Sharp Oracle Inequalities for Low-complexity Priors

Tung Duy Luu, Jalal Fadili, Christophe Chesneau

Sharp Oracle Inequalities for Low-complexity Priors

Tung Duy Luu* Jalal Fadili* Christophe Chesneau†

Abstract

In this paper, we consider a high-dimensional linear regression model with fixed design. We present a unified analysis of the performance guarantees of exponential weighted aggregation and penalized estimators with a general class of 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. The results are then applied to several instances including the Lasso, the group Lasso, their analysis-type counterparts, the ℓ_∞ and the nuclear norm penalties. When the noise is random, we provide oracle inequalities in probability under mild assumptions on the noise distribution. These estimators can be efficiently implemented using proximal splitting algorithms.

Key words. High-dimensional regression, exponential weighted aggregation, penalized estimation, oracle inequality, low-complexity models.

AMS subject classifications. 62G07 62G20

1 Introduction

1.1 Problem statement

Our statistical context is the following. We consider the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_0 + \boldsymbol{\xi}, \tag{1.1}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is a deterministic design matrix, and $\boldsymbol{\xi}$ are errors.

The idea of aggregating elements in a dictionary has been introduced in machine learning to combine different techniques (see [40, 66]) with some procedures such as bagging [9], boosting [30, 54] and random forests [1, 5–7, 10, 31]. In the recent years, there has been a flurry of research on the use of low-complexity regularization (among which sparsity and low-rank are the most popular) in various areas including statistics and machine learning in high dimension. The idea is that 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 an estimate $\boldsymbol{X}\widehat{\boldsymbol{\theta}}$ with good provable performance guarantees under appropriate conditions.

*Normandie Univ, ENSICAEN, CNRS, GREYC, France, Email: {duy-tung.luu, Jalal.Fadili}@ensicaen.fr.
†Normandie Univ, UNICAEN, CNRS, LMNO, France, Email: christophe.chesneau@unicaen.fr.

1.2 Variational/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

$$\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} \in \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\mathrm{Argmin}}\{V_n(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} \tfrac{1}{n}F(\boldsymbol{X}\boldsymbol{\theta},\boldsymbol{y}) + \lambda J(\boldsymbol{\theta})\}, \tag{1.2}$$

where $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a general loss function assumed to be a proper, convex and sufficiently smooth function of its first argument[1]. The regularizing penalty J is a proper closed convex function, and promotes some specific notion of simplicity/low-complexity, and $\lambda > 0$ is the regularization parameter. A prominent member covered by (1.2) is the Lasso [8, 11, 12, 16, 25, 46, 57] and its variants such the analysis/fused Lasso [53, 58] or group Lasso [2, 3, 67, 69]. 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 [14, 15, 29, 48]. See [11, 43, 62, 64] for generalizations and comprehensive reviews.

1.3 Exponential Weighted Aggregation (EWA)

An alternative to the the variational estimator (1.2) is the aggregation by exponential weighting, which consists in substituting averaging for minimization. The aggregators are defined via the probability density function

$$\mu_n(\boldsymbol{\theta}) = \frac{\exp\left(-V_n(\boldsymbol{\theta})/\beta\right)}{\int_\Theta \exp\left(-V_n(\boldsymbol{\omega})/\beta\right)d\boldsymbol{\omega}}, \tag{1.3}$$

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

$$\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} = \int_{\mathbb{R}^p} \boldsymbol{\theta}\mu_n(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{1.4}$$

Aggregation by exponential weighting has been widely considered in the statistical and machine learning literatures, see e.g. [18, 19, 22, 23, 27, 32, 38, 44, 49, 68] to name a few. $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ 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 $\pi(\boldsymbol{\theta}) \propto \exp\left(-\lambda J(\boldsymbol{\theta})/\beta\right)$.

1.3.1 Oracle inequalities

Oracle inequalities, which are at the heart of our work, quantify the quality of an estimator compared to the best possible one that could only be given with an oracle. 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 $\|\cdot\|_2$ is the ℓ_2 norm. R_n will serve as a measure of the performance of the estimator $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$. More precisely, we

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

aim to prove that $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ mimics as much as possible the best model of aggregation. This idea is materialized in the following type of inequalities

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\big) \leq C \inf_{\boldsymbol{\theta}\in\mathbb{R}^p} \big(R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \Delta_{n,p,\lambda,\beta}(\boldsymbol{\theta})\big), \tag{1.5}$$

where $C \geq 1$ is the leading constant of the oracle inequality and the remainder term $\Delta_{n,\lambda,\beta}(\boldsymbol{\theta})$ 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, 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 $\Delta_{n,p,\lambda,\beta}(\boldsymbol{\theta})$ is small and decreases rapidly to 0 as $n \to +\infty$.

1.4 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 [28, 60–63]. Our main contributions are summarized as follows:

- We show that the the EWA estimator $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ in (1.3) satisfies a sharp oracle inequality for prediction with optimal remainder term, for the general case where J is a proper finite-valued sublinear function, where sublinearity is equivalent to saying that J is convex and positively homogeneous (hence subadditive).
- We handle a more general data fidelity than the usual quadratic one.
- We prove a sharp prediction oracle inequality for the variational/penalized estimator $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ in (1.2). We highlight the differences between the two estimators in terms of the corresponding bounds.
- We then apply these two inequalities to several 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. When the noise is random (typically Gaussian or subgaussian), we provide oracle inequalities in probability. We show that we recover some known results as special cases and establish new ones.

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 (1.2) [62], 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 [27] to sample from log-semiconcave densities[2], see also [26] for log-concave densities.

1.5 Relation to previous work

Our oracle inequality for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ extends the work of [20] with an unprecedented level of generality, far beyond the Lasso and the nuclear norm. Our prediction sharp oracle inequality for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ specializes to that of [56] in the case of the Lasso (see also the discussion in [21] and references therein) and that of [37] for the case of the nuclear norm. Our work also goes much beyond that in [64] 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.

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

1.6 Paper organization

In Section 2, we provide some notations and preliminaries. Section 3 introduces some key concepts from convex analysis and low-complexity regularization which are behind this work, and then state our main assumptions on the data loss and the prior penalty. All these notions are exemplified on some penalties some of which are popular in the literature. In Section 4, we prove our main oracle inequalities, and then apply them to the previous penalty examples. A key intermediate result in the proof of our main results is established in the appendix with an elegant argument relying on Moreau-Yosida regularization.

2 Notations and Preliminaries

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 $\|\cdot\|_2$. 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 subset I of $\{1, \ldots, p\}$, we denote by I^c its complement, $\big|I\big|$ its cardinality. $\boldsymbol{\theta}_I$ 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 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.

Definition 2.1 (Polar set)**.** *Let $\mathcal{C}$ be a nonempty convex set. The set $\mathcal{C}^\circ$ given by*

$$\mathcal{C}^\circ = \big\{ \boldsymbol{\eta} \in \mathbb{R}^p \ : \ \langle \boldsymbol{\eta}, \boldsymbol{\theta} \rangle \leq 1 \textit{ for all } \boldsymbol{\theta} \in \mathcal{C} \big\}$$

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. $\mathcal{C}^{\circ\circ} = \mathcal{C}$.

Functions A function $f : \mathbb{R}^p \to \mathbb{R}$ is coercive, if $\lim_{\|\boldsymbol{\theta}\|_2 \to +\infty} f(\boldsymbol{\theta}) = +\infty$. The effective domain of f is defined by $\mathrm{dom}(f) = \big\{ \boldsymbol{\theta} \in \mathbb{R}^p \ : \ f(\boldsymbol{\theta}) < +\infty \big\}$ and f is proper if $f(\boldsymbol{\theta}) > -\infty$ for all $\boldsymbol{\theta}$ and $\mathrm{dom}(f) \neq \emptyset$ as is the case when it is finite-valued. A function is said sublinear if it is convex and positively homogeneous.

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

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

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

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

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

3 Estimation with Low-complexity Penalties

The estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ in (1.2) and (1.4) 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.

3.1 Choice of the data loss

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

(H.1) $F(\boldsymbol{u}, \boldsymbol{y}) = \varphi(\boldsymbol{u}) - \langle \boldsymbol{u}, \boldsymbol{y} \rangle$, where $\varphi \in C^1(\mathbb{R}^n)$ is strongly convex with modulus $\nu > 0$.

(H.2) For any $\overline{\boldsymbol{\theta}} \in \mathbb{R}^p$, $\int_{\mathbb{R}^p} \exp\left(-\|\boldsymbol{\theta}\|_2\right) \left|\langle \nabla\varphi(\boldsymbol{X}\boldsymbol{\theta}), \boldsymbol{X}(\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rangle\right| < +\infty$.

This is a fairly general class of data loss functions. It is reminiscent of the negative log-likelihood in the regular exponential family. The moment assumption **(H.2)** is satisfied in many situations of interest. For instance, one can immediately check that this is true if $\nabla\varphi$ is also Lipschitz continuous (as when φ is quadratic).

The following simple lemma gives useful bounds of the Bregman distance associated to F.

Lemma 3.1. *For F satisfying **(H.1)**, the following bounds hold*

$$D_{F(\cdot,\boldsymbol{y})}(\boldsymbol{v}, \boldsymbol{u}) \geq \tfrac{\nu}{2} \|\boldsymbol{v} - \boldsymbol{u}\|_2^2.$$

If $\nabla\varphi$ is κ-Lipschitz continuous, $\kappa > 0$, then

$$D_{F(\cdot,\boldsymbol{y})}(\boldsymbol{v}, \boldsymbol{u}) \leq \tfrac{\kappa}{2} \|\boldsymbol{v} - \boldsymbol{u}\|_2^2.$$

Proof. It is immediate to see that

$$D_{F(\cdot,\boldsymbol{y})}(\boldsymbol{v}, \boldsymbol{u}) = D_{\varphi}(\boldsymbol{v}, \boldsymbol{u}).$$

The lower-bound is then by definition of strong convexity of φ. The upper-bound is known as the descent lemma applied to φ, see e.g. [45, Theorem 2.1.5]. □

3.2 Choice of the prior penalty

Before stating our main assumption on J, we start by collecting some ingredients from convex analysis that are essential to our exposition.

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. [33, 51].

Lemma 3.2. *Let $\mathcal{C}$ be a non-empty set.*

(i) *$\sigma_{\mathcal{C}}$ is proper 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 $0 \in \operatorname{int}(\mathcal{C})$, then $\sigma_{\mathcal{C}}$ is finite-valued and coercive.*

Gauges and polars 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 3.3 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 3.3.

(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, $0 \in \operatorname{int}(\mathcal{C})$, in which case $\gamma_{\mathcal{C}}$ is 1-Lipschitz continuous.*

(iv) *$\gamma_{\mathcal{C}}$ is finite-valued and coercive if, and only if, $\mathcal{C}$ is compact and $0 \in \operatorname{int}(\mathcal{C})$.*

See [61] 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 3.1 (Polar Gauge)**.** *The polar of a gauge $\gamma_{\mathcal{C}}$ is the function $\gamma_{\mathcal{C}}^{\circ}$ defined by*

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

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

$$\langle \boldsymbol{\theta}, \boldsymbol{u} \rangle \leq \gamma_{\mathcal{C}}(\boldsymbol{\theta}) \gamma_{\mathcal{C}}^{\circ}(\boldsymbol{u}) \quad \forall (\boldsymbol{\theta}, \boldsymbol{u}) \in \operatorname{dom}(\gamma_{\mathcal{C}}) \times \operatorname{dom}(\gamma_{\mathcal{C}}^{\circ}), \tag{3.1}$$

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

Lemma 3.4. *Let $\mathcal{C} \subseteq \mathbb{R}^p$ be a closed convex set containing the origin. Then,*

(ii) $\gamma_{\mathcal{C}}^{\circ}$ *is a gauge function and* $\gamma_{\mathcal{C}}^{\circ\circ} = \gamma_{\mathcal{C}}$.

(iii) $\gamma_{\mathcal{C}}^{\circ} = \gamma_{\mathcal{C}^\circ}$, *or equivalently*

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

(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 [33, 51, 61] for the proof.

We are now ready to state our main assumption on J.

(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.

By Lemma 3.3, 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 3.4 and Lemma 3.2, the polar gauge $J^\circ \overset{\text{def}}{=} \gamma_{\mathcal{C}^\circ}$ enjoys the same properties as J in **(H.3)**.

3.3 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 [61].

Definition 3.2 (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

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

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

It can be shown, see [61, 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 3.1. *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

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

(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 [61, Theorem 1, Proposition 4 and Proposition 5(iii)].

(ii) From [61, 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 $\boldsymbol{\omega}$ [4, 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 3.1. *The coercivity assumption in* **(H.3)** *is not needed for Theorem 3.1 to hold.*

The decomposability of described in Theorem 3.1 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 [61, Definition 6], are precisely a class of gauges for which this situation occurs, and in this case, Theorem 3.1 has the simpler form

$$\partial J(\boldsymbol{\theta}) = \mathrm{aff}(\partial J(\boldsymbol{\theta})) \cap \mathcal{C}^\circ = \left\{ \boldsymbol{\eta} \in \mathbb{R}^n \ : \ \boldsymbol{\eta}_{T_{\boldsymbol{\theta}}} = e_{\boldsymbol{\theta}} \quad \text{and} \quad J^\circ(\boldsymbol{\eta}_{S_{\boldsymbol{\theta}}}) \leq 1 \right\}. \tag{3.2}$$

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.

3.4 Closure properties

A distinctive property of the class of penalties complying with **(H.3)** is that it enjoys important closure properties that we summarize in the next lemma.

Lemma 3.5. *The set of functions satisfying* **(H.3)** *is closed under addition*[3] *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 $H \overset{def}{=} J + G$ 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 $H \overset{def}{=} J \circ \boldsymbol{D}^\top$ also fulfills* **(H.3)**. *Moreover,*

(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;

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

(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 3.5 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 since our unified analysis in Section 4 will apply to them just as well.

Proof. (i) Convexity, positive homogeneity, coercivity and finite-valuedness are straightforward.

(a) This is [61, Proposition 8(i)-(ii)].

(b) We have from Lemma 3.4 and calculus rules on support functions,

$$\begin{aligned}
H^\circ(\boldsymbol{\omega}) = \sigma_{J(\boldsymbol{\theta})+G(\boldsymbol{\theta})\leq 1}(\boldsymbol{\omega}) &= \sup_{J(\boldsymbol{\theta})+G(\boldsymbol{\theta})\leq 1} \langle \boldsymbol{\omega}, \boldsymbol{\theta}\rangle = \max_{\rho\in[0,1]} \sup_{J(\boldsymbol{\theta})\leq\rho, G(\boldsymbol{\theta})\leq 1-\rho} \langle \boldsymbol{\omega}, \boldsymbol{\theta}\rangle \\
\text{([33, Theorem V.3.3.3])} &= \max_{\rho\in[0,1]} \overline{\operatorname{conv}}\left(\inf\left(\sigma_{J(\boldsymbol{\theta})\leq\rho}(\boldsymbol{\omega}), \sigma_{G(\boldsymbol{\theta})\leq 1-\rho}(\boldsymbol{\omega})\right)\right) \\
\text{(Positive homogeneity)} &= \max_{\rho\in[0,1]} \overline{\operatorname{conv}}\left(\inf\left(\rho\sigma_{J(\boldsymbol{\theta})\leq 1}(\boldsymbol{\omega}), (1-\rho)\sigma_{G(\boldsymbol{\theta})\leq 1}(\boldsymbol{\omega})\right)\right) \\
\text{(Lemma 3.4)} &= \max_{\rho\in[0,1]} \overline{\operatorname{conv}}\left(\inf\left(\rho J^\circ(\boldsymbol{\omega}), (1-\rho)G^\circ(\boldsymbol{\omega})\right)\right).
\end{aligned}$$

(ii) Again, Convexity, positive homogeneity and finite-valuedness are immediate. Coercivity holds by injectivity of $\boldsymbol{D}^\top$.

(a) This is [61, Proposition 10(i)-(ii)].

(b) We have

$$\begin{aligned}
H^\circ(\boldsymbol{\omega}) &= \sup_{J(\boldsymbol{D}^\top\boldsymbol{\theta})\leq 1} \langle \boldsymbol{\omega}, \boldsymbol{\theta}\rangle \\
(\boldsymbol{D}^\top \text{ is injective}) &= \sup_{J(\boldsymbol{D}^\top\boldsymbol{\theta})\leq 1} \langle \boldsymbol{D}^+\boldsymbol{\omega}, \boldsymbol{D}^\top\boldsymbol{\theta}\rangle \\
&= \sup_{J(\boldsymbol{u})\leq 1, \boldsymbol{u}\in\operatorname{Span}(\boldsymbol{D}^\top)} \langle \boldsymbol{D}^+\boldsymbol{\omega}, \boldsymbol{u}\rangle \\
\text{([33, Theorem V.3.3.3] and Lemma 3.4)} &= \overline{\operatorname{conv}}\left(\inf\left(J^\circ(\boldsymbol{D}^+\boldsymbol{\omega}), \iota_{\operatorname{Ker}(\boldsymbol{D})}(\boldsymbol{D}^+\boldsymbol{\omega})\right)\right) \\
&= J^\circ(\boldsymbol{D}^+\boldsymbol{\omega}).
\end{aligned}$$

where in the last equality, we used the fact that $\boldsymbol{D}^+\boldsymbol{\omega} \in \operatorname{Span}\left(\boldsymbol{D}^\top\right) = \operatorname{Ker}(\boldsymbol{D})^\perp$, and thus $\iota_{\operatorname{Ker}(\boldsymbol{D})}(\boldsymbol{D}^+\boldsymbol{\omega}) = +\infty$ unless $\boldsymbol{\omega} = 0$, and J° is continuous and convex by **(H.3)** and Lemma 3.4.

□

3.5 Examples

3.5.1 Lasso

The Lasso regularization is used to promote the sparsity of the minimizers, see [11] for a comphensive review. It corresponds to choosing J as the ℓ_1-norm

$$J(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1 = \sum_{i=1}^{p} |\boldsymbol{\theta}_i|. \tag{3.3}$$

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 $(a_i)_{1\leq i\leq p}$ 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,

$$T_{\boldsymbol{\theta}} = \mathrm{Span}\{(a_i)_{i\in\mathrm{supp}(\boldsymbol{\theta})}\}, \quad (e_{\boldsymbol{\theta}})_i = \begin{cases} \mathrm{sign}(\boldsymbol{\theta}_i) & \text{if } i \in \mathrm{supp}(\boldsymbol{\theta}) \\ 0 & \text{otherwise} \end{cases}, \quad \text{and} \quad J^\circ = \big\|\cdot\big\|_\infty. \tag{3.4}$$

3.5.2 Group Lasso

The group Lasso has been advocated to promotessparsity of the groups, i.e. it drives all the coefficients in one group to zero together hence leading to group selection, see [2, 3, 67, 69] to cite a few. The group Lasso penalty with L groups reads

$$J(\boldsymbol{\theta}) = \big\|\boldsymbol{\theta}\big\|_{1,2} \stackrel{\text{def}}{=} \sum_{i=1}^{L} \big\|\boldsymbol{\theta}_{b_i}\big\|_2. \tag{3.5}$$

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

$$T_{\boldsymbol{\theta}} = \mathrm{Span}\{(a_j)_{\left\{j \,:\, \exists i \in \mathrm{supp}_{\mathcal{B}}(\boldsymbol{\theta}), j \in b_i\right\}}\}, (e_{\boldsymbol{\theta}})_{b_i} = \begin{cases} \frac{\boldsymbol{\theta}_{b_i}}{\|\boldsymbol{\theta}_{b_i}\|_2} & \text{if } i \in \mathrm{supp}_{\mathcal{B}}(\boldsymbol{\theta}) \\ 0 & \text{otherwise} \end{cases}, \text{ and } J^\circ(\boldsymbol{\omega}) = \max_{i\in\{1,\ldots,L\}} \|\boldsymbol{\omega}_{b_i}\|_2 . \tag{3.6}$$

3.5.3 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

$$J(\boldsymbol{\theta}) = \big\|\boldsymbol{D}^\top\boldsymbol{\theta}\big\|_{1,2}. \tag{3.7}$$

This encompasses the 2-D isotropic total variation [53]. 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 [53], and the fused Lasso [58]. The overlapping group Lasso [35] is also a special case of (3.5) by taking $\boldsymbol{D}^\top$ to be an operator that exactract the blocks [17, 47] (in which case $\boldsymbol{D}$ has even orthogonal rows).

Let $\Lambda_{\boldsymbol{\theta}} = \bigcup_{i\in\mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top\boldsymbol{\theta})} b_i$ and $\Lambda_{\boldsymbol{\theta}}^c$ its complement. From Lemma 3.5(ii) and (3.6), we get

$$T_{\boldsymbol{\theta}} = \mathrm{Ker}(\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}^c}^\top), \quad e_{\boldsymbol{\theta}} = \mathrm{P}_{T_{\boldsymbol{\theta}}}\, \boldsymbol{D} e_{\boldsymbol{D}^\top\boldsymbol{\theta}}^{\|\|_{1,2}} \quad \text{where} \quad \Big(e_{\boldsymbol{D}^\top\boldsymbol{\theta}}^{\|\|_{1,2}}\Big)_{b_i} = \begin{cases} \frac{(\boldsymbol{D}^\top\boldsymbol{\theta})_{b_i}}{\|(\boldsymbol{D}^\top\boldsymbol{\theta})_{b_i}\|_2} & \text{if } i \in \mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top\boldsymbol{\theta}) \\ 0 & \text{otherwise.} \end{cases} \tag{3.8}$$

If, in addition, $\boldsymbol{D}$ is surjective, then by virtue of Lemma 3.5(ii) we also have

$$J^\circ(\boldsymbol{\omega}) = \big\|\boldsymbol{D}^+\boldsymbol{\omega}\big\|_{\infty,2} \stackrel{\text{def}}{=} \max_{i\in\{1,\ldots,L\}} \big\|(\boldsymbol{D}^+\boldsymbol{\omega})_{b_i}\big\|_2 \tag{3.9}$$

3.5.4 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

$$J(\boldsymbol{\theta}) = \big\|\boldsymbol{\theta}\big\|_\infty = \max_{i\in\{1,\dots,p\}} \big|\boldsymbol{\theta}_i\big|. \tag{3.10}$$

The ℓ_∞ regularization has found applications in several fields [36, 42, 55]. Suppose that $\boldsymbol{\theta} \neq 0$, and define the saturation support of $\boldsymbol{\theta}$ as $I_{\boldsymbol{\theta}}^{\mathrm{sat}} \stackrel{\text{def}}{=} \big\{i \in \{1,\dots,p\} \;:\; \big|\boldsymbol{\theta}_i\big| = \|\boldsymbol{\theta}\|_\infty \big\} \neq \emptyset$. From [61, Proposition 14], we have

$$T_{\boldsymbol{\theta}} = \big\{\overline{\boldsymbol{\theta}} \in \mathbb{R}^p \;:\; \overline{\boldsymbol{\theta}}_{I_{\boldsymbol{\theta}}^{\mathrm{sat}}} \in \mathbb{R}\operatorname{sign}(\boldsymbol{\theta}_{I_{\boldsymbol{\theta}}^{\mathrm{sat}}})\big\}, \quad (e_{\boldsymbol{\theta}})_i = \begin{cases} \operatorname{sign}(\boldsymbol{\theta}_i)/|I^{\mathrm{sat}}\boldsymbol{\theta}| & \text{if } i \in I_{\boldsymbol{\theta}}^{\mathrm{sat}} \\ 0 & \text{otherwise} \end{cases}, \quad \text{and} \quad J^\circ = \|\cdot\|_1 . \tag{3.11}$$

3.5.5 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

$$J(\boldsymbol{\theta}) = \big\|\boldsymbol{\theta}\big\|_* = \big\|\lambda(\boldsymbol{\theta})\big\|_1. \tag{3.12}$$

This penalty is the best convex surrogate to enforce a low-rank prior. It has been widely used for various applications [13–15, 29, 48].

Following e.g. [60, Example 21], we have

$$T_{\boldsymbol{\theta}} = \big\{\boldsymbol{U}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{V}^\top \;:\; \boldsymbol{A} \in \mathbb{R}^{p_2\times r}, \boldsymbol{B} \in \mathbb{R}^{p_1\times r}\big\}, \quad e_{\boldsymbol{\theta}} = \boldsymbol{U}\boldsymbol{V}^\top \quad \text{and} \quad J^\circ(\boldsymbol{\omega}) = \|\!|\boldsymbol{\omega}|\!\|_{2\to 2} = \big\|\lambda(\boldsymbol{\omega})\big\|_\infty. \tag{3.13}$$

4 Main results

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

Throughout this section, we recall $T_{\boldsymbol{\theta}}$ and $e_{\boldsymbol{\theta}}$ the model subspace and vector associated to $\boldsymbol{\theta}$ (see Definition 3.2). Denote $S_{\boldsymbol{\theta}} = T_{\boldsymbol{\theta}}^\perp$. Given two coercive finite-valued gauge J_1 and J_2, and a linear operator $\boldsymbol{A}$, we define $\|\!|\boldsymbol{A}|\!\|_{J_1\to J_2}$ the *operator bound* as

$$\|\!|\boldsymbol{A}|\!\|_{J_1\to J_2} = \sup_{\{\boldsymbol{\theta}\in\mathbb{R}^p \,:\, J_1(\boldsymbol{\theta})\leq 1\}} J_2(\boldsymbol{A}\boldsymbol{\theta}).$$

Note that $\|\!|\boldsymbol{A}|\!\|_{J_1\to J_2}$ is bounded (this follows from Lemma 3.3(v)). 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

$$\Upsilon(T, c) = \inf_{\{\boldsymbol{\omega}\in\mathbb{R}^p \,:\, J(\boldsymbol{\omega}_S) < cJ(\boldsymbol{\omega}_T)\}} \frac{\|\!|\mathrm{P}_T|\!\|_{2\to J}\big\|\boldsymbol{X}\boldsymbol{\omega}\big\|_2}{n^{1/2}(J(\boldsymbol{\omega}_T) - J(\boldsymbol{\omega}_S)/c)}. \tag{4.1}$$

This generalizes the compatibility factor introduced in [65] for the Lasso (and used in [20]). 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 variational/penalized estimators (1.2); see [28, 60–63]. One can see in particular that $\Upsilon(T, c)$ is larger than the smallest singular value of $\boldsymbol{X}_T$.

4.1 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 (1.4) under assumptions **(H.1)** and **(H.3)**. The oracle inequality is provided in terms of the loss

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

By Lemma 3.1, we indeed have $R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq \frac{\nu}{2n}\big\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\big\|_2^2$, and equality is attained when is quadratic. If $\nabla\varphi$ is also Lipschitz continuous, then Lemma 3.1 asserts that $R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big)$ is equivalent to the quadratic loss.

Theorem 4.1. *Consider the data generated by (1.1) and the EWA estimator $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ in (1.4) with the density (1.3), where F and J satisfy Assumptions **(H.1)**-**(H.2)** and **(H.3)**. Then, for any $0 < \epsilon < \nu$ and $\tau > 1$ such that $\lambda \geq \tau J^\circ\big(\boldsymbol{X}^\top \boldsymbol{\xi}\big)/n$, the following holds,*

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\big) \leq \inf_{\boldsymbol{\theta}\in\mathbb{R}^p} \left(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \frac{\lambda^2\big(\tau J^\circ(e_{\boldsymbol{\theta}}) + 1\big)^2 \|\mathrm{P}_{T_{\boldsymbol{\theta}}}\|_{2\to J}^2}{2\tau^2(\nu-\epsilon)\Upsilon\Big(T_{\boldsymbol{\theta}}, \frac{\tau J^\circ(e_{\boldsymbol{\theta}})+1}{\tau-1}\Big)^2} \right) + p\beta + \frac{1}{2n\epsilon}\big\|(\mathrm{Id} - \nabla\varphi)(\boldsymbol{X}\boldsymbol{\theta}_0)\big\|_2^2. \tag{4.2}$$

Remark 4.1.

1. *The oracle inequality is sharp. The remainder in it has several 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 $O(n^{-1})$. When φ is quadratic, the last term in (4.2) vanishes in which case we can set $\epsilon = 0$. We also point out that if $\nabla\varphi$ is also κ-Lipschitz continuous, then Lemma 3.1 asserts that $R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big)$ is equivalent to a quadratic loss. This means that the oracle inequaliy in Theorem 4.1 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.*

2. *If J is such that $e_{\boldsymbol{\theta}} \in \partial J(\boldsymbol{\theta}) \subset \mathcal{C}^\circ$ (typically for a strong gauge by (3.2)), 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 (4.2).*

3. *A close inspection of the proof of Theorem 4.1 reveals that the term $p\beta$ can be improved to the smaller bound*

$$p\beta + \Big(V_n(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}) - \mathbb{E}_{\mu_n}\left[V_n(\boldsymbol{\theta})\right]\Big) + \frac{\nu}{2n}\Big(\big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}\big\|_2^2 - \mathbb{E}_{\mu_n}\left[\big\|\boldsymbol{X}\boldsymbol{\theta}\big\|_2^2\right]\Big) \leq p\beta,$$

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

Proof. By convexity of J and Lemma 3.1, we have for any $\boldsymbol{\eta} \in \partial V_n(\boldsymbol{\theta})$ and any $\overline{\boldsymbol{\theta}} \in \mathbb{R}^p$,

$$D_{V_n}^{\boldsymbol{\eta}}\big(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta}\big) \geq \frac{\nu}{2n}\big\|\boldsymbol{X}\overline{\boldsymbol{\theta}} - \boldsymbol{X}\boldsymbol{\theta}\big\|_2^2.$$

Taking the expectation w.r.t. to μ_n on both sides, and using Jensen inequality, we get

$$\begin{aligned} V_n(\overline{\boldsymbol{\theta}}) &\geq \mathbb{E}_{\mu_n}\left[V_n(\boldsymbol{\theta})\right] + \mathbb{E}_{\mu_n}\left[\langle \boldsymbol{\eta}, \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\rangle\right] + \frac{\nu}{2n}\mathbb{E}_{\mu_n}\left[\left\|\boldsymbol{X}\overline{\boldsymbol{\theta}} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2\right] \\ &\geq V_n(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}) + \mathbb{E}_{\mu_n}\left[\langle \boldsymbol{\eta}, \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\rangle\right] + \frac{\nu}{2n}\left\|\boldsymbol{X}\overline{\boldsymbol{\theta}} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}\right\|_2^2. \end{aligned}$$

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 A for details). It then follows from Proposition A.1[4] that

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

We thus deduce the inequality

$$V_n(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}) - V_n(\overline{\boldsymbol{\theta}}) \leq p\beta - \frac{\nu}{2n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\overline{\boldsymbol{\theta}}\right\|_2^2, \quad \forall \overline{\boldsymbol{\theta}} \in \mathbb{R}^p. \tag{4.3}$$

By definition of the Bregman divergence and in view of **(H.1)**, we have

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

Using **(H.1)**, Young inequality and the duality inequality (3.1), we have for any $\epsilon > 0$

$$\begin{aligned} R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\big) - R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) &\leq \Big(V_n(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}) - V(\boldsymbol{\theta})\Big) + \frac{1}{2n\epsilon}\left\|(\mathrm{Id} - \nabla\varphi)(\boldsymbol{X}\boldsymbol{\theta}_0)\right\|_2^2 \\ &\quad + \frac{\epsilon}{2n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2 + \frac{1}{n}J^\circ\Big(\boldsymbol{X}^\top\boldsymbol{\xi}\Big)J(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{\theta}) - \lambda\big(J(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}) - J(\boldsymbol{\theta})\big) \\ &\leq \Big(V_n(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}) - V(\boldsymbol{\theta})\Big) + \frac{1}{2n\epsilon}\left\|(\mathrm{Id} - \nabla\varphi)(\boldsymbol{X}\boldsymbol{\theta}_0)\right\|_2^2 \\ &\quad + \frac{\epsilon}{2n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2 + \frac{\lambda}{\tau}\Big(J(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{\theta}) - \tau\big(J(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}) - J(\boldsymbol{\theta})\big)\Big). \end{aligned}$$

Denote $\boldsymbol{\omega} = \widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{\theta}$. By virtue of **(H.3)**, Lemma 3.1 and (3.1), we obtain

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

[4] In the appendix, we provide a self-contained proof based on a novel Moreau-Yosida regularization. In [20, Corollary 1 and 2], an alternative proof is given using an absolute continuity argument since μ_n is locally Lipschitz, hence a Sobolev function.

This inequality together with (4.3) (applied with $\overline{\boldsymbol{\theta}} = \boldsymbol{\theta}$) and (4.1) yield

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

where we applied Young inequality to get the last inequality. Taking the infimum over $\boldsymbol{\theta} \in \mathbb{R}^p$ yields the desired bound. □

Partly smooth functions Theorem 4.1 has a nice instanciation for the case of partly smooth functions convex functions. These functions have been thoroughly studied recently in [60–63] for various statistical and inverse problems. In particular, a finite-valued convex function J is partly smooth at a point $\boldsymbol{\theta}$ relative to an active set $\mathcal{M} \ni \boldsymbol{\theta}$, if $\mathcal{M}$ is a smooth submanifold of $\mathbb{R}^p$, and J behaves smoothly along $\mathcal{M}$ and sharply transverse to it. In addition, the partial smoothness submanifold is always unique. There are two consequences of partial smoothness. First, the sharpness property is equivalently characterized by $\mathcal{T}_{\boldsymbol{\theta}}(\mathcal{M}) = T_{\boldsymbol{\theta}}$, where $\mathcal{T}_{\boldsymbol{\theta}}(\mathcal{M})$ is the tangent space of $\mathcal{M}$ at $\boldsymbol{\theta}$. Second, $e_{\boldsymbol{\theta}}$ coincides with the Riemannian gradient of J alont $\mathcal{M}$. Moreover, both these two properties locally persists at all nearby points of $\boldsymbol{\theta}$ on $\mathcal{M}$; see [60, 62] for details. As an example, a smooth function at $\boldsymbol{\theta}$ is partly smooth relative to the whole space $\mathbb{R}^p$. All popular penalty functions discussed in Sections 3.5 and 4.3 are also partly smooth (see [60, 62]).

Let's denote $\mathscr{M}$ the set of all possible partial smoothness active submanifolds associated to J. An interesting fact is that there are many situations where $\mathscr{M}$ contains finitely many active submanifolds, as is the case for the examples of Section 3.5. With this notation at hand, the oracle inequality (4.2) now reads

$$
R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\big) \leq \inf_{\substack{\mathcal{M}\in\mathscr{M} \\ \boldsymbol{\theta}\in\mathcal{M}}} \left(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \frac{\lambda^2\big(\tau J^\circ(e_{\boldsymbol{\theta}})+1\big)^2\|\mathrm{P}_{\mathcal{T}_{\boldsymbol{\theta}}(\mathcal{M})}\|_{2\to J}^2}{2\tau^2(\nu-\epsilon)\Upsilon\Big(\mathcal{T}_{\boldsymbol{\theta}}(\mathcal{M}_{\boldsymbol{\theta}}), \frac{\tau J^\circ(e_{\boldsymbol{\theta}})+1}{\tau-1}\Big)^2} \right) + p\beta + \frac{1}{2n\epsilon}\big\|(\mathrm{Id} - \nabla\varphi)(\boldsymbol{X}\boldsymbol{\theta}_0)\big\|_2^2.
$$

4.2 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 (4.2).

Theorem 4.2. *Consider the data generated by (1.1) and the penalized estimator $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ in (1.2), where F and J satisfy Assumptions (H.1) and (H.3). Then, for any $0 < \epsilon < \nu$ and $\tau > 1$ such that $\lambda \geq \tau J^\circ\big(\boldsymbol{X}^\top\boldsymbol{\xi}\big)/n$, the following holds,*

$$
R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}, \boldsymbol{\theta}_0\big) \leq \inf_{\boldsymbol{\theta}\in\mathbb{R}^p} \left(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \frac{\lambda^2\big(\tau J^\circ(e_{\boldsymbol{\theta}}) + 1\big)^2\|\mathrm{P}_{T_{\boldsymbol{\theta}}}\|_{2\to J}^2}{2\tau^2(\nu-\epsilon)\Upsilon\Big(T_{\boldsymbol{\theta}}, \frac{\tau J^\circ(e_{\boldsymbol{\theta}})+1}{\tau-1}\Big)^2} \right) + \frac{1}{2n\epsilon}\big\|(\mathrm{Id} - \nabla\varphi)(\boldsymbol{X}\boldsymbol{\theta}_0)\big\|_2^2. \tag{4.4}
$$

In plain words, 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 4.1-3). Thus letting $p \to 0$ in (4.2), one recovers the oracle inequality (4.4) of penalized estimators. In particular, for $\beta = O\big((pn)^{-1}\big)$, this is at most of the same order as the first one in the remainder. Observe also that the penalized estimator $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ does not need the moment assumption **(H.2)** for (4.4) to hold.

Proof. The proof follows the same lines as that of Theorem 4.1 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 $\boldsymbol{\theta} \in \mathbb{R}^p$

$$V_n(\boldsymbol{\theta}) \geq V_n(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}) + \frac{\nu}{2n}\big\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}\big\|_2^2. \tag{4.5}$$

Continuing exactly as just after (4.3), replacing $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ with $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ and invoking (4.5) instead of (4.3), we arrive at the claimed result. □

4.3 Applications

In this section, we exemplify our oracle inequalities for the penalties described in Section 3.5.

4.3.1 Lasso

To lighten the notation, let $I_{\boldsymbol{\theta}} = \mathrm{supp}(\boldsymbol{\theta})$. From (3.4), it is easy to see that

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

where last bound holds as an equality whenever $\boldsymbol{\theta} \neq 0$. It remains to check whether the event $\{\lambda \geq \tau\|\boldsymbol{X}^\top\boldsymbol{\xi}\|_\infty/n\}$ holds with high probability when the noise is random under reasonable assumptions on $\boldsymbol{\xi}$. We have the following corollary.

Corollary 4.1. *Let the data generated by (1.1) with noise $\boldsymbol{\xi}$ whose entries are n iid subgaussian centered random variables 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 the Lasso penalty (3.3) and F satisfies Assumptions **(H.1)**-**(H.2)**. Let $0 < \epsilon < \nu$. Suppose that $\lambda \geq \tau\sigma\sqrt{\frac{2\log(p/\delta)}{n}}$, for some $\tau > 1$ and $\delta \in]0,1[$. Then, with probability at least $1-\delta$, the following holds*

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\big) \leq \inf_{\substack{I\subset\{1,\dots,p\}\\ \boldsymbol{\theta}:\ \mathrm{supp}(\boldsymbol{\theta})=I}} \left(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \frac{\lambda^2(\tau+1)^2|I|}{2\tau^2(\nu-\epsilon)\Upsilon\Big(\mathrm{Span}\{a_i\}_{i\in I}, \frac{\tau+1}{\tau-1}\Big)^2} \right) + p\beta + \frac{1}{2n\epsilon}\big\|(\mathrm{Id} - \nabla\varphi)(\boldsymbol{X}\boldsymbol{\theta}_0)\big\|_2^2, \tag{4.6}$$

and

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}, \boldsymbol{\theta}_0\big) \leq \inf_{\substack{I\subset\{1,\dots,p\}\\ \boldsymbol{\theta}:\ \mathrm{supp}(\boldsymbol{\theta})=I}} \left(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \frac{\lambda^2(\tau+1)^2|I|}{2\tau^2(\nu-\epsilon)\Upsilon\Big(\mathrm{Span}\{a_i\}_{i\in I}, \frac{\tau+1}{\tau-1}\Big)^2} \right) + \frac{1}{2n\epsilon}\big\|(\mathrm{Id} - \nabla\varphi)(\boldsymbol{X}\boldsymbol{\theta}_0)\big\|_2^2. \tag{4.7}$$

When φ is quadratic, the oracle inequality (4.7) recovers [20, Theorem 1] in the exactly sparse case. (4.7) also coincides in this case with the oracle inequaliy in [56, Theorem 4] (see also [21, Theorem2]). The first term in the remainder grows as $\frac{|I|\log(p)}{n}$ which is the classical scaling under the individual sparsity scenario.

Proof. By the union bound, we have

$$\mathbb{P}\left(\left\|\boldsymbol{X}^\top\boldsymbol{\xi}\right\|_\infty > \epsilon\right) \leq \mathbb{P}\left(\bigcup_{i=1}^{p} |\langle \boldsymbol{X}_i, \boldsymbol{\xi}\rangle| > \epsilon\right) \leq \sum_{i=1}^{p} \mathbb{P}\left(|\langle \boldsymbol{X}_i, \boldsymbol{\xi}\rangle| > \epsilon\right) \leq p \max_{i\in\{1,\dots,p\}} \mathbb{P}\left(|\langle \boldsymbol{X}_i, \boldsymbol{\xi}\rangle| > \epsilon\right).$$

Since $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n$ are n *iid* subgaussian centered random variables with parameter σ^2, for any $i \in \{1, \dots, p\}$, $\langle \boldsymbol{X}_i, \boldsymbol{\xi}\rangle$ is a sum of n random variables *iid* subgaussian centered random variables with parameter $\sigma^2 \boldsymbol{X}_{i,j}^2$. Owing to the Hoeffding bound [34], we obtain

$$\mathbb{P}\left(|\langle \boldsymbol{X}_i, \boldsymbol{\xi}\rangle| > \epsilon\right) \leq 2\exp\left(-\epsilon^2/\left(2\sigma^2 \left\|\boldsymbol{X}_i\right\|_2^2\right)\right).$$

Taking $\epsilon = \lambda n/\tau$, we conclude. □

4.3.2 Group Lasso

Recall the notations in Section 3.5.2, and denote $I_{\boldsymbol{\theta}} = \mathrm{supp}_{\mathcal{B}}(\boldsymbol{\theta})$. From (3.6), we have

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

where $|I_{\boldsymbol{\theta}}|$ is nothing but the number of active blocks in $\boldsymbol{\theta}$, and the last bound holds as an equality whenever $\boldsymbol{\theta} \neq 0$.

When the noise is *iid* Gaussian, we get the following oracle inequalities.

Corollary 4.2. *Let the data generated by (1.1) with noise $\boldsymbol{\xi}$ whose entries are n iid $\mathcal{N}(0, \sigma^2)$. 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 (3.5) with L non-overlapping blocks of equal size K. Assume that $\boldsymbol{X}$ is such that $\max_i \left\|\left\|\boldsymbol{X}_{b_i}^\top \boldsymbol{X}_{b_i}\right\|\right\|_{2\to 2} \leq n$. Let $0 < \epsilon < \nu$. Suppose that $\lambda \geq \tau\sigma\sqrt{\frac{K+2\left(2\delta\log(L)+\sqrt{K\delta\log(L)}\right)}{n}}$, for some $\tau > 1$ and $\delta > 1$. Then, with probability at least $1 - 2L^{1-\delta}$, the following holds*

$$R_n\left(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\right) \leq \inf_{\substack{I\subset\{1,\dots,L\} \\ \boldsymbol{\theta}:\ \mathrm{supp}_{\mathcal{B}}(\boldsymbol{\theta})=I}} \left(R_n\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0\right) + \frac{\lambda^2(\tau+1)^2|I|}{2\tau^2(\nu-\epsilon)\Upsilon\left(\mathrm{Span}\{a_j\}_{j\in b_i, i\in I}, \frac{\tau+1}{\tau-1}\right)^2}\right) + p\beta + \frac{1}{2n\epsilon}\left\|(\mathrm{Id} - \nabla\varphi)(\boldsymbol{X}\boldsymbol{\theta}_0)\right\|_2^2, \tag{4.8}$$

and

$$R_n\left(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}, \boldsymbol{\theta}_0\right) \leq \inf_{\substack{I\subset\{1,\dots,L\} \\ \boldsymbol{\theta}:\ \mathrm{supp}_{\mathcal{B}}(\boldsymbol{\theta})=I}} \left(R_n\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0\right) + \frac{\lambda^2(\tau+1)^2|I|}{2\tau^2(\nu-\epsilon)\Upsilon\left(\mathrm{Span}\{a_j\}_{j\in b_i, i\in I}, \frac{\tau+1}{\tau-1}\right)^2}\right) + \frac{1}{2n\epsilon}\left\|(\mathrm{Id} - \nabla\varphi)(\boldsymbol{X}\boldsymbol{\theta}_0)\right\|_2^2. \tag{4.9}$$

The first remainder term is on the order $\frac{|I|\left(\sqrt{K}+\sqrt{2\log(L)}\right)^2}{n}$, which is similar to the scaling that has been provided in the literature for EWA with other group sparsity priors [27, 50]. Similar rates were given for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ with group Lasso in [41, 43, 64].

Proof. The proof is essentially an adaptation of [41, Lemma 3.1], where we also used that for all $i = 1, \dots, L$,

$$n^{-1} \operatorname{tr}\left(\boldsymbol{X}_{b_i}^\top \boldsymbol{X}_{b_i}\right) \leq K n^{-1} \left|\!\left|\!\left| \boldsymbol{X}_{b_i}^\top \boldsymbol{X}_{b_i} \right|\!\right|\!\right|_{2\to 2} \leq K.$$

□

4.3.3 Analysis (group) Lasso

We now turn to the prior penalty (3.7). Recall the notations in Section 3.5.3, and let $I_{\boldsymbol{\theta}} = \mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top \boldsymbol{\theta})$ and $\Lambda_{\boldsymbol{\theta}} = \bigcup_{i \in I_{\boldsymbol{\theta}}} b_i$. Without loss of generality, we assume that $\boldsymbol{D}$ is a Parseval tight frame, meaning that $\boldsymbol{D}\boldsymbol{D}^\top = \mathrm{Id}$, and thus $\boldsymbol{D}^+ = \boldsymbol{D}^\top$. This together with (3.8)-(3.9) entail

$$\begin{aligned}
\left|\!\left|\!\left| \mathrm{P}_{T_{\boldsymbol{\theta}}} \right|\!\right|\!\right|_{2\to J} &= \sup_{\left\|\boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}\right\|_2 \leq 1} \left\|\boldsymbol{D}^\top \boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}\right\|_{1,2} = \sup_{\left\|\boldsymbol{D}^\top \boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}\right\|_2 \leq 1} \left\|\boldsymbol{D}^\top \boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}\right\|_{1,2} = \sup_{\left\|\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}}^\top \boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}\right\|_2 \leq 1} \left\|\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}}^\top \boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}\right\|_{1,2} \\
&= \sup_{\left\|\mathrm{P}_{\mathrm{Span}\{a_i\}_{i\in\Lambda_{\boldsymbol{\theta}}}} \boldsymbol{D}^\top \boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}\right\|_2 \leq 1} \left\|\mathrm{P}_{\mathrm{Span}\{a_i\}_{i\in\Lambda_{\boldsymbol{\theta}}}} \boldsymbol{D}^\top \boldsymbol{\omega}_{T_{\boldsymbol{\theta}}}\right\|_{1,2} \\
&= \sup_{\left\{\left\|\mathrm{P}_{\mathrm{Span}\{a_i\}_{i\in\Lambda_{\boldsymbol{\theta}}}} \boldsymbol{\omega}\right\|_2 \leq 1\right\} \cap \mathrm{Span}\left((\boldsymbol{D}^\top)_{T_{\boldsymbol{\theta}}}\right)} \left\|\mathrm{P}_{\mathrm{Span}\{a_i\}_{i\in\Lambda_{\boldsymbol{\theta}}}} \boldsymbol{\omega}\right\|_{1,2} \\
&\leq \left|\!\left|\!\left| \mathrm{P}_{\mathrm{Span}\{a_i\}_{i\in\Lambda_{\boldsymbol{\theta}}}} \right|\!\right|\!\right|_{2\to\|\cdot\|_{1,2}} = \sqrt{|I_{\boldsymbol{D}^\top\boldsymbol{\theta}}|}.
\end{aligned}$$

However, from (3.8), we do not have in general $\left\|\boldsymbol{D}^\top \mathrm{P}_{\mathrm{Ker}(\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}^c}^\top)} \boldsymbol{D} e^{\left|\!\left|\!\left|\cdot\right|\!\right|\!\right|_{1,2}}_{\boldsymbol{D}^\top\boldsymbol{\theta}}\right\|_{\infty,2} \leq 1$.

With arguments analogue to those for proving Corollary 4.2, we arrive at the following oracle inequalities.

Corollary 4.3. *Let the data generated by (1.1) with noise $\boldsymbol{\xi}$ whose entries are n iid $\mathcal{N}(0, \sigma^2)$. 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 (3.7) with L blocks of equal size K. Assume that $\boldsymbol{D}$ is a Parseval tight frame, and $\boldsymbol{X}$ is such that* $\max_i \left|\!\left|\!\left| \boldsymbol{D}_{b_i}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{D}_{b_i} \right|\!\right|\!\right|_{2\to 2} \leq n$. *Let $0 < \epsilon < \nu$. Suppose that $\lambda \geq \tau\sigma\sqrt{\frac{K + 2\left(2\delta\log(L) + \sqrt{K\delta\log(L)}\right)}{n}}$, for some $\tau > 1$ and $\delta > 1$. Then, with probability at least $1 - 2L^{1-\delta}$, the following holds*

$$\begin{aligned}
R_n\left(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\right) \leq \inf_{\substack{I\subset\{1,\dots,L\} \\ \boldsymbol{\theta}:\, \mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top\boldsymbol{\theta})=I}} \Bigg(& R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \frac{\lambda^2\left(\tau\left\|\boldsymbol{D}^\top \mathrm{P}_{\mathrm{Ker}(\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}^c}^\top)} \boldsymbol{D} e^{\left|\!\left|\!\left|\cdot\right|\!\right|\!\right|_{1,2}}_{\boldsymbol{D}^\top\boldsymbol{\theta}}\right\|_{\infty,2} + 1\right)^2 |I|}{2\tau^2(\nu-\epsilon)\Upsilon\left(\mathrm{Ker}(\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}^c}^\top), \frac{\tau\left\|\boldsymbol{D}^\top \mathrm{P}_{\mathrm{Ker}(\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}^c}^\top)} \boldsymbol{D} e^{\left|\!\left|\!\left|\cdot\right|\!\right|\!\right|_{1,2}}_{\boldsymbol{D}^\top\boldsymbol{\theta}}\right\|_{\infty,2} + 1}{\tau - 1}\right)^2} \Bigg) \\
& + p\beta + \frac{1}{2n\epsilon}\left\|(\mathrm{Id} - \nabla\varphi)(\boldsymbol{X}\boldsymbol{\theta}_0)\right\|_2^2, \qquad (4.10)
\end{aligned}$$

and

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}},\boldsymbol{\theta}_0\big) \leq \inf_{\substack{I\subset\{1,\dots,L\}\\ \boldsymbol{\theta}:\ \mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top\boldsymbol{\theta})=I}} \left(R_n\big(\boldsymbol{\theta},\boldsymbol{\theta}_0\big) + \frac{\lambda^2\Big(\tau\Big\|\boldsymbol{D}^\top \mathrm{P}_{\mathrm{Ker}(\boldsymbol{D}^\top_{\Lambda^c_{\boldsymbol{\theta}}})}\boldsymbol{D}e_{\boldsymbol{D}^\top\boldsymbol{\theta}}^{\|\|_{1,2}}\Big\|_{\infty,2}+1\Big)^2|I|}{2\tau^2(\nu-\epsilon)\Upsilon\left(\mathrm{Ker}(\boldsymbol{D}^\top_{\Lambda^c_{\boldsymbol{\theta}}}),\frac{\tau\Big\|\boldsymbol{D}^\top \mathrm{P}_{\mathrm{Ker}(\boldsymbol{D}^\top_{\Lambda^c_{\boldsymbol{\theta}}})}\boldsymbol{D}e_{\boldsymbol{D}^\top\boldsymbol{\theta}}^{\|\|_{1,2}}\Big\|_{\infty,2}+1}{\tau-1}\right)^2} \right) + \frac{1}{2n\epsilon}\big\|(\mathrm{Id}-\nabla\varphi)(\boldsymbol{X}\boldsymbol{\theta}_0)\big\|_2^2. \tag{4.11}$$

To the best of our knowledge, this result is new to the literature. The scaling of the remainder term is the same as in [27, Remark 6.2] and [50] with analysis sparsity priors different from ours (the authors in the latter also assume that $\boldsymbol{D}$ is invertible).

4.3.4 Anti-sparsity

From Section 3.5.4, recall the saturation support $I^{\mathrm{sat}}_{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. From (3.11), we get

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

with equality whenever $\boldsymbol{\theta}\neq 0$.

Assuming that the noise is *iid* zero-mean Gaussian, we get the following oracle inequalities.

Corollary 4.4. *Let the data generated by (1.1) with noise $\boldsymbol{\xi}$ whose entries are n iid $\mathcal{N}(0,\sigma^2)$. Assume that $\boldsymbol{X}$ is such that $\max_{i,j}|\boldsymbol{X}_{i,j}|\leq 1/\sqrt{n}$. 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 anti-sparsity penalty (3.10). Let $0<\epsilon<\nu$. Suppose that $\lambda\geq\sqrt{\frac{2}{\pi}}\tau(1+\delta)\sigma p/n$, for some $\tau>1$ and $\delta>0$. Then, with probability at least $1-e^{-\frac{\delta^2}{\pi}}$, the following holds*

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}},\boldsymbol{\theta}_0\big) \leq \inf_{\substack{I\subset\{1,\dots,p\}\\ \boldsymbol{\theta}:\ I^{\mathrm{sat}}_{\boldsymbol{\theta}}=I}} \left(R_n\big(\boldsymbol{\theta},\boldsymbol{\theta}_0\big) + \frac{\lambda^2(\tau+1)^2}{2\tau^2(\nu-\epsilon)\Upsilon\Big(\big\{\overline{\boldsymbol{\theta}}:\overline{\boldsymbol{\theta}}_I\in\mathbb{R}\,\mathrm{sign}(\boldsymbol{\theta}_I)\big\},\frac{\tau+1}{\tau-1}\Big)^2} \right) + p\beta + \frac{1}{2n\epsilon}\big\|(\mathrm{Id}-\nabla\varphi)(\boldsymbol{X}\boldsymbol{\theta}_0)\big\|_2^2, \tag{4.12}$$

and

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}},\boldsymbol{\theta}_0\big) \leq \inf_{\substack{I\subset\{1,\dots,p\}\\ \boldsymbol{\theta}:\ I^{\mathrm{sat}}_{\boldsymbol{\theta}}=I}} \left(R_n\big(\boldsymbol{\theta},\boldsymbol{\theta}_0\big) + \frac{\lambda^2(\tau+1)^2}{2\tau^2(\nu-\epsilon)\Upsilon\Big(\big\{\overline{\boldsymbol{\theta}}:\overline{\boldsymbol{\theta}}_I\in\mathbb{R}\,\mathrm{sign}(\boldsymbol{\theta}_I)\big\},\frac{\tau+1}{\tau-1}\Big)^2} \right) + \frac{1}{2n\epsilon}\big\|(\mathrm{Id}-\nabla\varphi)(\boldsymbol{X}\boldsymbol{\theta}_0)\big\|_2^2. \tag{4.13}$$

We are not aware of any result of this kind in the literature. The first remainder term scales as $\big(\frac{p}{n}\big)^2$. The bound imposed on $\boldsymbol{X}$ is similar to what is generally assumed in the vector quantization literature [42, 55].

Proof. The function $\boldsymbol{\xi} \mapsto \left\|\boldsymbol{X}^\top \boldsymbol{\xi}\right\|_1$ is Lipschitz continuous with Lipschitz constant $\sum_{i=1}^p \|\boldsymbol{X}_i\|_2$. Moreover, $\mathbb{E}\left[\left\|\boldsymbol{X}^\top \boldsymbol{\xi}\right\|_1\right] = \sqrt{\frac{2}{\pi}}\sigma \sum_{i=1}^p \|\boldsymbol{X}_i\|_2$. Setting $\epsilon = \lambda n/\tau - \sqrt{\frac{2}{\pi}}\sigma \sum_{i=1}^p \|\boldsymbol{X}_i\|_2 \geq \delta\sqrt{\frac{2}{\pi}}\sigma \sum_{i=1}^p \|\boldsymbol{X}_i\|_2$, it follows from the Gaussian concentration of Lipschitz functions [39] that

$$\begin{aligned}
\mathbb{P}\left(\left\|\boldsymbol{X}^\top \boldsymbol{\xi}\right\|_1 \geq \lambda n/\tau\right) &= \mathbb{P}\left(\left\|\boldsymbol{X}^\top \boldsymbol{\xi}\right\|_1 - \mathbb{E}\left[\left\|\boldsymbol{X}^\top \boldsymbol{\xi}\right\|_1\right] > \epsilon\right) \\
&\leq \exp\left(-\frac{\left(\lambda n/\tau - \sqrt{\frac{2}{\pi}}\sigma \sum_{i=1}^p \|\boldsymbol{X}_i\|_2\right)^2}{2\sigma^2\left(\sum_{i=1}^p \|\boldsymbol{X}_i\|_2\right)^2}\right) \\
&\leq \exp\left(-\frac{\delta^2}{\pi}\right).
\end{aligned}$$

□

4.3.5 Nuclear norm

We now turn to the nuclear norm case. Recall the notations of Section 3.5.5. 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

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

where $\boldsymbol{X}^i$ is a matrix in $\mathbb{R}^{p_1 \times p_2}$. We denote $\|\cdot\|_F$ the associated norm. From (3.13), it is immediate to see that whenever $\boldsymbol{\theta} \neq 0$,

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

Moreover, by Hölder and Von Neumann's trace inequalities, we have

$$\|\mathrm{P}_{T_{\boldsymbol{\theta}}}\|_{\mathrm{F}\to *} = \inf_{\boldsymbol{\omega} \in T_{\boldsymbol{\theta}}} \sqrt{\operatorname{rank}(\boldsymbol{\omega})} \leq \sqrt{\operatorname{rank}(\boldsymbol{\theta})} = \sqrt{r},$$

since $\boldsymbol{\theta} \in T_{\boldsymbol{\theta}}$ (see Section 3.3). For standard Gaussian noise $\boldsymbol{\xi}$, we need to bound

$$\|\boldsymbol{X}^*(\boldsymbol{\xi})\|_{2\to 2} = \left\|\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{\xi}_i\right\|_{2\to 2},$$

which is the operator norm of a random series with matrix coefficients. Thus arguing as in [20, Lemma 6] (which relies on [59, Theorem 4.1.1], we get the following oracle inequalities for the nuclear norm. Define

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

Corollary 4.5. *Let the data generated by (1.1) with a linear operator $\boldsymbol{X} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n$, with noise $\boldsymbol{\xi}$ whose entries are n iid $\mathcal{N}(0, \sigma^2)$. Assume that $v(\boldsymbol{X}) \leq n$. 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 (3.10). Let $0 < \epsilon < \nu$. Suppose that $\lambda \geq \tau\sigma\sqrt{\frac{2\log((p_1+p_2)/\delta)}{n}}$, for some $\tau > 1$ and $\delta > 0$. Then, with probability at least $1 - \delta$, the following*

holds

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\big) \leq \inf_{\substack{r\in\{1,\dots,\min(p_1,p_2)\} \\ \boldsymbol{\theta}:\ \mathrm{rank}\,\boldsymbol{\theta}=r}} \left(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \frac{\lambda^2(\tau+1)^2 r}{2\tau^2(\nu-\epsilon)\Upsilon\Big(T_{\boldsymbol{\theta}}, \frac{\tau+1}{\tau-1}\Big)^2} \right) + (p_1+p_2)\beta + \frac{1}{2n\epsilon}\big\|(\mathrm{Id}-\nabla\varphi)(\boldsymbol{X}\boldsymbol{\theta}_0)\big\|_2^2, \tag{4.14}$$

and

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}, \boldsymbol{\theta}_0\big) \leq \inf_{\substack{r\in\{1,\dots,\min(p_1,p_2)\} \\ \boldsymbol{\theta}:\ \mathrm{rank}\,\boldsymbol{\theta}=r}} \left(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \frac{\lambda^2(\tau+1)^2 r}{2\tau^2(\nu-\epsilon)\Upsilon\Big(T_{\boldsymbol{\theta}}, \frac{\tau+1}{\tau-1}\Big)^2} \right) + \frac{1}{2n\epsilon}\big\|(\mathrm{Id}-\nabla\varphi)(\boldsymbol{X}\boldsymbol{\theta}_0)\big\|_2^2. \tag{4.15}$$

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 $\mathbb{R}^{p_1\times p_2}$) [24, Theorem 3.19]. The first remainder term now scales as $\frac{r\log(p_1+p_2)}{n}$ and we recover the same rate as in [20, Theorem 3] for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and in [37, Theorem 2] for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$.

A 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 A.1. *Let f be a finite-valued and convex function. Then*

(i) $\big({}^{\gamma}f(\boldsymbol{\theta})\big)_{\gamma>0}$ *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* $\gamma\searrow 0$.

(iv) ${}^{\gamma}f$ *is coercive.*

Proof. (i) [4, Proposition 12.32]. (ii) [4, Proposition 12.29]. (iii) Since f is continuous, it is subdifferentiable everywhere, and the result follows from [4, Corollary 23.46(i)]. (iv) Combine [52, Theorem 3.31 and Exercise 3.29(b)]. □

We are now equipped to prove the following important result[5].

[5]The result will be proved using Moreau-Yosida regularization. Yet another alternative proof could be based on mollifiers for approximating subdifferentials.

Proposition A.1. *Let the density μ_n in (1.3), where F and J satisfy Assumptions (H.1) and (H.3). Then, $\forall \overline{\boldsymbol{\theta}} \in \mathbb{R}^p$,*

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

Proof. Let $V_n^\gamma(\boldsymbol{\theta}) \overset{\text{def}}{=} F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) + \lambda\, {}^\gamma J(\boldsymbol{\theta})$ and define $\mu_n^\gamma(\boldsymbol{\theta}) \overset{\text{def}}{=} \exp\left(-V_n^\gamma(\boldsymbol{\theta})\right)/Z$, where $0 < Z < +\infty$ is the normalizing constant of the density μ_n. Assumption (H.1) and Lemma A.1(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

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

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. Since F is bounded below (by (H.1)), and ${}^\gamma J$ is coercive by Lemma A.1-(iv), V_n^γ is also coercive. In turn, it follows from [52, Theorem 11.8(c)] that for some $a \in]0, +\infty[$, there exists $b \in]-\infty, +\infty[$ such that for all $\boldsymbol{\theta} \in \mathbb{R}^p$

$$\mu_n^\gamma(\boldsymbol{\theta}) \le \exp\left(-a\|\boldsymbol{\theta}\|_2 - b\right)/Z. \tag{A.1}$$

Moreover, for all $\boldsymbol{\theta} \in \mathbb{R}^p$, we have $\partial V_n(\boldsymbol{\theta}) \subset \mathcal{C}^\circ$ by Theorem 3.1(i). This together with Lemma A.1-(iii) yield

$$\|\nabla {}^\gamma J(\boldsymbol{\theta})\|_2 \le \|\big(\partial V_n(\boldsymbol{\theta})(\boldsymbol{\theta})\big)^0\|_2 \le \operatorname{diam}(\mathcal{C}^\circ),$$

where $\operatorname{diam}(\mathcal{C}^\circ)$ is the diameter of the compact set $\mathcal{C}^\circ$. Altogether, we have

$$\begin{aligned}
\left|\langle \mu_n^\gamma(\boldsymbol{\theta})\nabla V_n^\gamma(\boldsymbol{\theta}), \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\rangle\right| &\le \mu_n^\gamma(\boldsymbol{\theta})\Big(\left|\langle \boldsymbol{X}^\top F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}), \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\rangle\right| + \lambda \|\nabla {}^\gamma J(\boldsymbol{\theta})\|_2 \|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2\Big) \\
&\le (Z\exp(b))^{-1} \exp\left(-a\|\boldsymbol{\theta}\|_2\right)\left(\left|\langle F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}), \boldsymbol{X}(\overline{\boldsymbol{\theta}} - \boldsymbol{\theta})\rangle\right| + \lambda \operatorname{diam}(\mathcal{C}^\circ)\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2\right).
\end{aligned}$$

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

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

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

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

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

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

where we used coercivity of V_n^γ (see above) 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 A.1(i) that $\mu_n^\gamma \to \mu_n$ as $\gamma \to 0$. Thus, arguing again as in (A.1), 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. □

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

References

[1] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Comput.*, 9(7):1545–1588, Oct. 1997.

[2] F. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

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

[4] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.

[5] G. Biau. Analysis of a random forests model. *J. Mach. Learn. Res.*, 13(1):1063–1095, Apr. 2012.

[6] G. Biau and L. Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *J. Multivar. Anal.*, 101(10):2499–2518, Nov. 2010.

[7] G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.*, 9:2015–2033, June 2008.

[8] P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[9] L. Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, Aug. 1996.

[10] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, Oct. 2001.

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

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

[13] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11:1–11:37, June 2011.

[14] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

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

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

[17] X. Chen, Q. Lin, S. Kim, J. G. Carbonell, and E. P. Xing. An efficient proximal-gradient method for general structured sparse learning. *Preprint arXiv:1005.4717*, 2010.

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

[19] A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Mach. Learn.*, 72(1-2):39–61, Aug. 2008.

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

[21] A. S. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the lasso. *Bernoulli*, 23(1):552–581, 02 2017.

[22] A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Proceedings of the 20th Annual Conference on Learning Theory*, COLT'07, pages 97–111, Berlin, Heidelberg, 2007. Springer-Verlag.

[23] A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and langevin monte-carlo. *J. Comput. Syst. Sci.*, 78(5):1423–1443, Sept. 2012.

[24] A. Daniilidis, D. Drusvyatskiy, and A. S. Lewis. Orthogonal invariance and identifiability. Technical report, arXiv 1304.1198, 2013.

[25] D. Donoho. 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, 2006.

[26] A. Durmus, E. Moulines, and M. Pereyra. Sampling from convex non continuously differentiable functions, when Moreau meets Langevin. Preprint hal-01267115, Feb. 2016.

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

[28] M. J. Fadili, G. Peyré, S. Vaiter, C. Deledalle, and J. Salmon. Stable recovery with analysis decomposable priors. In *Proc. Sampta'13*, pages 113–116, 2013.

[29] M. Fazel, H. Hindi, and S. P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the American Control Conference*, volume 6, pages 4734–4739. IEEE, 2001.

[30] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256 – 285, 1995.

[31] R. Genuer. *Random Forests: elements of theory, variable selection and applications*. Theses, Université Paris Sud - Paris XI, Nov. 2010.

[32] B. Guedj and P. Alquier. Pac-bayesian estimation and prediction in sparse additive models. *Electron. J. Statist.*, 7:264–291, 2013.

[33] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis And Minimization Algorithms*, volume I and II. Springer, 2001.

[34] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[35] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In A. P. Danyluk, L. Bottou, and M. L. Littman, editors, *ICML'09*, volume 382, page 55, 2009.

[36] H. Jégou, T. Furon, and J.-J. Fuchs. Anti-sparse coding for approximate nearest neighbor search. In *IEEE ICASSP*, pages 2029–2032, 2012.

[37] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 10 2011.

[38] G. Lecué. Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.*, 35(4):1698–1721, 08 2007.

[39] M. Ledoux. *The concentration of measure phenomenon*. Mathematical surveys and monographs. American Mathematical Society, Providence (R.I.), 2001. L'ISSN figurant sur le substitut de la page de titre 0076-5376 correspond à la revue Mathematicals surveys. Le titre a changé en 1981 en Mathematicals surveys and monographs et porte le numéro 0885-4653.

[40] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Inf. Comput.*, 108(2):212–261, Feb. 1994.

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

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

[43] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, December 2012.

[44] A. Nemirovski. Topics in non-parametric statistics, 2000.

[45] Y. Nesterov and I. U. E. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer, 2004.

[46] M. Osborne, B. Presnell, and B. Turlach. A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3):389–403, 2000.

[47] G. Peyré, J. Fadili, and C. Chesneau. Adaptive Structured Block Sparsity Via Dyadic Partitioning. In *EUSIPCO*, Barcelona, Spain, Aug. 2011.

[48] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[49] P. Rigollet and A. Tsybakov. Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, 16(3):260–280, 2007.

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

[51] R. Rockafellar. *Convex analysis*, volume 28. Princeton University Press, 1996.

[52] R. T. Rockafellar and R. Wets. *Variational analysis*, volume 317. Springer Verlag, 1998.

[53] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[54] R. E. Schapire. The strength of weak learnability. *Mach. Learn.*, 5(2):197–227, July 1990.

[55] C. Studer, W. Yin, and R. G. Baraniuk. Signal representations with minimum ℓ_∞-norm. In *50th Annual Allerton Conference on Communication, Control, and Computing,*, 2012.

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

[57] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B. Methodological*, 58(1):267–288, 1996.

[58] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[59] J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

[60] S. Vaiter, C. Deledalle, M. J. Fadili, G. Peyré, and C. Dossal. The degrees of freedom of partly smooth regularizers. *Annals of the Institute of Statistical Mathematics*, 2016. in press.

[61] S. Vaiter, M. Golbabaee, M. J. Fadili, and G. Peyré. Model selection with low complexity priors. *Information and Inference: A Journal of the IMA (IMAIAI)*, 2015.

[62] S. Vaiter, G. Peyré, and M. J. Fadili. Low complexity regularization of linear inverse problems. In G. Pfander, editor, *Sampling Theory, a Renaissance*, Applied and Numerical Harmonic Analysis (ANHA). Birkhäuser/Springer, 2015.

[63] S. Vaiter, G. Peyré, and M. J. Fadili. Model consistency of partly smooth regularizers. Technical report, Preprint Hal-00987293, 2015. submitted.

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

[65] S. van de Geer and P. Buhlmann. On the conditions used to prove oracle results for the lasso. *Electron. J. Statist.*, 3:1360–1392, 2009.

[66] V. G. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, COLT '90, pages 371–386, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc.

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

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

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