theta}_0$ has some low-complexity structure (among which sparsity and low-rank are the most popular). That is, even if the ambient dimension p of $\boldsymbol{\theta}_0$ is very large, its intrinsic dimension is much smaller than the sample size n. This makes it possible to build estimates $\boldsymbol{X}\widehat{\boldsymbol{\theta}}$ with good provable performance guarantees under appropriate conditions. There has been a flurry of research on the use of low-complexity regularization in ill-posed recovery problems in various areas including statistics and machine learning.

1.2 Penalized Estimators

Regularization is now a central theme in many fields including statistics, machine learning and inverse problems. It allows one to impose on the set of candidate solutions some prior structure on the object to be estimated. This regularization ranges from squared Euclidean or Hilbertian norms to non-Hilbertian norms (e.g. ℓ_1 norm for sparse objects, or nuclear norm for low-rank matrices) that have sparked considerable interest in the recent years. In this paper, we consider the class of estimators obtained by solving the convex optimization problem[1]

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

where the regularizing penalty J is a proper closed convex function that promotes some specific notion of simplicity/low-complexity, and $\lambda_n > 0$ is the regularization parameter. A prominent member covered by (1) is the Lasso Chen et al (1999); Tibshirani (1996); Osborne et al (2000); Donoho (2006); Candès and Plan (2009); Bickel et al (2009); Bühlmann and van de Geer (2011); Koltchinskii (2008) and its variants such the analysis/fused Lasso Rudin et al (1992); Tibshirani et al (2005), SLOPE Bogdan et al (2014); Su and Candès (2015) or group Lasso Bakin (1999); Yuan and Lin (2006); Bach (2008); Wei and Huang (2010). Another example is the nuclear norm minimization for low rank matrix recovery motivated by various applications including robust PCA, phase retrieval, control and computer vision Recht et al (2010); Candès and Recht (2009); Fazel et al (2001); Candès et al (2013). See Negahban et al (2012); Bühlmann and van de Geer (2011); van de Geer (2014); Vaiter et al (2015b) for generalizations and comprehensive reviews.

[1] To avoid trivialities, the set of minimizers is assumed non-empty, which holds for instance if J is also coercive.

1.3 Exponential Weighted Aggregation (EWA)

An alternative to the penalized estimator (1) is the aggregation by exponential weighting, which consists in substituting averaging for minimization. The aggregators are defined via the probability density function

$$\mu_n(\boldsymbol{\theta}) = \frac{\exp\left(-V_n(\boldsymbol{\theta})/\beta\right)}{\int_{\Theta} \exp\left(-V_n(\boldsymbol{\omega})/\beta\right) d\boldsymbol{\omega}}, \tag{2}$$

where $\beta > 0$ is called temperature parameter. If all $\boldsymbol{\theta}$ are candidates to estimate the true vector $\boldsymbol{\theta}_0$, then $\Theta = \mathbb{R}^p$. The aggregate is thus defined by

$$\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} = \int_{\mathbb{R}^p} \boldsymbol{\theta}\mu_n(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{3}$$

Aggregation by exponential weighting has been widely considered in the statistical and machine learning literatures, see e.g., Dalalyan and Tsybakov (2007, 2008, 2009, 2012); Nemirovski (2000); Yang (2004); Rigollet and Tsybakov (2007); Lecué (2007); Guedj and Alquier (2013); Duy Luu et al (2016) to name a few. The technique used in these papers were initiated by Leung and Barron (2006) (use of Stein's identity to study an early version of EWA) and Catoni (2003, 2007) (PAC-Bayesian theory). $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ can also be interpreted as the posterior conditional mean in the Bayesian sense if $F/(n\beta)$ is the negative-loglikelihood associated to the noise $\boldsymbol{\xi}$ with the prior density $\pi(\boldsymbol{\theta}) \propto \exp\left(-\lambda_n J(\boldsymbol{\theta})/\beta\right)$.

1.4 Oracle inequalities

Oracle inequalities, which are at the heart of our work, quantify the quality of an estimator compared to the best possible one among a family of estimators. These inequalities are well adapted in the scenario where the prior penalty promotes some notion of low-complexity (e.g. sparsity, low rank, etc.). Given two vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, let $R_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ be a nonnegative error measure between their predictions, respectively $\boldsymbol{X}\boldsymbol{\theta}_1$ and $\boldsymbol{X}\boldsymbol{\theta}_2$. A popular example is the averaged prediction squared error $\frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{\theta}_1 - \boldsymbol{X}\boldsymbol{\theta}_2\right\|_2^2$, where $\left\|\cdot\right\|_2$ is the ℓ_2 norm. R_n will serve as a measure of the performance of the estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$. More precisely, we aim to prove that $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ mimic as much as possible the best possible model. This idea is materialized in the following type of inequalities (stated here for EWA)

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\big) \leq C \inf_{\boldsymbol{\theta}\in\mathbb{R}^p} \big(R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \Delta_{n,p,\lambda_n,\beta}(\boldsymbol{\theta})\big), \tag{4}$$

where $C \geq 1$ is the leading constant of the oracle inequality and the remainder term $\Delta_{n,\lambda_n,\beta}(\boldsymbol{\theta})$ depends on the performance of the estimator, the complexity of $\boldsymbol{\theta}$, the sample size n, the dimension p, and the regularization and temperature parameters (λ_n, β). An estimator with good oracle properties would

correspond to C close to 1 (ideally, $C = 1$, in which case the inequality is said "sharp"), and $\Delta_{n,p,\lambda_n,\beta}(\boldsymbol{\theta})$ is small and decreases rapidly to 0 as $n \to +\infty$.

1.5 Contributions

We provide a unified analysis where we capture the essential ingredients behind the low-complexity priors promoted by J, relying on sophisticated arguments from convex analysis and our previous work Fadili et al (2013); Vaiter et al (2015a, 2018, 2017, 2015b). Our main contributions are summarized as follows:

- We show that the EWA estimator $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ in (2) and the penalized estimator $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ in (1) satisfy (deterministic) sharp oracle inequalities for prediction with optimal remainder term, for general data losses F beyond the usual quadratic one, and J is a proper finite-valued sublinear function (i.e. J is finite-valued convex and positively homogeneous). We also highlight the differences between the two estimators in terms of the corresponding bounds.
- When the observations are random, we prove oracle inequalities in probability. The theory is non-asymptotic in nature, as it yields explicit bounds that hold with high probability for finite sample sizes, and reveals the dependence on dimension and other structural parameters of the model.
- For the standard linear model with Gaussian or sub-Gaussian noise, and a quadratic loss, we deliver refined versions of these oracle inequalities in probability. We underscore the role of the Gaussian width, a concept that captures important geometric characteristics of sets in $\mathbb{R}^n$.
- These results yield naturally a large number of corollaries when specialized to penalties routinely used in the literature, among which the Lasso, the group Lasso, their analysis-type counterparts (fused (group) Lasso), the ℓ_∞ and the nuclear norms. Some of these corollaries are known and others novel.

The estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ can be easily implemented thanks to the framework of proximal splitting methods, and more precisely forward-backward type splitting. While the latter is well-known to solve (1) Vaiter et al (2015b), its application within a proximal Langevin Monte-Carlo algorithm to compute $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ with provable guarantees has been recently developed by the authors in Duy Luu et al (2016) to sample from log-semiconcave densities[2], see also Durmus et al (2016) for log-concave densities.

1.6 Relation to previous work

Our oracle inequality for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ extends the work of Dalalyan et al (2018) with an unprecedented level of generality, far beyond the Lasso and the nuclear

[2] In a forthcoming paper, this framework was extended to cover the even more general class of prox-regular functions.

norm. Our prediction sharp oracle inequality for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ specializes to that of Sun and Zhang (2012) in the case of the Lasso (see also the discussion in Dalalyan et al (2017) and references therein) and that of Koltchinskii et al (2011) for the case of the nuclear norm. Our work also goes much beyond that in van de Geer (2014) on weakly decomposable priors, where we show in particular that there is no need to impose decomposability on the regularizer, since it is rather an intrinsic property of it.

1.7 Paper organization

Section 2 states our main assumptions on the data loss and the prior penalty. All the concepts and notions are exemplified on some penalties some of which are popular in the literature. In Section 3, we prove our main oracle inequalities, and their versions in probability. We then tackle the case of linear regression with quadratic data loss in Section 4. Concepts from convex analysis that are essential to this work are gathered in Section A. A key intermediate result in the proof of our main results is established in Section B with an elegant argument relying on Moreau-Yosida regularization.

1.8 Notations

Vectors and matrices For a d-dimensional Euclidean space $\mathbb{R}^d$, we endow it with its usual inner product $\langle \cdot, \cdot \rangle$ and associated norm $\|\cdot\|_2$. $\mathbf{Id}_d$ is the identity matrix on $\mathbb{R}^d$. For $p \geq 1$, $\|\cdot\|_p$ will denote the ℓ_p norm of a vector with the usual adaptation for $p = +\infty$.

In the following, if T is a vector space, P_T denotes the orthogonal projector on T, and

$$\boldsymbol{\theta}_T = \mathrm{P}_T\, \boldsymbol{\theta} \quad \text{and} \quad \boldsymbol{X}_T = \boldsymbol{X}\, \mathrm{P}_T\,.$$

For a finite set $\mathcal{C}$ we denote $\big|\mathcal{C}\big|$ its cardinality. For $I \subset \{1, \ldots, p\}$, we denote by I^c its complement. $\boldsymbol{\theta}_I$ is the subvector whose entries are those of $\boldsymbol{\theta}$ restricted to the indices in I, and $\boldsymbol{X}_I$ the submatrix whose columns are those of $\boldsymbol{X}$ indexed by I. For any matrix $\boldsymbol{X}$, $\boldsymbol{X}^\top$ denotes its transpose and $\boldsymbol{X}^+$ its Moore-Penrose pseudo-inverse. For a linear operator $\boldsymbol{A}$, $\boldsymbol{A}^*$ is its adjoint.

Sets For a nonempty set $\mathcal{C} \in \mathbb{R}^p$, we denote $\overline{\mathrm{conv}}\,(\mathcal{C})$ the closure of its convex hull, and $\iota_{\mathcal{C}}$ its indicator function, i.e. $\iota_{\mathcal{C}}(\boldsymbol{\theta}) = 0$ if $\boldsymbol{\theta} \in \mathcal{C}$ and $+\infty$ otherwise. For a nonempty convex set $\mathcal{C}$, its *affine hull* $\mathrm{aff}(\mathcal{C})$ is the smallest affine manifold containing it. It is a translate of its *parallel subspace* $\mathrm{par}(\mathcal{C})$, i.e. $\mathrm{par}(\mathcal{C}) = \mathrm{aff}(\mathcal{C}) - \boldsymbol{\theta} = \mathbb{R}(\mathcal{C} - \mathcal{C})$; for any $\boldsymbol{\theta} \in \mathcal{C}$. The *relative interior* $\mathrm{ri}(\mathcal{C})$ of a convex set $\mathcal{C}$ is the interior of $\mathcal{C}$ for the topology relative to its affine full.

Functions A function $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is closed (or lower semicontinuous (lsc)) if so is its epigraph. It is coercive if $\lim_{\|\boldsymbol{\theta}\|_2 \to +\infty} f(\boldsymbol{\theta}) = +\infty$, and strongly coercive if $\lim_{\|\boldsymbol{\theta}\|_2 \to +\infty} f(\boldsymbol{\theta}) / \|x\|_2 = +\infty$. The effective domain of f is $\mathrm{dom}(f) = \left\{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ f(\boldsymbol{\theta}) < +\infty\right\}$ and f is proper if $\mathrm{dom}(f) \neq \emptyset$ as is the case when it is finite-valued. A function is said sublinear if it is convex and positively homogeneous. The Legendre-Fenchel conjugate of f is $f^*(\boldsymbol{z}) = \sup_{\boldsymbol{\theta} \in \mathbb{R}^p} \langle \boldsymbol{z}, \boldsymbol{\theta} \rangle - f(\boldsymbol{\theta})$. For f proper, the functions (f, f^*) obey the Fenchel-Young inequality

$$f(\boldsymbol{\theta}) + f^*(\boldsymbol{z}) \geq \langle \boldsymbol{z}, \boldsymbol{\theta} \rangle, \quad \forall (\boldsymbol{\theta}, \boldsymbol{z}) \in \mathbb{R}^p \times \mathbb{R}^p. \tag{5}$$

When f is a proper lower semicontonuous and convex function, (f, f^*) is actually the best pair for which this inequality cannot be tightened. For a function g on $\mathbb{R}_+$, the function $g^+ : a \in \mathbb{R}_+ \mapsto g^+(a) = \sup_{t \geq 0} at - g(t)$ is called the monotone conjugate of g. The pair (g, g^+) obviously obeys (5) on $\mathbb{R}_+ \times \mathbb{R}_+$.

For a C^1-smooth function f, $\nabla f(\boldsymbol{\theta})$ is its (Euclidean) gradient. For a bivariate function $g : (\boldsymbol{\eta}, \boldsymbol{y}) \in \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ that is C^2 with respect to the first variable $\boldsymbol{\eta}$, for any $\boldsymbol{y}$, we will denote $\nabla g(\boldsymbol{\eta}, \boldsymbol{y})$ the gradient of g at $\boldsymbol{\eta}$ with respect to the first variable.

The *subdifferential* $\partial f(\boldsymbol{\theta})$ of a convex function f at $\boldsymbol{\theta}$ is the set

$$\partial f(\boldsymbol{\theta}) = \left\{\boldsymbol{\eta} \in \mathbb{R}^p \ : \ f(\boldsymbol{\theta}') \geq f(\boldsymbol{\theta}) + \langle \boldsymbol{\eta}, \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle, \quad \forall \boldsymbol{\theta}' \in \mathrm{dom}(f)\right\}.$$

An element of $\partial f(\boldsymbol{\theta})$ is a subgradient. If the convex function f is differentiable at $\boldsymbol{\theta}$, then its only subgradient is its gradient, i.e. $\partial f(\boldsymbol{\theta}) = \{\nabla f(\boldsymbol{\theta})\}$.

The *Bregman divergence* associated to a convex function f at $\boldsymbol{\theta}$ with respect to $\boldsymbol{\eta} \in \partial f(\boldsymbol{\theta}) \neq \emptyset$ is

$$D_f^{\boldsymbol{\eta}}\left(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta}\right) = f(\overline{\boldsymbol{\theta}}) - f(\boldsymbol{\theta}) - \langle \boldsymbol{\eta}, \overline{\boldsymbol{\theta}} - \boldsymbol{\theta} \rangle.$$

The Bregman divergence is in general nonsymmetric. It is also nonnegative by convexity. When f is differentiable at $\overline{\boldsymbol{\theta}}$, we simply write $D_f\left(\boldsymbol{\theta}, \overline{\boldsymbol{\theta}}\right)$ (which is, in this case, also known as the Taylor distance).

2 Estimation with low-complexity penalties

The estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ in (1) and (3) require two essential ingredients: the data loss term F and the prior penalty J. We here specify the class of such functions covered in our work, and provide illustrating examples.

2.1 Data loss

The class of loss functions F that we consider obey the following assumptions:

(H.1) $F(\cdot, \boldsymbol{y}) : \mathbb{R}^n \to \mathbb{R}$ is $C^1(\mathbb{R}^n)$ and uniformly convex for all $\boldsymbol{y}$ of modulus φ, i.e.

$$F(\boldsymbol{v}, \boldsymbol{y}) \geq F(\boldsymbol{u}, \boldsymbol{y}) + \langle \nabla F(\boldsymbol{u}, \boldsymbol{y}), \boldsymbol{v} - \boldsymbol{u} \rangle + \varphi(\|\boldsymbol{v} - \boldsymbol{u}\|_2),$$

where $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is a convex non-decreasing function that vanishes only at 0.

(H.2) For any $\overline{\boldsymbol{\theta}} \in \mathbb{R}^p$ and $\boldsymbol{y} \in \mathbb{R}^n$, $\int_{\mathbb{R}^p} \exp\left(-F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y})/(n\beta)\right) \left|\langle \nabla F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}), \boldsymbol{X}(\overline{\boldsymbol{\theta}} - \boldsymbol{\theta})\rangle\right| d\boldsymbol{\theta} < +\infty$.

Recall that by Lemma 2, the monotone conjugate φ^+ of φ is a proper, closed, convex, strongly coercive and non-decreasing function on $\mathbb{R}_+$ that vanishes at 0. Moreover, $\varphi^{++} = \varphi$. The function φ^+ is finite-valued on $\mathbb{R}_+$ if φ is strongly coercive, and it vanishes only at 0 under e.g. Lemma 2(iii).

The class of data loss functions in **(H.1)** is fairly general. It is reminiscent of the negative log-likelihood in the regular exponential family. For the moment assumption **(H.2)** to be satisfied, it is suffient that

$$\int_{\mathbb{R}^p} \exp\left(-\varphi\left(\left\|\boldsymbol{X}\boldsymbol{\theta}\right\|_2\right)/(n\beta)\right) \left\|\nabla F(\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{u}^\star, \boldsymbol{y})\right\|_2 \left\|\boldsymbol{X}\boldsymbol{\theta} + (\boldsymbol{u}^\star - \boldsymbol{X}\overline{\boldsymbol{\theta}})\right\|_2 d\boldsymbol{\theta} < +\infty,$$

where $\boldsymbol{u}^\star$ be a minimizer of $F(\cdot, \boldsymbol{y})$, which is unique by uniform convexity. We here provide an example.

Example 1 Consider the case where[3] $\varphi(t) = t^q/q$, $q \in]1, +\infty[$, or equivalently $\varphi^+(t) = t^{q_*}/q_*$ where $1/q + 1/q_* = 1$. For $q = q_* = 2$, **(H.1)** amounts to saying that $F(\cdot, \boldsymbol{y})$ is strongly convex for all $\boldsymbol{y}$. In particular, (Bauschke and Combettes 2011, Proposition 10.13) shows that $F(\boldsymbol{u}, \boldsymbol{y}) = \left\|\boldsymbol{u} - \boldsymbol{y}\right\|_2^q/q$ is uniformly convex for $q \in [2, +\infty[$ with modulus $\varphi(t) = C_q t^q/q$, where $C_q > 0$ is a constant that depends solely on q.

For **(H.2)** to be verified, it is suffient that

$$\int_{\mathbb{R}^p} \exp\left(-\left\|\boldsymbol{X}\boldsymbol{\theta}\right\|_2^q/(qn\beta)\right) \left\|\nabla F(\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{u}^\star, \boldsymbol{y})\right\|_2 \left\|(\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{u}^\star) - \boldsymbol{X}\overline{\boldsymbol{\theta}}\right\|_2 d\boldsymbol{\theta} < +\infty.$$

In particular, taking $F(\boldsymbol{u}, \boldsymbol{y}) = \left\|\boldsymbol{u} - \boldsymbol{y}\right\|_2^q/q$, $q \in [2, +\infty[$, we have $\left\|\nabla F(\boldsymbol{u}, \boldsymbol{y})\right\|_2 = \left\|\boldsymbol{u} - \boldsymbol{y}\right\|_2^{q-1}$, and thus **(H.2)** holds since

$$\int_{\mathbb{R}^p} \exp\left(-\left\|\boldsymbol{X}\boldsymbol{\theta}\right\|_2^q/(qn\beta)\right) \left\|\boldsymbol{y} - (\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{u}^\star)\right\|_2^{q-1} \left\|\boldsymbol{X}\overline{\boldsymbol{\theta}} - (\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{u}^\star)\right\|_2 d\boldsymbol{\theta} < +\infty.$$

2.2 Prior penalty

Recall the main definitions and results from convex analysis that are collected in Section A. Our main assumption on J is the following.

(H.3) $J : \mathbb{R}^p \to \mathbb{R}$ is the gauge of a non-empty convex compact set containing the origin as an interior point.

[3] We consider a scaled version of φ for simplicity, but the same conclusions remain valid if we take $\varphi(t) = Ct^q/q$, with $C > 0$.

By Lemma 4, this assumption is equivalent to saying that $J \overset{\text{def}}{=} \gamma_{\mathcal{C}}$ is proper, convex, positively homogeneous, finite-valued and coercive. In turn, J is locally Lipschitz continuous on $\mathbb{R}^p$. Observe also that by virtue of Lemma 5 and Lemma 3, the polar gauge $J^\circ \overset{\text{def}}{=} \gamma_{\mathcal{C}^\circ}$ enjoys the same properties as J in **(H.3)**.

2.3 Decomposability of the prior penalty

We are now in position to provide an important characterization of the subdifferential mapping of a function J satisfying **(H.3)**. This characterization will play a pivotal role in our proof of the oracle inequality.

We start by defining some essential geometrical objects that were introduced in Vaiter et al (2015a).

Definition 1 (Model Subspace) Let $\boldsymbol{\theta} \in \mathbb{R}^p$. We denote by $e_{\boldsymbol{\theta}}$ as

$$e_{\boldsymbol{\theta}} = \mathrm{P}_{\mathrm{aff}(\partial J(\boldsymbol{\theta}))}(0).$$

We denote

$$S_{\boldsymbol{\theta}} = \mathrm{par}(\partial J(\boldsymbol{\theta})) \quad \text{and} \quad T_{\boldsymbol{\theta}} = S_{\boldsymbol{\theta}}^{\perp}.$$

$T_{\boldsymbol{\theta}}$ is coined the *model subspace* of $\boldsymbol{\theta}$ associated to J.

It can be shown, see (Vaiter et al 2015a, Proposition 5), that $\boldsymbol{\theta} \in T_{\boldsymbol{\theta}}$, hence the name model subspace. When J is differentiable at $\boldsymbol{\theta}$, we have $e_{\boldsymbol{\theta}} = \nabla J(\boldsymbol{\theta})$ and $T_{\boldsymbol{\theta}} = \mathbb{R}^p$. When J is the ℓ_1-norm (Lasso), the vector $e_{\boldsymbol{\theta}}$ is nothing but the sign of $\boldsymbol{\theta}$. Thus, $e_{\boldsymbol{\theta}}$ can be viewed as a generalization of the sign vector. Observe also that $e_{\boldsymbol{\theta}} = \mathrm{P}_{T_{\boldsymbol{\theta}}}(\partial J(\boldsymbol{\theta}))$, and thus $e_{\boldsymbol{\theta}} \in T_{\boldsymbol{\theta}} \cap \mathrm{aff}(\partial J(\boldsymbol{\theta}))$. However, in general, $e_{\boldsymbol{\theta}} \notin \partial J(\boldsymbol{\theta})$.

We now provide a fundamental equivalent description of the subdifferential of J at $\boldsymbol{\theta}$ in terms of $e_{\boldsymbol{\theta}}$, $T_{\boldsymbol{\theta}}$, $S_{\boldsymbol{\theta}}$ and the polar gauge J°.

Theorem 1 *Let J satisfy* **(H.3)**. *Let $\boldsymbol{\theta} \in \mathbb{R}^p$ and $f_{\boldsymbol{\theta}} \in \mathrm{ri}(\partial J(\boldsymbol{\theta}))$.*

(i) The subdifferential of J at $\boldsymbol{\theta}$ reads

$$\begin{aligned}\partial J(\boldsymbol{\theta}) &= \mathrm{aff}(\partial J(\boldsymbol{\theta})) \cap \mathcal{C}^\circ \\ &= \{\boldsymbol{\eta} \in \mathbb{R}^n \ : \\ &\quad \boldsymbol{\eta}_{T_{\boldsymbol{\theta}}} = e_{\boldsymbol{\theta}} \ \textit{and} \ \inf_{\tau \geq 0} \max\left(J^\circ\left(\tau e_{\boldsymbol{\theta}} + \boldsymbol{\eta}_{S_{\boldsymbol{\theta}}} + (\tau - 1)\,\mathrm{P}_{S_{\boldsymbol{\theta}}} f_{\boldsymbol{\theta}}\right), \tau\right) \leq 1\right\}.\end{aligned}$$

(ii) For any $\boldsymbol{\omega} \in \mathbb{R}^p$, $\exists \boldsymbol{\eta} \in \partial J(\boldsymbol{\theta})$ such that

$$J(\boldsymbol{\omega}_{S_{\boldsymbol{\theta}}}) = \langle \boldsymbol{\eta}_{S_{\boldsymbol{\theta}}}, \boldsymbol{\omega}_{S_{\boldsymbol{\theta}}} \rangle.$$

Proof (i) This follows by piecing together (Vaiter et al 2015a, Theorem 1, Proposition 4 and Proposition 5(iii)).

(ii) From (Vaiter et al 2015a, Proposition 5(iv)), we have

$$\sigma_{\partial J(\boldsymbol{\theta})-f_{\boldsymbol{\theta}}}(\boldsymbol{\omega}) = J(\boldsymbol{\omega}_{S_{\boldsymbol{\theta}}}) - \langle \mathrm{P}_{S_{\boldsymbol{\theta}}} f_{\boldsymbol{\theta}}, \boldsymbol{\omega}_{S_{\boldsymbol{\theta}}} \rangle.$$

Thus there exists a supporting point $\boldsymbol{v} \in \partial J(\boldsymbol{\theta}) - f_{\boldsymbol{\theta}} \subset S_{\boldsymbol{\theta}}$ with normal vector $\boldsymbol{\omega}$ (Bauschke and Combettes 2011, Corollary 7.6(iii)), i.e.

$$\sigma_{\partial J(\boldsymbol{\theta})-f_{\boldsymbol{\theta}}}(\boldsymbol{\omega}) = \langle \boldsymbol{v}, \boldsymbol{\omega}_{S_{\boldsymbol{\theta}}} \rangle.$$

Taking $\boldsymbol{\eta} = \boldsymbol{v} + f_{\boldsymbol{\theta}}$ concludes the proof.

Remark 1 The coercivity assumption in **(H.3)** is not needed for Theorem 1 to hold.

The decomposability of described in Theorem 1(i) depends on the particular choice of the mapping $\boldsymbol{\theta} \mapsto f_{\boldsymbol{\theta}} \in \mathrm{ri}(\partial J(\boldsymbol{\theta}))$. An interesting situation is encountered when $e_{\boldsymbol{\theta}} \in \mathrm{ri}(J(\boldsymbol{\theta}))$, so that one can choose $f_{\boldsymbol{\theta}} = e_{\boldsymbol{\theta}}$. Strong gauges, see (Vaiter et al 2015a, Definition 6), are precisely a class of gauges for which this situation occurs, and in this case, Theorem 1(i) has the simpler form

$$\partial J(\boldsymbol{\theta}) = \mathrm{aff}(\partial J(\boldsymbol{\theta})) \cap \mathcal{C}^\circ = \left\{ \boldsymbol{\eta} \in \mathbb{R}^n \;:\; \boldsymbol{\eta}_{T_{\boldsymbol{\theta}}} = e_{\boldsymbol{\theta}} \quad \text{and} \quad J^\circ(\boldsymbol{\eta}_{S_{\boldsymbol{\theta}}}) \leq 1 \right\}. \quad (6)$$

The Lasso, group Lasso and nuclear norms are typical examples of (symmetric) strong gauges. However, analysis sparsity penalties (e.g. the fused Lasso) or the ℓ_∞-penalty are not strong gauges, though they obviously satisfy **(H.3)**. See the next section for a detailed discussion.

2.4 Calculus with the prior family

The family of penalties complying with **(H.3)** form a robust class enjoying important calculus rules. In particular it is closed under the sum and composition with an injective linear operator as we now prove.

Lemma 1 *The set of functions satisfying* **(H.3)** *is closed under addition*[4] *and pre-composition by an injective linear operator. More precisely, the following holds:*

(i) Let J and G be two gauges satisfying **(H.3)**. *Then $H \stackrel{def}{=} J + G$ also obeys* **(H.3)**. *Moreover,*
  (a) $T^H_{\boldsymbol{\theta}} = T^J_{\boldsymbol{\theta}} \cap T^G_{\boldsymbol{\theta}}$ and $e^H_{\boldsymbol{\theta}} = \mathrm{P}_{T^H_{\boldsymbol{\theta}}}(e^J_{\boldsymbol{\theta}} + e^G_{\boldsymbol{\theta}})$, where $T^J_{\boldsymbol{\theta}}$ and $e^J_{\boldsymbol{\theta}}$ (resp. $T^G_{\boldsymbol{\theta}}$ and $e^G_{\boldsymbol{\theta}}$) are the model subspace and vector at $\boldsymbol{\theta}$ associated to J (resp. G);
  (b) $H^\circ(\boldsymbol{\omega}) = \max_{\rho \in [0,1]} \overline{\mathrm{conv}}\left(\inf\left(\rho J^\circ(\boldsymbol{\omega}), (1-\rho) G^\circ(\boldsymbol{\omega})\right)\right)$.
(ii) Let J be a gauge satisfying **(H.3)**, *and $\boldsymbol{D} : \mathbb{R}^q \to \mathbb{R}^p$ be surjective. Then $H \stackrel{def}{=} J \circ \boldsymbol{D}^\top$ also fulfills* **(H.3)**. *Moreover,*

[4] It is obvious that the same holds with any positive linear combination.

(a) $T^H_{\boldsymbol{\theta}} = \mathrm{Ker}(\boldsymbol{D}^\top_{S^J_{\boldsymbol{u}}})$ *and* $e^H_{\boldsymbol{\theta}} = \mathrm{P}_{T^H_{\boldsymbol{\theta}}}\boldsymbol{D}e^J_{\boldsymbol{u}}$*, where* $T^J_{\boldsymbol{u}}$ *and* $e^J_{\boldsymbol{u}}$ *are the model subspace and vector at* $\boldsymbol{u} \overset{def}{=} \boldsymbol{D}^\top\boldsymbol{\theta}$ *associated to* J*;*
(b) $H^\circ(\boldsymbol{\omega}) = J^\circ(\boldsymbol{D}^+\boldsymbol{\omega})$*, where* $\boldsymbol{D}^+ = \boldsymbol{D}^\top\left(\boldsymbol{D}\boldsymbol{D}^\top\right)^{-1}$.

The outcome of Lemma 1 is naturally expected. For instance, assertion (i) states that combining several penalties/priors will promote objects living on the intersection of the respective low-complexity models. Similarly, for (ii), one promotes low-complexity in the image of the analysis operator $\boldsymbol{D}^\top$. It then follows that one has not to deploy an ad hoc analysis when linearly pre-composing or combining (or both) several penalties (e.g. ℓ_1+nuclear norms for recovering sparse and low-rank matrices) since our unified analysis in Section 3 will apply to them just as well.

Proof (i) Convexity, positive homogeneity, coercivity and finite-valuedness are straightforward.
(a) This is (Vaiter et al 2015a, Proposition 8(i)-(ii)).
(b) We have from Lemma 5 and calculus rules on support functions,

$$\begin{aligned}
H^\circ(\boldsymbol{\omega}) = \sigma_{J(\boldsymbol{\theta})+G(\boldsymbol{\theta})\leq 1}(\boldsymbol{\omega}) &= \sup_{J(\boldsymbol{\theta})+G(\boldsymbol{\theta})\leq 1} \langle \boldsymbol{\omega}, \boldsymbol{\theta}\rangle \\
&= \max_{\rho\in[0,1]} \sup_{J(\boldsymbol{\theta})\leq\rho, G(\boldsymbol{\theta})\leq 1-\rho} \langle \boldsymbol{\omega}, \boldsymbol{\theta}\rangle \\
&= \max_{\rho\in[0,1]} \overline{\mathrm{conv}}\left(\inf\left(\sigma_{J(\boldsymbol{\theta})\leq\rho}(\boldsymbol{\omega}), \sigma_{G(\boldsymbol{\theta})\leq 1-\rho}(\boldsymbol{\omega})\right)\right) \\
&= \max_{\rho\in[0,1]} \overline{\mathrm{conv}}\left(\inf\left(\rho\sigma_{J(\boldsymbol{\theta})\leq 1}(\boldsymbol{\omega}), (1-\rho)\sigma_{G(\boldsymbol{\theta})\leq 1}(\boldsymbol{\omega})\right)\right) \\
&= \max_{\rho\in[0,1]} \overline{\mathrm{conv}}\left(\inf\left(\rho J^\circ(\boldsymbol{\omega}), (1-\rho)G^\circ(\boldsymbol{\omega})\right)\right),
\end{aligned}$$

where we used (Hiriart-Urruty and Lemaréchal 2001, Theorem V.3.3.3) in third row, positive homogeneity in the fourth, and Lemma 5 in the fifth.

(ii) Again, convexity, positive homogeneity and finite-valuedness are immediate. Coercivity holds by injectivity of $\boldsymbol{D}^\top$.
(a) This is (Vaiter et al 2015a, Proposition 10(i)-(ii)).
(b) Denote $J = \gamma_{\mathcal{C}}$. We have

$$\begin{aligned}
H^\circ(\boldsymbol{\omega}) &= \sup_{\boldsymbol{D}^\top\boldsymbol{\theta}\in\mathcal{C}} \langle \boldsymbol{\omega}, \boldsymbol{\theta}\rangle \\
(\boldsymbol{D}^\top \text{ is injective}) &= \sup_{\boldsymbol{D}^\top\boldsymbol{\theta}\in\mathcal{C}} \langle \boldsymbol{D}^+\boldsymbol{\omega}, \boldsymbol{D}^\top\boldsymbol{\theta}\rangle \\
&= \sup_{\boldsymbol{u}\in\mathcal{C}\cap\mathrm{Span}(\boldsymbol{D}^\top)} \langle \boldsymbol{D}^+\boldsymbol{\omega}, \boldsymbol{u}\rangle \\
&= \overline{\mathrm{conv}}\left(\inf\left(J^\circ(\boldsymbol{D}^+\boldsymbol{\omega}), \iota_{\mathrm{Ker}(\boldsymbol{D})}(\boldsymbol{D}^+\boldsymbol{\omega})\right)\right) \\
&= J^\circ(\boldsymbol{D}^+\boldsymbol{\omega}).
\end{aligned}$$

where in the last equality, we used the fact that $\boldsymbol{D}^+\boldsymbol{\omega} \in \mathrm{Span}\left(\boldsymbol{D}^\top\right) = \mathrm{Ker}(\boldsymbol{D})^\perp$, and thus $\iota_{\mathrm{Ker}(\boldsymbol{D})}(\boldsymbol{D}^+\boldsymbol{\omega}) = +\infty$ unless $\boldsymbol{\omega} = 0$, and J° is

continuous and convex by **(H.3)** and Lemma 5. In the fourth equality, we invoked (Hiriart-Urruty and Lemaréchal 2001, Theorem V.3.3.3) and Lemma 5.

2.5 Examples

2.5.1 Lasso

The Lasso regularization is used to promote the sparsity of the minimizers, see Bühlmann and van de Geer (2011) for a comprehensive review. It corresponds to choosing J as the ℓ_1-norm

$$J(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1 = \sum_{i=1}^{p} |\boldsymbol{\theta}_i|. \tag{7}$$

It is also referred to as ℓ_1-synthesis in the signal processing community, in contrast to the more general ℓ_1-analysis sparsity penalty detailed below.
We denote $(\boldsymbol{a}_i)_{1\leq i\leq p}$ the canonical basis of $\mathbb{R}^p$ and $\mathrm{supp}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \big\{i \in \{1,\ldots,p\} \ : \ \boldsymbol{\theta}_i \neq 0\big\}$. Then,

$$T_{\boldsymbol{\theta}} = \mathrm{Span}\{(\boldsymbol{a}_i)_{i\in\mathrm{supp}(\boldsymbol{\theta})}\}, \ \ (e_{\boldsymbol{\theta}})_i = \begin{cases} \mathrm{sign}(\boldsymbol{\theta}_i) & \text{if } i \in \mathrm{supp}(\boldsymbol{\theta}) \\ 0 & \text{otherwise} \end{cases}, \text{ and } J^\circ = \|\cdot\|_\infty. \tag{8}$$

2.5.2 Group Lasso

The group Lasso has been advocated to promote sparsity by groups, i.e. it drives all the coefficients in one group to zero together hence leading to group selection, see Bakin (1999); Yuan and Lin (2006); Bach (2008); Wei and Huang (2010) to cite a few. The group Lasso penalty with L groups reads

$$J(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_{1,2} \stackrel{\text{def}}{=} \sum_{i=1}^{L} \|\boldsymbol{\theta}_{b_i}\|_2, \tag{9}$$

where $\bigcup_{i=1}^{L} b_i = \{1,\ldots,p\}$, $b_i, b_j \subset \{1,\ldots,p\}$, and $b_i \cap b_j = \emptyset$ whenever $i \neq j$. Define the group support as $\mathrm{supp}_{\mathcal{B}}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \big\{i \in \{1,\ldots,L\} \ : \ \boldsymbol{\theta}_{b_i} \neq 0\big\}$. Thus, one has

$$\begin{aligned} &T_{\boldsymbol{\theta}} = \mathrm{Span}\{(a_j)_{\big\{j \,:\, \exists i \in \mathrm{supp}_{\mathcal{B}}(\boldsymbol{\theta}), j \in b_i\big\}}\}, \\ &(e_{\boldsymbol{\theta}})_{b_i} = \begin{cases} \frac{\boldsymbol{\theta}_{b_i}}{\|\boldsymbol{\theta}_{b_i}\|_2} & \text{if } i \in \mathrm{supp}_{\mathcal{B}}(\boldsymbol{\theta}) \\ 0 & \text{otherwise} \end{cases}, \text{ and } J^\circ(\boldsymbol{\omega}) = \max_{i\in\{1,\ldots,L\}} \|\boldsymbol{\omega}_{b_i}\|_2 . \end{aligned} \tag{10}$$

2.5.3 Analysis (group) Lasso

One can push the structured sparsity idea one step further by promoting group/block sparsity through a linear operator, i.e. analysis-type sparsity. Given a linear operator $\boldsymbol{D} : \mathbb{R}^q \to \mathbb{R}^p$ (seen as a matrix), the analysis group sparsity penalty is

$$J(\boldsymbol{\theta}) = \left\|\boldsymbol{D}^\top \boldsymbol{\theta}\right\|_{1,2}. \tag{11}$$

This encompasses the 2-D isotropic total variation Rudin et al (1992). For when all groups of cardinality one, we have the analysis-ℓ_1 penalty (a.k.a. general Lasso), which encapsulates several important penalties including that of the 1-D total variation Rudin et al (1992), and the fused Lasso Tibshirani et al (2005). The overlapping group Lasso Jacob et al (2009) is also a special case of (9) by taking $\boldsymbol{D}^\top$ to be an operator that exactract the blocks Peyré et al (2011); Chen et al (2010) (in which case $\boldsymbol{D}$ has even orthogonal rows).

Let $\varLambda_{\boldsymbol{\theta}} = \bigcup_{i \in \mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top \boldsymbol{\theta})} b_i$ and $\varLambda_{\boldsymbol{\theta}}^c$ its complement. From Lemma 1(ii) and (10), we get

$$\begin{aligned} &T_{\boldsymbol{\theta}} = \mathrm{Ker}(\boldsymbol{D}_{\varLambda_{\boldsymbol{\theta}}^c}^\top), \quad e_{\boldsymbol{\theta}} = \mathrm{P}_{T_{\boldsymbol{\theta}}}\, \boldsymbol{D} e_{\boldsymbol{D}^\top \boldsymbol{\theta}}^{\|\|_{1,2}} \\ &\text{where} \quad \left(e_{\boldsymbol{D}^\top \boldsymbol{\theta}}^{\|\|_{1,2}}\right)_{b_i} = \begin{cases} \dfrac{(\boldsymbol{D}^\top \boldsymbol{\theta})_{b_i}}{\left\|(\boldsymbol{D}^\top \boldsymbol{\theta})_{b_i}\right\|_2} & \text{if } i \in \mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top \boldsymbol{\theta}) \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{12}$$

If, in addition, $\boldsymbol{D}$ is surjective, then by virtue of Lemma 1(ii) we also have

$$J^\circ(\boldsymbol{\omega}) = \left\|\boldsymbol{D}^+ \boldsymbol{\omega}\right\|_{\infty,2} \overset{\text{def}}{=} \max_{i \in \{1,\dots,L\}} \left\|(\boldsymbol{D}^+ \boldsymbol{\omega})_{b_i}\right\|_2. \tag{13}$$

2.5.4 Anti-sparsity

If the vector to be estimated is expected to be flat (anti-sparse), this can be captured using the ℓ_∞ norm (a.k.a. Tchebychev norm) as prior

$$J(\boldsymbol{\theta}) = \left\|\boldsymbol{\theta}\right\|_\infty = \max_{i \in \{1,\dots,p\}} \left|\boldsymbol{\theta}_i\right|. \tag{14}$$

The ℓ_∞ regularization has found applications in several fields Jégou et al (2012); Lyubarskii and Vershynin (2010); Studer et al (2012). Suppose that $\boldsymbol{\theta} \neq 0$, and define the saturation support of $\boldsymbol{\theta}$ as $I_{\boldsymbol{\theta}}^{\mathrm{sat}} \overset{\text{def}}{=} \left\{ i \in \{1,\dots,p\} \ : \ \left|\boldsymbol{\theta}_i\right| = \left\|\boldsymbol{\theta}\right\|_\infty \right\} \neq \emptyset$. From (Vaiter et al 2015a, Proposition 14), we have

$$\begin{aligned} &T_{\boldsymbol{\theta}} = \left\{ \overline{\boldsymbol{\theta}} \in \mathbb{R}^p \ : \ \overline{\boldsymbol{\theta}}_{I_{\boldsymbol{\theta}}^{\mathrm{sat}}} \in \mathbb{R}\, \mathrm{sign}(\boldsymbol{\theta}_{I_{\boldsymbol{\theta}}^{\mathrm{sat}}}) \right\}, \\ &(e_{\boldsymbol{\theta}})_i = \begin{cases} \mathrm{sign}(\boldsymbol{\theta}_i)/|I_{\boldsymbol{\theta}}^{\mathrm{sat}}| & \text{if } i \in I_{\boldsymbol{\theta}}^{\mathrm{sat}} \\ 0 & \text{otherwise} \end{cases}, \quad \text{and} \quad J^\circ = \left\|\cdot\right\|_1. \end{aligned} \tag{15}$$

2.5.5 Nuclear norm

The natural extension of low-complexity priors to matrices $\boldsymbol{\theta} \in \mathbb{R}^{p_1 \times p_2}$ is to penalize the singular values of the matrix. Let $\operatorname{rank}(\boldsymbol{\theta}) = r$, and $\boldsymbol{\theta} = \boldsymbol{U} \operatorname{diag}(\lambda(\boldsymbol{\theta}))\boldsymbol{V}^\top$ be a reduced rank-r SVD decomposition, where $\boldsymbol{U} \in \mathbb{R}^{p_1 \times r}$ and $\boldsymbol{V} \in \mathbb{R}^{p_2 \times r}$ have orthonormal columns, and $\lambda(\boldsymbol{\theta}) \in (\mathbb{R}_+ \setminus \{0\})^r$ is the vector of singular values $(\lambda_1(\boldsymbol{\theta}), \cdots, \lambda_r(\boldsymbol{\theta}))$ in non-increasing order. The nuclear norm of $\boldsymbol{\theta}$ is

$$J(\boldsymbol{\theta}) = \big\|\boldsymbol{\theta}\big\|_* = \big\|\lambda(\boldsymbol{\theta})\big\|_1. \tag{16}$$

This penalty is the best convex surrogate to enforce a low-rank prior. It has been widely used for various applications Recht et al (2010); Candès and Recht (2009); Candès et al (2011); Fazel et al (2001); Candès et al (2013).

Following e.g. (Vaiter et al 2017, Example 21), we have

$$\begin{aligned} &T_{\boldsymbol{\theta}} = \big\{\boldsymbol{U}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{V}^\top \;:\; \boldsymbol{A} \in \mathbb{R}^{p_2 \times r}, \boldsymbol{B} \in \mathbb{R}^{p_1 \times r}\big\}, \quad e_{\boldsymbol{\theta}} = \boldsymbol{U}\boldsymbol{V}^\top \\ &\text{and } J^\circ(\boldsymbol{\omega}) = \|\!|\!|\boldsymbol{\omega}|\!|\!|_{2\to 2} = \big\|\lambda(\boldsymbol{\omega})\big\|_\infty. \end{aligned} \tag{17}$$

3 Oracle inequalities for a general loss

Before delving into the details, in the sequel, we will need a bit of notations.

We recall $T_{\boldsymbol{\theta}}$ and $e_{\boldsymbol{\theta}}$ the model subspace and vector associated to $\boldsymbol{\theta}$ (see Definition 1). Denote $S_{\boldsymbol{\theta}} = T_{\boldsymbol{\theta}}^\perp$. Given two coercive finite-valued gauges $J_1 = \gamma_{\mathcal{C}_1}$ and $J_2 = \gamma_{\mathcal{C}_2}$, and a linear operator $\boldsymbol{A}$, we define $|\!|\!|\boldsymbol{A}|\!|\!|_{J_1 \to J_2}$ the *operator bound* as

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

Note that $|\!|\!|\boldsymbol{A}|\!|\!|_{J_1 \to J_2}$ is bounded (this follows from Lemma 4(v)). Furthermore, we have from Lemma 5 that

$$\begin{aligned} |\!|\!|\boldsymbol{A}|\!|\!|_{J_1 \to J_2} = \sup_{\boldsymbol{\theta} \in \mathcal{C}_1} \sup_{\boldsymbol{\omega} \in \mathcal{C}_2^\circ} \langle \boldsymbol{A}^\top \boldsymbol{\omega}, \boldsymbol{\theta} \rangle = \sup_{\boldsymbol{\omega} \in \mathcal{C}_2^\circ} \sup_{\boldsymbol{\theta} \in \mathcal{C}_1} \langle \boldsymbol{A}^\top \boldsymbol{\omega}, \boldsymbol{\theta} \rangle = \sup_{\boldsymbol{\omega} \in \mathcal{C}_2^\circ} J_1^\circ(\boldsymbol{A}^\top \boldsymbol{\omega}) \\ = |\!|\!|\boldsymbol{A}^\top|\!|\!|_{J_2^\circ \to J_1^\circ}. \end{aligned}$$

In the following, whenever it is clear from the context, to lighten notation when J_i is a norm, we write the subscript of the norm instead of J_i (e.g. p for the ℓ_p norm, $*$ for the nuclear norm, etc.).

Our main result will involve a measure of well-conditionedness of the design matrix $\boldsymbol{X}$ when restricted to some subspace T. More precisely, for $c > 0$, we introduce the coefficient

$$\Upsilon(T, c) = \inf_{\{\boldsymbol{\omega} \in \mathbb{R}^p \,:\, J(\boldsymbol{\omega}_S) < cJ(\boldsymbol{\omega}_T)\}} \frac{|\!|\!|\mathrm{P}_T|\!|\!|_{2 \to J} \big\|\boldsymbol{X}\boldsymbol{\omega}\big\|_2}{n^{1/2}(J(\boldsymbol{\omega}_T) - J(\boldsymbol{\omega}_S)/c)}. \tag{18}$$

This generalizes the compatibility factor introduced in van de Geer and Buhlmann (2009) for the Lasso (and used in Dalalyan et al (2018)). The experienced

reader may have recognized that this factor is reminescent of the null space property and restricted injectivity that play a central role in the analysis of the performance guarantees of penalized estimators (1); see Fadili et al (2013); Vaiter et al (2015a, 2018, 2017, 2015b). One can see in particular that $\Upsilon(T,c)$ is larger than the smallest singular value of $\boldsymbol{X}_T$.

The oracle inequalites will provided in terms of the loss

$$R_n\big(\boldsymbol{\theta},\boldsymbol{\theta}_0\big) = \tfrac{1}{n} D_F(\boldsymbol{X}\boldsymbol{\theta},\boldsymbol{X}\boldsymbol{\theta}_0).$$

3.1 Oracle inequality for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$

We are now ready to establish our first main result: an oracle inequality for the EWA estimator (3).

Theorem 2 *Consider the EWA estimator $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ in (3) with the density (2), where F and J satisfy Assumptions* **(H.1)**-**(H.2)** *and* **(H.3)**. *Then, for any $\tau > 1$ such that $\lambda_n \geq \tau J^\circ\big(-\boldsymbol{X}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0,\boldsymbol{y})\big)/n$, the following holds,*

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}},\boldsymbol{\theta}_0\big) \leq \inf_{\boldsymbol{\theta}\in\mathbb{R}^p}\left(R_n\big(\boldsymbol{\theta},\boldsymbol{\theta}_0\big) + \frac{1}{n}\varphi^+\left(\frac{\lambda_n\sqrt{n}\big(\tau J^\circ(e_{\boldsymbol{\theta}})+1\big)\|\mathrm{P}_{T_{\boldsymbol{\theta}}}\|_{2\to J}}{\tau\Upsilon\Big(T_{\boldsymbol{\theta}},\frac{\tau J^\circ(e_{\boldsymbol{\theta}})+1}{\tau-1}\Big)}\right)\right) + p\beta. \tag{19}$$

Remark 2

1. It should be emphasized that Theorem 2 is actually a deterministic statement for a fixed choice of λ_n. Probabilistic analysis will be required when the result is applied to particular statistical models as we will see later. For this, we will use concentration inequalities in order to provide bounds that hold with high probability over the data.
2. The oracle inequality is sharp. The remainder in it has two terms. The first one encodes the complexity of the model promoted by J. The second one, $p\beta$, captures the influence of the temperature parameter. In particular, taking β sufficiently small of the order $O\big((pn)^{-1}\big)$, this term becomes $O(n^{-1})$.
3. When $\varphi(t) = \nu t^2/2$, i.e. $F(\cdot,\boldsymbol{y})$ is ν-strongly convex, then $\varphi^+(t) = t^2/(2\nu)$, and the reminder term becomes

$$\frac{\lambda_n^2\big(\tau J^\circ(e_{\boldsymbol{\theta}})+1\big)^2\|\mathrm{P}_{T_{\boldsymbol{\theta}}}\|_{2\to J}^2}{2\tau^2\nu\Upsilon\Big(T_{\boldsymbol{\theta}},\frac{\tau J^\circ(e_{\boldsymbol{\theta}})+1}{\tau-1}\Big)^2}. \tag{20}$$

If, moreover, ∇F is also κ-Lipschitz continuous, then it can be shown that $R_n\big(\boldsymbol{\theta},\boldsymbol{\theta}_0\big)$ is equivalent to a quadratic loss. This means that the oracle inequality in Theorem 2 can be stated in terms of the quadratic prediction

error. However, the inequality is not anymore sharp in this case as a constant factor equal to the condition number $\kappa/\nu \geq 1$ naturally multiplies the right-hand side.

4. If J is such that $e_{\boldsymbol{\theta}} \in \partial J(\boldsymbol{\theta}) \subset \mathcal{C}^\circ$ (typically for a strong gauge by (6)), then $J^\circ(e_{\boldsymbol{\theta}}) \leq 1$ (in fact an equality if $\boldsymbol{\theta} \neq 0$). Thus the term $J^\circ(e_{\boldsymbol{\theta}})$ can be omitted in (19).
5. A close inspection of the proof of Theorem 2 reveals that the term $p\beta$ can be improved to the smaller bound

$$p\beta + \Big(V_n(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}) - \mathbb{E}_{\mu_n}\left[V_n(\boldsymbol{\theta})\right]\Big),$$

where the upper-bound is a consequence of Jensen inequality.

Proof By convexity of J and assumption **(H.1)**, we have for any $\boldsymbol{\eta} \in \partial V_n(\boldsymbol{\theta})$ and any $\overline{\boldsymbol{\theta}} \in \mathbb{R}^p$,

$$D_{V_n}^{\boldsymbol{\eta}}\big(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta}\big) \geq \frac{1}{n}\varphi\big(\big\|\boldsymbol{X}\overline{\boldsymbol{\theta}} - \boldsymbol{X}\boldsymbol{\theta}\big\|_2\big).$$

Since φ is non-decreasing and convex, $\varphi \circ \|\cdot\|_2$ is a convex function. Thus, taking the expectation w.r.t. to μ_n on both sides and using Jensen inequality, we get

$$\begin{aligned} V_n(\overline{\boldsymbol{\theta}}) &\geq \mathbb{E}_{\mu_n}\left[V_n(\boldsymbol{\theta})\right] + \mathbb{E}_{\mu_n}\left[\langle \boldsymbol{\eta}, \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\rangle\right] + \frac{1}{n}\mathbb{E}_{\mu_n}\left[\varphi\big(\big\|\boldsymbol{X}\overline{\boldsymbol{\theta}} - \boldsymbol{X}\boldsymbol{\theta}\big\|_2\big)\right] \\ &\geq V_n(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}) + \mathbb{E}_{\mu_n}\left[\langle \boldsymbol{\eta}, \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\rangle\right] + \frac{1}{n}\varphi\big(\big\|\boldsymbol{X}\overline{\boldsymbol{\theta}} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}\big\|_2\big). \end{aligned}$$

This holds for any $\boldsymbol{\eta} \in \partial V_n(\boldsymbol{\theta})$, and in particular at the minimal selection $\big(\partial V_n(\boldsymbol{\theta})\big)^0$ (see Section B for details). It then follows from the pillar result in Proposition 5[5] that

$$\mathbb{E}_{\mu_n}\left[\big\langle \big(\partial V_n(\boldsymbol{\theta})\big)^0, \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\big\rangle\right] = -p\beta.$$

We thus deduce the inequality

$$V_n(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}) - V_n(\overline{\boldsymbol{\theta}}) \leq p\beta - \frac{1}{n}\varphi\big(\big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\overline{\boldsymbol{\theta}}\big\|_2\big), \quad \forall \overline{\boldsymbol{\theta}} \in \mathbb{R}^p. \tag{21}$$

By definition of the Bregman divergence, we have

$$\begin{aligned} &R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\big) - R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) \\ &= \frac{1}{n}\Big(F(\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{y}) - F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) + \langle -\boldsymbol{X}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})), \widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{\theta}\rangle\Big) \\ &= \Big(V_n(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}) - V_n(\boldsymbol{\theta})\Big) + \frac{1}{n}\langle -\boldsymbol{X}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y}), \widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{\theta}\rangle \\ &\quad - \lambda_n\big(J(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}) - J(\boldsymbol{\theta})\big). \end{aligned}$$

[5] In the appendix, we provide a self-contained proof based on a novel Moreau-Yosida regularization argument. In (Dalalyan et al 2018, Corollary 1 and 2), an alternative proof is given using an absolute continuity argument since μ_n is locally Lipschitz, hence a Sobolev function.

By virtue of the duality inequality (42), we have

$$\begin{aligned}
&R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\big) - R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) \\
&\leq \Big(V_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}\big) - V_n(\boldsymbol{\theta})\Big) + \frac{1}{n} J^\circ\big(-\boldsymbol{X}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\big) J\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{\theta}\big) \\
&\quad - \lambda_n\big(J\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}\big) - J(\boldsymbol{\theta})\big) \\
&\leq \Big(V_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}\big) - V_n(\boldsymbol{\theta})\Big) + \frac{\lambda_n}{\tau}\Big(J\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{\theta}\big) - \tau\big(J\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}\big) - J(\boldsymbol{\theta})\big)\Big).
\end{aligned}$$

Denote $\boldsymbol{\omega} = \widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{\theta}$. By virtue of **(H.3)**, Theorem 1 and (42), we obtain

$$\begin{aligned}
J(\boldsymbol{\omega}) - \tau\big(J\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}\big) - J(\boldsymbol{\theta})\big) &\leq J(\boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}) + J(\boldsymbol{\omega}_{S_{\boldsymbol{\theta}}}) - \tau\langle e_{\boldsymbol{\theta}}, \boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}\rangle - \tau J(\boldsymbol{\omega}_{S_{\boldsymbol{\theta}}}) \\
&\leq J(\boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}) + J(\boldsymbol{\omega}_{S_{\boldsymbol{\theta}}}) + \tau J^\circ(e_{\boldsymbol{\theta}}) J(\boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}) - \tau J(\boldsymbol{\omega}_{S_{\boldsymbol{\theta}}}) \\
&= \big(\tau J^\circ(e_{\boldsymbol{\theta}}) + 1\big) J(\boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}) - (\tau - 1) J(\boldsymbol{\omega}_{S_{\boldsymbol{\theta}}}) \\
&\leq \big(\tau J^\circ(e_{\boldsymbol{\theta}}) + 1\big)\Big(J(\boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}) - \tfrac{\tau-1}{\tau J^\circ(e_{\boldsymbol{\theta}})+1} J(\boldsymbol{\omega}_{S_{\boldsymbol{\theta}}})\Big).
\end{aligned}$$

This inequality together with (21) (applied with $\overline{\boldsymbol{\theta}} = \boldsymbol{\theta}$) and (18) yield

$$\begin{aligned}
&R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\big) - R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) \\
&\leq p\beta - \frac{1}{n}\varphi\big(\big\|\boldsymbol{X}\boldsymbol{\omega}\big\|_2\big) + \frac{\lambda_n\big(\tau J^\circ(e_{\boldsymbol{\theta}}) + 1\big)\|\mathrm{P}_{T_{\boldsymbol{\theta}}}\|_{2\to J}\big\|\boldsymbol{X}\boldsymbol{\omega}\big\|_2}{n^{1/2}\tau\Upsilon\big(T_{\boldsymbol{\theta}}, \frac{\tau J^\circ(e_{\boldsymbol{\theta}})+1}{\tau-1}\big)} \\
&\leq p\beta + \frac{1}{n}\varphi^+\left(\frac{\lambda_n\sqrt{n}\big(\tau J^\circ(e_{\boldsymbol{\theta}}) + 1\big)\|\mathrm{P}_{T_{\boldsymbol{\theta}}}\|_{2\to J}}{\tau\Upsilon\big(T_{\boldsymbol{\theta}}, \frac{\tau J^\circ(e_{\boldsymbol{\theta}})+1}{\tau-1}\big)}\right),
\end{aligned}$$

where we applied Fenchel-Young inequality (5) to get the last bound. Taking the infimum over $\boldsymbol{\theta} \in \mathbb{R}^p$ yields the desired result.

Stratifiable functions Theorem 2 has a nice instanciation when $\mathbb{R}^p$ can be partitioned into a collection of subsets $\{\mathcal{M}_i\}_i$ that form a stratification of $\mathbb{R}^p$. That is, $\mathbb{R}^p$ is a finite disjoint union $\cup_i \mathcal{M}_i$ such that the partitioning sets $\mathcal{M}_i$ (called strata) must fit nicely together and the stratification is endowed with a partial ordering for the closure operation. For example, it is known that a polyhedral function has a polyhedral stratification, and more generally, semialgebraic functions induce stratifications into finite disjoint unions of manifolds; see, e.g., Coste (2002). Another example is that of partly smooth convex functions thoroughly studied in Vaiter et al (2015a, 2018, 2017, 2015b) for various statistical and inverse problems. These functions induce a stratification into strata that are C^2-smooth submanifolds of $\mathbb{R}^p$. In turns out that all popular penalty functions discussed in this paper are partly smooth (see Vaiter et al

(2017, 2015b)). Let's denote $\mathscr{M}$ the set of strata associated to J. With this notation at hand, the oracle inequality (19) now reads

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\big) \leq \inf_{\substack{\mathcal{M}\in\mathscr{M} \\ \boldsymbol{\theta}\in\mathcal{M}}} \left(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \frac{1}{n}\varphi^+\left(\frac{\lambda_n\sqrt{n}\big(\tau J^\circ(e_{\boldsymbol{\theta}})+1\big)\|\mathrm{P}_{T_{\boldsymbol{\theta}}}\|_{2\to J}}{\tau\Upsilon\Big(T_{\boldsymbol{\theta}}, \frac{\tau J^\circ(e_{\boldsymbol{\theta}})+1}{\tau-1}\Big)} \right) \right) + p\beta. \tag{22}$$

3.2 Oracle inequality for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$

The next result establishes that $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ satisfies a sharp prediction oracle inequality that we will compare to (19).

Theorem 3 *Consider the penalized estimator $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ in (1), where F and J satisfy Assumptions* **(H.1)** *and* **(H.3)**. *Then, for any $\tau > 1$ such that $\lambda_n \geq \tau J^\circ\big(-\boldsymbol{X}^\top\nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\big)/n$, the following holds,*

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}, \boldsymbol{\theta}_0\big) \leq \inf_{\boldsymbol{\theta}\in\mathbb{R}^p} \left(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \frac{1}{n}\varphi^+\left(\frac{\lambda_n\sqrt{n}\big(\tau J^\circ(e_{\boldsymbol{\theta}})+1\big)\|\mathrm{P}_{T_{\boldsymbol{\theta}}}\|_{2\to J}}{\tau\Upsilon\Big(T_{\boldsymbol{\theta}}, \frac{\tau J^\circ(e_{\boldsymbol{\theta}})+1}{\tau-1}\Big)} \right) \right). \tag{23}$$

Proof The proof follows the same lines as that of Theorem 2 except that we use the fact that $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ is a global minimizer of V_n, i.e. $0 \in \partial V_n(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}})$. Indeed, we have for any $\boldsymbol{\theta} \in \mathbb{R}^p$

$$V_n(\boldsymbol{\theta}) \geq V_n(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}) + \frac{1}{n}\varphi\big(\big\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}\big\|_2\big). \tag{24}$$

Continuing exactly as just after (21), replacing $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ with $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ and invoking (24) instead of (21), we arrive at the claimed result.

Remark 3

1. Observe that the penalized estimator $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ does not require the moment assumption **(H.2)** for (23) to hold. The convexity assumption on φ in **(H.1)**, which was important to apply Jensen's inequality in the proof of (19), is not needed either to get (23).
2. As we remarked for Theorem 2, Theorem 3 is also a deterministic statement for a fixed choice of λ_n that holds for any minimizer $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, which is not unique in general. The condition on λ_n is similar to the one in Negahban et al (2012) where authors established different guarantees for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$.

3.3 Discussion of $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ vs $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$

One clearly sees that the difference between the prediction performance of $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ lies in the term $p\beta$ (or rather its lower-bound in Remark 2-5). In particular, for $\beta = O\big((pn)^{-1}\big)$, this term is on the order $O(n^{-1})$. This choice can be refined in most situations. In particular, for the case of quadratic loss, one can take $\beta = O\Big(\lambda_n^2 \big|\!\big|\!\big| \mathrm{P}_{T_{\boldsymbol{\theta}_0}} \big|\!\big|\!\big|_{2\to J}^2 / p\Big)$, hence leading to remainder terms in (19) and (23) of the same order.

In view of this discussion, one may wonder what are the actual benefits of using $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ instead of $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$. Generalizing the arguments of Dalalyan et al (2018), we will show that $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ enjoys one main advantage compared to $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$. To simplify the discussion, we will focus on the case of linear regression (30) with Gaussian noise $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{Id}_n)$ and F is the quadratic loss.

The chief advantage of $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ is its stability as a function of the data and hyperparameters. It has been shown that for a large class of penalties J, including those studied here, the prediction $\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ is a smooth function of $\boldsymbol{y}$ outside a set of Lebesgue measure zero; see (Vaiter et al 2017, Theorem 2). Those authors also provided in (Vaiter et al 2017, Theorem 3) an expression of the Stein unbiased risk estimator (SURE). For instance, when J is the gauge of a polytope, the SURE is given by

$$\big\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}\big\|_2^2 + \sigma^2 \dim\big(T_{\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}}\big) - n\sigma^2.$$

The SURE was advocated as an automatic and objective way to choose λ. However, one can see that $\dim\big(T_{\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}}\big)$ is a non-smooth function of λ, which may lead to numerical instabilities in practice. In contrast, the SURE of $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$, whose closed-form is given in (Dalalyan et al 2018, (10)), is such that $\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ is a continuous function of $(\lambda, \beta) \in]0, +\infty[^2$ and $\boldsymbol{y} \in \mathbb{R}^n$. This better regularity suggests that it would be wiser to use the SURE associated to $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ for an automatic choice of λ.

3.4 Oracle inequalities in probability

It remains to check when the event $\mathscr{E} = \{\lambda_n \geq \tau J^\circ\big(-\boldsymbol{X}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\big)/n\}$ holds with high probability when $\boldsymbol{y}$ is random. We will use concentration inequalities in order to provide bounds that hold with high probability over the data. Toward this goal, we will need the following assumption.

(H.4) $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n)$ are independent random observations, and $F(\boldsymbol{u}, \boldsymbol{y}) = \sum_{i=1}^n f_i(\boldsymbol{u}_i, \boldsymbol{y}_i)$,
$f_i : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. Moreover,
(i) $\mathbb{E}\left[\big|f_i((\boldsymbol{X}\boldsymbol{\theta}_0)_i, \boldsymbol{y}_i)\big|\right] < +\infty$, $\forall 1 \leq i \leq n$;

(ii) $\left|f_i'((\boldsymbol{X}\boldsymbol{\theta}_0)_i, t)\right| \leq g(t)$, where $\mathbb{E}\left[g(\boldsymbol{y}_i)\right] < +\infty$, $\forall 1 \leq i \leq n$;
(iii) Bernstein moment condition: $\forall 1 \leq i \leq n$ and all integers $m \geq 2$, $\mathbb{E}\left[\left|f_i'((\boldsymbol{X}\boldsymbol{\theta}_0)_i, \boldsymbol{y}_i)\right|^m\right] \leq m!\kappa^{m-2}\sigma_i^2/2$ for some constants $\kappa > 0$, $\sigma_i > 0$ independent of n.

Let $\sigma^2 = \max_{1\leq i\leq n} \sigma_i^2$.

Observe that under **(H.4)**, and by virtue of Lemma 5(iv) and (Hiriart-Urruty and Lemaréchal 2001, Proposition V.3.3.4), we have

$$\begin{aligned} J^\circ\big(-\boldsymbol{X}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\big) &= \sigma_{\mathcal{C}}\big(-\boldsymbol{X}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\big) \\ &= \sup_{\boldsymbol{z}\in \boldsymbol{X}(\mathcal{C})} -\sum_{i=1}^n f_i'((\boldsymbol{X}\boldsymbol{\theta}_0)_i, \boldsymbol{y}_i)\boldsymbol{z}_i. \end{aligned} \tag{25}$$

Thus, checking the event $\mathscr{E}$ amounts to establishing a deviation inequality for the supremum of an empirical process[6] above its mean under the weak Bernstein moment condition **(H.4)**(iii), which essentially requires that the $f_i'((\boldsymbol{X}\boldsymbol{\theta}_0)_i, \boldsymbol{y}_i)$ have sub-exponential tails, We will first tackle the case where $\mathcal{C}$ is the convex hull of a finite set (i.e. $\mathcal{C}$ is a polytope).

3.4.1 Polyhedral penalty

We here suppose that J is a finite-valued gauge of $\mathcal{C} = \overline{\text{conv}}\,(\mathcal{V})$, where $\mathcal{V}$ is finite, i.e. $\mathcal{C}$ is a polytope with vertices (Rockafellar 1996, Corollary 19.1.1). Our first oracle inequality in probability is the following.

Proposition 1 *Consider the estimators* $\widehat{\boldsymbol{\theta}}_n^{\text{EWA}}$ *and* $\widehat{\boldsymbol{\theta}}_n^{\text{PEN}}$, *where* F *and* $J \overset{def}{=} \gamma_{\mathcal{C}}$ *satisfy Assumptions* **(H.1)**, **(H.2)**, **(H.3)** *and* **(H.4)**, *and* $\mathcal{C}$ *is a polytope with vertices* $\mathcal{V}$. *Suppose that* $\text{rank}(\boldsymbol{X}) = n$ *and let* $s(\boldsymbol{X}) = \max_{\boldsymbol{v}\in\mathcal{V}} \left\|\boldsymbol{X}\boldsymbol{v}\right\|_\infty$. *Choose*

$$\lambda_n \geq \tau\sigma s(\boldsymbol{X})\sqrt{\frac{2\delta\log(|\mathcal{V}|)}{n}}\left(1 + \sqrt{2}\kappa/\sigma\sqrt{\frac{\delta\log(|\mathcal{V}|)}{n}}\right),$$

for some $\tau > 1$ *and* $\delta > 1$. *Then* (19) *and* (23) *hold with probability at least* $1 - 2|\mathcal{V}|^{1-\delta}$.

Proof In view of Assumptions **(H.1)** and **(H.4)**, one can differentiate under the expectation sign (Leibniz rule) to conclude that $\mathbb{E}\left[F(\boldsymbol{X}\cdot, \boldsymbol{y})\right]$ is C^1 at $\boldsymbol{\theta}_0$ and $\nabla\mathbb{E}\left[F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\right] = \boldsymbol{X}^\top\mathbb{E}\left[\nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\right]$. As $\boldsymbol{\theta}_0$ minimizes the population risk, one has $\nabla\mathbb{E}\left[F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\right] = 0$. Using the rank assumption on $\boldsymbol{X}$, we deduce that

$$\mathbb{E}\left[f_i'((\boldsymbol{X}\boldsymbol{\theta}_0)_i, \boldsymbol{y}_i)\right] = 0, \quad \forall 1 \leq i \leq n.$$

Moreover, (25) specializes to

$$J^\circ\big(-\boldsymbol{X}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\big) = \sup_{\boldsymbol{z}\in \boldsymbol{X}(\mathcal{V})} -\sum_{i=1}^n f_i'((\boldsymbol{X}\boldsymbol{\theta}_0)_i, \boldsymbol{y}_i)\boldsymbol{z}_i.$$

[6] As $\boldsymbol{X}(\mathcal{C})$ is compact, it has a dense countable subset.

Let $t' = \lambda_n n/\tau$ and $t = t'/s(\boldsymbol{X})$. By the union bound and (25), we have

$$\begin{aligned}
\mathbb{P}\left(J^\circ\big(-\boldsymbol{X}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\big) \geq t'\right) &\leq \mathbb{P}\left(\max_{\boldsymbol{z}\in\boldsymbol{X}(\mathcal{V})} -\sum_{i=1}^n f_i'((\boldsymbol{X}\boldsymbol{\theta}_0)_i, \boldsymbol{y}_i)\boldsymbol{z}_i \geq t'\right) \\
&\leq |\mathcal{V}| \max_{\boldsymbol{z}\in\boldsymbol{X}(\mathcal{V})} \mathbb{P}\left(\Big|\sum_{i=1}^n f_i'((\boldsymbol{X}\boldsymbol{\theta}_0)_i, \boldsymbol{y}_i)\boldsymbol{z}_i\Big| \geq t'\right) \\
&\leq |\mathcal{V}|\mathbb{P}\left(s(\boldsymbol{X})\Big|\sum_{i=1}^n f_i'((\boldsymbol{X}\boldsymbol{\theta}_0)_i, \boldsymbol{y}_i)\Big| \geq t'\right) \\
&= |\mathcal{V}|\mathbb{P}\left(\Big|\sum_{i=1}^n f_i'((\boldsymbol{X}\boldsymbol{\theta}_0)_i, \boldsymbol{y}_i)\Big| \geq t\right).
\end{aligned}$$

Owing to assumption **(H.4)**(iii), we are in position to apply the Bernstein inequality to get

$$\mathbb{P}\left(J^\circ\big(-\boldsymbol{X}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\big) \geq t\right) \leq 2|\mathcal{V}| \exp\left(-\frac{t^2}{2(\kappa t + n\sigma^2)}\right).$$

Every t such that

$$t \geq \sqrt{\delta \log(|\mathcal{V}|)}\Big(\kappa\sqrt{\delta \log(|\mathcal{V}|)} + \sqrt{\kappa^2 \delta \log(|\mathcal{V}|) + 2n\sigma^2}\Big),$$

satisfies $t^2 \geq 2\delta \log(|\mathcal{V}|)(\kappa t + n\sigma^2)$. Applying the trivial inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ to the bound on t, we conclude.

Remark 4 In the monograph (Bühlmann and van de Geer 2011, Lemma 14.12), the authors derived an exponential deviation inequality for the supremum of an empirical process with finite $\mathcal{V}$ and possibly unbounded empirical processes under a Bernstein moment condition similar to ours (in fact ours implies theirs). The very last part of our proof can be obtained by applying their result. We detailed it here for the sake of completeness.

Lasso To lighten the notation, let $I_{\boldsymbol{\theta}} = \operatorname{supp}(\boldsymbol{\theta})$. From (8), it is easy to see that

$$\|\mathrm{P}_{T_{\boldsymbol{\theta}}}\|_{2\to 1} = \sqrt{|I_{\boldsymbol{\theta}}|} \quad \text{and} \quad J^\circ(e_{\boldsymbol{\theta}}) = \|\operatorname{sign}(\boldsymbol{\theta}_{I_{\boldsymbol{\theta}}})\|_\infty \leq 1,$$

where last bound holds as an equality whenever $\boldsymbol{\theta} \neq 0$. Further the ℓ_1 norm is the gauge of the cross-polytope (i.e. the unit ℓ_1 ball). Its vertex set $\mathcal{V}$ is the set of unit-norm one-sparse vectors $(\pm\boldsymbol{a}_i)_{1\leq i\leq p}$, where we recall $(\boldsymbol{a}_i)_{1\leq i\leq p}$ the canonical basis. Thus

$$|\mathcal{V}| = 2p \quad \text{and} \quad s(\boldsymbol{X}) = \max_{\boldsymbol{v}\in\mathcal{V}} \|\boldsymbol{X}\boldsymbol{v}\|_\infty = \max_{1\leq i\leq p} \|\boldsymbol{X}_i\|_\infty.$$

Inserting this into Proposition 1, we obtain the following corollary.

Corollary 1 *Consider the estimators* $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ *and* $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$*, where where* J *is the Lasso penalty and* F *satisfies Assumptions* **(H.1)**, **(H.2)** *and* **(H.4)**. *Suppose that* $\mathrm{rank}(\boldsymbol{X}) = n$ *and take*

$$\lambda_n \geq \tau\sigma s(\boldsymbol{X})\sqrt{\frac{2\delta\log(2p)}{n}}\left(1+\sqrt{2}\kappa/\sigma\sqrt{\frac{\delta\log(2p)}{n}}\right),$$

for some $\tau > 1$ *and* $\delta > 1$. *Then, with probability at least* $1-2(2p)^{1-\delta}$, *the following holds*

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}},\boldsymbol{\theta}_0\big) \leq \inf_{\substack{I\subset\{1,\dots,p\}\\ \boldsymbol{\theta}:\ \mathrm{supp}(\boldsymbol{\theta})=I}} \left(R_n\big(\boldsymbol{\theta},\boldsymbol{\theta}_0\big) + \tfrac{1}{n}\varphi^+\left(\frac{\lambda_n\sqrt{n}(\tau+1)\sqrt{|I|}}{\tau\Upsilon\Big(\mathrm{Span}\{\boldsymbol{a}_i\}_{i\in I},\frac{\tau+1}{\tau-1}\Big)}\right)\right) + p\beta, \tag{26}$$

and

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}},\boldsymbol{\theta}_0\big) \leq \inf_{\substack{I\subset\{1,\dots,p\}\\ \boldsymbol{\theta}:\ \mathrm{supp}(\boldsymbol{\theta})=I}} \left(R_n\big(\boldsymbol{\theta},\boldsymbol{\theta}_0\big) + \tfrac{1}{n}\varphi^+\left(\frac{\lambda_n\sqrt{n}(\tau+1)\sqrt{|I|}}{\tau\Upsilon\Big(\mathrm{Span}\{\boldsymbol{a}_i\}_{i\in I},\frac{\tau+1}{\tau-1}\Big)}\right)\right). \tag{27}$$

For $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, we recover a similar scaling for λ_n and the oracle inequality as in van de Geer (2008), though in the latter the oracle inequality is not sharp unlike ours. Note that the above oracle inequality extends readily to the case of analysis/fused Lasso $\big\|\boldsymbol{D}^\top\cdot\big\|_1$ where $\boldsymbol{D}$ is surjective. We leave the details to the interested reader (see also the analysis group Lasso example in Section 4).

Anti-sparsity From Section 2.5.4, recall the saturation support $I_{\boldsymbol{\theta}}^{\mathrm{sat}}$ of $\boldsymbol{\theta}$. From (15), we get

$$\|\!|\mathrm{P}_{T_{\boldsymbol{\theta}}}|\!\|_{2\to\infty} = 1 \quad \text{and} \quad J^\circ(e_{\boldsymbol{\theta}}) = \big\|\,\mathrm{sign}(\boldsymbol{\theta}_{I_{\boldsymbol{\theta}}^{\mathrm{sat}}})\big\|_1/|I_{\boldsymbol{\theta}}^{\mathrm{sat}}| \leq 1,$$

with equality whenever $\boldsymbol{\theta}\neq 0$. In addition, the ℓ_∞ norm is the gauge of the hypercube whose vertex set is $\mathcal{V} = \{\pm 1\}^p$. Thus

$$|\mathcal{V}| = 2^p.$$

We have the following oracle inequalities.

Corollary 2 *Consider the estimators* $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ *and* $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$*, where where* J *is anti-sparsity penalty* (14)*, and* F *satisfies Assumptions* **(H.1)**, **(H.2)** *and* **(H.4)**. *Suppose that* $\mathrm{rank}(\boldsymbol{X}) = n$ *and let* $s(\boldsymbol{X}) = \max_{i,j}|\boldsymbol{X}_{i,j}|$. *Choose*

$$\lambda_n \geq \tau\sigma s(\boldsymbol{X})\sqrt{2\delta\log(2)}\sqrt{\frac{p}{n}}\left(1+2\kappa/\sigma\sqrt{\delta\log(2)}\sqrt{\frac{p}{n}}\right),$$

for some $\tau > 1$ *and* $\delta > 1$. *Then, with probability at least* $1 - 2^{-p(\delta-1)+1}$, *the following holds*

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\big) \le \inf_{\substack{I\subset\{1,\dots,p\}\\ \boldsymbol{\theta}:\ I_{\boldsymbol{\theta}}^{\mathrm{sat}}=I}} \left(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \tfrac{1}{n}\varphi^+\left(\frac{\lambda_n\sqrt{n}(\tau+1)}{\tau\Upsilon\Big(\big\{\overline{\boldsymbol{\theta}} : \overline{\boldsymbol{\theta}}_I \in \mathbb{R}\,\mathrm{sign}(\boldsymbol{\theta}_I)\big\}, \frac{\tau+1}{\tau-1}\Big)} \right) \right) + p\beta, \tag{28}$$

and

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}, \boldsymbol{\theta}_0\big) \le \inf_{\substack{I\subset\{1,\dots,p\}\\ \boldsymbol{\theta}:\ I_{\boldsymbol{\theta}}^{\mathrm{sat}}=I}} \left(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \tfrac{1}{n}\varphi^+\left(\frac{\lambda_n\sqrt{n}(\tau+1)}{\tau\Upsilon\Big(\big\{\overline{\boldsymbol{\theta}} : \overline{\boldsymbol{\theta}}_I \in \mathbb{R}\,\mathrm{sign}(\boldsymbol{\theta}_I)\big\}, \frac{\tau+1}{\tau-1}\Big)} \right) \right). \tag{29}$$

We are not aware of any result of this kind in the literature. The bound imposed on $\boldsymbol{X}$ is similar to what is generally assumed in the vector quantization literature Lyubarskii and Vershynin (2010); Studer et al (2012).

3.4.2 General penalty

Extending the above reasoning to a general penalty requires a deviation inequality for the supremum of an empirical process in (25) under the Bernstein moment condition **(H.4)**(iii), but without the need of uniform boundedness. This can be achieved via generic chaining along a tree using entropy with bracketing; see (van de Geer and Lederer 2013, Theorem 8). The resulting deviation bound will thus depend on the entropies with bracketing. These quantities capture the complexity of the set $\boldsymbol{X}(\mathcal{C})$ but are intricate to compute in general. This subject deserves further investigation that we leave to a future work.

*Remark 5 (**Group Lasso**)* Using the union bound, we have

$$\mathbb{P}\left(\max_{i\in\{1,\dots,L\}} \left\|\boldsymbol{X}_{b_i}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\right\|_2 \ge \lambda_n n/\tau \right) \le L \max_i \mathbb{P}\left(\left\|\boldsymbol{X}_{b_i}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\right\|_2 \ge \lambda_n n/\tau \right).$$

This requires a concentration inequality for quadratic forms of independent random variables satisfying the Bernstein moment assumption above. We are not aware of any such result. But if our moment condition is strengthened to

$$\mathbb{E}\left[\left| f_i'((\boldsymbol{X}\boldsymbol{\theta}_0)_i, \boldsymbol{y}_i)\right|^{2m} \right] \le m!\kappa^{2(m-1)}\sigma_i^2/2, \quad \forall 1 \le i \le n, \forall m \ge 1,$$

then one can use (Bellec 2014, Theorem 3). Indeed, assuming $\max_i \|\boldsymbol{X}_i\|_2 \le \sqrt{n}$, which is a natural normalization on the design, we have by independence that

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{X}_{b_i}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\right\|_2 \right] &\le \mathbb{E}\left[\left\|\boldsymbol{X}_{b_i}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\right\|_2^2 \right]^{1/2} \\ &= \sigma\sqrt{\mathrm{tr}(\boldsymbol{X}_{b_i}^\top \boldsymbol{X}_{b_i})/2} = \sigma\sqrt{\sum_{j\in b_i} \left\|\boldsymbol{X}_j\right\|_2^2/2} \le \sigma\sqrt{Kn/2}.\end{aligned}$$

It then follows that taking

$$\lambda_n \geq \tau \frac{\sigma\sqrt{K} + 16\kappa\sqrt{\delta \log(L)}}{\sqrt{n}}, \quad \delta > 1,$$

the oracle inequalities (34) and (35) hold for the group Lasso with probability at least $1 - L^{1-\delta}$. A similar result can be proved for the analysis group Lasso just as well (see Section 4.3.3).

4 Oracle inequalities for low-complexity linear regression

In this section, we consider the classical linear regression problem where the n response-covariate pairs $(\boldsymbol{y}_i, \boldsymbol{X}_i)$ are linked as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_0 + \boldsymbol{\xi}, \tag{30}$$

where $\boldsymbol{\xi}$ is a noise vector. The data loss will be set to $F(\boldsymbol{u}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{u}\|_2^2$. This in turn entails that $\varphi = \varphi^+ = \frac{1}{2}(\cdot)^2$ on $\mathbb{R}_+$ and $R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\|_2^2$.

In this section, we assume that the noise $\boldsymbol{\xi}$ is a zero-mean sub-Gaussian vector in $\mathbb{R}^n$ with parameter σ. That is, its one-dimensional marginals $\langle \boldsymbol{\xi}, \boldsymbol{z} \rangle$ are sub-Gaussian random variables $\forall \boldsymbol{z} \in \mathbb{R}^n$, i.e. they satisfy

$$\mathbb{P}\left(\left|\langle \boldsymbol{\xi}, \boldsymbol{z} \rangle\right| \geq t\right) \leq 2e^{-t^2/(2\|\boldsymbol{z}\|_2^2\sigma^2)}, \quad \forall \boldsymbol{z} \in \mathbb{R}^n. \tag{31}$$

In this case, the bounds of Section 3.4 can be improved.

4.1 General penalty

As we will shortly show, the event $\mathscr{E}$ will depend on the Gaussian width, a summary geometric quantity which, informally speaking, measures the size of the bulk of a set in $\mathbb{R}^n$.

Definition 2 The Gaussian width of a subset $\mathcal{S} \subset \mathbb{R}^n$ is defined as

$$w(\mathcal{S}) \stackrel{\text{def}}{=} \mathbb{E}\left[\sigma_{\mathcal{S}}(\boldsymbol{g})\right], \quad \text{where} \quad \boldsymbol{g} \sim \mathcal{N}(0, \mathbf{Id}_n).$$

The concept of Gaussian width has appeared in the literature in different contexts. In particular, it has been used to establish sample complexity bounds to ensure exact recovery (noiseless case) and mean-square estimation stability (noisy case) for low-complexity penalized estimators from Gaussian measurements; see e.g. Rudelson and Vershynin (2008); Chandrasekaran et al (2012); Tropp (2015a); Vershynin (2015); Vaiter et al (2015b).

The Gaussian width has deep connections to convex geometry and it enjoys many useful properties. It is well-known that it is positively homogeneous, monotonic w.r.t. inclusion, and invariant under orthogonal transformations. Moreover, $w(\overline{\text{conv}}\,(\mathcal{S})) = w(\mathcal{S})$. From Lemma 3(ii)-(iii), $w(\mathcal{S})$ is a non-negative finite quantity whenever the set $\mathcal{S}$ is bounded and contains the origin.

We are now ready to state our oracle inequality in probability with sub-Gaussian noise.

Proposition 2 *Let the data generated by (30) where $\boldsymbol{\xi}$ is a zero-mean sub-Gaussian random vector with parameter σ. Consider the estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, where F and $J \overset{def}{=} \gamma_{\mathcal{C}}$ satisfy Assumptions* **(H.1)**-**(H.2)** *and* **(H.3)**. *Suppose that $\lambda_n \geq \frac{\tau\sigma c_1\sqrt{2\log(c_2/\delta)}w(\boldsymbol{X}(\mathcal{C}))}{n}$, for some $\tau > 1$ and $0 < \delta < \min(c_2, 1)$, where c_1 and c_2 are positive absolute constants. Then with probability at least $1-\delta$, (19) and (23) hold with the remainder term given by (20) with $\nu = 1$.*

The proof requires sophisticated ideas from the theory of generic chaining Talagrand (2005), but we only apply these results. The constants c_1 and c_2 can be traced back to the proof of these results as detailed in Talagrand (2005).

Proof First, from (31), we have the bound

$$\mathbb{P}\left(\left|\langle\boldsymbol{\xi}, \boldsymbol{z}-\boldsymbol{z}'\rangle\right| \geq t\right) \leq 2e^{-t^2/(2\|\boldsymbol{z}-\boldsymbol{z}'\|_2^2\sigma^2)}, \quad \forall \boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^n,$$

i.e. the increment condition (Talagrand 2005, (0.4)) is verified. Thus combining (25) with the probability bound in (Talagrand 2005, page 11), the generic chaining theorem (Talagrand 2005, Theorem 1.2.6) and the majorizing measure theorem (Talagrand 2005, Theorem 2.1.1), we have

$$\begin{aligned}\mathbb{P}\left(J^\circ(\boldsymbol{X}^\top\boldsymbol{\xi}) \geq \lambda_n n/\tau\right) &\leq \mathbb{P}\left(\sup_{\boldsymbol{z}\in\boldsymbol{X}(\mathcal{C})}\langle\boldsymbol{\xi}, \boldsymbol{z}\rangle \geq \sigma c_1\sqrt{2\log(c_2/\delta)}w(\boldsymbol{X}(\mathcal{C}))\right)\\ &\leq c_2\exp\left(-\frac{\sigma^2 2\log(c_2/\delta)}{2\sigma^2}\right) = \delta.\end{aligned}$$

If the noise is Gaussian, an enhanced version can be proved by invoking Gaussian concentration of Lipschitz functions Ledoux (2001).

Proposition 3 *Let the data generated by (30) with noise $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2\mathbf{Id}_n)$. Consider the estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, where F and $J \overset{def}{=} \gamma_{\mathcal{C}}$ satisfy Assumptions* **(H.1)**-**(H.2)** *and* **(H.3)**. *Suppose that $\lambda_n \geq \frac{(1+\delta)\tau\sigma w(\boldsymbol{X}(\mathcal{C}))}{n}$, for some $\tau > 1$ and $\delta > 0$. Then with probability at least $1 - \exp\left(-\frac{\delta^2 w(\boldsymbol{X}(\mathcal{C}))^2}{2\|\boldsymbol{X}\|_{J\to 2}^2}\right)$, (19) and (23) hold with the remainder term given by (20) with $\nu = 1$.*

Proof Thanks to sublinearity (see Lemma 4(i) and Lemma 5), the function $\boldsymbol{\xi} \mapsto J^\circ(\boldsymbol{X}^\top\boldsymbol{\xi})$ is Lipschitz continuous with Lipschitz constant $\left\|\boldsymbol{X}^\top\right\|_{2\to J^\circ} = \|\boldsymbol{X}\|_{J\to 2}$. From (25), we also have

$$\mathbb{E}\left[J^\circ\left(\boldsymbol{X}^\top\boldsymbol{\xi}\right)\right] = \sigma w(\boldsymbol{X}(\mathcal{C})).$$

Observe that $\boldsymbol{X}(\mathcal{C})$ is a convex compact set containing the origin. Setting $\epsilon = \lambda_n n/\tau - \sigma w(\boldsymbol{X}(\mathcal{C})) \geq \delta\sigma w(\boldsymbol{X}(\mathcal{C}))$, it follows from (25) and the Gaussian

concentration of Lipschitz functions Ledoux (2001) that

$$\begin{aligned}\mathbb{P}\left(J^\circ(\boldsymbol{X}^\top\boldsymbol{\xi}) \geq \lambda_n n/\tau\right) &\leq \mathbb{P}\left(J^\circ(\boldsymbol{X}^\top\boldsymbol{\xi}) - \mathbb{E}\left[J^\circ(\boldsymbol{X}^\top\boldsymbol{\xi})\right] \geq \epsilon\right)\\ &\leq \mathbb{P}\left(J^\circ(\boldsymbol{X}^\top\boldsymbol{\xi}/\sigma) - w(\boldsymbol{X}(\mathcal{C})) \geq \delta w(\boldsymbol{X}(\mathcal{C}))\right)\\ &\leq \exp\left(-\frac{\delta^2 w(\boldsymbol{X}(\mathcal{C}))^2}{2\|\boldsymbol{X}\|^2_{J\to 2}}\right).\end{aligned}$$

Estimating theoretically the Gaussian width of a set[7] is a non-trivial problem that has been extensively studied in the areas of probability in Banach spaces and stochastic processes. There are classical bounds on the Gaussian width (Sudakov's and Dudley's inequalities), but they are difficult to estimate in most cases and neither of these bounds is tight for all sets. When the set is a convex cone (intersected with a sphere), tractable estimates based on polarity arguments were proposed in, e.g., Chandrasekaran et al (2012).

4.2 Polyhedral penalty

When $\mathcal{C}$ and is polytope, enhanced oracle inequalities can be obtained by invoking a simple union bound argument.

Proposition 4 *Let the data generated by (30) where $\boldsymbol{\xi}$ is a zero-mean sub-Gaussian random vector with parameter σ. Consider the estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, where F and $J \overset{def}{=} \gamma_{\mathcal{C}}$ satisfy Assumptions* **(H.1)**-**(H.2)** *and* **(H.3)**, *and moreover $\mathcal{C}$ is a polytope with vertices $\mathcal{V}$. Suppose that $\lambda_n \geq \frac{\tau\sigma\left(\max_{\boldsymbol{v}\in\mathcal{V}}\|\boldsymbol{X}\boldsymbol{v}\|_2\right)\sqrt{2\delta\log(|\mathcal{V}|)}}{n}$, for some $\tau > 1$ and $\delta > 1$. Then with probability at least $1 - 2|\mathcal{V}|^{1-\delta}$, (19) and (23) hold with the remainder term given by (20) with $\nu = 1$.*

In particular, if $\max_{\boldsymbol{v}\in\mathcal{V}}\|\boldsymbol{X}\boldsymbol{v}\|_2 = C\sqrt{n}$, for a positive constant C, then one can take $\lambda_n \geq C\tau\sigma\sqrt{\frac{2\delta\log(|\mathcal{V}|)}{n}}$.

Proof From (25) we have

$$J^\circ\left(\boldsymbol{X}^\top\boldsymbol{\xi}\right) = \max_{\boldsymbol{v}\in\mathcal{C}} \langle\boldsymbol{X}\boldsymbol{v},\boldsymbol{\xi}\rangle = \max_{\boldsymbol{v}\in\mathcal{V}} \langle\boldsymbol{X}\boldsymbol{v},\boldsymbol{\xi}\rangle,$$

where in the last inequality, we used the fact that a convex function attains its maximum on $\mathcal{C}$ at an extreme point $\mathcal{V}$. Let $\epsilon = \sigma\left(\max_{\boldsymbol{v}\in\mathcal{V}}\|\boldsymbol{X}\boldsymbol{v}\|_2\right)\sqrt{2\delta\log(|\mathcal{V}|)}$.

[7] Not to mention its image with a linear operator as for $\boldsymbol{X}(\mathcal{C})$.

By the union bound, (31) and (25), we have

$$\begin{aligned}\mathbb{P}\left(J^\circ\left(\boldsymbol{X}^\top\boldsymbol{\xi}\right) \geq \lambda_n n/\tau\right) &\leq \mathbb{P}\left(\max_{\boldsymbol{v}\in\mathcal{V}} \langle \boldsymbol{X}\boldsymbol{v}, \boldsymbol{\xi}\rangle \geq \epsilon\right) \\ &\leq |\mathcal{V}| \max_{\boldsymbol{v}\in\mathcal{V}} \mathbb{P}\left(\langle \boldsymbol{X}\boldsymbol{v}, \boldsymbol{\xi}\rangle \geq \epsilon\right) \\ &\leq |\mathcal{V}| \max_{\boldsymbol{v}\in\mathcal{V}} \mathbb{P}\left(\left|\langle \boldsymbol{X}\boldsymbol{v}, \boldsymbol{\xi}\rangle\right| \geq \epsilon\right) \\ &\leq 2|\mathcal{V}| \exp\left(-\epsilon^2/\left(2\sigma^2 \max_{\boldsymbol{v}\in\mathcal{V}} \|\boldsymbol{X}\boldsymbol{v}\|_2^2\right)\right) \\ &\leq 2|\mathcal{V}|^{1-\delta}.\end{aligned}$$

4.3 Applications

In this section, we exemplify our oracle inequalities for the penalties described in Section 2.5.

4.3.1 Lasso

Recall the derivations for the Lasso in Section 3.4.1. We obtain the following corollary of Proposition 4.

Corollary 3 *Let the data generated by (30) where $\boldsymbol{\xi}$ is a zero-mean sub-Gaussian random vector with parameter σ. Assume that $\boldsymbol{X}$ is such that $\max_i \|\boldsymbol{X}_i\|_2 \leq \sqrt{n}$. Consider the estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, where J is the Lasso penalty (7) and F satisfies Assumptions* **(H.1)**-**(H.2)**. *Suppose that $\lambda_n \geq \tau\sigma\sqrt{\frac{2\delta\log(2p)}{n}}$, for some $\tau > 1$ and $\delta > 1$. Then, with probability at least $1 - 2(2p)^{1-\delta}$, the following holds*

$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 \leq \inf_{\substack{I\subset\{1,\dots,p\} \\ \boldsymbol{\theta}:\ \mathrm{supp}(\boldsymbol{\theta})=I}} \left(\frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 + \frac{\lambda_n^2(\tau+1)^2|I|}{\tau^2\Upsilon\left(\mathrm{Span}\{\boldsymbol{a}_i\}_{i\in I}, \frac{\tau+1}{\tau-1}\right)^2}\right) + 2p\beta, \tag{32}$$

and

$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 \leq \inf_{\substack{I\subset\{1,\dots,p\} \\ \boldsymbol{\theta}:\ \mathrm{supp}(\boldsymbol{\theta})=I}} \left(\frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 + \frac{\lambda_n^2(\tau+1)^2|I|}{\tau^2\Upsilon\left(\mathrm{Span}\{\boldsymbol{a}_i\}_{i\in I}, \frac{\tau+1}{\tau-1}\right)^2}\right). \tag{33}$$

The normalization on the design is natural. The remainder term grows as $\frac{|I|\log(p)}{n}$. The oracle inequality (33) recovers (Dalalyan et al 2018, Theorem 1) in the exactly sparse case, and (33) the one in (Sun and Zhang 2012, Theorem 4) (see also (Koltchinskii et al 2011, Theorem 11) and (Dalalyan et al 2017, Theorem 2)). It is worth mentioning, however, that (Dalalyan et al 2018,

Theorem 1) handles the inexactly sparse case while we do not. For the choice $\beta = O\big(\sigma^2|I|\log(2p)/(pn)\big)$, the remainder terms in (32) and (33) are of the same order. Observe that this choice of the temperature parameter is optimal in view of the results of Castillo et al (2015). These authors proved that for the ℓ_1 penalty, orthonormal design, noise $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{Id}_n)$, and all choices of the form $\lambda = C\sigma\sqrt{\log(n)/n}$, then the pseudo-posterior μ_n in (2) with temparature $\beta = \sigma^2/n$ puts asymptotically no mass on the ball centered at $\boldsymbol{\theta}_0$ of radius $\sim \sqrt{\log(n)/n}$.

4.3.2 Group Lasso

Recall the notations in Section 2.5.2, and denote $I_{\boldsymbol{\theta}} = \mathrm{supp}_{\mathcal{B}}(\boldsymbol{\theta})$ the set indexing active blocks in $\boldsymbol{\theta}$. From (10), we have

$$\|\mathrm{P}_{T_{\boldsymbol{\theta}}}\|_{2\to J} = \sqrt{|I_{\boldsymbol{\theta}}|} \quad \text{and} \quad J^\circ(e_{\boldsymbol{\theta}}) = \|e_{\boldsymbol{\theta}}\|_{\infty,2} \le 1,$$

where the last bound holds as an equality whenever $\boldsymbol{\theta} \neq 0$.

We have the following oracle inequalities as corollaries of Proposition 2 and Proposition 3.

Corollary 4 *Let the data generated by (30). Consider the estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, where F satisfies Assumptions* **(H.1)**-**(H.2)**, *and J is the group Lasso (9) with L non-overlapping blocks of equal size K. Let $s(\boldsymbol{X}) = \sqrt{\max_i \big\|\boldsymbol{X}_{b_i}^\top \boldsymbol{X}_{b_i}\big\|_{2\to 2}/n}$.*

(i) $\boldsymbol{\xi}$ is a zero-mean sub-Gaussian random vector with parameter σ: suppose that
$\lambda_n \ge 3\tau\sigma s(\boldsymbol{X}) c_1 \frac{\sqrt{2\log(c_2/\delta)}\left(\sqrt{K}+\sqrt{2\log(L)}\right)}{\sqrt{n}}$, for some $\tau > 1$ and $0 < \delta < \min(c_2, 1)$, where c_1 and c_2 are the positive absolute constants in Proposition 2. Then, with probability at least $1-\delta$, the following holds

$$\frac{1}{n}\big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\boldsymbol{\theta}_0\big\|_2^2 \le \inf_{\substack{I\subset\{1,\dots,L\}\\ \boldsymbol{\theta}:\ \mathrm{supp}_{\mathcal{B}}(\boldsymbol{\theta})=I}} \left(\frac{1}{n}\big\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\big\|_2^2 + \frac{\lambda_n^2(\tau+1)^2|I|}{\tau^2\Upsilon\Big(\mathrm{Span}\{a_j\}_{j\in b_i, i\in I}, \frac{\tau+1}{\tau-1}\Big)^2} \right) + 2p\beta, \tag{34}$$

and

$$\frac{1}{n}\big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\big\|_2^2 \le \inf_{\substack{I\subset\{1,\dots,L\}\\ \boldsymbol{\theta}:\ \mathrm{supp}_{\mathcal{B}}(\boldsymbol{\theta})=I}} \left(\frac{1}{n}\big\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\big\|_2^2 + \frac{\lambda_n^2(\tau+1)^2|I|}{\tau^2\Upsilon\Big(\mathrm{Span}\{a_j\}_{j\in b_i, i\in I}, \frac{\tau+1}{\tau-1}\Big)^2} \right). \tag{35}$$

(ii) $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{Id}_n)$*: suppose that* $\lambda_n \geq \tau\sigma s(\boldsymbol{X}) \frac{\sqrt{K}+\sqrt{2\delta \log(L)}}{\sqrt{n}}$*, for some* $\tau > 1$ *and* $\delta > 1$*. Then, with probability at least* $1 - L^{1-\delta}$*,* (34) *and* (35) *hold.*

When $s(\boldsymbol{X}) = O(1)$[8], the first remainder term is on the order $\frac{|I|\left(\sqrt{K}+\sqrt{2\log(L)}\right)^2}{n}$. This is similar to the scaling that has been provided in the literature for EWA with other group sparsity priors and noises Rigollet and Tsybakov (2012); Duy Luu et al (2016). Similar rates were given for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ with the group Lasso in Negahban et al (2012); Lounici et al (2011); van de Geer (2014).

Proof

(i) This is a consequence of Proposition 2, for which we need to bound

$$w(\boldsymbol{X}(\mathcal{C})) = \mathbb{E}\left[\max_{i\in\{1,\dots,L\}} \left\|\boldsymbol{X}_{b_i}^\top \boldsymbol{g}\right\|_2\right].$$

We first have, for any block b_i

$$\mathbb{E}\left[\left\|\boldsymbol{X}_{b_i}^\top \boldsymbol{g}\right\|_2\right] \leq \mathbb{E}\left[\left\|\boldsymbol{X}_{b_i}^\top \boldsymbol{g}\right\|_2^2\right]^{1/2} \leq s(\boldsymbol{X})\sqrt{Kn}.$$

Furthermore, $\left\|\boldsymbol{X}_{b_i}^\top \cdot\right\|_2$ is Lipschitz continuous with Lipschitz constant $s(\boldsymbol{X})\sqrt{n}$. Thus the union bound and Gaussian concentration of Lipschitz functions Ledoux (2001) yield, for any $t > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\max_{i\in\{1,\dots,L\}} \left\|\boldsymbol{X}_{b_i}^\top \boldsymbol{g}\right\|_2 \geq s(\boldsymbol{X})\sqrt{Kn} + t\right) \\
&\leq \sum_{i=1}^{L} \mathbb{P}\left(\left\|\boldsymbol{X}_{b_i}^\top \boldsymbol{\xi}\right\|_2 - \mathbb{E}\left[\left\|\boldsymbol{X}_{b_i}^\top \boldsymbol{\xi}\right\|_2\right] \geq t\right) \leq L \exp\left(-\frac{t^2}{2s(\boldsymbol{X})^2 n}\right).
\end{aligned}$$

Let $\kappa = s(\boldsymbol{X})\left(\sqrt{Kn} + \sqrt{2n\log(L)}\right)$. $w(\boldsymbol{X}(\mathcal{C}))$ can be expressed as

$$\begin{aligned}
w(\boldsymbol{X}(\mathcal{C})) &= \int_0^\infty \mathbb{P}\left(\max_{i\in\{1,\dots,L\}} \left\|\boldsymbol{X}_{b_i}^\top \boldsymbol{g}\right\|_2 \geq u\right) du \\
&\leq \int_0^\kappa du + \int_\kappa^\infty e^{-\frac{(u-s(\boldsymbol{X})\sqrt{Kn})^2 - 2s(\boldsymbol{X})^2 n\log(L)}{2n}} du \\
&= \kappa + s(\boldsymbol{X})\sqrt{n}\int_{\kappa/(s(\boldsymbol{X})\sqrt{n})}^\infty e^{-\frac{(s-\sqrt{K})^2 - 2\log(L)}{2}} du \\
&\leq \kappa + s(\boldsymbol{X})\sqrt{n}\int_{\kappa/(s(\boldsymbol{X})\sqrt{n})}^\infty e^{-\frac{s-\kappa/(s(\boldsymbol{X})\sqrt{n})}{2}} du = \kappa + 2s(\boldsymbol{X})\sqrt{n} \leq 3\kappa.
\end{aligned}$$

[8] This is for instance the case if $\boldsymbol{X}$ is drawn from the standard Gaussian ensemble and $K = O(n)$ (the $O(\cdot)$ is in fact even $o(\cdot)$ as the remainder term is supposed to go to 0 as $n \to +\infty$). In this case, classical concentration bounds of the largest eigenvalue of a Wishart matrix allow to conclude that $s(\boldsymbol{X}) = O(1 + \sqrt{K/n}) = O(1)$ with high probabilty.

(ii) The proof follows the lines of Proposition 3 where we additionally use the union bound. Indeed,

$$\begin{aligned}
&\mathbb{P}\left(\max_{i\in\{1,\dots,L\}} \left\|\boldsymbol{X}_{b_i}^\top\boldsymbol{\xi}\right\|_2 \geq \lambda_n n/\tau\right)\\
&\leq \sum_{i=1}^{L}\mathbb{P}\left(\left\|\boldsymbol{X}_{b_i}^\top\boldsymbol{\xi}\right\|_2 - \mathbb{E}\left[\left\|\boldsymbol{X}_{b_i}^\top\boldsymbol{\xi}\right\|_2\right] \geq \lambda_n n/\tau - \mathbb{E}\left[\left\|\boldsymbol{X}_{b_i}^\top\boldsymbol{\xi}\right\|_2\right]\right)\\
&\leq \sum_{i=1}^{L}\mathbb{P}\left(\left\|\boldsymbol{X}_{b_i}^\top\boldsymbol{\xi}\right\|_2 - \mathbb{E}\left[\left\|\boldsymbol{X}_{b_i}^\top\boldsymbol{\xi}\right\|_2\right] \geq \lambda_n n/\tau - \sigma s(\boldsymbol{X})\sqrt{Kn}\right)\\
&\leq \sum_{i=1}^{L}\mathbb{P}\left(\left\|\boldsymbol{X}_{b_i}^\top\boldsymbol{\xi}\right\|_2 - \mathbb{E}\left[\left\|\boldsymbol{X}_{b_i}^\top\boldsymbol{\xi}\right\|_2\right] \geq \sigma s(\boldsymbol{X})\sqrt{2\delta n\log(L)}\right)\\
&\leq L\exp\left(-\delta\log(L)\right) = L^{1-\delta},
\end{aligned}$$

where used the Gaussian concentration of Lipschitz functions Ledoux (2001) in the last inequality.

We observe in passing that another way to prove the oracle inequalities in the sub-Gaussian is to use Dudley's inequality on the sphere in $\mathbb{R}^K$ after applying a union bound on the L blocks. In addition, in the Gaussian case, the (similar) bound $\lambda_n \geq 3\delta\tau\sigma s(\boldsymbol{X})\frac{\sqrt{K}+\sqrt{2\log(L)}}{\sqrt{n}}$ can be obtained by combining Proposition 3 and the estimate $w(\boldsymbol{X}(\mathcal{C})) \leq 3s(\boldsymbol{X})(\sqrt{Kn}+\sqrt{2n\log(L)})$ in the proof of (i). The corresponding probability of success would be at least $1-L^{-9(\delta-1)^2}$.

4.3.3 Analysis group Lasso

We now turn to the prior penalty (11). Recall the notations in Section 2.5.3, and remind $\Lambda_{\boldsymbol{\theta}} = \bigcup_{i\in\mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top\boldsymbol{\theta})} b_i$. We assume that $\boldsymbol{D}$ is a frame of $\mathbb{R}^p$, hence surjective, meaning that there exist $c, d > 0$ such that for any $\boldsymbol{\omega}\in\mathbb{R}^p$

$$d\left\|\boldsymbol{\omega}\right\|_2^2 \leq \left\|\boldsymbol{D}^\top\boldsymbol{\omega}\right\|_2^2 \leq c\left\|\boldsymbol{\omega}\right\|_2^2.$$

This together with (12)-(13) and Cauchy-Schwarz inequality entail

$$\begin{aligned}
\left\|\mathrm{P}_{T_{\boldsymbol{\theta}}}\right\|_{2\to J} = \sup_{\left\|\boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}\right\|_2\leq 1}\left\|\boldsymbol{D}^\top\boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}\right\|_{1,2} &\leq \sqrt{c}\sup_{\left\|\boldsymbol{D}^\top\boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}\right\|_2\leq 1}\left\|\boldsymbol{D}^\top\boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}\right\|_{1,2}\\
&= \sqrt{c}\sup_{\left\|\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}}^\top\boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}\right\|_2\leq 1}\left\|\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}}^\top\boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}\right\|_{1,2}\\
&= \sqrt{c}\sqrt{|\mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top\boldsymbol{\theta})|}.
\end{aligned}$$

Note, however, that from (12), we do not have in general $C(\boldsymbol{D},\boldsymbol{\theta}) \overset{\text{def}}{=} \left\|\boldsymbol{D}^+\,\mathrm{P}_{\mathrm{Ker}(\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}^{\mathrm{c}}}^\top)}\,\boldsymbol{D}e^{|||_{1,2}}_{\boldsymbol{D}^\top\boldsymbol{\theta}}\right\|_{\infty,2} \leq 1$.

With exactly the same arguments to those for proving Corollary 4, replacing $\boldsymbol{X}$ by $\boldsymbol{X}\boldsymbol{D}$, we arrive at the following oracle inequalities.

Corollary 5 *Let the data generated by (30). Consider the estimators* $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ *and* $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$*, where* F *satisfies Assumptions* **(H.1)-(H.2)***, and* J *is the analysis group Lasso (11) with* L *blocks of equal size* K*. Assume that* $\boldsymbol{D}$ *is a frame, and le* $s(\boldsymbol{XD}) = \sqrt{\max_i \left\| \boldsymbol{D}_{b_i}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{D}_{b_i} \right\|_{2\to 2}}/n$.

(i) $\boldsymbol{\xi}$ *is a zero-mean sub-Gaussian random vector with parameter* σ*: suppose that* $\lambda_n \geq 3\tau\sigma s(\boldsymbol{XD}) c_1 \frac{\sqrt{\log(c_2/\delta)}\left(\sqrt{K}+\sqrt{2\log(L)}\right)}{\sqrt{n}}$*, for some* $\tau > 1$ *and* $0 < \delta < \min(c_2, 1)$*, where* c_1 *and* c_2 *are the positive absolute constants in Proposition 2. Then, with probability at least* $1-\delta$*, the following holds*

$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 \leq \inf_{\substack{I\subset\{1,\dots,L\} \\ \boldsymbol{\theta}:\ \mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top\boldsymbol{\theta})=I}} \left(\frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 + \frac{c\lambda_n^2\left(\tau C(\boldsymbol{D},\boldsymbol{\theta})+1\right)^2|I|}{\tau^2\Upsilon\left(\mathrm{Ker}(\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}^c}^\top), \frac{\tau C(\boldsymbol{D},\boldsymbol{\theta})+1}{\tau-1}\right)^2} \right) + 2p\beta, \tag{36}$$

and

$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 \leq \inf_{\substack{I\subset\{1,\dots,L\} \\ \boldsymbol{\theta}:\ \mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top\boldsymbol{\theta})=I}} \left(\frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 + \frac{c\lambda_n^2\left(\tau C(\boldsymbol{D},\boldsymbol{\theta})+1\right)^2|I|}{\tau^2\Upsilon\left(\mathrm{Ker}(\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}^c}^\top), \frac{\tau C(\boldsymbol{D},\boldsymbol{\theta})+1}{\tau-1}\right)^2} \right). \tag{37}$$

(ii) $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2\mathbf{Id}_n)$*: suppose that* $\lambda_n \geq \tau\sigma s(\boldsymbol{XD})\frac{\sqrt{K}+\sqrt{2\delta\log(L)}}{\sqrt{n}}$*, for some* $\tau > 1$ *and* $\delta > 1$*. Then, with probability at least* $1 - L^{1-\delta}$*, (36) and (37) hold.*

To the best of our knowledge, this result is new to the literature. The scaling of the remainder term is the same as in Duy Luu et al (2016) and Rigollet and Tsybakov (2012) with analysis sparsity priors different from ours (the authors in the latter also assume that $\boldsymbol{D}$ is invertible).

4.3.4 Anti-sparsity

Recall the derivations for the ℓ_∞ norm example in Section 3.4.1. We have the following oracle inequalities from Proposition 4.

Corollary 6 *Let the data generated by (30) where* $\boldsymbol{\xi}$ *is a zero-mean sub-Gaussian random vector with parameter* σ*. Assume that* $\boldsymbol{X}$ *is such that*

$\max_{i,j} |\boldsymbol{X}_{i,j}| \leq 1/p$. *Consider the estimators* $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ *and* $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, *where* F *satisfies Assumptions* **(H.1)**-**(H.2)**, *and* J *is the anti-sparsity penalty* (14). *Suppose that* $\lambda_n \geq \tau\sigma\sqrt{2\delta\log(2)}\sqrt{\frac{p}{n}}$, *for some* $\tau > 1$ *and* $\delta > 1$. *Then, with probability at least* $1 - 2^{-p(\delta-1)+1}$, *the following holds*

$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 \leq \inf_{\substack{I\subset\{1,\dots,p\} \\ \boldsymbol{\theta}:\ I_{\boldsymbol{\theta}}^{\mathrm{sat}}=I}} \left(\frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 + \frac{\lambda_n^2(\tau+1)^2}{\tau^2\Upsilon\left(\left\{\overline{\boldsymbol{\theta}} : \overline{\boldsymbol{\theta}}_I \in \mathbb{R}\,\mathrm{sign}(\boldsymbol{\theta}_I)\right\}, \frac{\tau+1}{\tau-1}\right)^2} \right) + 2p\beta, \tag{38}$$

and

$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 \leq \inf_{\substack{I\subset\{1,\dots,p\} \\ \boldsymbol{\theta}:\ I_{\boldsymbol{\theta}}^{\mathrm{sat}}=I}} \left(\frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 + \frac{\lambda_n^2(\tau+1)^2}{\tau^2\Upsilon\left(\left\{\overline{\boldsymbol{\theta}} : \overline{\boldsymbol{\theta}}_I \in \mathbb{R}\,\mathrm{sign}(\boldsymbol{\theta}_I)\right\}, \frac{\tau+1}{\tau-1}\right)^2} \right). \tag{39}$$

The first remainder term scales as $\frac{p}{n}$ which reflects that anti-sparsity regularization requires an overdetermined regime to ensure good stability performance. This is in agreement with (Vaiter et al 2015a, Theorem 7). This phenomenon was also observed by Donoho and Tanner (2010) who studied sample complexity thresholds for noiseless recovery from random projections of the hypercube.

4.3.5 Nuclear norm

We now turn to the nuclear norm case. Recall the notations of Section 2.5.5. For matrices $\boldsymbol{\theta} \in \mathbb{R}^{p_1\times p_2}$, a measurement map $\boldsymbol{X}$ takes the form of a linear operator whose ith component is given by the Frobenius scalar product

$$\boldsymbol{X}(\boldsymbol{\theta})_i = \mathrm{tr}((\boldsymbol{X}^i)^\top\boldsymbol{\theta}) = \langle \boldsymbol{X}^i, \boldsymbol{\theta}\rangle_{\mathrm{F}},$$

where $\boldsymbol{X}^i$ is a matrix in $\mathbb{R}^{p_1\times p_2}$. We denote $\|\cdot\|_{\mathrm{F}}$ the associated norm. From (17), it is immediate to see that whenever $\boldsymbol{\theta} \neq 0$,

$$J^\circ(e_{\boldsymbol{\theta}}) = \left\|\left|\boldsymbol{U}\boldsymbol{V}^\top\right|\right\|_{2\to 2} = 1.$$

Moreover, from (17), we have

$$\left\|\left|\mathrm{P}_{T_{\boldsymbol{\theta}}}\right|\right\|_{\mathrm{F}\to *} = \sup_{\boldsymbol{\theta}'\in T_{\boldsymbol{\theta}}} \frac{\|\boldsymbol{\theta}'\|_*}{\|\boldsymbol{\theta}'\|_{\mathrm{F}}} = \sup_{\boldsymbol{\theta}'\in T_{\boldsymbol{\theta}}} \frac{\|\lambda(\boldsymbol{\theta}')\|_1}{\|\lambda(\boldsymbol{\theta}')\|_2} \leq \sup_{\boldsymbol{\theta}'\in T_{\boldsymbol{\theta}}} \sqrt{\mathrm{rank}(\boldsymbol{\theta}')} \leq \sqrt{\min(r,p_1)+\min(r,p_2)} \leq \sqrt{2r}.$$

To apply Proposition 2 and Proposition 3, we need to bound $w(\boldsymbol{X}(\mathcal{C}))$ ($\mathcal{C}$ is the nuclear ball), or equivalently, to bound

$$\mathbb{E}\left[\|\boldsymbol{X}^*(\boldsymbol{g})\|_{2\to 2}\right] = \mathbb{E}\left[\left\|\sum_{i=1}^n \boldsymbol{X}^i \boldsymbol{g}_i\right\|_{2\to 2}\right], \quad \boldsymbol{g} \sim \mathcal{N}(0, \sigma^2 \mathbf{Id}_n),$$

which is the expectation of the operator norm of a random series with matrix coefficients. Thus using (Tropp 2015b, Theorem 4.1.1(4.1.5)) to get this bound, and inserting it into Proposition 2 and Proposition 3, we get the following oracle inequalities for the nuclear norm. Define

$$s(\boldsymbol{X}) = \sqrt{\max\left(\left\|\sum_{i=1}^n \boldsymbol{X}^i (\boldsymbol{X}^i)^\top\right\|_{2\to 2}, \left\|\sum_{i=1}^n (\boldsymbol{X}^i)^\top \boldsymbol{X}^i\right\|_{2\to 2}\right)/n}.$$

Corollary 7 *Let the data generated by (30) with a linear operator $\boldsymbol{X} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n$. Consider the estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, where F satisfies Assumptions* **(H.1)**-**(H.2)**, *and J is the nuclear norm (16).*

(i) $\boldsymbol{\xi}$ is a zero-mean sub-Gaussian random vector with parameter σ: suppose that
$\lambda_n \geq 2\tau\sigma s(\boldsymbol{X}) c_1 \sqrt{\frac{\log(c_2/\delta)\log(p_1+p_2)}{n}}$, for some $\tau > 1$ and $0 < \delta < \min(c_2, 1)$, where c_1 and c_2 are the positive absolute constants in Proposition 2. Then, with probability at least $1-\delta$, the following holds

$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 \leq \inf_{\substack{r\in\{1,\dots,\min(p_1,p_2)\} \\ \boldsymbol{\theta}:\ \mathrm{rank}(\boldsymbol{\theta})=r}} \left(\frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 + \frac{2\lambda_n^2(\tau+1)^2 r}{\tau^2 \Upsilon\left(T_{\boldsymbol{\theta}}, \frac{\tau+1}{\tau-1}\right)^2} \right) + 2p_1p_2\beta, \tag{40}$$

and

$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 \leq \inf_{\substack{r\in\{1,\dots,\min(p_1,p_2)\} \\ \boldsymbol{\theta}:\ \mathrm{rank}(\boldsymbol{\theta})=r}} \left(\frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 + \frac{2\lambda_n^2(\tau+1)^2 r}{\tau^2 \Upsilon\left(T_{\boldsymbol{\theta}}, \frac{\tau+1}{\tau-1}\right)^2} \right). \tag{41}$$

(ii) $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{Id}_n)$: suppose that $\lambda_n \geq (1+\delta)\tau\sigma s(\boldsymbol{X})\sqrt{\frac{2\log(p_1+p_2)}{n}}$, for some $\tau > 1$ and $\delta > 0$. Then, with probability at least $1 - (p_1+p_2)^{-\delta^2}$, (40) and (41) hold.

The set over which the infimum is taken just reminds us that the nuclear norm is partly smooth (see above) relative to the constant rank manifold

(which is a Riemannian submanifold of $\mathbb{R}^{p_1 \times p_2}$) (Daniilidis et al 2014, Theorem 3.19). In the iid Gaussian noise case, we recover the same rate as in (Dalalyan et al 2018, Theorem 3) for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and in (Koltchinskii et al 2011, Theorem 2) for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$. If $s(\boldsymbol{X}) = O(\sqrt{p_1 + p_2})$, then the first remainder term scales as $\frac{r(p_1+p_2)\log(p_1+p_2)}{n}$. For low-rank matrix recovery, the same rate was also independently proved in Mai and Alquier (2015); Suzuki (2015)[9] for EWA and the posterior conditional mean respectively, in the temperature regime $\beta = C/n$, though with completely different priors, but without requiring the compatibility factor assumption.

The assumption $s(\boldsymbol{X}) = O(\sqrt{p_1 + p_2})$ on the design is mild and verified in many situations. Indeed, by Jensen's inequality we have

$$s(\boldsymbol{X})^2 \leq n^{-1} \sum_{i=1}^{n} \max\left(\left|\!\left|\!\left| \boldsymbol{X}^i (\boldsymbol{X}^i)^\top \right|\!\right|\!\right|_{2\to 2}, \left|\!\left|\!\left| (\boldsymbol{X}^i)^\top \boldsymbol{X}^i \right|\!\right|\!\right|_{2\to 2} \right) \leq \left|\!\left|\!\left| \boldsymbol{X}^i \right|\!\right|\!\right|_{2\to 2}^2.$$

If, for example, $(\boldsymbol{X}^i)_i$ are independent copies of a standard random Gaussian matrix, then classical concentration bounds of the largest eigenvalue of a Wishart matrix entail that $\left|\!\left|\!\left| \boldsymbol{X}^i \right|\!\right|\!\right|_{2\to 2}$ concentrates around its mean $\mathbb{E}\left[\left|\!\left|\!\left| \boldsymbol{X}^i \right|\!\right|\!\right|_{2\to 2} \right] \leq \sqrt{p_1} + \sqrt{p_2} \leq \sqrt{2(p_1 + p_2)}$.

4.4 Discussion of minimax optimality

In this section, we discuss the optimality of the estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ (we remind the reader that the design $\boldsymbol{X}$ is fixed). Recall the discussion on stratification at the end of Section 3.1. Let $\mathcal{M}_0 \in \mathscr{M}$ be the stratum active at $\boldsymbol{\theta}_0 \in \mathcal{M}_0$. In this setting, choosing $\beta = C(1+\delta)^2\sigma^2 w(\boldsymbol{X}(\mathcal{C}))^2 \left|\!\left|\!\left| \mathrm{P}_{T_{\boldsymbol{\theta}_0}} \right|\!\right|\!\right|_{2\to J}^2 / (pn^2)$ for some constant $C > 0$, (22) and Proposition 3 ensure that

$$\begin{aligned}
&\tfrac{1}{n}\left\| \boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\boldsymbol{\theta}_0 \right\|_2^2 \\
&\leq \frac{(1+\delta)^2\sigma^2 w(\boldsymbol{X}(\mathcal{C}))^2 \left|\!\left|\!\left| \mathrm{P}_{T_{\boldsymbol{\theta}_0}} \right|\!\right|\!\right|_{2\to J}^2}{n^2} \left(\frac{\left(\tau J^\circ(e_{\boldsymbol{\theta}_0}) + 1\right)^2}{\Upsilon\left(T_{\boldsymbol{\theta}_0}, \frac{\tau J^\circ(e_{\boldsymbol{\theta}_0})+1}{\tau-1}\right)^2} + C \right) \\
&\tfrac{1}{n}\left\| \boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0 \right\|_2^2 \\
&\leq \frac{(1+\delta)^2\sigma^2 w(\boldsymbol{X}(\mathcal{C}))^2 \left|\!\left|\!\left| \mathrm{P}_{T_{\boldsymbol{\theta}_0}} \right|\!\right|\!\right|_{2\to J}^2}{n^2} \frac{\left(\tau J^\circ(e_{\boldsymbol{\theta}_0}) + 1\right)^2}{\Upsilon\left(T_{\boldsymbol{\theta}_0}, \frac{\tau J^\circ(e_{\boldsymbol{\theta}_0})+1}{\tau-1}\right)^2},
\end{aligned}$$

with high probability. In particular, for a polyhedral gauge penalty, in which case $\mathcal{M}_0 = T_{\boldsymbol{\theta}_0}$ (see Vaiter et al (2015a)), and under the normalization

[9] The noise is iid Gaussian in Suzuki (2015) and subexponential in Mai and Alquier (2015). The assumptions on the design are also different.

$\max_{\boldsymbol{v}\in\mathcal{V}} \|\boldsymbol{X}\boldsymbol{v}\|_2 \leq \sqrt{n}$ and with the choice $\beta = 2C\delta\sigma^2 \|\!|\mathrm{P}_{\mathcal{M}_0}|\!\|^2_{2\to J} \log(|\mathcal{V}|)/(pn)$, Proposition 4 entails

$$\frac{1}{n}\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\boldsymbol{\theta}_0\|_2^2 \leq \frac{2\delta\sigma^2 \|\!|\mathrm{P}_{\mathcal{M}_0}|\!\|^2_{2\to J} \log(|\mathcal{V}|)}{n} \left(\frac{\left(\tau J^\circ(e_{\boldsymbol{\theta}_0}) + 1\right)^2}{\Upsilon\left(\mathcal{M}_0, \frac{\tau J^\circ(e_{\boldsymbol{\theta}_0})+1}{\tau-1}\right)^2} + C \right)$$

$$\frac{1}{n}\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\|_2^2 \leq \frac{2\delta\sigma^2 \|\!|\mathrm{P}_{\mathcal{M}_0}|\!\|^2_{2\to J} \log(|\mathcal{V}|)}{n} \frac{\left(\tau J^\circ(e_{\boldsymbol{\theta}_0}) + 1\right)^2}{\Upsilon\left(\mathcal{M}_0, \frac{\tau J^\circ(e_{\boldsymbol{\theta}_0})+1}{\tau-1}\right)^2},$$

with high probability. Thus the risk bounds only depend on $\mathcal{M}_0$. A natural question that arises is whether the above bounds are optimal, i.e. whether an estimator can achieve a significantly better prediction risk than $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ uniformly on $\mathcal{M}_0$. A classical way to answer this question is the minimax point of view. This amounts to finding a lower bound on the minimax probabilities of the form

$$\inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}\in\mathcal{M}_0} \Pr\left(\tfrac{1}{n}\|\boldsymbol{X}\widehat{\boldsymbol{\theta}} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 \geq \psi_n\right),$$

where ψ_n is the rate, which ideally, should be comparable to the risk bounds above. A standard path to derive such a lower bound is to exhibit a subset of $\mathcal{M}_0$ of well-separated points while controlling its diameter, see (Tsybakov 2008, Chapter 2) or (Massart 2007, Section 4.3). This however must be worked out on a case-by-case basis.

Example 2 (Lasso) In this case, $\mathcal{M}_0 = T_{\boldsymbol{\theta}_0}$ is the subspace of vectors whose support is contained in that of $\boldsymbol{\theta}_0$. Let $I = \mathrm{supp}(\boldsymbol{\theta}_0)$ and $s = \|\boldsymbol{\theta}_0\|_0$. Define the set

$$\mathcal{B}_0 = \left\{\boldsymbol{\theta} \in \mathbb{R}^p \;:\; \boldsymbol{\theta}_I \in \{0,1\}^s \quad \text{and} \quad \boldsymbol{\theta}_{I^c} = 0\right\}.$$

We have $\mathcal{B}_0 \subset \mathcal{M}_0$ and $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_0 \leq 2s$ for all $(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \mathcal{B}_0$. Define $\mathcal{F}_0 \stackrel{\text{def}}{=} \left\{r\boldsymbol{X}\boldsymbol{\theta} \;:\; \boldsymbol{\theta} \in \mathcal{B}_0\right\}$, for $r > 0$ to be specified later. Due to the Varshamov-Gilbert lemma (Massart 2007, Lemma 4.7), given $a \in]0,1[$, there exists a subset $\mathcal{B} \subset \mathcal{B}_0$ with cardinality $|\mathcal{B}| \geq 2^{\rho s/2}$ such that for two distinct elements $\boldsymbol{X}\boldsymbol{\theta}$ and $\boldsymbol{X}\boldsymbol{\theta}'$ in $\mathcal{F}_0$

$$\|\boldsymbol{X}(\boldsymbol{\theta} - \boldsymbol{\theta}')\|_2^2 \geq \underline{\kappa} r^2 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 \geq 2(1-a)\underline{\kappa} r^2 s,$$
$$\|\boldsymbol{X}(\boldsymbol{\theta} - \boldsymbol{\theta}')\|_2^2 \leq \overline{\kappa} r^2 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 \leq 4\overline{\kappa} r^2 s,$$

where

$$\underline{\kappa} = \inf_{\boldsymbol{\theta}\in\mathcal{M}_0} \frac{\|\boldsymbol{X}\boldsymbol{\theta}\|_2^2}{\|\boldsymbol{\theta}\|_2^2} \leq \overline{\kappa} = \sup_{\boldsymbol{\theta}\in\mathcal{M}_0} \frac{\|\boldsymbol{X}\boldsymbol{\theta}\|_2^2}{\|\boldsymbol{\theta}\|_2^2}.$$

Standard results from random matrix theory, see Tropp (2015a), ensure that $\underline{\kappa} > 0$ for a Gaussian design with high probability as long as $n \geq s + C\sqrt{s}$ for some positive absolute constant C.

Then choosing $r^2 = \frac{c\rho\sigma^2}{4\overline{\kappa}}$, where $c \in]0, 1/8[$ and $\rho = (1+a)\log(1+a) + (1-a)\log(1-a)$, we get the bounds

$$\left\|\boldsymbol{X}(\boldsymbol{\theta}-\boldsymbol{\theta}')\right\|_2^2 \geq \frac{\sigma^2 c(1-a)\rho\underline{\kappa}}{2\overline{\kappa}} s,$$
$$\left\|\boldsymbol{X}(\boldsymbol{\theta}-\boldsymbol{\theta}')\right\|_2^2 \leq 2\sigma^2 c \log(|\mathcal{B}|).$$

We are now in position to apply (Tsybakov 2008, Theorem 2.5) to conclude that there exists $\eta \in]0, 1[$ (that depends on a) such that

$$\inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}\in\mathcal{M}_0} \Pr\left(\tfrac{1}{n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2 \geq \frac{\sigma^2 c(1-a)\rho\underline{\kappa}}{4\overline{\kappa}} \frac{s}{n}\right) \geq \eta.$$

This lower bound together with Corollary 3 show that $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ (with $\beta = O\big(\sigma^2 s\log(2p)/(pn)\big)$) and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ are nearly minimax (up to a logarithmic factor) over $\mathcal{M}_0$.

One can generalize this reasoning to get a minimax lower bound over the larger class of s-sparse vectors, i.e. $\bigcup\big\{V = \mathrm{Span}\{(\boldsymbol{a}_j)_{1\leq j\leq p}\} \,:\, \dim(V) = s\big\}$, which is a finite union of subspaces that contains $\mathcal{M}_0$. Let $(a, b) \in]0, 1[^2$ such that $1 \leq s \leq abp$ and $a(-1+b-\log(b)) \geq \log(2)$ [10], $c \in]0, 1/8[$. Then combining (Tsybakov 2008, Theorem 2.5) and (Massart 2007, Lemma 4.6 and Lemma 4.10), we have for $\eta \stackrel{\mathrm{def}}{=} \frac{1}{1+(ab)^{\rho s/2}}\Big(1 - 2c - \sqrt{\frac{2c}{-\rho\log(ab)}}\Big) \in]0, 1[$

$$\inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}\in\mathcal{M}_0} \Pr\left(\tfrac{1}{n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2 \geq \frac{\sigma^2 c\rho(1-\alpha)\underline{\kappa}}{2\overline{\kappa}} \frac{s\log(p/s)}{n}\right) \geq \eta,$$

where $\rho = -a(-1+b-\log(b))/\log(ab)$, and $\underline{\kappa}$ and $\overline{\kappa}$ are now the restricted isometry constants of $\boldsymbol{X}$ of degree $2s$, i.e.

$$\underline{\kappa} = \inf_{\|\boldsymbol{\theta}\|_0\leq 2s} \frac{\|\boldsymbol{X}\boldsymbol{\theta}\|_2^2}{\|\boldsymbol{\theta}\|_2^2} \leq \overline{\kappa} = \sup_{\|\boldsymbol{\theta}\|_0\leq 2s} \frac{\|\boldsymbol{X}\boldsymbol{\theta}\|_2^2}{\|\boldsymbol{\theta}\|_2^2}.$$

For this lower bound to be meaningful, $\underline{\kappa}$ should be positive. From the compressed sensing literature, many random designs are known to verify this condition for n large enough compared to s, e.g. sub-Gaussian designs with $n \gtrsim s\log(p)$.

One can see that the difference between this lower bound and the one on $\mathcal{M}_0$ lies in the $\log(p/s)$ factor, which basically derives from the control over the union of subspaces. The minimax prediction risk (in expectation) over the ℓ_0-ball were studied in Rigollet and Tsybakov (2011); Raskutti et al (2011); Verzelen (2012); Ye and Zhang (2010); Wang et al (2014), where similar lower bounds were obtained.

[10] E.g. take $b = 1/(1 + e\sqrt[a]{2})$.

Example 3 (Group Lasso) For the group Lasso with L groups of equal size K, $\mathcal{M}_0$ is the subspace group sparse vectors whose group support is included in that of $\boldsymbol{\theta}_0$. Let s be the number of non-zero (active) groups in $\boldsymbol{\theta}_0$. Following exactly the same reasoning as for the Lasso, one can show that the risk lower bound in probability scales as $C\sigma^2 sK/n$, which together with Corollary 4, shows that $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ (with $\beta = O\big(\sigma^2 s\big(\sqrt{K} + \sqrt{2\log(L)}\big)^2/(pn)\big)$) and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ are nearly minimax (up again to a logarithmic factor) over $\mathcal{M}_0$. One can also derive the lower bound $C\sigma^2 s(K + \log(L/s))/n$ over the set of s-block sparse vectors. Such minimax lower bound is comparable to the one in Lounici et al (2011).

Example 4 (Anti-sparsity) Denote the saturation support of $\boldsymbol{\theta}_0$ as I^{sat} and recall the subspace $T_{\boldsymbol{\theta}_0}$ form (15). Thus, $\mathcal{M}_0 = T_{\boldsymbol{\theta}_0}$ is the subspace of vectors which are collinear to $\mathrm{sign}(\boldsymbol{\theta}_0)$ on I^{sat} and free on its complement. Observe that $\dim(\mathcal{M}_0) = p - s + 1$, where $s = |I^{\mathrm{sat}}|$. Define the set

$$\mathcal{B}_0 = \big\{\boldsymbol{\theta} \in \mathbb{R}^p \;:\; \boldsymbol{\theta}_{I^{\mathrm{sat}}} = \mathrm{sign}(\boldsymbol{\theta}_{I^{\mathrm{sat}}}) \quad \text{and} \quad \boldsymbol{\theta}_{(I^{\mathrm{sat}})^c} \in \{0,1\}^{p-s}\big)\big\}.$$

By construction, $\mathcal{B}_0 \subset \mathcal{M}_0$, and $\big\|\boldsymbol{\theta} - \boldsymbol{\theta}'\big\|_0 \leq 2(p-s)$ for all $(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \mathcal{B}_0$. Thus following the same arguments as for the Lasso example (using again Varshamov-Gilbert lemma and (Tsybakov 2008, Theorem 2.5)), we conclude that there exists $\eta \in]0,1[$ (that depends on a) such that

$$\inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in \mathcal{M}_0} \Pr\left(\tfrac{1}{n}\big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}} - \boldsymbol{X}\boldsymbol{\theta}\big\|_2^2 \geq \frac{\sigma^2 c(1-a)\rho_{\underline{\kappa}}}{4\overline{\kappa}} \frac{p-s}{n}\right) \geq \eta,$$

where the restricted isometry constants are defined similarly to the Lasso but with respect to the model subspace $\mathcal{M}_0$ of the ℓ_∞ norm. Again, for a Gaussian design, $\underline{\kappa} > 0$ with high probability as long as $n \geq (p-s+1) + C\sqrt{p-s+1}$ Tropp (2015a).

The obtained minimax lower bound is consistent with the sample complexity thresholds derived in Donoho and Tanner (2010) for noiseless recovery from random projections of the hypercube. For a saturation support size s small compared to p, the bound of Corollary 6 (with $\beta = O\big(\sigma^2/n^2\big)$) comes close to the minimax lower bound.

Example 5 (Nuclear norm) Let $r = \mathrm{rank}(\boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0 \in \mathbb{R}^{p_1 \times p_2}$, and $p = \max(p_1, p_2)$. For the nuclear norm, $\mathcal{M}_0$ is the manifold of rank-r matrices. Thus arguing as in (Koltchinskii et al 2011, Theorem 5) (who use the Varshamov-Gilbert lemma Massart (2007) to find the covering set), one can show that the minimax risk lower bound over $\mathcal{M}_0$ is $C\sigma^2 r/n$. In view of Corollary 7, we deduce that $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ (with $\beta = O\big(\sigma^2 r \log(p_1+p_2)/(p_1 p_2 n)\big)$) and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ are nearly minimax over the constant rank manifolds.

A Pre-requisites from convex analysis

We here collect some ingredients from convex analysis that are essential to our exposition.

Monotone conjugate

Lemma 2 *Let g be a non-decreasing function on $\mathbb{R}_+$ that vanishes at 0. Then the following hold:*

(i) g^+ *is a proper closed convex and non-decreasing function on* $\mathbb{R}_+$ *that vanishes at* 0.
(ii) If g is also closed and convex, then $g^{+\,+} = g$.
(iii) Let $f : t \in \mathbb{R} \mapsto g(|t|)$ *such that f is differentiable on* $\mathbb{R}$*, where g is finite-valued, strictly convex and strongly coercive. Then* g^+ *is likewise finite-valued, strictly convex, strongly coercive, and* $f^* = g^+ \circ |\cdot|$ *is differentiable on* $\mathbb{R}$*. In particular, both g and* g^+ *are strictly increasing on* $\mathbb{R}_+$.

Proof (i) By (Bauschke and Combettes 2011, Proposition 13.11), g^+ is a closed convex function. We have $\inf_{t\geq 0} g(t) = -\sup_{t\geq 0} t\cdot 0 - g(t) = -g^+(0)$. Since g is non-decreasing and $g(0) = 0$, then $g^+(0) = -\inf_{t\geq 0} g(t) = -g(0) = 0$. In addition, by (5), we have $g^+(a) \geq a\cdot 0 - g(0) = 0, \forall a \in \mathbb{R}_+$. This shows that g^+ is non-negative and $\mathrm{dom}(g^+) \neq \emptyset$, and in turn, it is also proper.
Let a, b in $\mathbb{R}_+$ such that $a < b$. Then

$$g^+(a) - g^+(b) = (\sup_{t\geq 0} ta - g(t)) - (\sup_{t'\geq 0} t'b - g(t')) \leq \sup_{t\geq 0}(ta - g(t) - tb + g(t)) = \sup_{t\geq 0} t(a-b) = 0.$$

That is, g^+ is non-decreasing on $\mathbb{R}_+$.
(ii) This follows from (Rockafellar 1996, Theorem 12.4).
(iii) By definition of f, f is a finite-valued function on $\mathbb{R}$, strictly convex, differentiable and strongly coercive. It then follows from (Hiriart-Urruty and Lemaréchal 2001, Corollary X.4.1.4) that f^* enjoys the same properties. In turn, using the fact that both f and f^* are even, we have g^+ is strongly coercive, and strict convexity of f (resp. f^*) is equivalent to that of g (resp. g^+). Altogether, this shows the first claim. We now prove that g vanishes only at 0 (and similary for g^+). As g is non-decreasing and strictly convex, we have, for any $\rho \in]0, 1[$ and a, b in $\mathbb{R}_+$ such that $a < b$,

$$g(a) \leq g(\rho a + (1-\rho)b) < \rho g(a) + (1-\rho)g(b) \leq \rho g(b) + (1-\rho)g(b) = g(b).$$

Support function The *support function* of $\mathcal{C} \subset \mathbb{R}^p$ is

$$\sigma_{\mathcal{C}}(\boldsymbol{\omega}) = \sup_{\boldsymbol{\theta}\in\mathcal{C}} \langle \boldsymbol{\omega}, \boldsymbol{\theta} \rangle.$$

We recall the following properties whose proofs can be found in e.g. Rockafellar (1996); Hiriart-Urruty and Lemaréchal (2001).

Lemma 3 *Let $\mathcal{C}$ be a non-empty set.*

(i) $\sigma_{\mathcal{C}}$ *is proper lower semicontonuous (lsc) and sublinear.*
(ii) $\sigma_{\mathcal{C}}$ *is finite-valued if and only if $\mathcal{C}$ is bounded.*
(iii) If $0 \in \mathcal{C}$*, then* $\sigma_{\mathcal{C}}$ *is non-negative.*
(iv) If $\mathcal{C}$ is convex and compact with $0 \in \mathrm{int}(\mathcal{C})$*, then* $\sigma_{\mathcal{C}}$ *is finite-valued and coercive.*

Gauges and polars

Definition 3 (Polar set) Let $\mathcal{C}$ be a nonempty convex set. The set $\mathcal{C}^\circ$ given by

$$\mathcal{C}^\circ = \left\{ \boldsymbol{\eta} \in \mathbb{R}^p \ : \ \langle \boldsymbol{\eta}, \boldsymbol{\theta} \rangle \leq 1 \quad \forall \boldsymbol{\theta} \in \mathcal{C} \right\}$$

is called the *polar* of $\mathcal{C}$.

The set $\mathcal{C}^\circ$ is closed convex and contains the origin. When $\mathcal{C}$ is also closed and contains the origin, then it coincides with its bipolar, i.e. $\mathcal{C}^{\circ\circ} = \mathcal{C}$.

Let $\mathcal{C} \subseteq \mathbb{R}^p$ be a non-empty closed convex set containing the origin. The *gauge* of $\mathcal{C}$ is the function $\gamma_{\mathcal{C}}$ defined on $\mathbb{R}^p$ by

$$\gamma_{\mathcal{C}}(\boldsymbol{\theta}) = \inf\left\{\lambda > 0 \ : \ \boldsymbol{\theta} \in \lambda\mathcal{C}\right\}.$$

As usual, $\gamma_{\mathcal{C}}(\boldsymbol{\theta}) = +\infty$ if the infimum is not attained.

Lemma 4 hereafter recaps the main properties of a gauge that we need. In particular, (ii) is a fundamental result of convex analysis that states that there is a one-to-one correspondence between gauge functions and closed convex sets containing the origin. This allows to identify sets from their gauges, and vice versa.

Lemma 4

(i) $\gamma_{\mathcal{C}}$ *is a non-negative, lsc and sublinear function.*
(ii) $\mathcal{C}$ *is the unique closed convex set containing the origin such that*

$$\mathcal{C} = \left\{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \gamma_{\mathcal{C}}(\boldsymbol{\theta}) \leq 1\right\}.$$

(iii) $\gamma_{\mathcal{C}}$ *is finite-valued if, and only if,* $0 \in \mathrm{int}(\mathcal{C})$*, in which case* $\gamma_{\mathcal{C}}$ *is Lipschitz continuous.*
(iv) $\gamma_{\mathcal{C}}$ *is finite-valued and coercive if, and only if,* $\mathcal{C}$ *is compact and* $0 \in \mathrm{int}(\mathcal{C})$.

See Vaiter et al (2015a) for the proof.

Observe that thanks to sublinearity, local Lipschitz continuity valid for any finite-valued convex function is streghthned to global Lipschitz continuity. Moreover, $\gamma_{\mathcal{C}}$ is a norm, having $\mathcal{C}$ as its unit ball, if and only if $\mathcal{C}$ is bounded with nonempty interior and symmetric.

We now define the polar gauge.

Definition 4 (Polar Gauge) The polar of a gauge $\gamma_{\mathcal{C}}$ is the function $\gamma^\circ_{\mathcal{C}}$ defined by

$$\gamma^\circ_{\mathcal{C}}(\boldsymbol{\omega}) = \inf\left\{\mu \geq 0 \ : \ \langle\boldsymbol{\theta}, \boldsymbol{\omega}\rangle \leq \mu\gamma_{\mathcal{C}}(\boldsymbol{\theta}), \forall\boldsymbol{\theta}\right\}.$$

An immediate consequence is that gauges polar to each other have the property

$$\langle\boldsymbol{\theta}, \boldsymbol{u}\rangle \leq \gamma_{\mathcal{C}}(\boldsymbol{\theta})\gamma^\circ_{\mathcal{C}}(\boldsymbol{u}) \quad \forall(\boldsymbol{\theta}, \boldsymbol{u}) \in \mathrm{dom}(\gamma_{\mathcal{C}}) \times \mathrm{dom}(\gamma^\circ_{\mathcal{C}}), \tag{42}$$

just as dual norms satisfy a duality inequality. In fact, polar pairs of gauges correspond to the best inequalities of this type.

Lemma 5 *Let* $\mathcal{C} \subseteq \mathbb{R}^p$ *be a closed convex set containing the origin. Then,*

(ii) $\gamma^\circ_{\mathcal{C}}$ *is a gauge function and* $\gamma^{\circ\circ}_{\mathcal{C}} = \gamma_{\mathcal{C}}$.
(iii) $\gamma^\circ_{\mathcal{C}} = \gamma_{\mathcal{C}^\circ}$*, or equivalently*

$$\mathcal{C}^\circ = \left\{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \gamma^\circ_{\mathcal{C}}(\boldsymbol{\theta}) \leq 1\right\}.$$

(iv) The gauge of $\mathcal{C}$ *and the support function of* $\mathcal{C}$ *are mutually polar, i.e.*

$$\gamma_{\mathcal{C}} = \sigma_{\mathcal{C}^\circ} \quad \textit{and} \quad \gamma_{\mathcal{C}^\circ} = \sigma_{\mathcal{C}} \ .$$

See Rockafellar (1996); Hiriart-Urruty and Lemaréchal (2001); Vaiter et al (2015a) for the proof.

B Expectation of the inner product

We start with some definitions and notations that will be used in the proof. For a non-empty closed convex set $\mathcal{C} \in \mathbb{R}^p$, we denote $\big(\mathcal{C}\big)^0$ its minimal selection, i.e. the element of minimal norm in $\mathcal{C}$. This element is of course unique. For a proper lsc and convex function f and $\gamma > 0$, its Moreau envelope (or Moreau-Yosida regularization) is defined by

$$^{\gamma}f(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \min_{\overline{\boldsymbol{\theta}} \in \mathbb{R}^p} \frac{1}{2\gamma} \big\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\big\|_2^2 + f(\overline{\boldsymbol{\theta}}).$$

The Moreau envelope enjoys several important properties that we collect in the following lemma.

Lemma 6 *Let f be a finite-valued and convex function. Then*
(i) $(^{\gamma}f(\boldsymbol{\theta}))_{\gamma>0}$ is a decreasing net, and $\forall \boldsymbol{\theta} \in \mathbb{R}^p$, $^{\gamma}f(\boldsymbol{\theta}) \nearrow f(\boldsymbol{\theta})$ as $\gamma \searrow 0$.
(ii) $^{\gamma}f \in C^1(\mathbb{R}^p)$ with γ^{-1}-Lipschitz continuous gradient.
(iii) $\forall \boldsymbol{\theta} \in \mathbb{R}^p$, $\nabla\, ^{\gamma}f(\boldsymbol{\theta}) \to \big(\partial f(\boldsymbol{\theta})\big)^0$ and $\big\|\nabla\, ^{\gamma}f(\boldsymbol{\theta})\big\|_2 \nearrow \big\|\big(\partial f(\boldsymbol{\theta})\big)^0\big\|_2$ as $\gamma \searrow 0$.

Proof (i) (Bauschke and Combettes 2011, Proposition 12.32). (ii) (Bauschke and Combettes 2011, Proposition 12.29). (iii) Since f is finite-valued and convex, it is is subdifferentiable everywhere and its subdifferential is a maximal monotone operator with full domain $\mathbb{R}^p$, and the result follows from (Bauschke and Combettes 2011, Corollary 23.46(i)).

We are now equipped to prove the following important result[11]. To study EWA, Leung and Barron (2006) used Stein's identity and Catoni (2003, 2007); Dalalyan and Tsybakov (2008); Mai and Alquier (2015) used PAC-Bayesian bounds. Our result hereafter turns out to be instrumental to study EWA in the low-temperature regime for general penalties.

Proposition 5 *Let the density μ_n in (2), where*
(a) F satisfies Assumptions **(H.1)**-**(H.2)**;
(b) J is a finite-valued lower-bounded convex function, and $\exists R > 0$ and $\rho \geq 0$, such that $\forall \boldsymbol{\theta} \in \mathbb{R}^p$, $\big\|\big(\partial J(\boldsymbol{\theta})\big)^0\big\|_2 \leq R\,\|\boldsymbol{\theta}\|_2^{\rho}$;
(c) and V_n is coercive.
Then, $\forall \overline{\boldsymbol{\theta}} \in \mathbb{R}^p$,

$$\mathbb{E}_{\mu_n}\Big[\big\langle \big(\partial V_n(\boldsymbol{\theta})\big)^0, \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\big\rangle\Big] = -p\beta.$$

This result covers of course the situation where J fulfills **(H.3)**. In this case, since $\partial J(\boldsymbol{\theta}) \subset \mathcal{C}^\circ$ by Theorem 1(i), we have $\rho = 0$ and $R = \mathrm{diam}(\mathcal{C}^\circ)$, the diameter of the convex compact set $\mathcal{C}^\circ$ containing the origin. It can be shown that, when $F(\cdot, \boldsymbol{y})$ is strongly coercive, the coercivity assumption (c) can be equivalently stated as $J_\infty(\boldsymbol{\theta}) > 0$, $\forall \boldsymbol{\theta} \in \ker(\boldsymbol{X}) \setminus \{0\}$, where J_∞ is the recession/asymptotic function of J; see e.g. Rockafellar and Wets (1998).

Proof Let $V_n^{\gamma}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \frac{1}{n}F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) + \lambda_n\, ^{\gamma}J(\boldsymbol{\theta})$ and define $\mu_n^{\gamma}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \exp\big(-V_n^{\gamma}(\boldsymbol{\theta})/\beta\big)/Z$, where $0 < Z < +\infty$ is the normalizing constant of the density μ_n. Assumption **(H.1)** and Lemma 6(ii)-(iii) tell us that $V_n^{\gamma} \in C^1(\mathbb{R}^p)$ and $\nabla V_n^{\gamma}(\boldsymbol{\theta}) \to \big(\partial V_n(\boldsymbol{\theta})\big)^0$ as $\gamma \to 0$. Thus

$$\mathbb{E}_{\mu_n}\Big[\big\langle \big(\partial V_n(\boldsymbol{\theta})\big)^0, \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\big\rangle\Big] = \int_{\mathbb{R}^p} \lim_{\gamma \to 0} \langle \mu_n^{\gamma}(\boldsymbol{\theta}) \nabla V_n^{\gamma}(\boldsymbol{\theta}), \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\rangle d\boldsymbol{\theta}.$$

We now check that $\langle \mu_n^{\gamma}(\boldsymbol{\theta}) \nabla V_n^{\gamma}(\boldsymbol{\theta}), \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\rangle$ is dominated by an integrable function. From the definition of the Moreau envelope, we have

$$V_n^{\gamma}(\boldsymbol{\theta}) = \min_{\overline{\boldsymbol{\theta}} \in \mathbb{R}^p} \tfrac{1}{n}F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) + \lambda_n \Big(J(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}) + \frac{1}{2\gamma}\big\|\overline{\boldsymbol{\theta}}\big\|_2^2\Big).$$

[11] It will be proved here using Moreau-Yosida regularization. Yet another alternative proof could be based on mollifiers for approximating subdifferentials.

From coercivity of V_n, the objective in the min is also coercive in $(\boldsymbol{\theta}, \overline{\boldsymbol{\theta}})$ by (Rockafellar and Wets 1998, Exercise 3.29(b)). It then follows from (Rockafellar and Wets 1998, Theorem 3.31) that V_n^γ is also coercive. In turn, (Rockafellar and Wets 1998, Theorem 11.8(c) and 3.26(a)) allow to assert that for some $a \in]0, +\infty[$, $\exists b \in]-\infty, +\infty[$ such that for all $\gamma > 0$ and $\boldsymbol{\theta} \in \mathbb{R}^p$

$$\mu_n^\gamma(\boldsymbol{\theta}) \leq \exp\left(-a \left\|\boldsymbol{\theta}\right\|_2 - b\right)/Z. \tag{43}$$

Lemma 6-(iii) and assumption (b) on J entail that for any $\boldsymbol{\theta} \in \mathbb{R}^p$,

$$\left\|\nabla^\gamma J(\boldsymbol{\theta})\right\|_2 \leq \left\|\left(\partial J(\boldsymbol{\theta})\right)^0\right\|_2 \leq R \left\|\boldsymbol{\theta}\right\|_2^\rho .$$

Altogether, we have

$$\begin{aligned}
\left|\langle \mu_n^\gamma(\boldsymbol{\theta}) \nabla V_n^\gamma(\boldsymbol{\theta}), \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\rangle\right| &\leq \mu_n^\gamma(\boldsymbol{\theta}) \left(\left|\langle \boldsymbol{X}^\top \tfrac{1}{n} \nabla F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}), \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\rangle\right| + \lambda_n \left\|\nabla^\gamma J(\boldsymbol{\theta})\right\|_2 \left\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_2\right) \\
&\leq CZ^{-1} \exp\left(-F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y})/(n\beta)\right) \left|\langle \tfrac{1}{n} \nabla F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}), \boldsymbol{X}(\overline{\boldsymbol{\theta}} - \boldsymbol{\theta})\rangle\right| \\
&\quad + (Z \exp b)^{-1} \lambda_n R \exp\left(-a \left\|\boldsymbol{\theta}\right\|_2\right) \left\|\boldsymbol{\theta}\right\|_2^\rho \left\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_2,
\end{aligned}$$

where the constant $C > 0$ reflects the lower-boundedness of J. It is easy to see that the function in this upper-bound is integrable, where we also use **(H.2)**. Hence, we can apply the dominated convergence theorem to get

$$\mathbb{E}_{\mu_n}\left[\langle \left(\partial V_n(\boldsymbol{\theta})\right)^0, \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\rangle\right] = \lim_{\gamma \to 0} \int_{\mathbb{R}^p} \langle \mu_n^\gamma(\boldsymbol{\theta}) \nabla V_n^\gamma(\boldsymbol{\theta}), \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\rangle d\boldsymbol{\theta}.$$

Now, by simple differential calculus (chain and product rules), we have

$$\begin{aligned}
\langle \mu_n^\gamma(\boldsymbol{\theta}) \nabla V_n^\gamma(\boldsymbol{\theta}), \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\rangle &= -\beta \langle \nabla \mu_n^\gamma(\boldsymbol{\theta}), \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\rangle \\
&= -\beta \sum_{i=1}^p \frac{\partial}{\partial \boldsymbol{\theta}_i} \left(\mu_n^\gamma(\boldsymbol{\theta})(\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\right) - p\beta \mu_n^\gamma(\boldsymbol{\theta}).
\end{aligned}$$

Integrating the first term, we get by Fubini theorem and the Newton-Leibniz formula

$$\begin{aligned}
&\int_{\mathbb{R}^{p-1}} \left(\int_{\mathbb{R}} \frac{\partial}{\partial \boldsymbol{\theta}_i} \left(\mu_n^\gamma(\boldsymbol{\theta})(\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\right) d\boldsymbol{\theta}_i\right) d\boldsymbol{\theta}_1 \cdots d\boldsymbol{\theta}_{i-1} d\boldsymbol{\theta}_{i+1} \cdots d\boldsymbol{\theta}_p \\
&= \int_{\mathbb{R}^{p-1}} \left[\mu_n^\gamma(\boldsymbol{\theta})(\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\right]_{\mathbb{R}} d\boldsymbol{\theta}_1 \cdots d\boldsymbol{\theta}_{i-1} d\boldsymbol{\theta}_{i+1} \cdots d\boldsymbol{\theta}_p = 0,
\end{aligned}$$

where we used coercivity of V_n^γ (see (43)) to conclude that $\lim_{|\boldsymbol{\theta}_i| \to +\infty} \mu_n^\gamma(\boldsymbol{\theta})(\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) = 0$. For the second term, we have from Lemma 6(i) that $\mu_n^\gamma \to \mu_n$ as $\gamma \to 0$. Thus, arguing again as in (43), we can apply the dominated convergence theorem to conclude that

$$\lim_{\gamma \to 0} \int_{\mathbb{R}^p} \mu_n^\gamma(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\mathbb{R}^p} \mu_n(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1.$$

This concludes the proof.

Acknowledgements This work was supported by Conseil Régional de Basse-Normandie and partly by Institut Universitaire de France.

References

Bach F (2008) Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research* 9:1179–1225

Bakin S (1999) Adaptive regression and model selection in data mining problems. Thesis (Ph.D.)–Australian National University, 1999

Bauschke HH, Combettes PL (2011) *Convex analysis and monotone operator theory in Hilbert spaces*. Springer

Bellec P (2014) Concentration of quadratic forms under a bernstein moment assumption. Technical report, Ecole Polytechnique

Bickel PJ, Ritov Y, Tsybakov A (2009) Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics* 37(4):1705–1732

Bogdan M, van den Berg E, Sabatti C, Su W, Candès EJ (2014) Slope – adaptive variable selection via convex optimization. *Annals of Applied Statistics* 9(3):1103–1140

Bühlmann P, van de Geer S (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics, Springer-Verlag Berlin Heidelberg

Candès E, Plan Y (2009) Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics* 37(5A):2145–2177

Candès EJ, Recht B (2009) Exact matrix completion via convex optimization. *Foundations of Computational mathematics* 9(6):717–772

Candès EJ, Li X, Ma Y, Wright J (2011) Robust principal component analysis? *Journal of the ACM* 58(3):11:1–11:37

Candès EJ, Strohmer T, Voroninski V (2013) Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics* 66(8):1241–1274

Castillo I, Schmidt-Hieber J, van der Vaart A (2015) Bayesian linear regression with sparse priors. *The Annals of Statistics* 43(5):1986–2018

Catoni O (2003) A PAC-bayesian approach to adaptive classification

Catoni O (2007) *PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning)*, Lecture Notes-Monograph Series, vol 56. IMS

Chandrasekaran V, Recht B, Parrilo PA, Willsky A (2012) The convex geometry of linear inverse problems. *Foundations of Computational Mathematics* 12(6):805–849

Chen S, Donoho D, Saunders M (1999) Atomic decomposition by basis pursuit. *SIAM journal on scientific computing* 20(1):33–61

Chen X, Lin Q, Kim S, Carbonell JG, Xing EP (2010) An efficient proximal-gradient method for general structured sparse learning. *Preprint arXiv:10054717*

Coste M (2002) An introduction to semialgebraic geometry. Technical report, Institut de Recherche Mathematiques de Rennes

Dalalyan A, Tsybakov A (2009) pac-bayesian bounds for the expected error of aggregation by exponential weights. Technical report, Université Paris 6, CREST and CERTIS, Ecole des Ponts ParisTech, personal communication

Dalalyan A, Tsybakov AB (2008) Aggregation by exponential weighting, sharp PAC-bayesian bounds and sparsity. *Machine Learning* 72(1-2):39–61, DOI 10.1007/s10994-008-5051-0, URL http://dx.doi.org/10.1007/s10994-008-5051-0

Dalalyan AS, Tsybakov AB (2007) Aggregation by exponential weighting and sharp oracle inequalities. In: Proceedings of the 20th Annual Conference on Learning Theory, Springer-Verlag, Berlin, Heidelberg, COLT'07, pp 97–111, URL http://dl.acm.org/citation.cfm?id=1768841.1768854

Dalalyan AS, Tsybakov AB (2012) Sparse regression learning by aggregation and langevin monte-carlo. *Journal of Computer and System Sciences* 78(5):1423–1443, DOI 10.1016/j.jcss.2011.12.023, URL http://dx.doi.org/10.1016/j.jcss.2011.12.023

Dalalyan AS, Hebiri M, Lederer J (2017) On the prediction performance of the lasso. *Bernoulli* 23(1):552–581, DOI 10.3150/15-BEJ756, URL http://dx.doi.org/10.3150/15-BEJ756

Dalalyan AS, Grappin E, Paris Q (2018) On the Exponentially Weighted Aggregate with the Laplace Prior. *The Annals of Statistics* 46(5):2452–2478

Daniilidis A, Drusvyatskiy D, Lewis AS (2014) Orthogonal invariance and identifiability. *SIAM Journal on Matrix Analysis and Applications* 35(2):580–598, DOI 10.1137/130916710

Donoho D (2006) For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics* 59(6):797–829

Donoho D, Tanner J (2010) Counting the faces of randomly-projected hypercubes and orthants. *Discrete and Computational Geometry* 43(3):522–541

Durmus A, Moulines E, Pereyra M (2016) Sampling from convex non continuously differentiable functions, when Moreau meets Langevin, URL https://hal.archives-ouvertes.fr/hal-01267115, preprint hal-01267115

Duy Luu T, Fadili JM, Chesneau C (2016) PAC-Bayesian risk bounds for group-analysis sparse regression by exponential weighting. Technical report, hal-01367742, URL https://hal.archives-ouvertes.fr/hal-01367742

Fadili MJ, Peyré G, Vaiter S, Deledalle C, Salmon J (2013) Stable recovery with analysis decomposable priors. In: Proc. Sampta'13, pp 113–116

Fazel M, Hindi H, Boyd SP (2001) A rank minimization heuristic with application to minimum order system approximation. In: Proceedings of the American Control Conference, IEEE, vol 6, pp 4734–4739

van de Geer S (2008) High-dimensional generalized linear models and the Lasso. *The Annals of Statistics* 36:614–645

van de Geer S (2014) Weakly decomposable regularization penalties and structured sparsity. *Scandinavian Journal of Statistics* 41(1):72–86, DOI 10.1111/sjos.12032, URL http://dx.doi.org/10.1111/sjos.12032

van de Geer S, Buhlmann P (2009) On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics* 3:1360–1392, DOI 10.1214/09-EJS506, URL http://dx.doi.org/10.1214/09-EJS506

van de Geer S, Lederer J (2013) The bernstein–orlicz norm and deviation inequalities. *Probab Theory Relat Fields* 157:225–250

Guedj B, Alquier P (2013) PAC-bayesian estimation and prediction in sparse additive models. *Electronic Journal of Statistics* 7:264–291, DOI 10.1214/13-EJS771, URL http://dx.doi.org/10.1214/13-EJS771

Hiriart-Urruty JB, Lemaréchal C (2001) *Convex Analysis And Minimization Algorithms*, vol I and II. Springer

Jacob L, Obozinski G, Vert JP (2009) Group lasso with overlap and graph lasso. In: Danyluk AP, Bottou L, Littman ML (eds) ICML'09, vol 382, p 55

Jégou H, Furon T, Fuchs JJ (2012) Anti-sparse coding for approximate nearest neighbor search. In: IEEE ICASSP, pp 2029–2032

Koltchinskii V (2008) Oracle inequalities in empirical risk minimization and sparse recovery problems. In: Lectures from the 38th Probability Summer School held in Saint-Flour, Lecture Notes in Mathematics, vol 2033, Springer

Koltchinskii V, Lounici K, Tsybakov AB (2011) Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics* 39(5):2302–2329, DOI 10.1214/11-AOS894, URL http://dx.doi.org/10.1214/11-AOS894

Lecué G (2007) Simultaneous adaptation to the margin and to complexity in classification. *The Annals of Statistics* 35(4):1698–1721, DOI 10.1214/009053607000000055, URL http://dx.doi.org/10.1214/009053607000000055

Ledoux M (2001) *The concentration of measure phenomenon*. Mathematical surveys and monographs, American Mathematical Society, Providence (R.I.)

Leung G, Barron AR (2006) Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory* 52(8):3396–3410

Lounici K, Pontil M, van de Geer S, Tsybakov AB (2011) Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics* 39(4):2164–2204, DOI 10.1214/11-AOS896, URL http://dx.doi.org/10.1214/11-AOS896

Lyubarskii Y, Vershynin R (2010) Uncertainty principles and vector quantization. *IEEE Transactions on Information Theory* 56(7):3491–3501

Mai TT, Alquier P (2015) A bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electronic Journal of Statistics* 9(1):823–841, DOI

10.1214/15-EJS1020, URL https://doi.org/10.1214/15-EJS1020
Massart P (2007) *Concentration inequalities and model selection*. Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003, Springer Verlag
Negahban S, Ravikumar P, Wainwright MJ, Yu B (2012) A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science* 27(4):538–557
Nemirovski A (2000) Topics in non-parametric statistics. In: Emery M, Nemirovski A, Voiculescu D (eds) Ecole d'Ete de Probabilites de Saint-Flour XXVIII-1998, Springer, Lecture Notes in Mathematics, vol 1738, pp 87–285
Osborne M, Presnell B, Turlach B (2000) A new approach to variable selection in least squares problems. *IMA journal of numerical analysis* 20(3):389–403
Peyré G, Fadili J, Chesneau C (2011) Adaptive Structured Block Sparsity Via Dyadic Partitioning. In: EUSIPCO, Barcelona, Spain
Raskutti G, Wainwright MJ, Yu B (2011) Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. *IEEE Transactions on Information Theory* 57(10):6976–6994, DOI 10.1109/TIT.2011.2165799
Recht B, Fazel M, Parrilo PA (2010) Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review* 52(3):471–501
Rigollet P, Tsybakov A (2007) Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics* 16(3):260–280, DOI 10.3103/S1066530707030052, URL http://dx.doi.org/10.3103/S1066530707030052
Rigollet P, Tsybakov A (2011) Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics* 39(2):731–771, DOI 10.1214/10-AOS854, URL http://dx.doi.org/10.1214/10-AOS854
Rigollet P, Tsybakov AB (2012) Sparse estimation by exponential weighting. *Statistical Science* 27(4):558–575, DOI 10.1214/12-STS393, URL http://dx.doi.org/10.1214/12-STS393
Rockafellar R (1996) *Convex analysis*, vol 28. Princeton University Press
Rockafellar RT, Wets R (1998) *Variational analysis*, vol 317. Springer Verlag
Rudelson M, Vershynin R (2008) On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics* 61(8):1025–1045
Rudin L, Osher S, Fatemi E (1992) Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena* 60(1-4):259–268
Studer C, Yin W, Baraniuk RG (2012) Signal representations with minimum ℓ_∞-norm. In: 50th Annual Allerton Conference on Communication, Control, and Computing,
Su W, Candès EJ (2015) Slope is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics* 44(3):1038–1068
Sun T, Zhang CH (2012) Scaled sparse linear regression. *Biometrika* 99(4):879, DOI 10.1093/biomet/ass043
Suzuki T (2015) Convergence rate of bayesian tensor estimator and its minimax optimality. In: Proceedings of the 32nd International Conference on Machine Learning (ICML), vol 37, pp 1273–1282
Talagrand M (2005) *The generic chaining. Upper and lower bounds of stochastic processes*. Springer-Verlag, Berlin
Tibshirani R (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B Methodological* 58(1):267–288
Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(1):91–108
Tropp J (2015a) Convex recovery of a structured signal from independent random linear measurements. In: Pfander G (ed) Sampling Theory, a Renaissance, Applied and Numerical Harmonic Analysis (ANHA), Birkhäuser/Springer
Tropp JA (2015b) An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning* 8(1-2):1–230, DOI 10.1561/2200000048, URL http://dx.doi.org/10.1561/2200000048
Tsybakov AB (2008) *Introduction to Nonparametric Estimation*, 1st edn. Springer
Vaiter S, Golbabaee M, Fadili J, Peyré G (2015a) Model selection with low complexity priors. *Information and Inference: A Journal of the IMA* 4(3):230

Vaiter S, Peyré G, Fadili MJ (2015b) Low complexity regularization of linear inverse problems. In: Pfander G (ed) Sampling Theory, a Renaissance, Applied and Numerical Harmonic Analysis (ANHA), Birkhäuser/Springer

Vaiter S, Deledalle C, Fadili MJ, Peyré G, Dossal C (2017) The degrees of freedom of partly smooth regularizers. *Annals of the Institute of Statistical Mathematics* 69(4):791–832

Vaiter S, Peyré G, Fadili MJ (2018) Model consistency of partly smooth regularizers. *IEEE Transactions on Information Theory* 64(3):1725 – 1737

Vershynin R (2015) Estimation in high dimension : A geometric perspective. In: Pfander G (ed) Sampling Theory, a Renaissance, Applied and Numerical Harmonic Analysis (ANHA), Birkhäuser/Springer

Verzelen N (2012) Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics* 6(0):38–90, DOI 10.1214/12-ejs666, URL http://dx.doi.org/10.1214/12-ejs666

Wang Z, Paterlini S, Gao F, Yang Y (2014) Adaptive minimax regression estimation over sparse lq-hulls. *Journal of Machine Learning Research* 15(1):1675–1711, URL http://dl.acm.org/citation.cfm?id=2627435.2638589

Wei F, Huang J (2010) Consistent group selection in high-dimensional linear regression. *Bernoulli* 16(4):1369–1384

Yang Y (2004) Aggregating regression procedures to improve performance. *Bernoulli* 10(1):25–47

Ye F, Zhang CH (2010) Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *Journal of Machine Learning Research* 11:3519–3540, URL http://dl.acm.org/citation.cfm?id=1756006.1953043

Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(1):49–67