Aggregation for low complexity recovery: Oracle inequalities and algorithms

Duy Tung Luu

Normandie Université

THESE

Pour obtenir le diplôme de doctorat

Spécialité Mathématiques

Préparée au sein de l'Université de Caen Normandie

Aggregation for low complexity recovery: Oracle inequalities and algorithms

Présentée et soutenue par

Duy Tung LUU

Thèse soutenue publiquement le 23 Novembre 2017 devant le jury composé de			
Pierre	ALQUIER	Professeur, ENSAE	Examinateur
Christophe	CHESNEAU	Maître de Conférences HDR, Université de Caen Normandie	Codirecteur de thèse
Arnak	DALALYAN	Professeur, ENSAE	Examinateur
Jalal	FADILI	Professeur des Universités, ENSICAEN	Directeur de thèse
Guillaume	LECUE	Chargé de Recherche CNRS, ENSAE	Rapporteur
Erwan	LE PENNEC	Professeur, École Polytechnique	Rapporteur
Gabriel	PEYRE	Directeur de Recherche CNRS, ENS Paris	Examinateur

Thèse dirigée par Jalal FADILI et Christophe CHESNEAU, laboratoire GREYC

UNIVERSITÉ CAEN NORMANDIE

ENSI CAEN ÉCOLE PUBLIQUE D'INGÉNIEURS CENTRE DE RECHERCHE

ed MIIS

GREYC

Remerciements

En premier lieu, je tiens à remercier Jalal Fadili et Christophe Chesneau, mon directeur et codirecteur de thèse, qui ont accepté de m'encadrer pendant le stage et la période doctorale. Tout au long de ces années de travail, ils m'ont apporté une disponibilité, des connaissances approfondies de plusieurs domaines mathématiques, des conseils précieux et des corrections précises pour la rédaction de mes produits scientifiques. En plus du travail de recherche, ils m'ont également aidé à débloquer des problèmes dans la vie quotidienne et à trouver des orientations pour l'avenir. Leur compétence, leur méthodologie et attitude de travail m'ont appris beaucoup.

Je suis très honoré par la présence dans mon jury de thèse de Messieurs Pierre Alquier, Arnak Dalalyan, Guillaume Lecué, Erwan Le Pennec et Gabriel Peyré. Je tiens à remercier en particulier Erwan Le Pennec et Guillaume Lecué d'avoir accepté de rapporter sur mon travail de thèse.

Mes remerciements vont aussi aux membres de l'équipe Image du GREYC qui m'ont offert un environnement de recherche agréable et de grande qualité. Je remercie particulièrement François Lozes pour ses discussions quotidiennes et ses aides dans le domain informatique, Jingwei Liang pour les échanges scientifiques, et Nicole Delamotte et Gaëlle Lenogue pour leur aide précieuse dans les méandres des procédures administratives.

Je tiens à remercier le Conseil Régional de Basse Normandie pour avoir financé cette thèse.

Je remercie aussi Marie-Anne Poursat, Christine Kéribin, Michel Prenat, Erwan Le Pennec, Vincent Brault et l'ensemble de mes enseignants de mes cinq années d'étude à la faculté d'Orsay, pour m'avoir transmis leur passion des mathématiques, statistiques et informatique, et qui m'ont permis d'effectuer cette thèse. Je remercie particulièrment monsieur Erwan Le Pennec de m'avoir recommandé auprès de mon directeur de thèse, Jalal Fadili.

Enfin, J'adresse toute ma gratitude à ma famille et tous mes chers amis pour leur encouragement permanent et leur indéfectible soutien.

Abstract

In many areas of statistics, including signal and image processing, high-dimensional estimation is an important task to recover an object of interest. However, in the overwhelming majority of cases, the recovery problem is ill-posed. Fortunately, even if the ambient dimension of the object to be restored (signal, image, video) is very large, its intrinsic "complexity" is generally small. The introduction of this prior information can be done through two approaches: (i) penalization (very popular) and (ii) aggregation by exponential weighting (EWA). The penalized approach aims at finding an estimator that minimizes a data loss function penalized by a term promoting objects of low (simple) complexity. The EWA combines a family of pre-estimators, each associated with a weight exponentially promoting the same objects of low complexity.

This manuscript consists of two parts: a theoretical part and an algorithmic part. In the theoretical part, we first propose the EWA with a new family of priors promoting analysis-group sparse signals whose performance is guaranteed by oracle inequalities. Next, we will analysis the penalized estimator and EWA, with a general prior promoting simple objects, in a unified framework for establishing some theoretical guarantees. Two types of guarantees will be established: (i) prediction oracle inequalities, and (ii) estimation bounds. We will exemplify them for particular cases some of which studied in the literature. In the algorithmic part, we will propose an implementation of these estimators by combining Monte-Carlo simulation (Langevin diffusion process) and proximal splitting algorithms, and show their guarantees of convergence. Several numerical experiments will be considered for illustrating our theoretical guarantees and our algorithms.

Keywords: High-dimensional estimation, low-complexity prior, exponential weighted aggregation, penalized estimation, oracle inequality, Langevin diffusion, forward-backward algorithm, consistency.

Résumé

Dans plusieurs domaines des statistiques, y compris le traitement du signal et des images, l'estimation en grande dimension est une tâche importante pour recouvrer un objet d'intérêt. Toutefois, dans la grande majorité de situations, ce problème est mal-posé. Cependant, bien que la dimension ambiante de l'objet à restaurer (signal, image, vidéo) est très grande, sa "complexité" intrinsèque est généralement petite. La prise en compte de cette information a priori peut se faire au travers de deux approches: (i) la pénalisation (très populaire) et (ii) l'agrégation à poids exponentiels (EWA). L'approche penalisée vise à chercher un estimateur qui minimise une attache aux données pénalisée par un terme promouvant des objets de faible complexité (simples). L'EWA combine une famille des pré-estimateurs, chacun associé à un poids favorisant exponentiellement des pré-estimateurs, lesquels privilègent les mêmes objets de faible complexité.

Ce manuscrit se divise en deux grandes parties: une partie théorique et une partie algorithmique. Dans la partie théorique, on propose l'EWA avec une nouvelle famille d'a priori favorisant les signaux parcimonieux à l'analyse par group dont la performance est garantie par des inégalités oracle. Ensuite, on analysera l'estimateur pénalisé et EWA, avec des a prioris généraux favorisant des objets simples, dans un cardre unifié pour établir des garanties théoriques. Deux types de garanties seront montrés: (i) inégalités oracle en prédiction, et (ii) bornes en estimation. On les déclinera ensuite pour des cas particuliers dont certains ont été étudiés dans littérature. Quant à la partie algorithmique, on y proposera une implémentation de ces estimateurs en alliant simulation Monte-Carlo (processus de diffusion de Langevin) et algorithmes d'éclatement proximaux, et montrera leurs garanties de conver-

gence. Plusieurs expériences numériques seront décrites pour illustrer nos garanties théoriques et nos algorithmes.

Mots-clés: Estimation en grande dimension, a priori de faible complexité, agrégation à poids exponentiels, estimation pénalisée, inégalité d'oracle, diffusion de Langevin, algorithme explicite-implicite, consistence.

Table of contents

Chapter 1

Introduction

Contents

1.1 Overview

1.1.1 Problem statement

1.1.1.1 Regression problem

This manuscript focuses on the fundamental problem of high-dimensional estimation in statistics. We are given a sample $(x_i, y_i) \in (\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R})$, $i = 1, \dots, n$ generated from two random variables (X, Y) defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. X and Y are respectively called design and response. We aim to exploit the link between them. This link is expressed via the regression function

$$f : \mathcal{X} \to \mathbb{R}, \quad x \mapsto f(x) = \mathbb{E}\left[Y \mid X = x\right]. \tag{1.1}$$

For any function $g : \mathcal{X} \to \mathbb{R}$, let us define $\boldsymbol{g} \stackrel{\text{def}}{=} (g(x_1), \dots, g(x_n))^\top$, and

$$\left\|g\right\|_n \stackrel{\text{def}}{=} \sqrt{\frac{1}{n}\sum_{i=1}^{n} g^2(x_i)}. \tag{1.2}$$

.

1.1.1.2 Aggregation approach

The approach of aggregation has been introduced in machine learning to combine different techniques (see [154, 96]) with some procedures such as bagging [17], boosting [71, 129] and random forests [2, 18, 13, 12, 74, 11]. It assumes that there exists a dictionary $\mathcal{H} = \{f_j : \mathcal{X} \to \mathbb{R},\ j \in \{1, \dots, p\}\}$ such that f is well approximated by a linear combination of elements in $\mathcal{H}$. Here, the f_j are known and may be either fixed atoms in a basis or pre-estimators. More precisely, let

$$f_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \sum_{j=1}^{p} \boldsymbol{\theta}_j f_j, \quad \forall \boldsymbol{\theta} \in \mathbb{R}^p.$$

We approximate f by $f_{\boldsymbol{\theta}_0}$, with $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{0,1}, \dots, \boldsymbol{\theta}_{0,p})^\top \in \mathbb{R}^p$ is a reference vector defined as any solution to the following minimization problem

$$\boldsymbol{\theta}_0 \in \operatorname*{Argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathbb{E}\left[F(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{y})\right], \tag{1.3}$$

where $\boldsymbol{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ and $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a loss function that assigns to each $\boldsymbol{\theta} \in \mathbb{R}^p$ a cost $F(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{y})$.

Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $[\boldsymbol{X}]_{i,j} \stackrel{\text{def}}{=} f_j(x_i)$ be the design matrix. A usual instance of this statistical setting is the standard linear regression model, i.e.

$$\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\xi}, \quad \text{with } \boldsymbol{f} = \boldsymbol{X}\boldsymbol{\theta}^*, \tag{1.4}$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ called design matrix, $\boldsymbol{\theta}^* \in \mathbb{R}^p$ called regression vector, and $\boldsymbol{\xi} \in \mathbb{R}^n$ are i.i.d. zero-mean random errors. Then, with

$$F(\boldsymbol{u}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{u}\|_2^2,$$

$\boldsymbol{\theta}_0$ coincides with $\boldsymbol{\theta}^*$, hence $f_{\boldsymbol{\theta}_0}$ coincides with f. The loss F defined above is called quadratic loss.

1.1.1.3 Low-complexity

Our goal is to provide estimators of $\boldsymbol{\theta}_0$ in a high-dimensional context whose performance is certified by theoretical guarantees. Namely, the estimation (c.f. from (1.4)) is ill-posed. To circumvent this difficulty, we will exploit the prior that $\boldsymbol{\theta}_0$ has some low-complexity structure which is manifested through the fact that $\boldsymbol{\theta}_0$ belongs to a low-dimensional model subset. That is, even if the ambient dimension p of $\boldsymbol{\theta}_0$ is very large, the intrinsic dimension of the model subset is much smaller than the sample size n. This makes it possible to build estimates $\widehat{\boldsymbol{\theta}}$ with good provable performance guarantees under appropriate conditions.

There has been a flurry of research on the use of low-complexity regularization in ill-posed recovery problems in various areas including statistics and machine learning. Among which sparsity and low-rank are the most popular.

(i) Sparsity prior states that the non-zero components of $\boldsymbol{\theta}_0$ is much smaller that n.

(ii) Low-rank prior considers $\boldsymbol{\theta}_0$ as a matrix whose rank is much smaller that n.

The sparsity prior can be generalized in several ways:

(i) Analysis sparsity: let $q \geq p$ and $\boldsymbol{D}^\top \in \mathbb{R}^{q \times p}$ be a linear analysis operator. The analysis sparsity means that $\|\boldsymbol{D}^\top \boldsymbol{\theta}\|_0 \ll n$. A typical example is total variation [128] where the operator $\boldsymbol{D}^\top$ corresponds to "finite differences" (i.e. $(\boldsymbol{D}^\top \boldsymbol{\theta})_1 = \boldsymbol{\theta}_1$, $(\boldsymbol{D}^\top \boldsymbol{\theta})_j = \boldsymbol{\theta}_j - \boldsymbol{\theta}_{j-1}$, $\forall j \geq 2$). Another example is the fused Lasso [128] where $\boldsymbol{D}^\top$ is a positive combination of the identity and finite differences.

(ii) Group sparsity: that corresponds to saying that the aggregator $\boldsymbol{\theta}$ is group sparse. Group sparsity is at the heart of the group Lasso and related methods [161, 81, 89, 106, 117, 105]. In the EWA

context, the group sparsity prior is considered in [123] as an application of the aggregation of orthogonal projectors.

(iii) Analysis-group sparsity: that combines (i) and (ii). Some popular applications are: estimation of 2-D piecewise constant images with the isotropic total variation (see [128]), and estimation of signals with group-sparsity where the groups are overlapping (see [113, 39]).

In the manuscript, we will illustrate our results on the sparsity prior and its generalized versions and also give some analogous results under low-rank prior.

1.1.2 Oracle inequalities

This type of guarantees dates back, for instance, to the work [60, 57, 61] on orthogonal wavelet thresholding estimators. Oracle inequalities (according to the terminology introduced in e.g. [57]), which are at the heart of many of our theoretical guarantees, quantify the quality of an estimator compared to the best possible one among a collection of candidates. Formally, when the dictionary is deterministic (resp. random), the performance of a candidate is measured by a function $R(\cdot, \boldsymbol{f})$ (resp. $\mathbb{E}\left[R(\cdot, \boldsymbol{f})\right]$) whose definition depends on the estimation guarantee that one is targeting. Generally, there are three main guarantees:

(i) Correct model selection: The goal is to ensure that the estimated vector identifies the same low-dimensional model subset as that of $\boldsymbol{\theta}_0$.

(ii) Estimation guarantee: The goal is to assess the performance of estimating $\boldsymbol{\theta}_0$ directly, e.g. $R(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{f}) = \left\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\right\|_2^2$.

(iii) Prediction guarantee: The goal is to assess the performance of estimating $\boldsymbol{f}$, e.g. $R(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{f}) = \left\|\boldsymbol{f}_{\boldsymbol{\theta}} - \boldsymbol{f}\right\|_2^2$.

In the manuscript, we consider the oracle inequalities for deterministic dictionaries. Let Θ be the set of candidates. A candidate $\boldsymbol{f}_{\bar{\boldsymbol{\theta}}}$ is called *oracle* if it has the best performance in Θ, i.e.

$$\boldsymbol{f}_{\bar{\boldsymbol{\theta}}} = \operatorname*{Arginf}_{\boldsymbol{f}_{\boldsymbol{\theta}} \in \left\{\boldsymbol{f}_{\boldsymbol{\theta}'} \,:\, \boldsymbol{\theta}' \in \Theta\right\}} R(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{f}).$$

Since $\boldsymbol{f}$ is unknown, $\bar{\boldsymbol{\theta}}$ is not accessible. However, one can find an estimator $\widehat{\boldsymbol{f}}$ that mimics as much as possible the performance of the best model of aggregation in a given class Θ. This idea is expressed in the following type of inequalities:

$$\mathbb{E}\left[R(\widehat{\boldsymbol{f}}, \boldsymbol{f})\right] \leq C \inf_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{f}) + \Delta_{n,p}(\Theta), \tag{1.5}$$

or

$$R(\widehat{\boldsymbol{f}}, \boldsymbol{f}) \leq C \inf_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{f}) + \Delta_{n,p}(\Theta) \quad \text{with a high probability}, \tag{1.6}$$

where $C \geq 1$ and the remainder term $\Delta_{n,p}(\Theta)$ depends on the performance of the estimator, the complexity of Θ, the dimension p and the sample size n. Inequality (1.5) (resp. (1.6)) is called oracle inequality in expectation (resp. probability). Namely, in some context, we prefer to measure the quality of estimators by the fluctuations in $R(\widehat{\boldsymbol{f}}, \boldsymbol{f})$ under a large probability instead of summarizing it by the expectation $\mathbb{E}\left[R(\widehat{\boldsymbol{f}}, \boldsymbol{f})\right]$. In general, the oracle inequality in probability is harder to achieve than its counterpart in expectation.

An estimator with good oracle properties would correspond to C close to 1 (ideally, $C = 1$, in which case the inequality is coined "sharp"), $\Delta_{n,p}(\Theta)$ should be small even if $n \ll p$ and decreases rapidly to 0 as $n \to +\infty$. When $C = 1$, the results are more interesting as $\widehat{\boldsymbol{f}}$ mimics exactly the performance of the oracle.

Besides, the choice of Θ is crucial: on the one hand, a non suitable choice can lead to a large bias. On the other hand, if Θ is too complex, the remainder term becomes large. Then, a suitable choice

for Θ must achieve a good bias-complexity trade-off. The works in [141] and [119] have proven the optimal rate of the remainder term for several choices of Θ.

In this manuscript, we consider the more general version of oracle inequalities (1.5) and (1.6) which are defined as

$$\mathbb{E}\left[R(\widehat{\boldsymbol{f}}, \boldsymbol{f})\right] \leq C \inf_{\boldsymbol{\theta}\in\Theta'} \{R(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{f}) + \Delta_{n,p}(\boldsymbol{\theta})\}, \tag{1.7}$$

$$R(\widehat{\boldsymbol{f}}, \boldsymbol{f}) \leq C \inf_{\boldsymbol{\theta}\in\Theta'} \{R(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{f}) + \Delta_{n,p}(\boldsymbol{\theta})\} \quad \text{with a high probability.} \tag{1.8}$$

Indeed, when $\Theta \subseteq \Theta'$, (1.7) (resp. (1.8)) implies directly (1.5) (resp. (1.6)). Such type of inequalities are for instance well adapted under the sparsity scenario. Namely, the complexity of $\boldsymbol{\theta}$ in the remainder term is characterized by the sparsity parameters (like the number of its non-zero components), in which case these inequalities are called sparse oracle inequalities (SOI).

1.1.3 Variational/Penalized approach

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 manuscript, we consider the class of estimators obtained by solving the convex optimization problem[1]

$$\widehat{\boldsymbol{\theta}}_n^{\text{PEN}} \in \operatorname*{Argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \{V_n(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \frac{1}{n} F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) + J_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta})\}, \tag{1.9}$$

where the regularizing penalty $J_{\boldsymbol{\lambda}_n}$ is a proper closed convex function that promotes some specific notion of simplicity/low-complexity, and $\boldsymbol{\lambda}_n$ is the vector of parameters.

The ℓ_0 penalty has been studied as a regularizer for sparse recovery (see [132, 20, 57, 61]). This type of penalties yields the optimal oracle inequalities in several problems with no assumption on the dictionary. However, its numerical computation is an NP-hard problem which becomes impossible in a high-dimensional context.

To deal with this issue, several works consist on convexifying the optimization problem which becomes computable by convex programming solvers, see [55] for a comprehensive review. A prominent member is the Lasso [38, 137, 110, 56, 22, 14, 19, 87] and its variants such the analysis/fused Lasso [128, 138], SLOPE [15, 134] or group Lasso [5, 161, 4, 156].

Besides, for low rank matrix recovery, the nuclear norm minimization is motivated by various applications including robust PCA, phase retrieval, control and computer vision [118, 31, 70, 33]. See [108, 19, 149, 146] for generalizations and comprehensive reviews.

An interesting result was introduced in [9] which generalizes the works in [49, 95] based on aggregation of pre-estimators by exponential weighting. Using a penalized procedure inspired from Q-aggregation (see [43, 120]), the authors relaxed the conditions of affine pre-estimators and established oracle inequalities (in probability) of model selection with an optimal remainder term for linear regression with i.i.d. Gaussian and sub-Gaussian noises.

1.1.4 Exponential Weighted Aggregation (EWA)

Let $(\Theta, \mathcal{A})$ be a space equipped with a σ-algebra and

$$\mathcal{F}_\Theta = \{f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R} : \boldsymbol{\theta} \in \Theta\}$$

[1] To avoid trivialities, the set of minimizers is assumed non-empty.

be a given dictionary where $\boldsymbol{\theta} \to f_{\boldsymbol{\theta}}(x)$ is measurable $\forall x \in \mathcal{X}$. The functions $f_{\boldsymbol{\theta}}$ may be deterministic or random. To distinguish from the initial dictionary $\mathcal{H}$ (see Section 1.1.1.2), we call $\mathcal{F}_\Theta$ as EWA-dictionary. The aggregators depend on the nature of $f_{\boldsymbol{\theta}}$ if the latter is random. Otherwise, the aggregators are defined via the probability measure

$$\widehat{\mu}_n(d\boldsymbol{\theta}) = \frac{\exp\left(-F(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{y})/\beta\right)\pi(d\boldsymbol{\theta})}{\int_\Theta \exp\left(-F(\boldsymbol{f}_{\boldsymbol{\omega}}, \boldsymbol{y})/\beta\right)\pi(d\boldsymbol{\omega})} = \frac{\exp\left(-F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y})/\beta\right)\pi(d\boldsymbol{\theta})}{\int_\Theta \exp\left(-F(\boldsymbol{X}\boldsymbol{\omega}, \boldsymbol{y})/\beta\right)\pi(d\boldsymbol{\omega})}, \tag{1.10}$$

where $\beta > 0$ called temperature parameter and π called prior which is a probability measure on Θ. The choice of β weights the contribution of the data loss/prior. When β is small (resp. large), the data loss risk (resp. prior) dominates. Remind that, in the standard linear regression, we can set $F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) = \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2$. Next, we define the aggregate by

$$\widehat{f}_n = f_{\widehat{\boldsymbol{\theta}}_n^{\text{EWA}}}, \quad \text{with} \quad \widehat{\boldsymbol{\theta}}_n^{\text{EWA}} = \int_\Theta \boldsymbol{\theta}\widehat{\mu}_n(d\boldsymbol{\theta}). \tag{1.11}$$

This idea was initially proposed in [154, 96, 85] with a uniform prior on a finite set Θ.

The references below consider EWA in the standard linear regression (1.4) with EWA-dictionaries are either deterministic or random.

In the individual sparsity context, the works in [50, 45, 44, 52] consider deterministic EWA-dictionaries. These papers proposed several PAC-Bayesian type of oracle inequalities under different assumptions. Especially, the assumptions in [52] depend only on the noise and turns out to be fulfilled for a large class of noises. This serves to construct, for a suitable prior and dictionary, a SOI with a remainder term of order $O\left(\left\|\boldsymbol{\theta}\right\|_0 \log(p)/n\right)$, which scales linearly with the sparsity level and increases in p only logarithmically.

The random EWA-dictionary case is tackled in [109]. The initial idea is to obtain two independent samples from the initial sample by randomization or sample splitting (see [159, 121, 92]). The first sample is used to construct the pre-estimators, and the aggregation is performed on the second sample conditionally on the first one. However this idea does not work when the observations are not i.i.d.. Several authors have proposed exponentially aggregating linear pre-estimators without splitting, and with discrete priors on the weights. Typical cases of linear pre-estimators are orthogonal projectors on all possible linear subspaces that are in the model set (e.g. in the sparsity context, linear subspaces spanned by the standard basis restricted to supports of increasing size). This was introduced in [95] and generalized to the high-dimensional context in [122] (with Exponential Screening) where F is the Stein's unbiased risk estimate (SURE) and the noise is Gaussian. More recent works such as [49] generalizes the idea where the pre-estimators are affine and the priors are continuous. Moreover, the work in [107] enlarges the family of noise where $\boldsymbol{\xi}$ is a sub-Gaussian and its components are non i.i.d. using a penalized version of the SURE.

A shortcoming of EWA is the suboptimality in deviation. The work in [43, Section 2] show that the EWA lead a suboptimal remainder term for oracle inequalities in probability. To deal with that, the authors in [47] consider a modified version of EWA where

$$\widehat{f}_n = f_{\widehat{\boldsymbol{\theta}}_n^{\text{EWA}}}, \quad \text{with} \quad \widehat{\boldsymbol{\theta}}_n^{\text{EWA}} = \int_\Theta \boldsymbol{\theta}\widehat{\mu}_n(d\boldsymbol{\theta}), \tag{1.12}$$

where

$$\widehat{\mu}_n(d\boldsymbol{\theta}) = \frac{\exp\left(-V_n(\boldsymbol{\theta})/\beta\right)d\boldsymbol{\theta}}{\int_\Theta \exp\left(-V_n(\boldsymbol{\omega})/\beta\right)d\boldsymbol{\omega}}, \quad \beta > 0.$$

We remind that $V_n(\boldsymbol{\theta}) \overset{\text{def}}{=} \frac{1}{n}F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) + J_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta})$. Compared to (1.10), the difference in (1.12) is that we use the same scale for risk term and prior term. When β is close to 0, the candidate minimizing V_n dominates, the EWA becomes a penalized/regularization procedure. With the interpretation (1.12), $\widehat{\boldsymbol{\theta}}_n^{\text{EWA}}$ 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 $\pi(\boldsymbol{\theta}) \propto \exp\left(-J_{\boldsymbol{\lambda}_n}(\boldsymbol{\theta})/\beta\right)$. The authors in [47] have established oracle inequalities in probability with the optimal remainder term for EWA with ℓ_1 and the nuclear norm penalties.

1.1.5 Numerical implementation of EWA

When the components in the EWA-dictionary are orthogonal projectors on all possible linear subspaces, the work in [122] proposed an algorithm of type Metropolis-Hastings for the Exponential Screening estimator where the transition kernel is symmetric which simplifies the calculation of acceptance rate. This algorithm avoids the computation of 2^{R-1} least square estimators, where R is the rank of the design matrix.

In the manuscript, we focus on the computation of EWA when the prior is continuous. That corresponds to an integration problem which becomes very involved to solve analytically or even numerically in high dimension. A classical alternative is to approximate $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ via a Markov chain Monte-Carlo (MCMC) method which consists in sampling from $\widehat{\mu}_n$ by constructing an appropriate Markov chain whose stationary distribution is $\widehat{\mu}_n$, and to compute sample path averages based on the output of the Markov chain. The theory of MCMC methods is based on that of Markov chains on continuous state space. As in [52], we here use the Langevin diffusion process; see [124].

1.1.5.1 Langevin diffusion

Continuous dynamics A Langevin diffusion $\boldsymbol{L}$ in $\mathbb{R}^p$, $p \geq 1$ is a homogeneous Markov process defined by the stochastic differential equation (SDE)

$$d\boldsymbol{L}(t) = \frac{1}{2}\boldsymbol{\rho}(\boldsymbol{L}(t))dt + d\boldsymbol{W}(t), \quad t > 0,\ \boldsymbol{L}(0) = \boldsymbol{l}_0, \tag{1.13}$$

where $\boldsymbol{\rho} = \nabla \log \mu$, μ is everywhere non-zero and suitably smooth target density function on $\mathbb{R}^p$, $\boldsymbol{W}$ is a p-dimensional Brownian process and $\boldsymbol{l}_0 \in \mathbb{R}^p$ is the initial value. Under mild assumptions, the SDE (1.13) has a unique strong solution and, $\boldsymbol{L}(t)$ has a stationary distribution with density precisely μ [124, Theorem 2.1]. $\boldsymbol{L}(t)$ is therefore interesting for sampling from μ. In particular, this opens the door to approximating integrals

$$\int_{\mathbb{R}^p} \boldsymbol{\theta}\mu(\boldsymbol{\theta})d\boldsymbol{\theta}$$

by the average value of a Langevin diffusion, i.e.

$$\frac{1}{T}\int_0^T \boldsymbol{L}(t)dt$$

for a large enough T. Under additional assumptions on μ, the expected squared error of the approximation can be controlled [158].

Forward Euler discretization In practice, in simulating the diffusion sample path, we cannot follow exactly the dynamic defined by the SDE (1.13). Instead, we must discretize it. A popular discretization is given by the forward (Euler) scheme, which reads

$$\boldsymbol{L}_{k+1} = \boldsymbol{L}_k + \frac{\delta}{2}\boldsymbol{\rho}(\boldsymbol{L}_k) + \sqrt{\delta}\boldsymbol{Z}_k, \quad t > 0,\ \boldsymbol{L}_0 = \boldsymbol{l}_0,$$

where $\delta > 0$ is a sufficiently small constant discretization step-size and $\{\boldsymbol{Z}_k\}_k$ are i.i.d. $\sim \mathcal{N}(0, \mathbf{I}_p)$. The average value $\frac{1}{T}\int_0^T \boldsymbol{L}(t)dt$ can then be naturally approximated via the Riemann sum

$$\frac{\delta}{T}\sum_{k=0}^{\lfloor T/\delta \rfloor - 1} \boldsymbol{L}_k, \tag{1.14}$$

where $\lfloor T/\delta \rfloor$ denotes the interger part of T/δ. It is then natural to approximate $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ by applying this discretization strategy to the Langevin diffusion with $\widehat{\mu}_n$ as the target density. However, quantitative

consistency guarantees of this discretization require μ (hence $\boldsymbol{\rho}$) to be sufficiently smooth. A more comprehensive account on Langevin diffusion can be found in [82, 158, 86].

1.2 Contributions

1.2.1 Chapter 3: PAC-Bayesian Risk Bounds for Analysis-group Sparse Regression by EWA

Consider an additive non-parametric regression model, i.e.

$$\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\xi},$$

with $\boldsymbol{\xi}$ is a zero-mean Gaussian noise or any bounded symmetric noise. Suppose that $F(\boldsymbol{u}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{y}\|_2^2$, we impose an analysis-group sparsity prior (see also Section 1.1.1.3) where

(i) vectors are group-sparse in the domain of a transform $\boldsymbol{D}^\top$, where $\boldsymbol{D}$ is a frame, hence surjective but not invertible;

(ii) the groups $\{b_1, \ldots, b_L\}$ have the same size, denoted by K, and satisfy $b_1 \oplus \cdots \oplus b_L = \{1, \ldots, q\}$.

For any $\boldsymbol{x} \in \mathbb{R}^q$ and $\mathcal{I} \subseteq \{1, \ldots, q\}$, denote $\boldsymbol{x}_\mathcal{I}$ is the subvector whose entries are those of $\boldsymbol{x}$ restricted to the indices in $\mathcal{I}$. We establish an EWA by procedure (1.11) with the prior $\pi(d\boldsymbol{\theta})$ as the form

$$\pi(d\boldsymbol{\theta}) \propto \prod_{l=1}^L \exp\left(-\alpha^a \|[\boldsymbol{D}^\top\boldsymbol{\theta}]_{b_l}\|_2^a\right) g\left(\|[\boldsymbol{D}^\top\boldsymbol{\theta}]_{b_l}\|_2\right) I_\Theta(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad \alpha \geq 0,\ a \in (0, 1], \tag{1.15}$$

where I_Θ is the characteristic function of the set Θ (= 1 in Θ and 0 otherwise), and $g : \mathbb{R}_+ \to \mathbb{R}_+$ satisfies some assumptions which allow us establishing a general analysis-group SOI where the remainder term depends on the number of active groups.

For an appropriate choice of g and Θ, this remainder term scales as (see Corollary 3.4.6).

$$O\left(\frac{\sum_{l=1}^L I_{\left\{\boldsymbol{u} \in \mathbb{R}^K\,:\, \|\boldsymbol{u}\|_2 \neq 0\right\}}(\boldsymbol{\theta}_{b_l}) \log(L)}{n}\right). \tag{1.16}$$

This rate coincides with the classical one $O\left(\|\boldsymbol{\theta}\|_0 \log(p)/n\right)$ under the sparsity scenario, i.e. $\boldsymbol{D} = \mathbf{I}_p$ and $K = 1$.

Moreover, we implement our EWA with the forward-backward proximal Langevin Monte-Carlo (LMC) algorithms proposed in Chapter 6, and illustrate the performance of our estimator on some numerical examples in Chapter 7.

Relation to previous work Our results is a non trivial generalization of the work in [52] to the analysis-group sparsity context. In fact, we consider the class of prior (1.15) with some wise assumptions on the function g well suited to the analysis-group context, and then we exhibit an analysis-group SOI. In the sparsity context, the prior in [52] coincides with (1.15) for the special choice

$$g(x) = \frac{1}{(\tau^2 + x^2)^2}, \quad x \in \mathbb{R}_+.$$

However, with effects of groups, this prior does not satisfy our assumptions anymore. In this case, we propose another prior, with

$$g(x) = \frac{1}{(\tau^b + x^b)^c}, \quad x \in \mathbb{R}_+,\ b \in]0, 1],\ c \in]2 + K/b, +\infty[$$

which is more feasible hence adaptive in analysis-group context. Its performance is guaranteed by the analysis-group SOI with the remainder term of order (1.16). We also emphasize the fact that, in our results, $\boldsymbol{D}$ is a frame and thus is not necessarily invertible unlike the previous work [123].

1.2.2 Chapter 4: Sharp Oracle Inequalities for Low-complexity Priors

Consider a general regression problem (1.1). By the aggregation approach, the regression function f is approximated by $f_{\boldsymbol{\theta}_0}$ (see Section 1.1.1) with

$$\boldsymbol{\theta}_0 \in \operatorname*{Argmin}_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathbb{E}\left[F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y})\right]. \tag{1.17}$$

Here, the loss function F is supposed to be smooth and convex.

This chapter provides a unified analysis where we capture the essential ingredients behind the low-complexity priors promoted by $J_{\boldsymbol{\lambda}_n}$ in (1.12) and (1.9), relying on sophisticated arguments from convex analysis and previous work [67, 145, 147, 144, 146]. Our main contributions are summarized as follows.

(i) We show in Theorem 4.3.1 and Theorem 4.3.3 that the EWA $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ in (1.12) and the variational/penalized estimator $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ in (1.9) satisfy (deterministic) sharp oracle inequalities for prediction with optimal remainder term, for general data losses F beyond the usual quadratic one, and $J_{\boldsymbol{\lambda}_n}$ is a proper finite-valued sublinear function (i.e. $J_{\boldsymbol{\lambda}_n}$ is finite-valued convex and positively homogeneous). For both estimators, the remainder contains a term that encodes the complexity of the model promoted by $J_{\boldsymbol{\lambda}_n}$. For EWA, there is an additional overhead term, $p\beta$, which captures the influence of temperature parameter.

(ii) When the observations are random, we prove oracle inequalities in probability (see Section 4.3.3). The theory is non-asymptotic in nature, as it yields explicit bounds that hold with high probability for finite sample sizes, and reveals the dependence on the dimension and other structural parameters of the model.

(iii) For the standard linear model with Gaussian or sub-Gaussian noise, and a quadratic loss, we deliver refined versions of these oracle inequalities in probability (see Section 4.4). We underscore the role of the Gaussian width, a concept that captures important geometric characteristics of sets in $\mathbb{R}^n$.

(iv) These results yield naturally a large number of corollaries when specialized to penalties routinely used in the literature, among which the Lasso, the group Lasso, their analysis-type counterparts (fused (group) Lasso), the ℓ_∞ and the nuclear norms. Some of these corollaries are known and others novel.

(v) We finally discuss minimax optimality and provide lower-bounds, showing that our estimator are indeed nearly minimax over low-complexity model subsets.

Relation to previous work Our oracle inequality for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ extends the work of [47] 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 [135] in the case of the Lasso (see also the discussion in [48] and references therein) and that of [88] for the case of the nuclear norm. Our work also goes much beyond that in [149] on weakly decomposable priors, where we show in particular that there is no need to impose decomposability on the regularizer, since it is rather an intrinsic property of it.

1.2.3 Chapter 5: Estimation Bounds with Low-concave Priors

The context in Chapter 5 is similar to Chapter 4 where we consider the general regression problem with f is approximated by $f_{\boldsymbol{\theta}_0}$, $\boldsymbol{\theta}_0$ defined in (1.17). However, in this chapter, our primary interest lies in consistently estimating $\boldsymbol{\theta}_0$ itself (inverse problem) and studying the corresponding loss function $\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|_2$, for $\widehat{\boldsymbol{\theta}}$ either $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ in (1.12) and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ in (1.9). We will thus assess how the estimation loss function decays as a function of the noise level.

We thus provide a unified analysis and deliver bounds on the estimation loss for the parameter

estimates for both $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, where the penalty J is any finite-valued convex function. More precisely, we develop bounds guaranteeing that both $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ stably estimate $\boldsymbol{\theta}_0$ from the noisy measurements $\boldsymbol{y}$, and quantify the rate of convergence. Our framework allows to handle more general data losses beyond the usual strongly convex one. In the case of a Gaussian design, we provide sample complexity bounds that guarantee that our consistency bounds hold with high probability. We exemplify our bounds on 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. We also discuss extension beyond Gaussian designs.

Relation to previous work While bounds on the estimation loss function $\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{\theta}_0\right\|_2$ have been well studied in the literature in a more or less general setting, we are not aware of any work in this direction for the EWA estimator. Even for the penalized estimator, our bounds are new as they handle more general data losses beyond the usual strongly convex one. Our review hereafter is only partial and we refer the reader to e.g. [146] for a comprehensive treatment.

A large body of literature from the inverse problems community has been devoted to study how $\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{\theta}_0\right\|_2$ in the regression setting with a quadratic data loss and different penalties, see [131, 97, 77, 76].

In the compressed sensing literature, bounds on $\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{\theta}_0\right\|_2$ with the Lasso, analysis Lasso, group Lasso or nuclear norm were shown under the restricted isometry property (RIP) and its variants; see e.g. [24, 32, 34, 26, 118, 30, 14].

RIP-based guarantees are uniform. Non-uniform bounds with RIP-less arguments were derived for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ with a quadratic data loss (actually a constrained version of it) in [35, 28, 78, 37, 29, 139].

1.2.4 Chapter 6: EWA for Non-smooth Priors through Langevin Diffusion and Proximal Splitting

We aim to enlarge the family of μ covered by [52, 111, 62, 63] by relaxing some underlying conditions. Especially, in our study, μ is structured as $\widehat{\mu}_n$ in (1.12), and it is not necessarily differentiable nor log-concave. From the Langevin diffusion smoothed by Moreau-Yosida regularization, we propose two algorithms based on forward-backward proximal splitting for which we prove theoretical consistency guarantees. They are named Forward-Backward Langevin Monte-Carlo (FBLMC) and Semi-Forward-Backward Langevin Monte-Carlo (semi-FBLMC). These algorithms are established from the Langevin diffusions defined by the SDE of type (1.13) with the Moreau-regularized version of ρ. Under mild assumptions, we prove that:

(i) Each SDE has a unique solution which is strongly Markovian, non explosive and admits an (unique) invariant measure whose density converges to $\widehat{\mu}_n$ in total variation.

(ii) The Moreau-regularized terms are continuously differentiable and their gradients can be expressed through the proximal mappings which are computable.

(iii) The algorithms are introduced by discretizing these SDE by the forward Euler scheme. We also prove the consistency of each discretized SDE (see Theorem 6.2.4).

Besides, these algorithms are applied to compute EWA with several popular penalties in the literature, and illustrated on some numerical problems in Chapter 7.

Relation to previous work For a comprehensive review of sampling by Langevin diffusion from smooth and log-concave densities, we refer the reader to e.g. [46]. To cope with non-smooth densities, several works have proposed to replace $\log \mu$ with a smoothed version (typically involving the Moreau-Yosida regularization/envelope) [52, 111, 62, 63]. In [111, 63] for instance, the authors proposed

proximal-type algorithms to sample from possibly non-smooth log-concave densities μ using the forward Euler discretization and the Moreau-Yosida regularization. In [111][2], $-\log\mu$ is replaced with its Moreau envelope, while in [63], it is assumed that $-\log\mu = L + H$, L is convex Lipschitz continuously differentiable, and H is a proper closed convex function replaced by its Moreau envelope. In both these works, convexity plays a crucial role to get quantitative convergence guarantees. Proximal steps within MCMC methods have been recently proposed for some simple (convex) signal processing problems [36], though without any guarantees.

1.3 Reading guide

The manuscript consists of 8 chapters. The chapters containing the contributions of this work are collected into 2 parts: theoretical guarantees part (Chapters 3-5) and algorithmic part (Chapters 6-7).

- Chapter 2 provides some pre-requisites on the main mathematical tools on which this manuscript relies.
- In Chapter 3, we remind the PAC-Bayesian type oracle inequalities proposed in [52] which are a classical starting point in the literature for EWA in the deterministic case. From that, we define our family of EWA in analysis-group sparsity context which yields a general analysis-group SOI. For an appropriate choice of the prior, its remainder term is of order (1.16) that guarantees the performance of the estimator.
- Chapter 4 provides a unified analysis for EWA (1.12) and the penalized estimator (1.9). For a large class of data losses and penalties, we establish a general (deterministic) oracle inequality for prediction with optimal remainder terms which yields the oracle inequalities in probability when the observations are random. These inequalities are refined in the standard linear regression with Gaussian or sub-Gaussian noise and specialized to several penalties. We also discuss minimax optimality.
- Chapter 5 considers the same framework as in Chapter 4. However, it consists in establishing bounds guaranteeing the quality of estimators for estimating $\boldsymbol{\theta}_0$. These bounds are instantiated in the case of Gaussian design and exemplified on several penalties. We also discuss extension beyond Gaussian designs.
- Chapter 6 proposes two algorithms of type forward-backward proximal splitting for sampling from a distribution whose density is not smooth nor log-concave with consistency guarantees. They are applied to compute numerically EWA estimator.
- Chapter 7 collects the numerical experiments of EWA computed by the algorithms proposed in Chapter 6 on three problems: Compressed Sensing, Deconvolution and Inpainting.
- The conclusions and perspectives are drawn in Chapter 8.

[2]The author however applied it to problems where $-\log\mu = L + H$. But the gradient of the Moreau envelope of a sum, which amounts to computing the proximity operator of $-\log\mu$ does not have an easily implementable expression even if those of L and H do.

Chapter 2

Mathematical Background

Contents

In this chapter, we collect the necessary mathematical material used in the manuscript. Denote $\mathbb{R}$ the set of real numbers, $\mathbb{R}_+ = [0, +\infty[$ the set of non-negative real numbers, $\mathbb{R}^d$ the d-dimensional real Euclidean space, $\mathbb{R}^{d\times r}$ the set of $d \times r$ real matrices and $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ the extended real line. For a function f, its effective domain is $\mathrm{dom}(f) = \big\{\boldsymbol{x} \in \mathbb{R}^d \ : \ f(\boldsymbol{x}) < +\infty\big\}$. We denote $\langle \cdot, \cdot \rangle$ the Euclidean scalar inner product and $\left\| \cdot \right\|_2$ the associated norm.

2.1 Basics of analysis

For any $x \in \mathbb{R}$, we define $\mathrm{sgn}(x)$ the sign operator, $x_+ \stackrel{\text{def}}{=} \max(x, 0)$ its positive part, and $\lfloor x \rfloor$ its stands for the integer part. Define

$$\Gamma : x \in]0, +\infty[\mapsto \int_0^{+\infty} u^{x-1} \exp(-u) du$$

the Gamma function.

For $\mathcal{C}$ a non-empty set, denote $|\mathcal{C}|$ its cardinality, $\mathcal{C}^c$ its complement, and $\mathrm{bd}(\mathcal{C})$ its boundary. Moreover, we denote $\mathrm{P}_{\mathcal{C}}$ the orthogonal projector on $\mathcal{C}$, $\iota_{\mathcal{C}}$ its indicator function, i.e.

$$\iota_{\mathcal{C}}(\boldsymbol{x}) = 0 \text{ if } \boldsymbol{x} \in \mathcal{C} \quad \text{and} \quad \iota_{\mathcal{C}}(\boldsymbol{x}) = +\infty \text{ otherwise,}$$

and $I_{\mathcal{C}}$ its characteristic function, i.e.

$$I_{\mathcal{C}}(\boldsymbol{x}) = 1 \text{ if } \boldsymbol{x} \in \mathcal{C} \quad \text{and} \quad I_{\mathcal{C}}(\boldsymbol{x}) = 0 \text{ otherwise.}$$

2.1.1 Mappings

Definition 2.1.1 (Proper functions). A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$ is proper if $f(\boldsymbol{x}) > -\infty$ for any $\boldsymbol{x} \in \mathcal{C}$, and $\mathrm{dom}(f) \neq \emptyset$.

Definition 2.1.2 (Coercive functions). A function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is coercive if

$$\lim_{\|\boldsymbol{x}\|_2 \to +\infty} f(\boldsymbol{x}) = +\infty.$$

Definition 2.1.3 (Positively homogeneous functions). A function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is positively homogeneous if $0 \in \mathrm{dom}(f)$ and

$$f(\lambda \boldsymbol{x}) = \lambda f(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathrm{dom}(f),\ \forall \lambda > 0.$$

Definition 2.1.4 (Sublinear functions). A function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is sublinear if it is positively homogeneous and subadditive, i.e.

$$f(\boldsymbol{x} + \boldsymbol{x}') \leq f(\boldsymbol{x}) + f(\boldsymbol{x}'), \quad \forall \boldsymbol{x},\ \boldsymbol{x}' \in \mathrm{dom}(f).$$

A convex and positively homogeneous function is sublinear.

Definition 2.1.5 (Set-valued mappings). Let $\mathcal{Z}$ and $\mathcal{C}$ be two non-empty sets. An operator $S : \mathcal{Z} \rightrightarrows \mathcal{C}$ is called a set-valued mapping if S maps every $\boldsymbol{x} \in \mathcal{Z}$ to a set $S(\boldsymbol{x}) \subseteq \mathcal{C}$. The graph of S is defined by $\mathrm{gph}(S) = \big\{(\boldsymbol{x}, \boldsymbol{v}) \in \mathcal{Z} \times \mathcal{C} \ : \ \boldsymbol{v} \in S(\boldsymbol{x})\big\}$.

2.1.2 Continuity and differentiability

Definition 2.1.6 (Lower semi-continuity). A function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is lower semi-continuous (lsc) at $\overline{\boldsymbol{x}} \in \mathbb{R}^d$ if

$$\liminf_{\boldsymbol{x} \to \overline{\boldsymbol{x}}} f(\boldsymbol{x}) \geq f(\overline{\boldsymbol{x}}),$$

and f is lsc if that holds for any $\overline{\boldsymbol{x}} \in \mathrm{dom}(f)$.

Definition 2.1.7 (Lipschitz continuity and local Lipschitz continuity). Let $\boldsymbol{f} : \mathcal{Z} \subseteq \mathbb{R}^d \to \mathbb{R}^r$, $\boldsymbol{f}$ is Lipschitz continuous on a set $\mathcal{C} \subseteq \mathcal{Z}$ if

$$\big\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}')\big\|_2 \leq K \big\|\boldsymbol{x} - \boldsymbol{x}'\big\|_2, \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{C}, \text{ for some } K \in [0, +\infty[.$$

(i) A function $\boldsymbol{f} : \mathcal{Z} \subseteq \mathbb{R}^d \to \mathbb{R}^r$ is Lipschitz continuous if $\boldsymbol{f}$ is Lipschitz continuous on $\mathcal{Z}$.

(ii) A function $\boldsymbol{f} : \mathcal{Z} \subseteq \mathbb{R}^d \to \mathbb{R}^r$ is locally Lipschitz continuous if for any $\overline{\boldsymbol{x}} \in \mathcal{Z}$, there exists a $\epsilon > 0$ such that $\boldsymbol{f}$ is Lipschitz continuous on $\mathbb{B}(\overline{\boldsymbol{x}}, \epsilon) \stackrel{\text{def}}{=} \left\{\boldsymbol{x} \in \mathbb{R}^d \ : \ \left\|\overline{\boldsymbol{x}} - \boldsymbol{x}\right\|_2 \leq \epsilon\right\}$.

Lemma 2.1.8. *Assume that $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous, then there exists $K > 0$ such that*

$$\langle \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{x} \rangle \leq K(1 + \left\|\boldsymbol{x}\right\|_2^2), \quad \forall \boldsymbol{x} \in \mathbb{R}^d.$$

Proof. Let $\boldsymbol{x}^* \in \mathcal{C}$, a bounded subset of $\mathbb{R}^d$. Using Young and Jensen inequalities as well as $\widetilde{K}$-Lipschitz continuity of $\boldsymbol{f}$, we obtain

$$\begin{aligned}
\langle \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{x} \rangle &\leq \|\boldsymbol{f}(\boldsymbol{x})\|_2^2 /2 + \|\boldsymbol{x}\|_2^2 /2 \\
&\leq \|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}^*)\|_2^2 + \|\boldsymbol{f}(\boldsymbol{x}^*)\|_2^2 + \|\boldsymbol{x}\|_2^2 /2 \\
&\leq \widetilde{K} \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 + \|\boldsymbol{f}(\boldsymbol{x}^*)\|_2^2 + \|\boldsymbol{x}\|_2^2 /2 \\
&\leq \left(2\widetilde{K} + 1/2\right) \|\boldsymbol{x}\|_2^2 + \left(2\widetilde{K} \|\boldsymbol{x}^*\|_2^2 + \|\boldsymbol{f}(\boldsymbol{x}^*)\|_2^2\right) \\
&\leq K(1 + \|\boldsymbol{x}\|_2^2),
\end{aligned}$$

with $K \geq \max\left\{2\widetilde{K} + 1/2, 2\widetilde{K} \|\boldsymbol{x}^*\|_2^2 + \|\boldsymbol{f}(\boldsymbol{x}^*)\|_2^2\right\}$. Recalling that $\boldsymbol{f}$ is bounded on bounded sets concludes the proof. □

Denote $C^k(\mathbb{R}^d)$ the class of functions $f : \mathbb{R}^d \to \mathbb{R}$ such that its first k derivatives all exist and are continuous (with k a non-negative integer), and $C^\infty(\mathbb{R}^d)$ if that holds for any non-negative integer k. For a function $f \in C^1(\mathbb{R}^d)$, ∇f denotes its (Euclidean) gradient. For a bivariate function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that $g(\cdot, \boldsymbol{x}) \in C^1(\mathbb{R}^d)$ for any $\boldsymbol{x} \in \mathbb{R}^d$, ∇g denotes the gradient of g w.r.t. the first variable.

2.1.3 Lebesgue's convergence theorems

Theorem 2.1.9 (Lebesgue's monotone convergence theorem). *Let $f_n : \mathbb{R}^d \to [0, +\infty]$ be a monotone increasing sequence of measurable functions such that $f_n \to f$ pointwise almost everywhere as $n \to +\infty$. Then f_n and f are measurable, and*

$$\lim_{n \to +\infty} \int_{\mathbb{R}^d} f_n d\mu = \int_{\mathbb{R}^d} f d\mu.$$

Theorem 2.1.10 (Lebesgue's dominated convergence theorem). *Let $f_n : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$ are measurable functions such that*

(i) $f_n \to f$ pointwise almost everywhere as $n \to +\infty$,

(ii) there exists an integrable function $g : \mathbb{R}^d \to [0, +\infty]$ with $|f_n| \leq g$ for any n.

Then f and f_n are integrable, and

$$\lim_{n \to +\infty} \int_{\mathbb{R}^d} f_n d\mu = \int_{\mathbb{R}^d} f d\mu.$$

Theorem 2.1.11 (Scheffé's lemma (see [130, 90])). *Let $f_n : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$ are integrable functions such that f_n converges almost everywhere to an integrable function f as $n \to +\infty$, then*

$$\int_{\mathbb{R}^d} |f_n - f| \, d\mu \underset{n \to +\infty}{\to} 0,$$

if and only if

$$\int_{\mathbb{R}^d} |f_n| \, d\mu \underset{n \to +\infty}{\to} \int_{\mathbb{R}^d} |f| \, d\mu.$$

Theorem 2.1.12 (Leibniz's rule). *Let $\mathcal{C} = [x_0, x_1] \times [y_0, y_1] \subset \mathbb{R}^2$. Suppose that $f(x, y)$ and $\frac{\partial f}{\partial y}(x, y)$ are continuous on $\mathcal{C}$. Then*

$$\frac{d}{dy}\left(\int_{x_0}^{x_1} f(x, y) dx\right) = \int_{x_0}^{x_1} \frac{\partial f}{\partial y}(x, y) dx.$$

2.2 Linear algebra

In this manuscript, we denote bold uppercase letters and bold lowercase letters respectively for matrices and vectors in Euclidean space. The identity matrix on $\mathbb{R}^d$ is denoted by $\mathbf{I}_d$.

For a matrix $\boldsymbol{M} \in \mathbb{R}^{d\times r}$. Denote $\boldsymbol{M}^\top$ the transpose of $\boldsymbol{M}$. For a linear operator $\boldsymbol{A}$, $\boldsymbol{A}^*$ is its adjoint. We denote $\boldsymbol{M}_I$ (resp. $\boldsymbol{M}_{I,J}$), with $I \subseteq \{1,\dots,r\}$ and $J \subseteq \{1,\dots,d\}$, the submatrix whose columns (resp. columns and rows) are those of $\boldsymbol{M}$ indexed by the index set I (resp. I and J). We define $\boldsymbol{M}_T \stackrel{\text{def}}{=} \boldsymbol{M}\,\mathrm{P}_T$, where P_T is the orthogonal projector onto the linear subspace T. Denote $\mathrm{vec}(\boldsymbol{M})$ the vectorization opertator, i.e. the operator which stacks the columns of its arguments. We also denote $\boldsymbol{\sigma}(\boldsymbol{M}) = (\boldsymbol{\sigma}_1(\boldsymbol{M}),\dots,\boldsymbol{\sigma}_r(\boldsymbol{M}))^\top \in \mathbb{R}^r$ the vector of singular values of $\boldsymbol{M}$ in non-increasing order, and $\sigma_{\min}(\boldsymbol{M}) \stackrel{\text{def}}{=} \boldsymbol{\sigma}_r(\boldsymbol{M})$ the smallest singular value.

For a square matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$, denote $\mathrm{tr}(\boldsymbol{M})$ and $\det(\boldsymbol{M})$ respectively the trace and the determinant of $\boldsymbol{M}$. The Frobenius scalar product is defined by $\langle \boldsymbol{A},\boldsymbol{B}\rangle_F \stackrel{\text{def}}{=} \mathrm{tr}(\boldsymbol{A}^\top\boldsymbol{B})$ for any $\boldsymbol{A},\ \boldsymbol{B} \in \mathbb{R}^{d\times d}$.

For a vector $\boldsymbol{x} \in \mathbb{R}^d$. Denote $\boldsymbol{x}_I$, with $I \subseteq \{1,\dots,d\}$, the subvector whose entries are those of $\boldsymbol{x}$ indexed by a index set I. With T a subspace, $\boldsymbol{x}_T$ denotes the orthogonal projection of a vector $\boldsymbol{x}$ on T. We denote $\mathrm{supp}(\boldsymbol{x})$ the support of $\boldsymbol{x}$, i.e. $\mathrm{supp}(\boldsymbol{x}) = \big\{i \in \{1,\dots,d\} \ :\ \boldsymbol{x}_i \neq 0\big\}$. $\mathrm{diag}(\boldsymbol{x})$ denotes the diagonal matrix whose diagonal entries are the components of $\boldsymbol{x}$.

For a linear subspace T. Denote $\dim(T)$ the dimension of T, and $T^\perp$ its orthogonal subspace.

2.2.1 Norms

2.2.1.1 ℓ_p-norms

Definition 2.2.1 (ℓ_p-norms). For any $p \geq 1$ and $\boldsymbol{x} \in \mathbb{R}^d$, the ℓ_p norm is defined by

$$\big\|\boldsymbol{x}\big\|_p \stackrel{\text{def}}{=} \left(\sum_{j=1}^{d} |\boldsymbol{x}_j|^p\right)^{1/p},$$

with the adaptation $\big\|\boldsymbol{x}\big\|_\infty = \max_{j\in\{1,\dots,d\}} |\boldsymbol{x}_j|$.

With $p \in]0,1[$, $\big\|\boldsymbol{x}\big\|_p$ is a quasi-norm. $\big\|\boldsymbol{x}\big\|_0$ is the ℓ_0 pseudo-norm which counts the number of non-zero elements in $\boldsymbol{x}$.

2.2.1.2 $\ell_{p,2}$-group norms

We partition the index set $\{1,\dots,d\}$ into L groups/blocks of indices $\{b_l\}_{1\leq l\leq L}$ such that $b_l \subseteq \{1,\dots,d\}$.

Definition 2.2.2 ($\ell_{p,2}$-group norms). For any $p \geq 1$ and $\boldsymbol{x} \in \mathbb{R}^d$, the $\ell_{p,2}$- group norm is defined by

$$\big\|\boldsymbol{x}\big\|_{p,2} = \left(\sum_{l=1}^{L} \big\|\boldsymbol{x}_{b_l}\big\|_2^p\right)^{1/p},$$

with the adaption $\big\|\boldsymbol{x}\big\|_{\infty,2} = \max_{l\in\{1,\dots,L\}} \big\|\boldsymbol{x}_{b_l}\big\|_2$.

With $p \in]0,1[$, $\big\|\boldsymbol{x}\big\|_{p,2}$ is a quasi-norm. $\big\|\boldsymbol{x}\big\|_{0,2}$ is a pseudo-norm which counts the number of active (i.e. non-zero) groups in $\boldsymbol{x}$.

2.2.1.3 Schatten p-norms

Definition 2.2.3 (Schatten p-norms). For any $p \geq 1$ and $\boldsymbol{M} \in \mathbb{R}^{d\times d}$, the Schatten p-norm is defined as

$$\|\boldsymbol{M}\|_{S_p} \stackrel{\text{def}}{=} \big\|\boldsymbol{\sigma}(\boldsymbol{M})\big\|_p.$$

Let us detail some special cases of Schatten p-norm

(i) Nuclear norm ($p = 1$)

$$\|\boldsymbol{M}\|_* \overset{\text{def}}{=} \|\boldsymbol{M}\|_{S_1} = \operatorname{tr}\left(\sqrt{\boldsymbol{M}^\top \boldsymbol{M}}\right), \quad \forall \boldsymbol{M} \in \mathbb{R}^{d\times r}.$$

(ii) Frobenius norm ($p = 2$)

$$\|\boldsymbol{M}\|_F \overset{\text{def}}{=} \|\boldsymbol{M}\|_{S_2} = \sqrt{\operatorname{tr}\left(\boldsymbol{M}^\top \boldsymbol{M}\right)}, \quad \forall \boldsymbol{M} \in \mathbb{R}^{d\times r}.$$

(iii) Operator (Spectral) norm ($p = \infty$)

$$\|\boldsymbol{M}\|_{2\to 2} \overset{\text{def}}{=} \|\boldsymbol{M}\|_{S_\infty} = \boldsymbol{\sigma}_1(\boldsymbol{M}), \quad \forall \boldsymbol{M} \in \mathbb{R}^{d\times r}.$$

2.2.1.4 Inner products and norms in a metric

Definition 2.2.4 (Inner products and norms in a metric). Let $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ symmetric positive definite be a metric matrix. The inner product on $\mathbb{R}^d$ in the metric $\boldsymbol{M}$ is defined by

$$\langle \cdot, \cdot \rangle_{\boldsymbol{M}} \overset{\text{def}}{=} \langle \cdot, \boldsymbol{M} \cdot \rangle,$$

and $\|\cdot\|_{\boldsymbol{M}}$ its associated norm.

For any $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ symmetric positive definite, we obviously have the equivalence between the norm $\|.\|_{\boldsymbol{M}}$ and the ℓ_2-norm via the inequality

$$\sigma_{\min}(\boldsymbol{M})\|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_{\boldsymbol{M}} \leq \|\boldsymbol{M}\|_{2\to 2}\|\boldsymbol{x}\|_2, \quad \forall \boldsymbol{x} \in \mathbb{R}^d.$$

2.2.2 Frame

Definition 2.2.5 (Frame). A matrix $\boldsymbol{M} \in \mathbb{R}^{d\times r}$ is a frame if there exist two constants ν and μ with $\nu \geq \mu > 0$, called frame bounds, such that the generalized Parseval relation is satisfied, i.e.

$$\mu\|\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{M}^\top \boldsymbol{x}\|_2^2 \leq \nu\|\boldsymbol{x}\|_2^2, \quad \forall \boldsymbol{x} \in \mathbb{R}^d.$$

By the Courant-Fischer theorem, Definition 2.2.5 is equivalent to the fact that μ (resp. ν) is a lower (resp. upper) bound of the eigenvalues of $\boldsymbol{M}\boldsymbol{M}^\top$. Moreover, since $\mu > 0$, we have that $\boldsymbol{M}\boldsymbol{M}^\top$ is bijective and $\boldsymbol{M}$ is surjective. The frame is said tight when $\mu = \nu$. Typical examples of (tight) frames that have been used in statistics are translation invariant wavelets [41], ridgelets [25] and curvelets [21] (example of groups and what they represent for wavelets/ridgelets/curvelets in applications are discussed in [40]). Let $\widetilde{\boldsymbol{M}} \in \mathbb{R}^{d\times r}$ be the canonical dual frame associated to $\boldsymbol{M}$, i.e.

$$\widetilde{\boldsymbol{M}} = (\boldsymbol{M}\boldsymbol{M}^\top)^{-1}\boldsymbol{M}.$$

We know that

$$\widetilde{\boldsymbol{M}}\boldsymbol{M}^\top = \mathbf{I}_d \tag{2.1}$$

and

$$\frac{1}{\mu} \geq \boldsymbol{\sigma}_1\left(\widetilde{\boldsymbol{M}}^\top \widetilde{\boldsymbol{M}}\right) \geq \cdots \geq \boldsymbol{\sigma}_d\left(\widetilde{\boldsymbol{M}}^\top \widetilde{\boldsymbol{M}}\right) \geq \frac{1}{\nu}. \tag{2.2}$$

Note that we focus on the canonical dual frame for the sake of simplicity. In fact, our exposition remains unchanged if any other dual frame is used instead of the canonical one.

2.3 Convex analysis

2.3.1 Convex sets and functions

Let $\mathcal{C}$ be a non-empty convex set in $\mathbb{R}^d$. The smallest linear manifold containing $\mathcal{C}$ is denoted by $\operatorname{Span}(\mathcal{C})$. The smallest affine subspace that contains $\mathcal{C}$ is denoted by $\operatorname{aff}(\mathcal{C})$ which is also called the

affine hull of $\mathcal{C}$, and $\mathrm{par}(\mathcal{C})$ denotes the subspace parallel to $\mathrm{aff}(\mathcal{C})$. We denote $\mathrm{ri}(\mathcal{C})$ its relative interior to its affine hull. The convex hull of $\mathcal{C}$ is $\mathrm{conv}\,(\mathcal{C})$ and its closure is $\overline{\mathrm{conv}}\,(\mathcal{C})$.

Definition 2.3.1 (Convex sets and functions).

(i) A set $\mathcal{C} \subseteq \mathbb{R}^d$ is convex if

$$(1-\tau)\boldsymbol{x} + \tau\boldsymbol{x}' \in \mathcal{C}, \quad \forall \boldsymbol{x},\ \boldsymbol{x}' \in \mathcal{C},\ \forall \tau \in]0,1[.$$

(ii) A function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is convex if $\mathrm{dom}(f)$ is convex, and

$$f\left((1-\tau)\boldsymbol{x} + \tau\boldsymbol{x}'\right) \leq (1-\tau)f(\boldsymbol{x}) + \tau f(\boldsymbol{x}'), \quad \forall \boldsymbol{x},\ \boldsymbol{x}' \in \mathrm{dom}(f),\ \forall \tau \in]0,1[.$$

Definition 2.3.2 ((Fenchel) Subdifferential). Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$. Given a point $\boldsymbol{x} \in \mathrm{dom}(f)$, the Fenchel subdifferential of f at $\boldsymbol{x}$ is defined as

$$\partial f(\boldsymbol{x}) = \left\{\boldsymbol{v} \in \mathbb{R}^d \ :\ f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{y} - \boldsymbol{x}\rangle, \quad \forall \boldsymbol{y} \in \mathbb{R}^d\right\}.$$

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

Definition 2.3.3 (Bregman divergence). The *Bregman divergence* associated to a convex function f at $\boldsymbol{x}$ with respect to $\boldsymbol{\eta} \in \partial f(\boldsymbol{x}) \neq \emptyset$ is

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

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

2.3.2 Gauges

Definition 2.3.4 (Gauges). Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a non-empty closed convex set containing the origin. The *gauge* of $\mathcal{C}$ is the function $\gamma_{\mathcal{C}}$ defined on $\mathbb{R}^d$ by

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

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

Lemma 2.3.5 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 2.3.5.

(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*

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

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

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

See [145] for the proof.

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

2.3.3 Dualization

2.3.3.1 Legendre-Fenchel conjugate and support functions

Definition 2.3.6 (Legendre-Fenchel conjugate). The Legendre-Fenchel conjugate of f is

$$f^*(\boldsymbol{z}) = \sup_{\boldsymbol{x}\in\mathbb{R}^d} \langle \boldsymbol{z}, \boldsymbol{x}\rangle - f(\boldsymbol{x}).$$

For any f, f^* is always a lsc convex function.

Theorem 2.3.7 (Fenchel-Moreau theorem). *Let f be proper. Then f is lsc and convex if and only if $f^{**} = f$, in which case f^* is also proper.*

For f proper, the functions (f, f^*) obey the Fenchel-Young inequality

$$f(\boldsymbol{x}) + f^*(\boldsymbol{z}) \geq \langle \boldsymbol{z}, \boldsymbol{x}\rangle, \quad \forall(\boldsymbol{x}, \boldsymbol{z}). \tag{2.3}$$

It is actually the best pair for which this inequality cannot be tightened.

For a function g on $\mathbb{R}_+$, the function $g^+ : a \in \mathbb{R}_+ \mapsto g^+(a) = \sup_{t\geq 0} at - g(t)$ is called the monotone conjugate of g. The pair (g, g^+) obviously obeys (2.3) on $\mathbb{R}_+ \times \mathbb{R}_+$.

We collect some properties of the monotone conjugate, some of which are new, in the following lemma.

Lemma 2.3.8 (Monotone conjugate). *Let g be a non-decreasing function on $\mathbb{R}_+$ that vanishes at 0. Then the following hold:*

(i) g^+ is a proper closed convex and non-decreasing function on $\mathbb{R}_+$ that vanishes at 0.

(ii) If g is also closed and convex, then $g^{++} = g$.

(iii) Let $f : t \in \mathbb{R} \mapsto g(|t|)$ such that f is differentiable on $\mathbb{R}$, where g is finite-valued, strictly convex and strongly coercive. Then g^+ is likewise finite-valued, strictly convex, strongly coercive, and $f^* = g^+ \circ |\cdot|$ is differentiable on $\mathbb{R}$. In particular, both g and g^+ are strictly increasing on $\mathbb{R}_+$.

Proof.

(i) By [7, Proposition 13.11], g^+ is a closed convex function. We have $\inf_{t\geq 0} g(t) = -\sup_{t\geq 0} t\cdot 0 - g(t) = -g^+(0)$. Since g is non-decreasing and $g(0) = 0$, then $g^+(0) = -\inf_{t\geq 0} g(t) = -g(0) = 0$. In addition, by (2.3), we have $g^+(a) \geq a\cdot 0 - g(0) = 0$, $\forall a \in \mathbb{R}_+$. This shows that g^+ is non-negative and $\mathrm{dom}(g^+) \neq \emptyset$, and in turn, it is also proper.
Let a, b in $\mathbb{R}_+$ such that $a < b$. Then

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

That is, g^+ is non-decreasing on $\mathbb{R}_+$.

(ii) This follows from [125, Theorem 12.4].

(iii) By definition of f, f is a finite-valued function on $\mathbb{R}$, strictly convex, differentiable and strognly coercive. It then follows from [80, Corollary X.4.1.4] that f^* enjoys the same properties. In turn, using the fact that both f and f^* are even, we have g^+ is strongly coercive, and strict convexity of f (resp. f^*) is equivalent to that of g (resp. g^+). Altogether, this shows the first claim. We now prove that g vanishes only at 0 (and similary for g^+). As g is non-decreasing and strictly convex, we have, for any $\rho \in]0,1[$ and a, b in $\mathbb{R}_+$ such that $a < b$,

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

□

Definition 2.3.9 (Support functions). The *support function* of $\mathcal{C} \subset \mathbb{R}^d$ is

$$\sigma_{\mathcal{C}}(\boldsymbol{\omega}) = \sup_{\boldsymbol{x}\in\mathcal{C}} \langle \boldsymbol{\omega}, \boldsymbol{x}\rangle = \iota^*_{\mathcal{C}}(\boldsymbol{v}).$$

We recall the following properties whose proofs can be found in e.g. [125, 80].

Lemma 2.3.10. *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 $0 \in \mathrm{int}(\mathcal{C})$, then $\sigma_{\mathcal{C}}$ is finite-valued and coercive.

2.3.3.2 Polar sets and polar gauges

Definition 2.3.11 (Polar set). Let $\mathcal{C}$ be a non-empty convex set. The set $\mathcal{C}^\circ$ given by

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

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

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

We now define the polar gauge.

Definition 2.3.12 (Polar Gauge). The polar of a gauge $\gamma_{\mathcal{C}}$ is the function $\gamma^\circ_{\mathcal{C}}$ defined by

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

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

$$\langle \boldsymbol{x}, \boldsymbol{u}\rangle \leq \gamma_{\mathcal{C}}(\boldsymbol{x})\gamma^\circ_{\mathcal{C}}(\boldsymbol{u}) \quad \forall (\boldsymbol{x}, \boldsymbol{u}) \in \mathrm{dom}(\gamma_{\mathcal{C}}) \times \mathrm{dom}(\gamma^\circ_{\mathcal{C}}), \tag{2.4}$$

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

Lemma 2.3.13. *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a closed convex set containing the origin. Then,*

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

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

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

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

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

See [125, 80, 145] for the proof.

2.3.3.3 Subdifferential gauge and its polar

Definition 2.3.14 (Subdifferential gauge and its polar). Let J be proper lsc and convex. Let $f_{\boldsymbol{x}} \in \mathrm{ri}(\partial J(\boldsymbol{x}))$. The *subdifferential gauge* associated to $f_{\boldsymbol{x}}$ is the gauge $J^\circ_{f_{\boldsymbol{x}}} = \gamma_{\partial J(\boldsymbol{x})-f_{\boldsymbol{x}}}$. Its polar gauge is $J_{f_{\boldsymbol{x}}}$, which is also the support function of $\partial J(\boldsymbol{x}) - f_{\boldsymbol{x}}$.

The following lemma gathers the main properties of these gauges that we will need in the sequel.

Lemma 2.3.15. *Let J proper lsc convex. Let $f_{\boldsymbol{x}} \in \mathrm{ri}(\partial J(\boldsymbol{x}))$. Then,*

(i) $\mathrm{dom}(J^\circ_{f_{\boldsymbol{x}}}) = S_{\boldsymbol{x}}$.

(ii) $J^\circ_{f_{\boldsymbol{x}}}$ *is coercive on* $S_{\boldsymbol{x}}$.

(iii) $J_{f_{\boldsymbol{x}}}(\boldsymbol{u}) = 0$ *if, and only if,* $\boldsymbol{u} \in T_{\boldsymbol{x}}$.

(iv) $J_{f_{\boldsymbol{x}}}(\boldsymbol{u})$ *is finite everywhere, and coercive on* $S_{\boldsymbol{x}}$.

Proof. (i)-(ii) [145, Proposition 2]. (iii)-(iv) [145, Proposition 3] □

We denote $\Gamma_0(\mathbb{R}^d)$ the class of proper lsc convex functions, and $\mathcal{G}(\mathbb{R}^d)$ the class of finite-valued gauges $\gamma_{\mathcal{C}}$, i.e. $\mathcal{C}$ is a non-empty convex compact set containing the origin as an interior point.

2.4 Geometrical decomposability

2.4.1 Decomposability

We start by defining some essential geometrical objects that were introduced in [145].

Definition 2.4.1 (Model Subspace). Assume that $J \in \Gamma_0(\mathbb{R}^d)$. Let $\boldsymbol{x} \in \mathbb{R}^d$. We denote by $e_{\boldsymbol{x}}$ as

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

We denote

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

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

It can be shown, see [145, Proposition 5], that $\boldsymbol{x} \in T_{\boldsymbol{x}}$ when J is a gauge, hence the name model subspace. When J is differentiable at $\boldsymbol{x}$, we have $e_{\boldsymbol{x}} = \nabla J(\boldsymbol{x})$ and $T_{\boldsymbol{x}} = \mathbb{R}^d$. When J is the ℓ_1-norm (Lasso), the vector $e_{\boldsymbol{x}}$ is nothing but the sign of $\boldsymbol{x}_I$, where $I = \mathrm{supp}(\boldsymbol{x})$. $e_{\boldsymbol{x}}$ will be detailed for many examples later. Observe also that $e_{\boldsymbol{x}} = \mathrm{P}_{T_{\boldsymbol{x}}}(\partial J(\boldsymbol{x}))$, and thus $e_{\boldsymbol{x}} \in T_{\boldsymbol{x}} \cap \mathrm{aff}(\partial J(\boldsymbol{x}))$. However, in general, $e_{\boldsymbol{x}} \notin \partial J(\boldsymbol{x})$.

We also provide a fundamental equivalent description of the subdifferential of a function $J \in \Gamma_0(\mathbb{R}^d)$ at $\boldsymbol{x}$ in terms of $e_{\boldsymbol{x}}$, $T_{\boldsymbol{x}}$, $S_{\boldsymbol{x}}$ and the gauge $J^\circ_{f_{\boldsymbol{x}}}$.

Theorem 2.4.2. *Assume that* $J \in \Gamma_0(\mathbb{R}^d)$. *Let* $f_{\boldsymbol{x}} \in \mathrm{ri}(\partial J(\boldsymbol{x}))$.

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

$$\partial J(\boldsymbol{x}) = \left\{\boldsymbol{\eta} \in \mathbb{R}^d \;:\; \boldsymbol{\eta}_{T_{\boldsymbol{x}}} = e_{\boldsymbol{x}} \quad \text{and} \quad J^\circ_{f_{\boldsymbol{x}}}(\mathrm{P}_{S_{\boldsymbol{x}}}(\boldsymbol{\eta} - f_{\boldsymbol{x}})) \leq 1\right\}.$$

Moreover, $\boldsymbol{\eta} \in \mathrm{ri}(\partial J(\boldsymbol{x}))$ *if, and only if,* $J^\circ_{f_{\boldsymbol{x}}}(\mathrm{P}_{S_{\boldsymbol{x}}}(\boldsymbol{\eta} - f_{\boldsymbol{x}})) < 1$.

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

$$J_{f_{\boldsymbol{x}}}(\boldsymbol{\omega}_{S_{\boldsymbol{x}}}) = \left\langle \boldsymbol{\eta}_{S_{\boldsymbol{x}}} - \mathrm{P}_{S_{\boldsymbol{x}}} f_{\boldsymbol{x}}, \boldsymbol{\omega}_{S_{\boldsymbol{x}}} \right\rangle.$$

Proof.

(i) This follows from [145, Theorem 1].

(ii) By definition, $J_{f_{\boldsymbol{x}}}$ is the gauge polar to $J^\circ_{f_{\boldsymbol{x}}}$ whose domain is $S_{\boldsymbol{x}}$ by Lemma 2.3.15. It is also the support function of $\partial J(\boldsymbol{x}) - f_{\boldsymbol{x}} = \left\{\boldsymbol{\eta} \;:\; J^\circ_{f_{\boldsymbol{x}}}(\boldsymbol{\eta}) \leq 1\right\} \subset S_{\boldsymbol{x}}$, i.e.

$$J_{f_{\boldsymbol{x}}}(\boldsymbol{\omega}) = \max_{J^\circ_{f_{\boldsymbol{x}}}(\boldsymbol{\eta}) \leq 1} \langle \boldsymbol{\eta}, \boldsymbol{\omega} \rangle = \max_{J^\circ_{f_{\boldsymbol{x}}}(\boldsymbol{\eta}) \leq 1} \langle \boldsymbol{\eta}, \boldsymbol{\omega}_{S_{\boldsymbol{x}}} \rangle = J_{f_{\boldsymbol{x}}}(\boldsymbol{\omega}_{S_{\boldsymbol{x}}}).$$

Thus there exists a supporting point $\boldsymbol{v} \in S_{\boldsymbol{x}} \cap \mathrm{bd}\left(\left\{\boldsymbol{\eta} \;:\; J^\circ_{f_{\boldsymbol{x}}}(\boldsymbol{\eta}) \leq 1\right\}\right)$ with normal vector $\boldsymbol{\omega}$ [7, Corollary 7.6(iii)], i.e.

$$\exists \boldsymbol{v} \in S_{\boldsymbol{x}} \text{ such that } J^\circ_{f_{\boldsymbol{x}}}(\boldsymbol{v}) \leq 1 \quad \text{and} \quad J_{f_{\boldsymbol{x}}}(\boldsymbol{\omega}_{S_{\boldsymbol{x}}}) = \langle \boldsymbol{v}, \boldsymbol{\omega}_{S_{\boldsymbol{x}}} \rangle.$$

Taking $\boldsymbol{\eta} = \boldsymbol{v} + f_{\boldsymbol{x}} \in \partial J(\boldsymbol{x})$ concludes the proof.

□

When J is a finite gauge, we obtain the following corollary.

Corollary 2.4.3. *Let $J \in \mathcal{G}(\mathbb{R}^d)$. Let $\boldsymbol{x} \in \mathbb{R}^d$ and $f_{\boldsymbol{x}} \in \mathrm{ri}(\partial J(\boldsymbol{x}))$.*

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

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

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

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

Proof.

(i) This follows by piecing together [145, Theorem 1, Proposition 4 and Proposition 5(iii)].

(ii) From [145, Proposition 5(iv)], we have

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

Thus there exists a supporting point $\boldsymbol{v} \in \partial J(\boldsymbol{x}) - f_{\boldsymbol{x}} \subset S_{\boldsymbol{x}}$ with normal vector $\boldsymbol{\omega}$ [7, Corollary 7.6(iii)], i.e.

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

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

□

Remark 2.4.4. The coercivity assumption in the class $\mathcal{G}(\mathbb{R}^d)$ is not needed for Corollary 2.4.3 to hold.

The decomposability of described in Theorem 2.4.2(i) depends on the particular choice of the mapping $\boldsymbol{x} \mapsto f_{\boldsymbol{x}} \in \mathrm{ri}(\partial J(\boldsymbol{x}))$. An interesting situation is encountered when $e_{\boldsymbol{x}} \in \mathrm{ri}(J(\boldsymbol{x}))$, so that one can choose $f_{\boldsymbol{x}} = e_{\boldsymbol{x}}$. Strong gauges, see [145, Definition 6], are precisely a class of gauges for which this situation occurs, and in this case, Corollary 2.4.3(i) and Theorem 2.4.2(i) has the simpler form

$$\partial J(\boldsymbol{x}) = \mathrm{aff}(\partial J(\boldsymbol{x})) \cap \mathcal{C}^\circ = \left\{\boldsymbol{\eta} \in \mathbb{R}^d \ : \ \boldsymbol{\eta}_{T_{\boldsymbol{x}}} = e_{\boldsymbol{x}} \quad \text{and} \quad J^\circ(\boldsymbol{\eta}_{S_{\boldsymbol{x}}}) \leq 1\right\}. \tag{2.5}$$

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 are obviously in $\mathcal{G}(\mathbb{R}^d)$ and $\Gamma_0(\mathbb{R}^d)$.

2.4.2 Stability

The gauges in $\mathcal{G}(\mathbb{R}^d)$ form a robust class enjoying important calculus rules. In particular it is closed under the sum and composition with an injective linear operator as we now prove.

Lemma 2.4.5. *The set of functions in $\mathcal{G}(\mathbb{R}^d)$ is closed under addition[1] and pre-composition by an injective linear operator. More precisely, the following holds:*

(i) Let J and G be two gauges in $\mathcal{G}(\mathbb{R}^d)$. Then $H \overset{def}{=} J + G$ is also in $\mathcal{G}(\mathbb{R}^d)$. Moreover,

(a) $T^H_{\boldsymbol{x}} = T^J_{\boldsymbol{x}} \cap T^G_{\boldsymbol{x}}$ and $e^H_{\boldsymbol{x}} = \mathrm{P}_{T^H_{\boldsymbol{x}}}(e^J_{\boldsymbol{x}} + e^G_{\boldsymbol{x}})$, where $T^J_{\boldsymbol{x}}$ and $e^J_{\boldsymbol{x}}$ (resp. $T^G_{\boldsymbol{x}}$ and $e^G_{\boldsymbol{x}}$) are the model subspace and vector at $\boldsymbol{x}$ 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 $J \in \mathcal{G}(\mathbb{R}^d)$, and $\boldsymbol{D} : \mathbb{R}^r \to \mathbb{R}^d$ be surjective. Then $H \overset{def}{=} J \circ \boldsymbol{D}^\top$ is also in $\mathcal{G}(\mathbb{R}^d)$. Moreover,

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

(a) $T_{\boldsymbol{x}}^H = \mathrm{Ker}(\boldsymbol{D}_{S_{\boldsymbol{u}}^J}^\top)$ *and* $e_{\boldsymbol{x}}^H = \mathrm{P}_{T_{\boldsymbol{x}}^H}\boldsymbol{D}e_{\boldsymbol{u}}^J$*, where* $T_{\boldsymbol{u}}^J$ *and* $e_{\boldsymbol{u}}^J$ *are the model subspace and vector at* $\boldsymbol{u} \overset{def}{=} \boldsymbol{D}^\top \boldsymbol{x}$ *associated to* J*;*

(b) $H^\circ(\boldsymbol{\omega}) = J^\circ(\boldsymbol{D}^+\boldsymbol{\omega})$*, where* $\boldsymbol{D}^+ = \boldsymbol{D}^\top\left(\boldsymbol{D}\boldsymbol{D}^\top\right)^{-1}$.

The outcome of Lemma 2.4.5 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 (e.g. ℓ_1+nuclear norms for recovering sparse and low-rank matrices).

Proof.

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

(a) This is [145, Proposition 8(i)-(ii)].

(b) We have from Lemma 2.3.13 and calculus rules on support functions,

$$\begin{aligned}
H^\circ(\boldsymbol{\omega}) = \sigma_{J(\boldsymbol{x})+G(\boldsymbol{x})\leq 1}(\boldsymbol{\omega}) &= \sup_{J(\boldsymbol{x})+G(\boldsymbol{x})\leq 1} \langle \boldsymbol{\omega}, \boldsymbol{x}\rangle = \max_{\rho\in[0,1]} \sup_{J(\boldsymbol{x})\leq\rho, G(\boldsymbol{x})\leq 1-\rho} \langle \boldsymbol{\omega}, \boldsymbol{x}\rangle \\
\text{([80, Theorem V.3.3.3])} &= \max_{\rho\in[0,1]} \overline{\mathrm{conv}}\left(\inf\left(\sigma_{J(\boldsymbol{x})\leq\rho}(\boldsymbol{\omega}), \sigma_{G(\boldsymbol{x})\leq 1-\rho}(\boldsymbol{\omega})\right)\right) \\
\text{(Positive homogeneity)} &= \max_{\rho\in[0,1]} \overline{\mathrm{conv}}\left(\inf\left(\rho\sigma_{J(\boldsymbol{x})\leq 1}(\boldsymbol{\omega}), (1-\rho)\sigma_{G(\boldsymbol{x})\leq 1}(\boldsymbol{\omega})\right)\right) \\
\text{(Lemma 2.3.13)} &= \max_{\rho\in[0,1]} \overline{\mathrm{conv}}\left(\inf\left(\rho J^\circ(\boldsymbol{\omega}), (1-\rho)G^\circ(\boldsymbol{\omega})\right)\right).
\end{aligned}$$

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

(a) This is [145, Proposition 10(i)-(ii)].

(b) Denote $J = \gamma_{\mathcal{C}}$. We have

$$\begin{aligned}
H^\circ(\boldsymbol{\omega}) &= \sup_{\boldsymbol{D}^\top\boldsymbol{x}\in\mathcal{C}} \langle \boldsymbol{\omega}, \boldsymbol{x}\rangle \\
\text{($\boldsymbol{D}^\top$ is injective)} &= \sup_{\boldsymbol{D}^\top\boldsymbol{x}\in\mathcal{C}} \left\langle \boldsymbol{D}^+\boldsymbol{\omega}, \boldsymbol{D}^\top\boldsymbol{x}\right\rangle \\
&= \sup_{\boldsymbol{u}\in\mathcal{C}\cap\mathrm{Span}(\boldsymbol{D}^\top)} \left\langle \boldsymbol{D}^+\boldsymbol{\omega}, \boldsymbol{u}\right\rangle \\
\text{([80, Theorem V.3.3.3] and Lemma 2.3.13)} &= \overline{\mathrm{conv}}\left(\inf\left(J^\circ(\boldsymbol{D}^+\boldsymbol{\omega}), \iota_{\mathrm{Ker}(\boldsymbol{D})}(\boldsymbol{D}^+\boldsymbol{\omega})\right)\right) \\
&= J^\circ(\boldsymbol{D}^+\boldsymbol{\omega}).
\end{aligned}$$

where in the last equality, we used the fact that $\boldsymbol{D}^+\boldsymbol{\omega} \in \mathrm{Span}\left(\boldsymbol{D}^\top\right) = \mathrm{Ker}(\boldsymbol{D})^\perp$, and thus $\iota_{\mathrm{Ker}(\boldsymbol{D})}(\boldsymbol{D}^+\boldsymbol{\omega}) = +\infty$ unless $\boldsymbol{\omega} = 0$, and J° is continuous and convex (as $J^\circ \in \mathcal{G}(\mathbb{R}^d)$) and Lemma 2.3.13.

□

2.4.3 Examples

2.4.3.1 Lasso

The Lasso regularization is used to promote the sparsity of the minimizers, see [19] for a comphensive review. It corresponds to choosing J as the ℓ_1-norm

$$J(\boldsymbol{x}) = \left\|\boldsymbol{x}\right\|_1 = \sum_{i=1}^{p} \left|\boldsymbol{x}_i\right|. \tag{2.6}$$

It is also referred to as ℓ_1-synthesis in the signal processing community, in contrast to the more general ℓ_1-analysis sparsity penalty detailed below.

We denote $(\boldsymbol{a}_i)_{1\leq i\leq p}$ the canonical basis of $\mathbb{R}^p$ and $\mathrm{supp}(\boldsymbol{x}) \stackrel{\text{def}}{=} \big\{i \in \{1,\ldots,p\} \;:\; \boldsymbol{x}_i \neq 0\big\}$. Then,

$$T_{\boldsymbol{x}} = \mathrm{Span}\{(\boldsymbol{a}_i)_{i\in\mathrm{supp}(\boldsymbol{x})}\}, \quad (e_{\boldsymbol{x}})_i = \begin{cases} \mathrm{sgn}(\boldsymbol{x}_i) & \text{if } i \in \mathrm{supp}(\boldsymbol{x}) \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad J^\circ = \big\|\cdot\big\|_\infty. \tag{2.7}$$

2.4.3.2 Group Lasso

The group Lasso has been advocated to promote sparsity by groups, i.e. it drives all the coefficients in one group to zero together hence leading to group selection, see [5, 161, 4, 156] to cite a few. The group Lasso penalty with L groups reads

$$J(\boldsymbol{x}) = \big\|\boldsymbol{x}\big\|_{1,2} \stackrel{\text{def}}{=} \sum_{i=1}^{L} \big\|\boldsymbol{x}_{b_i}\big\|_2. \tag{2.8}$$

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 $\mathrm{supp}_{\mathcal{B}}(\boldsymbol{x}) \stackrel{\text{def}}{=} \big\{i \in \{1,\ldots,L\} \;:\; \boldsymbol{x}_{b_i} \neq 0\big\}$. Thus, one has

$$T_{\boldsymbol{x}} = \mathrm{Span}\{(a_j)_{\left\{j \,:\, \exists i\in\mathrm{supp}_{\mathcal{B}}(\boldsymbol{x}), j\in b_i\right\}}\}, (e_{\boldsymbol{x}})_{b_i} = \begin{cases} \frac{\boldsymbol{x}_{b_i}}{\|\boldsymbol{x}_{b_i}\|_2} & \text{if } i \in \mathrm{supp}_{\mathcal{B}}(\boldsymbol{x}) \\ 0 & \text{otherwise} \end{cases}, \text{ and } J^\circ(\boldsymbol{\omega}) = \max_{i\in\{1,\ldots,L\}} \|\boldsymbol{\omega}_{b_i}\|_2 \,. \tag{2.9}$$

2.4.3.3 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

$$J(\boldsymbol{x}) = \big\|\boldsymbol{D}^\top \boldsymbol{x}\big\|_{1,2}. \tag{2.10}$$

This encompasses the 2-D isotropic total variation [128]. 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 [128], and the fused Lasso [138]. The overlapping group Lasso [83] is also a special case of (2.8) by taking $\boldsymbol{D}^\top$ to be an operator that exactract the blocks [112, 39] (in which case $\boldsymbol{D}$ has even orthogonal rows).

Let $\Lambda_{\boldsymbol{x}} = \bigcup_{i\in\mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top\boldsymbol{x})} b_i$ and $\Lambda_{\boldsymbol{x}}^c$ its complement. From Lemma 2.4.5(ii) and (2.9), we get

$$T_{\boldsymbol{x}} = \mathrm{Ker}(\boldsymbol{D}_{\Lambda_{\boldsymbol{x}}^c}^\top), \quad e_{\boldsymbol{x}} = \mathrm{P}_{T_{\boldsymbol{x}}}\, \boldsymbol{D} e_{\boldsymbol{D}^\top\boldsymbol{x}}^{\|\|_{1,2}}, \quad \text{where} \quad \left(e_{\boldsymbol{D}^\top\boldsymbol{x}}^{\|\|_{1,2}}\right)_{b_i} = \begin{cases} \frac{(\boldsymbol{D}^\top\boldsymbol{x})_{b_i}}{\left\|(\boldsymbol{D}^\top\boldsymbol{x})_{b_i}\right\|_2} & \text{if } i \in \mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top\boldsymbol{x}) \\ 0 & \text{otherwise.} \end{cases} \tag{2.11}$$

If, in addition, $\boldsymbol{D}$ is surjective, then by virtue of Lemma 2.4.5(ii) we also have

$$J^\circ(\boldsymbol{\omega}) = \big\|\boldsymbol{D}^+\boldsymbol{\omega}\big\|_{\infty,2} \stackrel{\text{def}}{=} \max_{i\in\{1,\ldots,L\}} \big\|(\boldsymbol{D}^+\boldsymbol{\omega})_{b_i}\big\|_2 \,. \tag{2.12}$$

2.4.3.4 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

$$J(\boldsymbol{x}) = \big\|\boldsymbol{x}\big\|_\infty = \max_{i\in\{1,\ldots,p\}} |\boldsymbol{x}_i| \,. \tag{2.13}$$

The ℓ_∞ regularization has found applications in several fields [84, 102, 133]. Suppose that $\boldsymbol{x} \neq 0$, and define the saturation support of $\boldsymbol{x}$ as $I_{\boldsymbol{x}}^{\text{sat}} \overset{\text{def}}{=} \left\{ i \in \{1, \dots, p\} \ : \ |\boldsymbol{x}_i| = \|\boldsymbol{x}\|_\infty \right\} \neq \emptyset$. From [145, Proposition 14], we have

$$T_{\boldsymbol{x}} = \left\{ \bar{\boldsymbol{\theta}} \in \mathbb{R}^p \ : \ \bar{\boldsymbol{\theta}}_{I_{\boldsymbol{x}}^{\text{sat}}} \in \mathbb{R} \operatorname{sgn}(\boldsymbol{x}_{I_{\boldsymbol{x}}^{\text{sat}}}) \right\}, \quad (e_{\boldsymbol{x}})_i = \begin{cases} \operatorname{sgn}(\boldsymbol{x}_i)/|I_{\boldsymbol{x}}^{\text{sat}}| & \text{if } i \in I_{\boldsymbol{x}}^{\text{sat}} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad J^\circ = \|\cdot\|_1 . \tag{2.14}$$

2.4.3.5 Nuclear norm

The natural extension of low-complexity priors to matrices $\boldsymbol{x} \in \mathbb{R}^{p_1 \times p_2}$ is to penalize the singular values of the matrix. Let $\operatorname{rank}(\boldsymbol{x}) = r$, and $\boldsymbol{x} = \boldsymbol{U} \operatorname{diag}(\boldsymbol{\sigma}(\boldsymbol{x})) \boldsymbol{V}^\top$ 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 $\boldsymbol{\sigma}(\boldsymbol{x}) \in (\mathbb{R}_+ \setminus \{0\})^r$ is the vector of singular values $(\boldsymbol{\sigma}_1(\boldsymbol{x}), \cdots, \boldsymbol{\sigma}_r(\boldsymbol{x}))^\top$ in non-increasing order. The nuclear norm of $\boldsymbol{x}$ is

$$J(\boldsymbol{x}) = \left\|\boldsymbol{x}\right\|_* = \left\|\boldsymbol{\sigma}(\boldsymbol{x})\right\|_1. \tag{2.15}$$

This penalty is the best convex surrogate to enforce a low-rank prior. It has been widely used for various applications [118, 31, 27, 70, 33].

Following e.g. [144, Example 21], we have

$$T_{\boldsymbol{x}} = \left\{ \boldsymbol{U}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{V}^\top \ : \ \boldsymbol{A} \in \mathbb{R}^{p_2 \times r}, \boldsymbol{B} \in \mathbb{R}^{p_1 \times r} \right\}, \quad e_{\boldsymbol{x}} = \boldsymbol{U}\boldsymbol{V}^\top \quad \text{and} \quad J^\circ(\boldsymbol{\omega}) = \left\|\boldsymbol{\omega}\right\|_{2\to 2} = \left\|\boldsymbol{\sigma}(\boldsymbol{\omega})\right\|_\infty. \tag{2.16}$$

2.5 Variational analysis

Denote $\mathcal{J}(\mathbb{R}^d)$ the class of functions which are proper, lsc and bounded from below.

2.5.1 Differentiability

Definition 2.5.1 ((Limiting) Subdifferential). Given a point $\boldsymbol{x} \in \mathbb{R}^d$ where a function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is finite, the subdifferential of f at $\boldsymbol{x}$ is defined as

$$\partial f(\boldsymbol{x}) = \left\{ \boldsymbol{v} \in \mathbb{R}^d \ : \ \exists \boldsymbol{x}_k \to \boldsymbol{x}, f(\boldsymbol{x}_k) \to f(\boldsymbol{x}), \boldsymbol{v} \leftarrow \boldsymbol{v}_k \in \partial^F f(\boldsymbol{x}_k) \right\},$$

where the Fréchet subdifferential $\partial^F f(\boldsymbol{x})$ of f at $\boldsymbol{x}$, is the set of vectors $\boldsymbol{v}$ such that

$$f(\boldsymbol{w}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{w} - \boldsymbol{x} \rangle + o\left(\left\|\boldsymbol{w} - \boldsymbol{x}\right\|_2\right), \quad \forall \boldsymbol{w} \in \mathbb{R}^d.$$

We say that f is subdifferentially regular at $\boldsymbol{x}$ if and only if f is locally lsc there with $\partial f(\boldsymbol{x}) = \partial^F f(\boldsymbol{x})$.

Let us note that $\partial f(\boldsymbol{x})$ and $\partial^F f(\boldsymbol{x})$ are closed, with $\partial^F f(\boldsymbol{x})$ convex and $\partial^F f(\boldsymbol{x}) \subset \partial f(\boldsymbol{x})$ [126, Theorem 8.6]. In particular, if f is a proper lsc convex function, $\partial^F f(\boldsymbol{x}) = \partial f(\boldsymbol{x})$ and ∂f is actually the Fenchel subdifferential (see Definition 2.3.2). This is why in the rest of the manuscript, we will call the limiting subdifferential simply the subdifferential.

Theorem 2.5.2 (Fermat's rule). *If a proper function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ has a local minimum at $\overline{\boldsymbol{x}}$, then*

$$0 \in \partial J(\overline{\boldsymbol{x}}).$$

If f is convex, then the above inclusion is also sufficient for $\overline{\boldsymbol{x}}$ to be a global minimizer.

2.5.2 Proximal mapping and Moreau envelope

Definition 2.5.3 (Proximal mapping and Moreau envelope). Let $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ symmetric positive definite. For a proper lsc function f and $\gamma > 0$, the proximal mapping and Moreau envelope in the metric $\boldsymbol{M}$ are defined respectively by

$$\operatorname{prox}_{\gamma f}^{\boldsymbol{M}}(\boldsymbol{x}) \overset{\text{def}}{=} \operatorname*{Argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \left\{ \frac{1}{2\gamma}\|\boldsymbol{w}-\boldsymbol{x}\|_{\boldsymbol{M}}^2 + f(\boldsymbol{w}) \right\},$$
$$^{\boldsymbol{M},\gamma}f(\boldsymbol{x}) \overset{\text{def}}{=} \inf_{\boldsymbol{w}\in\mathbb{R}^d} \left\{ \frac{1}{2\gamma}\|\boldsymbol{w}-\boldsymbol{x}\|_{\boldsymbol{M}}^2 + f(\boldsymbol{w}) \right\},$$

$\operatorname{prox}_{\gamma f}^{\boldsymbol{M}}$ here is a set-valued operator since the minimizer, if it exists, is not necessarily unique. When $\boldsymbol{M} = \mathbf{I}_p$, we simply write $\operatorname{prox}_{\gamma f}$ and $^{\gamma}f$.

Some key properties of the Moreau envelope and proximal mapping are detailed below.

Lemma 2.5.4. *Let $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ depending on $\gamma \in]0,\gamma_0[$ with $\gamma_0 > 0$, we denote it $\boldsymbol{M}_\gamma$, such that $\boldsymbol{M}_\gamma$ is symmetric positive definite for any $\gamma \in]0,\gamma_0[$, and $\gamma \mapsto \|\boldsymbol{\theta}\|_{\boldsymbol{M}_\gamma}$, $\forall \boldsymbol{\theta} \in \mathbb{R}^d$, is a decreasing mapping on $]0,\gamma_0[$. Assume that $f \in \mathcal{J}(\mathbb{R}^d)$.*

(i) $\operatorname{prox}_{\gamma f}^{\boldsymbol{M}_\gamma}(\boldsymbol{x})$ *are non-empty compact sets for any $\boldsymbol{x}$, and*

$$\boldsymbol{x} \in \operatorname{Argmin}(f) \Rightarrow \boldsymbol{x} \in \operatorname{prox}_{\gamma f}^{\boldsymbol{M}_\gamma}(\boldsymbol{x}).$$

(ii) $^{\boldsymbol{M}_\gamma,\gamma}f(\theta)$ *is finite and depends continuously on $(\boldsymbol{x},\gamma) \in \mathbb{R}^d\times]0,\gamma_0[$, and $\left(^{\boldsymbol{M}_\gamma,\gamma}f(\boldsymbol{x})\right)_{\gamma\in]0,\gamma_0[}$ is a decreasing net. More precisely,*

$$^{\boldsymbol{M}_\gamma,\gamma}f(\boldsymbol{x}) \nearrow f(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \text{ as } \gamma \searrow 0.$$

Proof.

(i) Since $f \in \mathcal{J}(\mathbb{R}^d)$, f is prox-bounded by [126, Exercise 1.24] for any $\gamma \in]0,\gamma_0[$, and then for any $\boldsymbol{x}$, $\frac{1}{2\gamma}\|\boldsymbol{x}-\cdot\|_{\boldsymbol{M}_\gamma}^2 + f$ is proper lsc and level-bounded uniformly in $(\boldsymbol{x},\gamma) \in \mathbb{R}^d\times]0,\gamma_0[$. This entails that the set of minimizers of this function, i.e. $\operatorname{prox}_{\gamma f}^{\boldsymbol{M}_\gamma}(\boldsymbol{x})$, is a non-empty compact set. For the last claim, suppose that $\boldsymbol{x} \in \operatorname{Argmin}(f) \neq \emptyset$ and bounded but $\boldsymbol{x} \notin \operatorname{prox}_{\gamma f}^{\boldsymbol{M}_\gamma}(\boldsymbol{x})$. Thus, for any $\boldsymbol{p} \in \operatorname{prox}_{\gamma f}^{\boldsymbol{M}_\gamma}(\boldsymbol{x})$, we have $\boldsymbol{p} \neq \boldsymbol{x}$ and

$$f(\boldsymbol{p}) < \frac{1}{2\gamma}\|\boldsymbol{p}-\boldsymbol{x}\|_{\boldsymbol{M}_\gamma}^2 + f(\boldsymbol{p}) \leq f(\boldsymbol{x}),$$

leading to a contradiction with $\boldsymbol{x}$ is a minimizer of f.

(ii) Since $f \in \mathcal{J}(\mathbb{R}^d)$, the continuity and finiteness properties follow from [126, Theorem 1.17(c)] (see also [126, Theorem 1.25]). For the second claim, we have $\forall \boldsymbol{x} \in \mathbb{R}^d$

$$-\infty < \inf f \leq {}^{\boldsymbol{M}_\gamma,\gamma}f(\boldsymbol{x}) \leq f(\boldsymbol{x}).$$

Moreover, let $\boldsymbol{p} \in \operatorname{prox}_{\gamma f}^{\boldsymbol{M}_\gamma}(\boldsymbol{x})$. Then, $\forall \delta > \gamma$,

$$\begin{aligned} {}^{\boldsymbol{M}_\delta,\delta}f(\boldsymbol{x}) &= \inf_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{2\delta}\|\boldsymbol{w}-\boldsymbol{x}\|_{\boldsymbol{M}_\delta}^2 + f(\boldsymbol{w}) \\ &\leq \frac{1}{2\delta}\|\boldsymbol{p}-\boldsymbol{x}\|_{\boldsymbol{M}_\delta}^2 + f(\boldsymbol{p}) \\ &\leq \frac{1}{2\gamma}\|\boldsymbol{p}-\boldsymbol{x}\|_{\boldsymbol{M}_\gamma}^2 + f(\boldsymbol{p}) \\ &= {}^{\boldsymbol{M}_\gamma,\gamma}f(\boldsymbol{x}). \end{aligned}$$

This together with continuity concludes the proof of Assertion (ii).

□

The fixed points of this proximal mapping include minimizers of f. They are not equal however in general, unless for instance f is convex.

Lemma 2.5.5. *Let $\boldsymbol{M}_\gamma \in \mathbb{R}^{d\times d}$ symmetric positive definite, assume that $f \in \mathcal{J}(\mathbb{R}^d)$, and $\operatorname{prox}_{\gamma f}^{\boldsymbol{M}_\gamma}$ is single-valued. Then $\operatorname{prox}_{\gamma f}^{\boldsymbol{M}_\gamma}$ is continuous on $(\boldsymbol{x},\gamma) \in \mathbb{R}^d\times]0,\gamma_0[$, and ${}^{\boldsymbol{M}_\gamma,\gamma}f \in C^1(\mathbb{R}^d)$ with gradient*

$$\nabla\, {}^{\boldsymbol{M}_\gamma,\gamma}f = \gamma^{-1}\boldsymbol{M}_\gamma\left(\mathbf{I}_d - \operatorname{prox}_{\gamma f}^{\boldsymbol{M}_\gamma}\right).$$

In plain words, Lemma 2.5.5 tells us that under mild and fairly general conditions, the Moreau envelope is a smooth function, hence the name Moreau-Yosida regularization. Moreover, the action of the operator $\operatorname{prox}_{\gamma f}^{\boldsymbol{M}_\gamma}$ is equivalent to a gradient descent on the Moreau envelope of f in the metric $\boldsymbol{M}_\gamma$ with step-size γ.

Proof. By virtue of Lemma 2.5.4-(i) and the single-valuedness, $\operatorname{prox}_{\gamma f}^{\boldsymbol{M}_\gamma}$ is clearly non-empty and single-valued. The continuity property follows from [126, Theorem 1.17(b)] (see also [126, Theorem 1.25]) and single-valuedness. By Lemma 2.5.4-(ii), ${}^{\boldsymbol{M}_\gamma,\gamma}f(\boldsymbol{\theta})$ is finite. Since $f \in \mathcal{J}(\mathbb{R}^d)$, f is prox-bounded with threshold $+\infty$ by [126, Exercise 1.24]. Invoking [126, Example 10.32], we get that $-\,{}^{\boldsymbol{M}_\gamma,\gamma}f$ is locally Lipschitz continuous, subdifferentially regular and

$$\partial\left(-\,{}^{\boldsymbol{M}_\gamma,\gamma}f\right)(\boldsymbol{\theta}) = \left\{\gamma^{-1}\boldsymbol{M}_\gamma\left(\operatorname{prox}_{\gamma f}^{\boldsymbol{M}_\gamma}(\boldsymbol{\theta}) - \boldsymbol{\theta}\right)\right\}, \quad \forall \boldsymbol{\theta} \in \mathbb{R}^p.$$

Combining this with [126, Theorem 9.18] applied to $-\,{}^{\boldsymbol{M}_\gamma,\gamma}f$, we obtain that ${}^{\boldsymbol{M}_\gamma,\gamma}f$ is differentiable and its gradient is precisely as given. □

Remark 2.5.6. When the metric matrix does not depend on γ, Lemmas 2.5.4 and 2.5.5 hold with $\gamma_0 = +\infty$.

2.5.3 Monotonicity

Definition 2.5.7 (Hypomonotone and monotone operators). A set-valued operator $S : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is hypomonotone of modulus $r > 0$ if

$$\left\langle \boldsymbol{x}' - \boldsymbol{x}, \boldsymbol{v}' - \boldsymbol{v}\right\rangle \geq -r\left\|\boldsymbol{x}' - \boldsymbol{x}\right\|_2^2, \quad \forall (\boldsymbol{x},\boldsymbol{v}) \in \operatorname{gph}(S), (\boldsymbol{x}',\boldsymbol{v}') \in \operatorname{gph}(S).$$

It is monotone if the inequality holds with $r = 0$, and strongly monotone with $r > 0$. Moreover, it is maximal monotone if there is no enlargement of its graph without destroying monotonicity.

The subdifferential of a function in $\Gamma_0(\mathbb{R}^d)$ is a typical example of a maximal monotone operator.

2.5.4 Prox-regularity

Roughly speaking, a lsc function f is prox-regular at $\overline{\boldsymbol{x}} \in \operatorname{dom}(f)$ if it has a "local quadratic support" at $\overline{\boldsymbol{x}}$ for all $(\boldsymbol{x},\boldsymbol{v}) \in \operatorname{gph}(\partial f)$ close enough to $(\overline{\boldsymbol{x}},\overline{\boldsymbol{v}}) \in \operatorname{gph}(\partial f)$ with $f(\boldsymbol{x})$ nearby $f(\overline{\boldsymbol{x}})$. This is formalized in the following definition.

Definition 2.5.8 (Prox-regularity). Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$, given a point $\overline{\boldsymbol{x}} \in \operatorname{dom}(f)$. f is prox-regular at $\overline{\boldsymbol{x}}$ for $\overline{\boldsymbol{v}}$, with $\overline{\boldsymbol{v}} \in \partial f(\overline{\boldsymbol{x}})$ if f is locally lsc at $\overline{\boldsymbol{x}}$, there exist $\epsilon > 0$ and $r > 0$ such that

$$f(\boldsymbol{x}') > f(\boldsymbol{x}) + (\boldsymbol{x}' - \boldsymbol{x})^\top \boldsymbol{v} - \frac{1}{2r}\left\|\boldsymbol{x}' - \boldsymbol{x}\right\|_2^2,$$

when $\left\|\boldsymbol{x}' - \overline{\boldsymbol{x}}\right\|_2 < \epsilon$ and $\left\|\boldsymbol{x} - \overline{\boldsymbol{x}}\right\|_2 < \epsilon$ with $\boldsymbol{x}' \neq \boldsymbol{x}$ and $\left\|f(\boldsymbol{x}) - f(\overline{\boldsymbol{x}})\right\|_2 < \epsilon$ while $\left\|\boldsymbol{v} - \overline{\boldsymbol{v}}\right\|_2 < \epsilon$ with $\boldsymbol{v} \in \partial f(\boldsymbol{x})$. When this holds for all $\overline{\boldsymbol{v}} \in \partial f(\overline{\boldsymbol{x}})$, f is said prox-regular at $\overline{\boldsymbol{x}}$. When f is prox-regular at every $\boldsymbol{x} \in \operatorname{dom}(f)$, f is said prox-regular.

Example 2.5.9. The class of prox-regular functions is large enough to include many of those used in statistics. For instance, here examples where prox-regularity is fullfilled (see [126, Chapter 13, Section F] and [115]):

(i) Proper lsc convex functions.

(ii) Proper lsc lower-C^2 (or semi-convex) functions, i.e. f is such that $f + \frac{1}{2r}\|\cdot\|_2^2$ is convex, $r > 0$.

(iii) Strongly amenable functions, i.e. $f = g \circ \boldsymbol{R}$, $\boldsymbol{R} : \mathbb{R}^d \to \mathbb{R}^d \in C^2(\mathbb{R}^d)$ and $g : \mathbb{R}^d \to \overline{\mathbb{R}}$ proper lsc convex.

(iv) A closed set $\mathcal{C} \subset \mathbb{R}^d$ is prox-regular if, and only if, ι_C is a prox-regular function. This is also equivalent to: for any $\boldsymbol{x} \in \mathbb{R}^d$ and for any $\gamma > 0$,

$$\mathrm{P}_{\mathcal{C}}(\boldsymbol{x}) = \operatorname*{Argmin}_{\boldsymbol{v}\in\mathbb{R}^d} \left\{ \frac{1}{\gamma}\|\boldsymbol{x}-\boldsymbol{v}\|_2^2 + \iota_{\mathcal{C}}(\boldsymbol{v}) \right\} = \mathrm{prox}_{\gamma\iota_{\mathcal{C}}}(\boldsymbol{x})$$

is single-valued and continuous, or equivalently, to

$$d_{\mathcal{C}}^2 = \min_{\boldsymbol{v}\in\mathbb{R}^d} \left\{ \frac{1}{\gamma}\|\cdot-\boldsymbol{v}\|_2^2 + \iota_{\mathcal{C}}(\boldsymbol{v}) \right\} = {}^{\gamma}\iota_{\mathcal{C}} \in C^{1,+}(\mathbb{R}^d).$$

The following lemma summarizes a fundamental property of prox-regular functions.

Lemma 2.5.10 ([114, Theorem 3.2]). *When $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is locally lsc at $\overline{\boldsymbol{x}} \in \mathbb{R}^d$, the following are equivalent*

(i) f is prox-regular at $\overline{\boldsymbol{x}}$ for $\overline{\boldsymbol{v}} \in \partial f(\overline{\boldsymbol{x}})$.

(ii) $\overline{\boldsymbol{v}}$ is a proximal subgradient to f at $\overline{\boldsymbol{x}}$, i.e. there exist $r > 0$ and $\epsilon > 0$ such that

$$f(\boldsymbol{x}) \geq f(\overline{\boldsymbol{x}}) + \langle \overline{\boldsymbol{v}}, \boldsymbol{x}-\overline{\boldsymbol{x}}\rangle - \frac{r}{2}\|\boldsymbol{x}-\overline{\boldsymbol{x}}\|_2^2, \quad \forall \boldsymbol{x} \quad \text{such that} \quad \|\boldsymbol{x}-\overline{\boldsymbol{x}}\|_2 < \epsilon.$$

Moreover, there exist $r > 0$ and an f-attentive ϵ-localization (with $\epsilon > 0$) of ∂f around $(\overline{\boldsymbol{x}}, \overline{\boldsymbol{v}})$ defined by

$$\mathrm{T}^f_{\epsilon,\overline{\boldsymbol{x}},\overline{\boldsymbol{v}}}(\boldsymbol{x}) = \begin{cases} \{\boldsymbol{v} \in \partial f(\boldsymbol{x}) \ : \ \|\boldsymbol{v}-\overline{\boldsymbol{v}}\|_2 < \epsilon\} & \text{if } \|\boldsymbol{x}-\overline{\boldsymbol{x}}\|_2 < \epsilon \text{ and } \|f(\boldsymbol{x})-f(\overline{\boldsymbol{x}})\|_2 < \epsilon, \\ \emptyset & \text{otherwise,} \end{cases}$$

such that $\mathrm{T}^f_{\epsilon,\overline{\boldsymbol{x}},\overline{\boldsymbol{v}}} + r\mathbf{I}_d$ is monotone.

2.6 Some useful integration formulas

Lemma 2.6.1 ([75, 3.251.11]). *Let $p, \gamma, \nu, \eta > 0$. If $\gamma/\nu < \eta + 1$ we have*

$$\int_0^{+\infty} \frac{x^{\gamma-1}}{(p+x^{\nu})^{\eta+1}} dx = \frac{1}{\nu p^{\eta+1-\gamma/\nu}} \frac{\Gamma(\gamma/\nu)\,\Gamma(1+\eta-\gamma/\nu)}{\Gamma(1+\eta)}, \tag{2.17}$$

otherwise this integral is not definite.

Lemma 2.6.2 (Cartesian to spherical coordinates [69]). *Let $d \geq 1$ and a mapping $h : \mathbb{R}_+ \to \mathbb{R}$ such that $\boldsymbol{u} \to h(\|\boldsymbol{u}\|_2)$ is measurable in $\mathbb{R}^d$. We obtain*

$$\int_{\mathbb{R}^d} h(\|\boldsymbol{u}\|_2) d\boldsymbol{u} = C_d \int_0^{+\infty} x^{d-1} h(x) dx, \tag{2.18}$$

where $C_d = 2\pi^{d/2}/\Gamma(d/2)$ is the surface area of a d-dimensional ball of radius 1.

Lemma 2.6.3 (Change of variables in the case of frame). *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a measurable set. Suppose that $\boldsymbol{M} \in \mathbb{R}^{d\times r}$ is a frame, let $u : \mathbb{R}^r \to \mathbb{R}$ such that the mapping $\boldsymbol{x} \mapsto u(\boldsymbol{M}^\top \boldsymbol{x})$ is measurable on $\mathcal{C}$. We have*

$$\int_{\mathcal{C}} u(\boldsymbol{M}^\top \boldsymbol{x}) d\boldsymbol{x} = \frac{1}{\sqrt{\det(\boldsymbol{M}\boldsymbol{M}^\top)}} \int_{\boldsymbol{M}^\top \mathcal{C}} u(\boldsymbol{v}) d\boldsymbol{v} \tag{2.19}$$

provided either u is non-negative valued or the integral on the left converges.

Though quite natural, proving Lemma 2.6.3 rigorously requires nontrivial arguments from geometric measure theory.

Proof. Consider the linear mapping $\boldsymbol{M}^\top : \boldsymbol{x} \in \mathbb{R}^d \mapsto \boldsymbol{M}^\top \boldsymbol{x} \in \mathbb{R}^r$, $r \geq d$. The Jacobian matrix of this mapping is obviously $\boldsymbol{M}^\top$ for any $\boldsymbol{x} \in \mathbb{R}^d$. Since $\boldsymbol{M}$ is a frame, $\boldsymbol{M}^\top$ is injective, hence so-called d-regular (see [103, Section 1.5]). In particular, $\det(\boldsymbol{M}\boldsymbol{M}^\top) > 0$. Thus combining [103, Theorems 1.12 and 3.4] and the Cauchy-Binet formula [103, Theorem 3.3]), we have the change of variables formula

$$\int_{\mathcal{C}} u(\boldsymbol{M}^\top \boldsymbol{x}) d\boldsymbol{x} = \frac{\int_{\mathbb{R}^r} \sum_{\boldsymbol{x} \in \mathcal{C} \cap \left\{ \boldsymbol{\omega} \,:\, \boldsymbol{M}^\top \boldsymbol{\omega} = \boldsymbol{v} \right\}} u(\boldsymbol{M}^\top \boldsymbol{x}) d\boldsymbol{v}}{\sqrt{\det(\boldsymbol{M}\boldsymbol{M}^\top)}} = \frac{\int_{\mathrm{Span}(\boldsymbol{M}^\top)} \sum_{\boldsymbol{x} \in \mathcal{C} \cap \left\{ \boldsymbol{\omega} \,:\, \boldsymbol{M}^\top \boldsymbol{\omega} = \boldsymbol{v} \right\}} u(\boldsymbol{M}^\top \boldsymbol{x}) d\boldsymbol{v}}{\sqrt{\det(\boldsymbol{M}\boldsymbol{M}^\top)}}. \tag{2.20}$$

Using once again that $\boldsymbol{M}^\top$ is injective, i.e. it is bijective on its image $\mathrm{Span}(\boldsymbol{M}^\top)$, the result follows. This concludes the proof. □

2.7 Inequalities from probability theory

Our work strongly relies on several important deviation inequalities from probability and concentration of measure. In particular, we will repeatedly use classical inequalities (Hoeffding, Bernstein), Gaussian concentration of Lipschitz functions, and its implications on e.g. concentration of the extreme singular values of a matrix, deviation inequalities for quadratic forms, and suprema of empirical processes.

For the sake of conciseness in this manuscript, we will not state them here. Rather we refer to e.g. [16, 94] for a comprehensive treatment.

Part I

Theoretical Guarantees

Chapter 3

PAC-Bayesian Risk Bounds for Analysis-group Sparse Regression by EWA

Main contributions of this chapter

- Propose an EWA of type (1.11) (see Section 3.3), with a deterministic dictionary, under an analysis-group sparsity prior where the analysis operator is associated to a frame $\boldsymbol{D}$. The performance of our estimator is guaranteed by an analysis-group SOI.
- Exhibit, for an appropriate choice of prior, an analysis-group SOI whose remainder term scales as $O\left(\left\|\boldsymbol{D}^\top\boldsymbol{\theta}\right\|_{0,2}\log(L)/n\right)$, where $\left\|\boldsymbol{D}^\top\boldsymbol{\theta}\right\|_{0,2}$ is the number of active groups in $\boldsymbol{D}^\top\boldsymbol{\theta}$, and L is the total number of groups (see Corollary 3.4.5). This rate coincides with the classical one $O\left(\|\boldsymbol{\theta}\|_0\log(p)/n\right)$ under the individual sparsity context.

The results in this chapter can be found in [64].

Contents

3.1 Introduction

In this chapter, we consider a high-dimensional non-parametric regression model with fixed design and random errors. We propose a powerful estimator by exponential weighted aggregation (EWA) with an analysis-group sparsity promoting prior on the weights. We prove that our estimator satisfies a sharp analysis-group sparse oracle inequality with a small remainder term ensuring its good theoretical performances.

3.1.1 Problem statement

Let us briefly recall our statistical context. Assume that the given data (x_i, y_i), $i = 1, \dots, n$, is generated according to the high-dimensional non-parametric regression model

$$y_i = f(x_i) + \xi_i, \quad i \in \{1, \dots, n\}, \tag{3.1}$$

where $x_1, \dots, x_n$ are deterministic in an arbitrary set $\mathcal{X}$, $f : \mathcal{X} \to \mathbb{R}$ is the unknown regression function and $(\xi_1, \dots, \xi_n)$ are random errors. (3.1) is equivalently written in vector form

$$\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\xi}.$$

Let $F(\boldsymbol{u}, \boldsymbol{y}) = \frac{1}{2} \|\boldsymbol{u} - \boldsymbol{y}\|_2^2$ for any $\boldsymbol{u} \in \mathbb{R}^n$. By the aggregation approach, we approximate f by $f_{\boldsymbol{\theta}_0}$ (see Section 1.1.1) where

$$\boldsymbol{\theta}_0 \in \operatorname*{Argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathbb{E}\left[F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y})\right].$$

Assume that $\boldsymbol{\theta}_0$ is analysis-group sparse. This chapter consists in estimating $\boldsymbol{\theta}_0$ by a class of EWA (1.11) whose performance is guaranteed by proven oracle inequalities.

3.1.2 Chapter organization

Section 3.2 reminds the PAC-Bayesian type oracle inequalities proposed in [52] which are a classical starting point in the literature for EWA in the deterministic case. In Section 3.3, we describe our EWA procedure after specifying the aggregation dictionary and our prior family. In Section 3.4, we establish our main results, namely analysis-group SOI. The proof of our general analysis-group SOI (c.f. Theorem 3.4.1) is reported in Section 3.5.

3.2 PAC-Bayesian type oracle inequalities

This section recalls a PAC-Bayesian type oracle inequality which holds for the EWA (1.11) with any deterministic aggregation dictionary, any prior and a large class of noises. Such type of oracle inequalities was introduced in [52] for i.i.d. noise. We adapt it to the non i.i.d. case. Indeed, let us start with the two following assumptions.

(P.1) The noise vector $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n)^\top$ has zero mean.

(P.2) For any $\gamma > 0$ small enough, there exist a probability space and two random variables $\boldsymbol{\xi}'$ and $\boldsymbol{\zeta}$ defined on this probability space satisfying the three following points:

(a) $\boldsymbol{\xi}'$ has the same distribution as $\boldsymbol{\xi}$.

(b) $\boldsymbol{\xi}' + \boldsymbol{\zeta}$ has the same distribution as $(1+\gamma)\boldsymbol{\xi}'$ and the conditional expectation satisfies $\mathbb{E}\left[\boldsymbol{\zeta}|\boldsymbol{\xi}'\right] = 0$.

(c) There exist $t_0 \in]0, +\infty]$ and a bounded Borel function $v : \mathbb{R}^n \to \mathbb{R}_+$ such that

$$\limsup_{\gamma \to 0} \sup_{(\boldsymbol{t},\boldsymbol{a}) \in \mathbb{R}^p \times \mathbb{R}^p : (\left\|\boldsymbol{t}\right\|_2, \boldsymbol{a}) \in [-t_0, t_0] \times \mathrm{supp}(\boldsymbol{\xi}')} \frac{\log \mathbb{E}\left[\exp\left(\boldsymbol{t}^\top \boldsymbol{\zeta}\right) | \boldsymbol{\xi}' = \boldsymbol{a}\right]}{\left\|\boldsymbol{t}\right\|_2^2 \gamma v(\boldsymbol{a})} \leq 1.$$

Assumption (P.2) is based on [52, Assumption N], and can be shown to be fulfilled for a large class of noises.

Proposition 3.2.1. *Assume that $\boldsymbol{\xi}$ has zero mean. Assumption (P.2) is fulfilled when:*

(i) $\boldsymbol{\xi}$ is a Gaussian noise having covariance matrix $\boldsymbol{\Sigma}$, with $t_0 = +\infty$ and $v(\boldsymbol{a}) \equiv \left\|\boldsymbol{\Sigma}\right\|_{2\to 2}$,

(ii) $\boldsymbol{\xi}$ is a Laplace noise having covariance matrix $\boldsymbol{\Sigma}$, with $t_0 < \sqrt{2/\left\|\boldsymbol{\Sigma}\right\|_{2\to 2}}$ and $v(\boldsymbol{a}) \equiv \frac{\left\|\boldsymbol{\Sigma}\right\|_{2\to 2}}{1 - t_0^2 \left\|\boldsymbol{\Sigma}\right\|_{2\to 2}^2 / 2}$,

(iii) $\boldsymbol{\xi}$ is a bounded symmetric noise, i.e. $\mathbb{P}\left[|\boldsymbol{\xi}_i| \leq \boldsymbol{B}_i\right]$ for some $\boldsymbol{B} \in \mathbb{R}^n$, with $t_0 = +\infty$ and $v(\boldsymbol{a}) = \left\|\boldsymbol{a}\right\|_2 \leq \left\|\boldsymbol{B}\right\|_2$.

Proof.

(i) **Gaussian noise:** Let $\boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, we set

$$\boldsymbol{\zeta} \sim \mathcal{N}(0, (2\gamma + \gamma^2)\boldsymbol{\Sigma}),$$

the conditions (a) and (b) in Assumption (P.2) are then verified. We check now the condition (c). Let $\boldsymbol{t} \in \left\{\boldsymbol{x} \in \mathbb{R}^n \ : \ \left\|\boldsymbol{x}\right\|_2 \in [-\infty, +\infty]\right\}$, $\boldsymbol{u} = \left(\boldsymbol{t}^\top \sqrt{2\gamma + \gamma^2}\boldsymbol{\Sigma}^{1/2}\right)^\top$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_n)$, we get that

$$\mathbb{E}\left[e^{\boldsymbol{t}^\top \boldsymbol{\zeta}}\right] = \mathbb{E}\left[e^{\boldsymbol{u}^\top \boldsymbol{\epsilon}}\right] = \prod_{j=1}^n \mathbb{E}\left[e^{\boldsymbol{u}_j \boldsymbol{\epsilon}_j}\right] = e^{\frac{\left\|\boldsymbol{u}\right\|_2^2}{2}} \leq e^{\frac{1}{2}\left\|\sqrt{2\gamma+\gamma^2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{t}\right\|_2^2} \leq e^{\left(\gamma + \frac{\gamma^2}{2}\right)\left\|\boldsymbol{\Sigma}\right\|_{2\to 2}\left\|\boldsymbol{t}\right\|_2^2}.$$

Thus, let $\boldsymbol{a} \in \mathbb{R}^n$ and $v(\boldsymbol{a}) \equiv \left\|\boldsymbol{\Sigma}\right\|_{2\to 2}$, we get

$$\frac{\log \mathbb{E}\left[e^{\boldsymbol{t}^\top \boldsymbol{\zeta}} | \boldsymbol{\xi} = \boldsymbol{a}\right]}{\left\|\boldsymbol{t}\right\|_2^2 \gamma v(\boldsymbol{a})} \leq 1 + \frac{\gamma}{2} \underset{\gamma \to 0}{\to} 1 \leq 1.$$

(ii) **Laplace noise:** Let $\boldsymbol{\xi} \sim \mathcal{L}(0, \boldsymbol{\Sigma})$, i.e. its associated characteristic function is $\varphi_{\boldsymbol{\xi}}(\boldsymbol{t}) = \frac{1}{1 + \boldsymbol{t}^\top \boldsymbol{\Sigma} \boldsymbol{t}/2}$, we choose $\boldsymbol{\zeta}$ according to the distribution associated to the characteristic function

$$\varphi_{\boldsymbol{\zeta}}(\boldsymbol{t}) = \frac{1}{(1+\gamma)^2}\left(1 + \frac{2\gamma + \gamma^2}{1 + (1+\gamma)^2 \boldsymbol{t}^\top \boldsymbol{\Sigma} \boldsymbol{t}/2}\right).$$

For any $\boldsymbol{t} \in \mathbb{R}^n$, we get that

$$\varphi_{\boldsymbol{\xi}+\boldsymbol{\zeta}}(\boldsymbol{t}) = \varphi_{\boldsymbol{\xi}}(\boldsymbol{t})\varphi_{\boldsymbol{\zeta}}(\boldsymbol{t}) = \frac{1}{1 + (1+\gamma)^2 \boldsymbol{t}^\top \boldsymbol{\Sigma} \boldsymbol{t}/2} = \varphi_{(1+\gamma)\boldsymbol{\xi}}(\boldsymbol{t}). \tag{3.2}$$

Thus, $\boldsymbol{\zeta}+\boldsymbol{\xi}$ has the same distribution as $(1+\gamma)\boldsymbol{\xi}$. We also obtain

$$\mathbb{E}\left[\boldsymbol{\zeta}|\boldsymbol{\xi}\right]=\mathbb{E}\left[\boldsymbol{\zeta}\right]=(-i)\nabla\varphi_{\boldsymbol{\zeta}}(0)=0.$$

It suffices to check the condition (3) of Assumption (**P.2**). We know that

$$\mathbb{E}\left[e^{\boldsymbol{t}^\top\boldsymbol{\zeta}}|\boldsymbol{\xi}\right]=\mathbb{E}\left[e^{\boldsymbol{t}^\top\boldsymbol{\zeta}}\right]=\varphi_{\boldsymbol{\zeta}}(-i\boldsymbol{t})=\frac{1}{(1+\gamma)^2}\left(1+\frac{2\gamma+\gamma^2}{1-(1+\gamma)^2\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}/2}\right).$$

Using Taylor's formula, we have

$$\log\left(\mathbb{E}\left[e^{t\boldsymbol{\zeta}}|\boldsymbol{\xi}\right]\right)=\frac{\gamma\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}}{1-\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}/2}+O(\gamma^2).$$

Thus, let $\boldsymbol{t}\in\left\{\boldsymbol{x}\in\mathbb{R}^n\ :\ \|\boldsymbol{x}\|_2\in[-t_0,t_0]\right\}$, $\boldsymbol{a}\in\mathbb{R}^n$ and $v(\boldsymbol{a})\equiv\frac{\|\boldsymbol{\Sigma}\|_{2\to2}}{1-t_0^2\|\boldsymbol{\Sigma}\|_{2\to2}^{\ 2}/2}$, we get

$$\begin{aligned}\frac{\log\mathbb{E}\left[e^{\boldsymbol{t}^\top\boldsymbol{\zeta}}|\boldsymbol{\xi}=\boldsymbol{a}\right]}{\|\boldsymbol{t}\|_2^\top\gamma v(\boldsymbol{a})}\underset{\gamma\to0}{\to}\frac{\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}/(1-\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}/2)}{\|\boldsymbol{t}\|_2^2\|\boldsymbol{\Sigma}\|_{2\to2}/(1-t_0^2\|\boldsymbol{\Sigma}\|_{2\to2}/2)}&\leq\frac{1-t_0^2\|\boldsymbol{\Sigma}\|_{2\to2}/2}{1-\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}/2}\\&\leq\frac{1-t_0^2\|\boldsymbol{\Sigma}\|_{2\to2}/2}{1-\|\boldsymbol{t}\|_2^2\|\boldsymbol{\Sigma}\|_{2\to2}/2}\leq1.\end{aligned}$$

We get two last inequalities under the condition $1-t_0^2\|\boldsymbol{\Sigma}\|_{2\to2}/2>0$ equivalent $t_0<\sqrt{2/\|\boldsymbol{\Sigma}\|_{2\to2}}$.

(iii) **Bounded symmetric noise:** Let $\boldsymbol{\xi}$ are symmetric and $\mathbb{P}\left[|\boldsymbol{\xi}_i|\leq B_i\right]=1$ for some $\boldsymbol{B}\in\mathbb{R}^n$, we set $\boldsymbol{\zeta}=(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_n)^\top$ such that

$$\boldsymbol{\zeta}_i=(1+\gamma)\,|\boldsymbol{\xi}_i|\,\mathrm{sgn}(\mathrm{sgn}(\boldsymbol{\xi}_i)-(1+\gamma)U_i)-\boldsymbol{\xi}_i,\quad U_i\sim\mathcal{U}([-1,1]),\ \forall i\in\{1,\ldots,n\}.$$

Using [52, Equation (22)], for any $\boldsymbol{t}\in\mathbb{R}^n$ and $\boldsymbol{a}\in\left\{\boldsymbol{x}\in\mathbb{R}^n\ :\ \boldsymbol{x}_i\in[-B_i,B_i],\ \forall i\in\{1,\ldots,n\}\right\}$, we get that

$$\mathbb{E}\left[e^{\boldsymbol{t}^\top\boldsymbol{\zeta}}|\boldsymbol{\xi}=\boldsymbol{a}\right]=\prod_{j=1}^n\mathbb{E}\left[e^{\boldsymbol{t}_i\boldsymbol{\zeta}_i}|\boldsymbol{\xi}_i=\boldsymbol{a}_i\right]=e^{-\boldsymbol{t}^\top\boldsymbol{a}}\left(e^{(1+\gamma)\boldsymbol{t}^\top\boldsymbol{a}}\frac{2+\gamma}{2+2\gamma}+e^{-(1+\gamma)\boldsymbol{t}^\top\boldsymbol{a}}\frac{\gamma}{2+2\gamma}\right).\quad(3.3)$$

From (3.3) and the symmetry of $\boldsymbol{\xi}$, we obtain

$$\mathbb{E}\left[e^{\boldsymbol{t}^\top(\boldsymbol{\zeta}+\boldsymbol{\xi})}\right]=\mathbb{E}\left[\mathbb{E}\left[e^{\boldsymbol{t}^\top(\boldsymbol{\zeta}+\boldsymbol{\xi})}|\boldsymbol{\xi}\right]\right]=\mathbb{E}\left[e^{(1+\gamma)\boldsymbol{t}^\top\boldsymbol{\xi}}\frac{2+\gamma}{2+2\gamma}+e^{-(1+\gamma)\boldsymbol{t}^\top\boldsymbol{\xi}}\frac{\gamma}{2+2\gamma}\right]=\mathbb{E}\left[e^{(1+\gamma)\boldsymbol{t}^\top\boldsymbol{\xi}}\right].$$

Thus, $\boldsymbol{\zeta}+\boldsymbol{\xi}$ has the same distribution as $(1+\gamma)\boldsymbol{\xi}$. Since $\mathbb{E}\left[\boldsymbol{\zeta}|\boldsymbol{\xi}=\boldsymbol{a}\right]$ equals to the gradient of $\mathbb{E}\left[e^{\boldsymbol{t}^\top\boldsymbol{\zeta}}|\boldsymbol{\xi}=\boldsymbol{a}\right]$ at $\boldsymbol{t}=0$, from (3.3) we have then $\mathbb{E}\left[\boldsymbol{\zeta}|\boldsymbol{\xi}=\boldsymbol{a}\right]=0$, $\forall\boldsymbol{a}\in[-\boldsymbol{B},\boldsymbol{B}]$. It suffices to check the condition (c) of Assumption (**P.2**). Owing to [45, Lemma 3] and [52, Equation (22)], we get that $\log\left(\mathbb{E}\left[e^{\boldsymbol{t}_i\boldsymbol{\zeta}_i}|\boldsymbol{\xi}_i=\boldsymbol{a}_i\right]\right)\leq(\boldsymbol{t}_i\boldsymbol{a}_i)^2\gamma(1+\gamma)$. Thus, let $\boldsymbol{t}\in\left\{\boldsymbol{x}\in\mathbb{R}^n\ :\ \|\boldsymbol{x}\|_2\in[-\infty,+\infty]\right\}$ and $v(\boldsymbol{a})=\|\boldsymbol{a}\|_2^2$, we get that

$$\frac{\log\mathbb{E}\left[e^{\boldsymbol{t}^\top\boldsymbol{\zeta}}|\boldsymbol{\xi}=\boldsymbol{a}\right]}{\|\boldsymbol{t}\|_2^2\gamma v(\boldsymbol{a})}=\frac{\sum_{i=1}^n\log\mathbb{E}\left[e^{\boldsymbol{t}_i\boldsymbol{\zeta}_i}|\boldsymbol{\xi}_i=\boldsymbol{a}_i\right]}{\|\boldsymbol{t}\|_2^2\gamma\|\boldsymbol{a}\|_2^2}\leq\frac{\sum_{i=1}^n\boldsymbol{t}_i^2\boldsymbol{a}_i^2\gamma(1+\gamma)}{\|\boldsymbol{t}\|_2^2\gamma\|\boldsymbol{a}\|_2^2}\leq1+\gamma\underset{\gamma\to0}{\to}1\leq1.$$

□

Besides, let $H\in]0,+\infty]$ such that

$$\sup_{(\boldsymbol{\theta},\boldsymbol{\theta}')\in\Theta^2}\left\|\boldsymbol{f}_{\boldsymbol{\theta}}-\boldsymbol{f}_{\boldsymbol{\theta}'}\right\|_2\leq H.\quad(3.4)$$

Note that (3.4) is always satisfied since H is allowed to be infinite. However, for the sake of sharpness in our theoretical results, we wish to choose H as small as possible.

Remind the definition of $\|\cdot\|_n$ in (1.2), and define

$$\|v\|_\infty = \sup_{\boldsymbol{x}\in\mathbb{R}^n} |v(\boldsymbol{x})|,$$

for any function on $\mathbb{R}^n$. We are now ready to state the PAC-Bayesian type oracle inequalities.

Theorem 3.2.2. *Let Assumptions* (**P.1**) *and* (**P.2**) *be satisfied with some function v and let* (3.4) *holds. Then for any prior π, any probability measure p on Θ and any $\beta \geq \max(4\|v\|_\infty, 2H/t_0)$ or $\beta \geq 4\|v\|_\infty$ when $H = +\infty$, $t_0 = +\infty$, we have*

$$\mathbb{E}\left[\|\widehat{f}_n - f\|_n^2\right] \leq \int_\Theta \|f - f_{\boldsymbol{\theta}}\|_n^2 p(d\boldsymbol{\theta}) + \frac{\beta\,\mathrm{KL}(p,\pi)}{n}, \tag{3.5}$$

where $\widehat{f}_n$ is the aggregate defined in (1.11) *and* $\mathrm{KL}(p,\pi) = \int_\Theta \log\left(p(d\boldsymbol{\theta})/\pi(d\boldsymbol{\theta})\right) p(d\boldsymbol{\theta})$ *is the Kullback-Leibler divergence.*

The proof of Theorem 3.2.2 is a mild adaptation of the original one in [52, Section 2], where we used directly Assumption (**P.2**)-(c) in the vector $\boldsymbol{\zeta}$ instead of splitting it into $\boldsymbol{\zeta}_{i,i\in\{1,\dots,n\}}$ (that are no longer i.i.d.).

Related work The work of [45] proposed three types of oracle inequalities which are similar to (3.5) under different assumptions. The first type (see [45, Theorem 1]) holds under a restrictive condition on the noise. The second (see [45, Theorem 2]) involves conditions depending on the noise and also on the dictionary. The last (see [45, Theorem 4]) works for all symmetric noises without conditions on the dictionary. However, an additional term appears in the remainder term which has a low rate for some types of noise. Therefore, Theorem 3.2.2 (with Assumption (**P.2**)) is a good trade-off between these types of oracle inequalities.

Moreover, there exist some related forms of (3.5) in different frameworks. For example, when $\boldsymbol{\xi}_i \sim \mathcal{N}(0,\sigma_i^2)$, $i = 1,\dots,n$, the following aggregate was proposed in [49]:

$$\widehat{f} = \int_\Theta \widehat{f}_{\boldsymbol{\theta}} p(d\boldsymbol{\theta}), \quad p(d\boldsymbol{\theta}) = \frac{\exp\left(-\frac{n}{\beta}\widehat{r}_{\boldsymbol{\theta}}\right)\pi(d\boldsymbol{\theta})}{\int_\Theta \exp\left(-\frac{n}{\beta}\widehat{r}_{\boldsymbol{\omega}}\right)\pi(d\boldsymbol{\omega})},$$

where $\widehat{f}_{\boldsymbol{\theta}}$, $\boldsymbol{\theta}\in\Theta$ are affine estimators satisfying some conditions imposed in [49, Theorem 1] which yield the definition of $\widehat{r}_{\boldsymbol{\theta}}$, $\boldsymbol{\theta}\in\Theta$. This aggregate satisfies oracle inequalities defined therein which are the counterparts of (3.5) for the aggregation of estimators. In addition, in the case of random design (i.e. $x_1,\dots,x_n$ are random and i.i.d.), the work in [51] constructed a mirror averaging aggregate to obtain a generalized type of oracle inequalities where the performance is measured by any loss instead of the averaged square loss.

3.3 EWA

3.3.1 Analysis-group sparsity

We now describe formally what is intended by analysis-group sparsity, which measures group sparsity of the image of a vector with an analysis linear sparsifying transform. Let $q \geq p$. We partition the index set $\{1,\dots,q\}$ into L non-overlapping groups/blocks of indices $\{b_l\}_{1\leq l\leq L}$ such that

$$\bigcup_{l=1}^{L} b_l = \{1,\dots,q\} \quad \text{and} \quad b_l \cap b_k = \emptyset, \quad \forall l \neq k.$$

For the sake of simplicity, and without loss of generality, the groups are assumed to have the same size $|b_l| = K \geq 1$ and the total number of blocks L is supposed to be an integer. With these notations, the analysis-group sparsity assumption is formalized as follows.

(**H.1**) There exists $\boldsymbol{D} \in \mathbb{R}^{p\times q}$ such that $\left\|\boldsymbol{D}^\top \boldsymbol{\theta}_0\right\|_{0,2} \ll n$.

In plain words, Assumption (H.1) says that the number of active groups of $\boldsymbol{D}^\top\boldsymbol{\theta}_0$ is much smaller than the sample size. Note that this is a strict notion of analysis-group sparsity, and a weaker one could be also considered where most $\left(\boldsymbol{D}^\top\boldsymbol{\theta}_0\right)_{b_l}$ are nearly zero. We also impose the following assumption on $\boldsymbol{D}$.

(**H.2**) $\boldsymbol{D}$ is a frame (see Definition 2.2.5).

Let us now introduce some applications in the literature in which our sparsity context is applicable.

Example 3.3.1 (2-D piecewise constant image). Let $\boldsymbol{\theta}_0 \in \mathbb{R}^{\sqrt{p}\times\sqrt{p}}$ be a 2-D piecewise constant image. In this framework, a popular regularization is the isotropic total variation associated to analysis operator denoted $\boldsymbol{D}_{\text{TV}}$ (see [128]). Namely, let $\boldsymbol{D}_c : \mathbb{R}^{\sqrt{p}\times\sqrt{p}} \to \mathbb{R}^{\sqrt{p}\times\sqrt{p}}$ and $\boldsymbol{D}_r : \mathbb{R}^{\sqrt{p}\times\sqrt{p}} \to \mathbb{R}^{\sqrt{p}\times\sqrt{p}}$ the finite difference operators along, respectively, the columns and rows of an image. We define $\boldsymbol{D}_{\text{TV}}$ as

$$\boldsymbol{D}_{\text{TV}} \colon \boldsymbol{\theta} \in \mathbb{R}^{\sqrt{p}\times\sqrt{p}} \mapsto \text{vec}\left(\left(\text{vec}(\boldsymbol{D}_r(\boldsymbol{\theta})), \text{vec}(\boldsymbol{D}_c(\boldsymbol{\theta}))\right)^\top\right)^\top \in \mathbb{R}^{2p}.$$

By vectorizing $\boldsymbol{\theta}_0$, $\boldsymbol{D}^\top$ would correspond to $\boldsymbol{D}_{\text{TV}}$. Here, $\boldsymbol{D}^\top = [\boldsymbol{D}_l^{\text{LIN}\top}\boldsymbol{D}_c^{\text{LIN}\top}]^\top$ where $\boldsymbol{D}_l^{\text{LIN}} \in \mathbb{R}^{p\times p}$ (resp. $\boldsymbol{D}_c^{\text{LIN}} \in \mathbb{R}^{p\times p}$) is the linearized counterpart of $\boldsymbol{D}_r$ (resp. $\boldsymbol{D}_c$). With Neumann boundary conditions, $\boldsymbol{D}_l^{\text{LIN}}$ and $\boldsymbol{D}_c^{\text{LIN}}$ are bijective implying injectivity of $\boldsymbol{D}^\top$. Thus, $\boldsymbol{D}$ is a frame in view of Courant-Fisher theorem.

Example 3.3.2 (Signal with overlapping groups). Consider $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ generated from L groups which overlap. The analysis operator acts as a group extractor (see [113, 39]). In this framework, $\boldsymbol{D}\boldsymbol{D}^\top = \sum_{l=1}^L \boldsymbol{B}_l^\top\boldsymbol{B}_l$ where $\boldsymbol{B}_l \in \mathbb{R}^{q_l\times p}$, for $l \in \{1,\dots,L\}$, is a countable collection of localization operators, and then $q = \sum_{l=1}^L q_l$. In which case, $\boldsymbol{D}\boldsymbol{D}^\top$ is a block diagonal matrix, each block is invertible. Hence, $\boldsymbol{D}\boldsymbol{D}^\top$ is invertible. Thus $\boldsymbol{D}$ is a frame in view of Courant-Fisher theorem.

To design an aggregation by exponential weighting, two ingredients are essential: the aggregation dictionary and the prior which promotes analysis-group sparsity. We specify them below.

3.3.2 Choice of dictionary

We impose the following standard normalization assumption on $\boldsymbol{X}$.

(**H.3**) $\boldsymbol{X}$ is normalized such that all the diagonal entries of $\boldsymbol{X}^\top\boldsymbol{X}/n$ are 1.

Now, let us introduce our dictionary of aggregation:

$$\mathcal{F}_\Theta = \left\{f_{\boldsymbol{\theta}} = \ell\left(\sum_{j=1}^p \boldsymbol{\theta}_j f_j\right) : \boldsymbol{\theta} \in \Theta = \left\{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \left\|\boldsymbol{D}^\top\boldsymbol{\theta}\right\|_{a,2}^a \le R\right\}\right\}, \tag{3.6}$$

where $a \in]0,1]$, $R \in]0,+\infty]$ and $\ell : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable and known depending on the regression problem (for example: $\ell(x) = e^x$ for the exponential regression, $\ell(x) = e^x/(e^x+1)$ for the logistic regression and $\ell(x) = x$ for the linear regression). This dictionary of aggregation is similar to the one proposed in [52, 51, 50]. However, the set of indices is modified to adapt to the analysis-group sparsity and the exponent a is varied in $]0,1]$ instead of a fixed $a = 1$. The bound H in (3.4) for $\mathcal{F}_\Theta$ in (3.6) is established in the following result.

Proposition 3.3.3. *Let $\mathcal{F}_\Theta = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ defined in (3.6) with some $R > 0$, $a \in]0,1]$ and $\ell : \mathbb{R} \to \mathbb{R}$ twice continuously differentiable. Let Assumption (H.2) hold for some $\mu > 0$. We get that*

$$\sup_{(\boldsymbol{\theta},\boldsymbol{\theta}')\in\Theta^2} \left\|\boldsymbol{f}_{\boldsymbol{\theta}} - \boldsymbol{f}_{\boldsymbol{\theta}'}\right\|_2 \le 2\max_{\boldsymbol{x}\in\mathcal{B}} \left\|\boldsymbol{L}(\boldsymbol{x})\right\|_2,$$

where $\mathcal{B} = \left\{\boldsymbol{x} \in \mathbb{R}^n \ : \ \left\|\boldsymbol{x}\right\|_2 \le \frac{\left\|\boldsymbol{X}\right\|_{2\to 2} R^{1/a}}{\sqrt{\mu}}\right\}$ and $\boldsymbol{L} : \boldsymbol{x} \in \mathbb{R}^n \to (\ell(\boldsymbol{x}_1),\dots,\ell(\boldsymbol{x}_n))$.

From Proposition 3.3.3, one can choose $H = 2\max_{\boldsymbol{x}\in\mathcal{B}} \|\boldsymbol{L}(\boldsymbol{x})\|_2$.

Proof. Let $\boldsymbol{\theta} \in \Theta$ and $i \in \{1,\dots,n\}$. Setting $\boldsymbol{u}_i^{\boldsymbol{\theta}} = \sum_{j=1}^p \boldsymbol{\theta}_j f_j(x_i)$ and $\boldsymbol{u}^{\boldsymbol{\theta}} = (\boldsymbol{u}_1^{\boldsymbol{\theta}},\dots,\boldsymbol{u}_n^{\boldsymbol{\theta}})^\top$, and by virtue of (2.1), (2.2), (3.6) and the fact that $a \in]0,1]$, we have

$$\|\boldsymbol{u}^{\boldsymbol{\theta}}\|_2 = \|\boldsymbol{X}\boldsymbol{\theta}\|_2 \le \|\boldsymbol{X}\|_{2\to 2}\|\widetilde{\boldsymbol{D}}\|_{2\to 2}\|\boldsymbol{D}^\top\boldsymbol{\theta}\|_2 \le \frac{\|\boldsymbol{X}\|_{2\to 2}\|\boldsymbol{D}^\top\boldsymbol{\theta}\|_{a,2}}{\sqrt{\mu}} \le \frac{\|\boldsymbol{X}\|_{2\to 2}R^{1/a}}{\sqrt{\mu}}.$$

Which in turn implies $\boldsymbol{u}^{\boldsymbol{\theta}} \in \mathcal{B}$. Therefore, for any $(\boldsymbol{\theta},\boldsymbol{\theta}') \in \Theta^2$,

$$\|\boldsymbol{f}_{\boldsymbol{\theta}} - \boldsymbol{f}_{\boldsymbol{\theta}'}\|_2 = \|\boldsymbol{L}(\boldsymbol{u}^{\boldsymbol{\theta}}) - \boldsymbol{L}(\boldsymbol{u}^{\boldsymbol{\theta}'})\|_2 \le 2\max_{\boldsymbol{x}\in\mathcal{B}} \|\boldsymbol{L}(\boldsymbol{x})\|_2.$$

□

Remark 3.3.4. By choosing $H = 2\max_{\boldsymbol{x}\in\mathcal{B}} \|\boldsymbol{L}(\boldsymbol{x})\|_2$ with $\mathcal{B} = \{\boldsymbol{x}\in\mathbb{R}^n : \|\boldsymbol{x}\|_2 \le \frac{\|\boldsymbol{X}\|_{2\to 2}R^{1/a}}{\sqrt{\mu}}\}$, H depends on $\boldsymbol{X}$ and then on n under Assumption (H.3). So $\beta \ge \max(4\|v\|_\infty, 2H/t_0)$ also depends on n. For this isssue, $\boldsymbol{\xi}$ must satisfy Assumption (P.2) with $t_0 = +\infty$. In view of Proposition 3.2.1, we can consider $\boldsymbol{\xi}$ as a Gaussian or a bounded symmetric noise.

3.3.3 Choice of prior

We choose a general prior of the form

$$\pi(d\boldsymbol{\theta}) = \frac{1}{C_{\alpha,g,R}}\prod_{l=1}^L \exp\left(-\alpha^a \left\|[\boldsymbol{D}^\top\boldsymbol{\theta}]_{b_l}\right\|_2^a\right) g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{\theta}]_{b_l}\right\|_2\right) I_\Theta(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{3.7}$$

where $\alpha \ge 0$ and g satisfies the following requirements:

(**G.1**) $g : \mathbb{R}_+ \to \mathbb{R}_+$ is a bounded function such that $g \not\equiv 0$, $\boldsymbol{\theta} \mapsto g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{\theta}]_{b_l}\right\|_2\right)$ is measurable on $\mathbb{R}^p$, for any $l \in \{1,\dots,L\}$.

(**G.2**) The integrability condition:

$$\int_{\mathbb{R}^p} \prod_{l=1}^L g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{u}]_{b_l}\right\|_2\right) d\boldsymbol{u} < +\infty.$$

(**G.3**) The moment condition:

$$\int_{\mathbb{R}^p} \left\|[\boldsymbol{D}^\top\boldsymbol{u}]_{b_l}\right\|_2^2 \prod_{k=1}^L g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{u}]_{b_k}\right\|_2\right) d\boldsymbol{u} < +\infty, \quad \forall l \in \{1,\dots,L\}.$$

(**G.4**) There exist $\lambda \ge 0$ and $h : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\frac{g\left(\|\boldsymbol{t}-\boldsymbol{t}^*\|_2\right)}{g\left(\|\boldsymbol{t}\|_2\right)} \le h\left(\|\boldsymbol{t}^*\|_2\right)^\lambda, \quad \forall(\boldsymbol{t},\ \boldsymbol{t}^*) \in \mathbb{R}^K \times \mathbb{R}^K.$$

.

Assumptions (G.2) and (G.3) lead the following remark that is a core part for the construction of the general analysis-group SOI in Theorem 3.4.1.

Remark 3.3.5. Let $K \ge 1$ and $\boldsymbol{D} \in \mathbb{R}^{p\times q}$ satisfying Assumption (H.2). For any function g satisfying Assumptions (G.1), (G.2) and (G.3), and any $a \in]0,1]$, there exists $K_{a,g}^{\boldsymbol{D}} \in]0,+\infty[$ such that

$$\frac{\int_{\mathbb{R}^p} \left\|[\boldsymbol{D}^\top\boldsymbol{u}]_{b_l}\right\|_2^{2a} \prod_{k=1}^L g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{u}]_{b_k}\right\|_2\right) d\boldsymbol{u}}{\int_{\mathbb{R}^p} \prod_{k=1}^L g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{v}]_{b_k}\right\|_2\right) d\boldsymbol{v}} \le K_{a,g}^{\boldsymbol{D}}, \quad \forall l \in \{1,\dots,L\}. \tag{3.8}$$

Proof. Remark 3.3.5 is briefly proved as follows. From Assumption (G.3) and the fact that $g \not\equiv 0$, one can show that (3.8) holds for $a = 1$. Moreover, since $g \not\equiv 0$ and g satisfies Assumption (G.2), we have

$$\boldsymbol{u} \to \frac{\prod_{l=1}^L g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{u}]_{b_l}\right\|_2\right) d\boldsymbol{u}}{\int_{\mathbb{R}^p} \prod_{k=1}^L g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{v}]_{b_k}\right\|_2\right) d\boldsymbol{v}}$$

is a probability measure. Therefore, (3.8) holds for any a in $]0,1]$ by Hölder's inequality. □

At first glance, Assumptions (**G.2**) and (**G.3**) may seem cumbersome. However the following lemma gives a simple condition on g that implies them.

Lemma 3.3.6. *Let $K \geq 1$ and $\boldsymbol{D} \in \mathbb{R}^{p\times q}$ satisfying Assumption (**H.2**). Suppose that g satisfies Assumption (**G.1**) and*

$$\int_0^{+\infty} z^{K+1} g(z) dz < +\infty. \tag{3.9}$$

*Then Assumptions (**G.2**) and (**G.3**) are fulfilled.*

Proof. Let us first check the integrability condition (**G.2**). By Lemmas 2.6.3 and 2.6.2, we obtain

$$\begin{aligned}\int_{\mathbb{R}^p} \prod_{l=1}^{L} g\left(\left\|[\boldsymbol{D}^\top \boldsymbol{u}]_{b_l}\right\|_2\right) d\boldsymbol{u} = \frac{\int_{\mathrm{Span}(\boldsymbol{D}^\top)} \prod_{l=1}^{L} g\left(\left\|\boldsymbol{v}_{b_l}\right\|_2\right) d\boldsymbol{v}}{\sqrt{\det\left(\boldsymbol{D}\boldsymbol{D}^\top\right)}} &\leq \frac{\int_{\mathbb{R}^q} \prod_{l=1}^{L} g\left(\left\|\boldsymbol{v}_{b_l}\right\|_2\right) d\boldsymbol{v}}{\sqrt{\det\left(\boldsymbol{D}\boldsymbol{D}^\top\right)}} \\ &= \frac{\left(\int_{\mathbb{R}^K} g\left(\left\|\boldsymbol{u}\right\|_2\right) d\boldsymbol{u}\right)^L}{\sqrt{\det\left(\boldsymbol{D}\boldsymbol{D}^\top\right)}} \\ &= \frac{C_K^L \left(\int_0^{+\infty} z^{K-1} g(z) dz\right)^L}{\sqrt{\det\left(\boldsymbol{D}\boldsymbol{D}^\top\right)}}.\end{aligned}$$

Since $K \geq 1$ and $g : \mathbb{R}_+ \to \mathbb{R}_+$, by (3.9), we get

$$\begin{aligned}\int_{\mathbb{R}^p} \prod_{l=1}^{L} g\left(\left\|[\boldsymbol{D}^\top \boldsymbol{u}]_{b_l}\right\|_2\right) d\boldsymbol{u} &\leq \frac{C_K^L \left(\int_0^1 z^{K-1} g(z) dz + \int_1^{+\infty} w^{K-1} g(w) dw\right)^L}{\sqrt{\det\left(\boldsymbol{D}\boldsymbol{D}^\top\right)}} \\ &\leq \frac{C_K^L \left(\sup_{z\in[0,1]} g(z) + \int_1^{+\infty} w^{K+1} g(w) dw\right)^L}{\sqrt{\det\left(\boldsymbol{D}\boldsymbol{D}^\top\right)}} < +\infty.\end{aligned} \tag{3.10}$$

Therefore, g satisfies Assumption (**G.2**). Now, we check the moment condition (**G.3**). Using similar arguments to the bound (3.10), we have

$$\begin{aligned}\int_{\mathbb{R}^p} \left\|[\boldsymbol{D}^\top \boldsymbol{u}]_{b_l}\right\|_2^2 \prod_{k=1}^{L} g\left(\left\|[\boldsymbol{D}^\top \boldsymbol{u}]_{b_k}\right\|_2\right) d\boldsymbol{u} &\leq \frac{\int_{\mathbb{R}^q} \left\|\boldsymbol{v}_{b_l}\right\|_2^2 \prod_{k=1}^{L} g\left(\left\|\boldsymbol{v}_{b_k}\right\|_2\right) d\boldsymbol{v}}{\sqrt{\det\left(\boldsymbol{D}\boldsymbol{D}^\top\right)}} \\ &= \frac{\left(\int_{\mathbb{R}^K} \left\|\boldsymbol{u}\right\|_2^2 g\left(\left\|\boldsymbol{u}\right\|_2\right) d\boldsymbol{u}\right) \left(\int_{\mathbb{R}^K} g\left(\left\|\boldsymbol{v}\right\|_2\right) d\boldsymbol{v}\right)^{L-1}}{\sqrt{\det\left(\boldsymbol{D}\boldsymbol{D}^\top\right)}} \\ &= \frac{C_K^L \int_0^{+\infty} z^{K+1} g(z) dz \left(\int_0^{+\infty} w^{K-1} g(w) dw\right)^{L-1}}{\sqrt{\det\left(\boldsymbol{D}\boldsymbol{D}^\top\right)}} < +\infty,\end{aligned} \tag{3.11}$$

whence we conclude that g satisfies Assumption (**G.3**). □

In the two following remarks, we consider the case where $\boldsymbol{D}$ is invertible. Remark 3.3.7 provides a simple and explicit form for $K_{a,g}^{\boldsymbol{D}}$ and Remark 3.3.8 shows that condition (3.9) is necessary for g to obey Assumption (**G.3**).

Remark 3.3.7. Let $K \geq 1$ and $\boldsymbol{D} \in \mathbb{R}^{p\times p}$ be invertible. For any function g satisfying Assumptions (G.1), (G.2) and (G.3), and for any $a \in]0,1]$, one can choose $K_{a,g}^{\boldsymbol{D}}$ in (3.8) as

$$K_{a,g}^{\boldsymbol{D}} = \frac{\int_0^{+\infty} x^{K-1+2a} g(x)dx}{\int_0^{+\infty} z^{K-1} g(z)dz}. \tag{3.12}$$

Proof. The proof follows by combining Lemmas 2.6.3 and 2.6.2, i.e.

$$\begin{aligned}\frac{\int_{\mathbb{R}^p} \left\|[\boldsymbol{D}^\top \boldsymbol{u}]_{b_l}\right\|_2^{2a} \prod_{k=1}^L g\left(\left\|[\boldsymbol{D}^\top \boldsymbol{u}]_{b_k}\right\|_2\right) d\boldsymbol{u}}{\int_{\mathbb{R}^p} \prod_{k=1}^L g\left(\left\|[\boldsymbol{D}^\top \boldsymbol{v}]_{b_k}\right\|_2\right) d\boldsymbol{v}} &= \frac{\int_{\mathbb{R}^K} \|\boldsymbol{u}\|_2^{2a} g\left(\|\boldsymbol{u}\|_2\right) d\boldsymbol{u} \left(\int_{\mathbb{R}^K} g\left(\|\boldsymbol{v}\|_2\right) d\boldsymbol{v}\right)^{L-1}}{\left(\int_{\mathbb{R}^K} g\left(\|\boldsymbol{w}\|_2\right) d\boldsymbol{w}\right)^L} \\ &= \frac{\int_{\mathbb{R}^K} \|\boldsymbol{u}\|_2^{2a} g\left(\|\boldsymbol{u}\|_2\right) d\boldsymbol{u}}{\int_{\mathbb{R}^K} g\left(\|\boldsymbol{v}\|_2\right) d\boldsymbol{v}} \\ &= \frac{\int_0^{+\infty} x^{K-1+2a} g(x)dx}{\int_0^{+\infty} z^{K-1} g(z)dz}.\end{aligned}$$

□

Remark 3.3.8. When $\boldsymbol{D}$ is invertible, if g does not satisfy (3.9) then g cannot fulfill Assumption (G.3). Consequently, Assumption (G.3) and condition (3.9) are equivalent in the invertible case.

Proof. By Lemmas 2.6.3 and 2.6.2, we get

$$\int_{\mathbb{R}^p} \left\|[\boldsymbol{D}^\top \boldsymbol{u}]_{b_l}\right\|_2^2 \prod_{k=1}^L g\left(\left\|[\boldsymbol{D}^\top \boldsymbol{u}]_{b_k}\right\|_2\right) d\boldsymbol{u} = \frac{C_K^L \int_0^{+\infty} z^{K+1} g(z)dz}{\sqrt{\det\left(\boldsymbol{D}\boldsymbol{D}^\top\right)}} \left(\int_0^{+\infty} w^{K-1} g(w)dw\right)^{L-1}.$$

□

Let us now discuss some choices of g. Recall that the goal is to find a prior leading an oracle inequality with a small remainder term while promoting analysis-group sparsity.

Example 3.3.9. Consider $g : \mathbb{R}_+ \to \mathbb{R}_+$ defined by

$$g(x) = \frac{1}{(\tau^2 + x^2)^2}, \quad \tau > 0.$$

This choice of g yields a prior that specializes to the one in [52] for the individual sparsity scenario, i.e. with $\boldsymbol{D} = \mathbf{I}_p$, $K = 1$ and $a = 1$.

Example 3.3.10. Consider $g : \mathbb{R}_+ \to \mathbb{R}_+$ defined by

$$g(x) = \frac{1}{(\tau^b + x^b)^c},$$

where $\tau > 0$, $b \in]0,1]$ and $c > (2+K)/b$. The choice of c guarantees the validity of Assumptions (G.2) and (G.3). Thanks to the parameters b and c, this choice of g offers more flexibility than the one in the previous example. This allows for example to optimize the performance of EWA by tuning these parameters for the particular dataset at hand.

3.4 Analysis-group sparse oracle inequality

Once a suitable dictionary and prior are chosen according to the above, our goal now is to provide a theoretical guarantee for the aggregates by constructing an analysis-group SOI. First of all, based on PAC-Bayesian type oracle inequalities in Section 3.2, we establish our first main result: an analysis-group SOI for the dictionary (3.6) and the prior (3.7) with a function g obeying Assumptions (G.1), (G.2), (G.3) and (G.4).

Theorem 3.4.1 (General analysis-group sparse oracle inequality). *Let $K \geq 1$, $\boldsymbol{X}$ satisfying Assumption (H.3), and $\boldsymbol{D}$ satisfying Assumption (H.2) with $\mu > 0$. Let Assumptions (P.1) and (P.2) be satisfied with some function v, (3.4) holds and $\beta \geq \max\left(4\|v\|_\infty, 2H/t_0\right)$. For some $a \in]0,1]$, take the dictionary (3.6) and the prior (3.7) with g satisfying Assumptions (G.1), (G.2), (G.3) and (G.4). Let $K^{\boldsymbol{D}}_{a,g}$, as defined in (3.8), and assume that $R > 3\sqrt{K^{\boldsymbol{D}}_{a,g}}L$. Then the following analysis-group SOI holds*

$$\mathbb{E}\left[\|\widehat{f}_n - f\|_n^2\right] \leq \inf_{\Theta^{\boldsymbol{D}}_{\widehat{\mu}_n,L,R}} \left(\|f_{\boldsymbol{\theta}} - f\|_n^2 + \Phi^{\boldsymbol{D}}_{\widehat{\mu}_n,n,L}(\boldsymbol{\theta}) + \Omega^{\boldsymbol{D}}_{\widehat{\mu}_n,n,L,\lambda}(\boldsymbol{\theta})\right) + \Psi^{\boldsymbol{D}}_{\widehat{\mu}_n,L,p}, \tag{3.13}$$

with

$$\Theta^{\boldsymbol{D}}_{\widehat{\mu}_n,L,R} = \{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \|\boldsymbol{D}^\top\boldsymbol{\theta}\|_{a,2}^a \leq R - 3\sqrt{K^{\boldsymbol{D}}_{a,g}}L\},$$

$$\Phi^{\boldsymbol{D}}_{\widehat{\mu}_n,n,L}(\boldsymbol{\theta}) = \frac{\beta}{n}\left(1 + 3\sqrt{K^{\boldsymbol{D}}_{a,g}}L\alpha^a + \alpha^a\|\boldsymbol{D}^\top\boldsymbol{\theta}\|_{a,2}^a\right),$$

$$\Omega^{\boldsymbol{D}}_{\widehat{\mu}_n,n,L,\lambda}(\boldsymbol{\theta}) = \frac{\lambda\beta}{n}\sum_{l=1}^{L}\log h\left(\|\left[\boldsymbol{D}^\top\boldsymbol{\theta}\right]_{b_l}\|_2\right),$$

$$\Psi^{\boldsymbol{D}}_{\widehat{\mu}_n,L,p} = \frac{2K^{\boldsymbol{D}}_{1,g}\exp\left(3\sqrt{K^{\boldsymbol{D}}_{a,g}}L\alpha^a\right)pC_{f,\ell}}{\mu},$$

and $C_{f,\ell} = \|\ell'\|_\infty^2 + \|\ell''\|_\infty\left(\|\ell\|_\infty + \|f\|_\infty\right)$.

Remark 3.4.2. The analysis-group SOI (3.13) is sharp. It depends on several parameters discussed below.

1. The parameter R appears in the dictionary. Namely, $\widehat{f}_n$ mimics the best aggregate $f_{\boldsymbol{\theta}}$ for all possible weights belonging to $\left\{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{D}^\top\boldsymbol{\theta}\|_{a,2}^a \leq R - 3\sqrt{K^{\boldsymbol{D}}_{a,g}}L\right\}$. Then R must be sufficiently large to cover the "good" model $f_{\boldsymbol{\theta}_0}$ in Assumption (H.1). Moreover, since $R > 3\sqrt{K^{\boldsymbol{D}}_{a,g}}L$, $R \sim \sqrt{K^{\boldsymbol{D}}_{a,g}}L$ is the smallest rate we can choose to reduce the rate of $\Phi^{\boldsymbol{D}}_{\widehat{\mu}_n,n,L}(\boldsymbol{\theta})$ as $\|\boldsymbol{D}^\top\boldsymbol{\theta}\|_{a,2}^a \leq R$.
2. The parameter α is used to cancel the effect of $L\sqrt{K^{\boldsymbol{D}}_{a,g}}$ in the remainder terms. By choosing $\alpha \leq (3K\sqrt{K^{\boldsymbol{D}}_{a,g}})^{-1/a}$, we get that
$$\Phi^{\boldsymbol{D}}_{\widehat{\mu}_n,n,L}(\boldsymbol{\theta}) \leq \frac{\beta}{n}\left(1 + 3\sqrt{K^{\boldsymbol{D}}_{a,g}}\alpha^a + \alpha^a R\right) \sim \frac{1}{n} \quad \text{and} \quad \Psi^{\boldsymbol{D}}_{\widehat{\mu}_n,L,p} \leq \frac{2eC_{f,\ell}}{\mu}K^{\boldsymbol{D}}_{1,g}p.$$
3. The parameter $K^{\boldsymbol{D}}_{a,g}$ and the function h depend on the choice of g. They also control the rate of $\Omega^{\boldsymbol{D}}_{\widehat{\mu}_n,n,L,\lambda}(\boldsymbol{\theta})$ and $\Psi^{\boldsymbol{D}}_{\widehat{\mu}_n,L,p}$.

In what follows, let us state the consequences of Theorem 3.4.1 with the choices of g in Example 3.3.9 and 3.3.10. Especially, we will discuss the rate of $\Omega^{\boldsymbol{D}}_{\widehat{\mu}_n,n,L,\lambda}(\boldsymbol{\theta})$ and $\Psi^{\boldsymbol{D}}_{\widehat{\mu}_n,L,p}$.

We first consider the prior (3.7) in Example 3.3.9, under the individual sparsity scenario ($\boldsymbol{D} = \mathbf{I}_p$, $K = 1$) and the choice $a = 1$ (i.e. $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta}\|_1 \leq R\}$). This is the setting considered in [52]. We obtain the following SOI as a corollary of our main result.

Corollary 3.4.3. *Let $\boldsymbol{X}$ satisfying Assumption (H.3), $\boldsymbol{D} = \mathbf{I}_p$ and fix $K = 1$. Suppose that Assumptions (P.1) and (P.2) hold with some function v, (3.4) holds and $\beta \geq \max\left(4\|v\|_\infty, 2H/t_0\right)$. Fix $a = 1$, take the dictionary (3.6) and the prior (3.7) with g defined in Example 3.3.9 and $\alpha \leq 1/(3p\tau)$. Assume that $R > 3p\tau$. Choosing $\tau^2 \sim 1/(pn)$ and $R \sim p\tau$, SOI (3.13) holds with $\Theta^{\boldsymbol{D}}_{\widehat{\mu}_n,L,R} = \left\{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \|\boldsymbol{\theta}\|_1 \leq R - 3p\tau\right\}$,*

$$\Omega^{\boldsymbol{D}}_{\widehat{\mu}_n,n,L,\lambda}(\boldsymbol{\theta}) \sim \frac{\|\boldsymbol{\theta}\|_0\log(p)}{n} \quad \textit{and} \quad \Psi^{\boldsymbol{D}}_{\widehat{\mu}_n,L,p} \sim \frac{1}{n}.$$

Proof. Let $\gamma = 3$, $\nu = 2$ and $\eta = 1$. We have $\gamma/\nu < \eta + 1$ so that Lemma 2.6.1 applies. We thus obtain

$$\int_0^{+\infty} z^{K+1} g(z) dz = \int_0^{+\infty} \frac{z^2}{(z^2+\tau^2)^2} dz < +\infty.$$

From Lemma 3.3.6, g satisfies Assumptions (**G.2**) and (**G.3**). Moreover, taking $h(t) = 1 + t/\tau$ and $\lambda = 4$, for all $(t, t^*) \in \mathbb{R}^2$, we have by Young's inequality

$$\frac{g(|t-t^*|)}{g(|t|)} = \left[\frac{\tau^2+t^2}{\tau^2+(t-t^*)^2}\right]^2 = \left[1 + \frac{2\tau(t-t^*)t^*/\tau + t^{*2}}{\tau^2+(t-t^*)^2}\right]^2 \leq \left[1 + \frac{|t^*|}{\tau} + \frac{t^{*2}}{\tau^2}\right]^2 \leq h(|t^*|)^\lambda.$$

Therefore, g satisfies Assumptions (**G.1**), (**G.2**), (**G.3**) and (**G.4**) for $K = 1$. Owing to Remark 3.3.7 and Lemma 2.6.1, we obtain

$$K_{1,g}^{\boldsymbol{D}} = \frac{\int_0^{+\infty} \frac{x^2}{(\tau^2+x^2)^2} dx}{\int_0^{+\infty} \frac{1}{(\tau^2+y^2)^2} dy} = \tau^2.$$

We are now in position to apply Theorem 3.4.1 with $\boldsymbol{D} = \mathbf{I}_p$ (then $q = p$), $K = 1$ (then $L = q$), $a = 1$ and $\alpha \leq 1/(3p\tau)$ to conclude. Namely, since $\tau^2 \sim \frac{1}{pn}$ and $R \sim p\tau$, we get that

$$\Psi_{\widehat{\mu}_n, L, p}^{\boldsymbol{D}} \leq 2eC_{f,\ell}\tau^2 p \sim \frac{1}{n},$$

and

$$\Omega_{\widehat{\mu}_n, n, L, \lambda}^{\boldsymbol{D}}(\boldsymbol{\theta}) = \frac{4\beta}{n} \sum_{j=1}^{p} \log\left(1 + \frac{|\theta_j|}{\tau}\right) \leq \frac{4\beta}{n} \|\boldsymbol{\theta}\|_0 \log\left(1 + \frac{R}{\tau}\right) \sim \frac{\|\boldsymbol{\theta}\|_0 \log(p)}{n}.$$

This completes the proof. □

The order of $\Omega_{\widehat{\mu}_n, n, L, \lambda}^{\boldsymbol{D}}(\boldsymbol{\theta})$ is the classical rate under the sparsity scenario. This scaling is similar to the one in [52] with the same prior. However, the following remark shows that this prior is not adapted in the analysis-group case for any group size strictly larger than 1.

Remark 3.4.4. Suppose that $K \geq 2$, and let $\gamma = K + 2$, $\nu = 2$ and $\eta = 1$. We have $\gamma/\nu \geq \eta + 1$, and thus Lemma 2.6.1 yields $\int_0^{+\infty} \frac{x^{K+1}}{(\tau^2+x^2)^2} dx$ is not definite. Consequently, condition (3.9) is not fulfilled with g defined in Example 3.3.9 when $K \geq 2$.

According to Remark 3.3.8, Remark 3.4.4 implies that Assumption (**G.3**) is not fulfilled for g in Example 3.3.9 when the group size $K \geq 2$ and $\boldsymbol{D}$ invertible. Thus one cannot construct an analysis-group SOI from Theorem 3.4.1 to guarantee the quality of the corresponding estimator. Overcoming this limitation was yet another motivation behind the choice of g in Example 3.3.10, which turns out to work well under the analysis-group sparsity scenario. In a nutshell, an aggregate with g in Example 3.3.10 exhibits the analysis-group SOI defined in the following corollary with any $K \geq 1$, any $\boldsymbol{D} \in \mathbb{R}^{p \times q}$ satisfying Assumption (**H.2**) and any $a \in]0, 1]$.

Corollary 3.4.5. *Let $\boldsymbol{X}$ satisfying Assumption (**H.3**), $K \geq 1$ and $\boldsymbol{D}$ satisfying Assumption (**H.2**) with $\mu > 0$. Let Assumptions (**P.1**) and (**P.2**) be satisfied with some function v, (3.4) holds and $\beta \geq \max\left(4\|v\|_\infty, 2H/t_0\right)$. Take the dictionary (3.6) and the prior (3.7) with $a \in]0, 1]$, $\alpha \geq 0$ and g defined in Example 3.3.10. We get that g satisfies Assumptions (**G.1**), (**G.2**), (**G.3**) and (**G.4**). Then, let $K_{a,g}^{\boldsymbol{D}}$ as defined in (3.8), and assume that $R > 3\sqrt{K_{a,g}^{\boldsymbol{D}}} L$. Then the analysis-group SOI (3.13) holds, with $\lambda = c$ and $h(x) = 1 + (x/\tau)^b$.*

Proof. Let $\gamma = 2 + K$, $\nu = b$ and $\eta = c - 1$. We have $\gamma/\nu < \eta + 1$ and thus Lemma 2.6.1 applies, whence we obtain

$$\int_0^{+\infty} x^{K+1} g(x) dx = \int_0^{+\infty} \frac{x^{K+1}}{(\tau^b + x^b)^c} dx < +\infty.$$

From Lemma 3.3.6, g satisfies Assumptions (**G.2**) and (**G.3**). Recall that $b \in]0,1]$. Taking $h(x) = 1 + (x/\tau)^b$ and $\lambda = c$, for all $(\boldsymbol{t}, \boldsymbol{t}^*) \in \mathbb{R}^K \times \mathbb{R}^K$, we have

$$\frac{g(\|\boldsymbol{t}-\boldsymbol{t}^*\|_2)}{g(\|\boldsymbol{t}\|_2)} = \left[\frac{\tau^b + \|\boldsymbol{t}\|_2^b}{\tau^b + \|\boldsymbol{t}-\boldsymbol{t}^*\|_2^b}\right]^c \leq \left[\frac{\tau^b + \|\boldsymbol{t}-\boldsymbol{t}^*\|_2^b + \|\boldsymbol{t}^*\|_2^b}{\tau^b + \|\boldsymbol{t}-\boldsymbol{t}^*\|_2^b}\right]^c \leq \left[1 + \frac{\|\boldsymbol{t}^*\|_2^b}{\tau^b + \|\boldsymbol{t}-\boldsymbol{t}^*\|_2^b}\right]^c \leq h(\|\boldsymbol{t}^*\|_2)^\lambda.$$

Therefore, g satisfies Assumptions (**G.1**), (**G.2**), (**G.3**) and (**G.4**) with any $K \geq 1$. Applying Theorem 3.4.1, we conclude the proof. □

To get an explicit control of the remainder term, it is instructive to have a closed-form of $K_{a,g}^{\boldsymbol{D}}$. This can be done for instance when $\boldsymbol{D}$ is invertible, see (3.12). The obtained analysis-group SOI is stated as follows.

Corollary 3.4.6. *Consider the same framework as Corollary 3.4.5 with $\boldsymbol{D}$ invertible. For $a \in]0,1]$, let $\widetilde{K}_{a,g}^{\boldsymbol{D}} = \frac{\Gamma((2a+K)/b)\Gamma(c-(2a+K)/b)}{\Gamma(K/b)\Gamma(c-K/b)}$, and set $\alpha \leq 1/\left(3\tau^a\sqrt{\widetilde{K}_{a,g}^{\boldsymbol{D}}L}\right)^{1/a}$. Choosing $\tau^2 \sim 1/(pn)$ and $R \sim L\tau^a$, the analysis-group SOI (3.13) holds with $\Theta_{\widehat{\mu}_n,L,R}^{\boldsymbol{D}} = \left\{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \|\boldsymbol{D}^\top\boldsymbol{\theta}\|_{a,2}^a \leq R - 3\tau^a\sqrt{\widetilde{K}_{a,g}^{\boldsymbol{D}}L}\right\}$,*

$$\Omega_{\widehat{\mu}_n,n,L,\lambda}^{\boldsymbol{D}}(\boldsymbol{\theta}) \sim \frac{\|\boldsymbol{D}^\top\boldsymbol{\theta}\|_{0,2}\log(L)}{n} \quad \textit{and} \quad \Psi_{\widehat{\mu}_n,L,p}^{\boldsymbol{D}} \sim \frac{1}{n}.$$

Proof. Since g satisfies Assumptions (**G.2**), (**G.3**) and $\boldsymbol{D}$ is invertible, by Remark 3.3.7 and Lemma 2.6.1, we get

$$K_{a,g}^{\boldsymbol{D}} = \frac{\int_0^{+\infty} \frac{r^{K-1+2a}}{\left(\tau^b+r^b\right)^c}dr}{\int_0^{+\infty} \frac{q^{K-1}}{\left(\tau^b+q^b\right)^c}dq} = \tau^{2a}\frac{\Gamma\left(\frac{2a+K}{b}\right)\Gamma\left(c-\frac{2a+K}{b}\right)}{\Gamma\left(\frac{K}{b}\right)\Gamma\left(c-\frac{K}{b}\right)} = \widetilde{K}_{a,g}^{\boldsymbol{D}}\tau^{2a}.$$

Since $\tau^2 \sim \frac{1}{pn}$ and $R \sim L\tau^a$, we get that

$$\Psi_{\widehat{\mu}_n,L,p}^{\boldsymbol{D}} \leq \frac{2C_{f,\ell}\widetilde{K}_{1,g}^{\boldsymbol{D}}e}{\mu}p\tau^2 \sim \frac{1}{n},$$

and

$$\Omega_{\widehat{\mu}_n,n,L,\lambda}^{\boldsymbol{D}}(\boldsymbol{\theta}) = \frac{c\beta}{n}\sum_{l=1}^{L}\log\left(1+\left[\frac{\|[\boldsymbol{D}^\top\boldsymbol{\theta}]_{b_l}\|_2}{\tau}\right]^b\right) \leq \frac{c\beta}{n}\|\boldsymbol{D}^\top\boldsymbol{\theta}\|_{0,2}\log\left(1+\left[\frac{R^{1/a}}{\tau}\right]^b\right) \sim \frac{\|\boldsymbol{D}^\top\boldsymbol{\theta}\|_{0,2}\log(L)}{n}.$$

This ends the proof. □

By Assumption (**H.1**), $\|\boldsymbol{D}^\top\boldsymbol{\theta}\|_{0,2}$ is small when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ (with R must be sufficiently large to cover $\boldsymbol{\theta}_0$). Thus, $\|\boldsymbol{D}^\top\boldsymbol{\theta}_0\|_{0,2}\log(L)$ is small compared to n. Under the sparsity scenario, the order of $\Omega_{\widehat{\mu}_n,n,L,\lambda}^{\boldsymbol{D}}(\boldsymbol{\theta})$ becomes $O\left(\|\boldsymbol{\theta}\|_0\log(p)/n\right)$ which is the same rate as the aggregate with g in Example 3.3.9.

3.5 Proof of SOI results

3.5.1 Proof of Theorem 3.4.1

Proof. Remind the prior $\pi(d\boldsymbol{\theta})$ from (3.7), where $\Theta = \left\{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \|\boldsymbol{D}^\top\boldsymbol{\theta}\|_{a,2}^a \leq R\right\}$. Let $r_L = 3\sqrt{K_{a,g}^{\boldsymbol{D}}L}$, $\Theta_{p_0\boldsymbol{D}} = \left\{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \|\boldsymbol{D}^\top\boldsymbol{\theta} - \boldsymbol{D}^\top\boldsymbol{\theta}^*\|_{a,2}^a \leq r_L\right\}$ and

$$\boldsymbol{\theta}^* \in \left\{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \|\boldsymbol{D}^\top\boldsymbol{\theta}\|_{a,2}^a \leq R - 3\sqrt{K_{a,g}^{\boldsymbol{D}}L} = R - r_L\right\}. \tag{3.14}$$

We define the probability measure

$$p_0{}^{\boldsymbol{D}}(d\boldsymbol{\theta}) = \frac{1}{C_L}\left(\frac{d\pi}{d\boldsymbol{\theta}}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\right) I_{\Theta_{p_0{}^{\boldsymbol{D}}}}(\boldsymbol{\theta})d\boldsymbol{\theta},$$

where $C_L > 0$ is the normalization factor for $p_0{}^{\boldsymbol{D}}$. Since $r_L < R$, $\boldsymbol{\theta} \in \Theta_{p_0{}^{\boldsymbol{D}}}$ implies that $\boldsymbol{\theta} - \boldsymbol{\theta}^* \in \Theta$. Therefore,

$$\begin{aligned} p_0{}^{\boldsymbol{D}}(d\boldsymbol{\theta}) &= \frac{1}{C_L}\prod_{l=1}^{L} \exp\left(-\alpha^a \left\|[\boldsymbol{D}^\top\boldsymbol{\theta} - \boldsymbol{D}^\top\boldsymbol{\theta}^*]_{b_l}\right\|_2^a\right) g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{\theta} - \boldsymbol{D}^\top\boldsymbol{\theta}^*]_{b_l}\right\|_2\right) I_\Theta(\boldsymbol{\theta}-\boldsymbol{\theta}^*) I_{\Theta_{p_0{}^{\boldsymbol{D}}}}(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &= \frac{1}{C_L}\prod_{l=1}^{L} \exp\left(-\alpha^a \left\|[\boldsymbol{D}^\top\boldsymbol{\theta} - \boldsymbol{D}^\top\boldsymbol{\theta}^*]_{b_l}\right\|_2^a\right) g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{\theta} - \boldsymbol{D}^\top\boldsymbol{\theta}^*]_{b_l}\right\|_2\right) I_{\Theta_{p_0{}^{\boldsymbol{D}}}}(\boldsymbol{\theta})d\boldsymbol{\theta}. \end{aligned}$$

For any $i \in \{1,\dots,n\}$, with $\boldsymbol{X}_i = (f_1(x_i),\dots,f_p(x_i))^\top$, one can write $f_{\boldsymbol{\theta}}(x_i) = \ell\left(\sum_{j=1}^p \boldsymbol{\theta}_j f_j(x_i)\right) = \ell\left(\boldsymbol{X}_i^\top\boldsymbol{\theta}\right)$. Taylor-Lagrange formula then gives us

$$(f_{\boldsymbol{\theta}}(x_i) - f(x_i))^2 \le (f_{\boldsymbol{\theta}^*}(x_i) - f(x_i))^2 + C_{f,\ell}\left[\boldsymbol{X}_i^\top(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\right]^2 + 2\left(f_{\boldsymbol{\theta}^*}(x_i) - f(x_i)\right)\ell'\left(\boldsymbol{X}_i^\top\boldsymbol{\theta}^*\right)\boldsymbol{X}_i^\top(\boldsymbol{\theta}-\boldsymbol{\theta}^*), \tag{3.15}$$

where $C_{f,\ell} = \|\ell'\|_\infty^2 + \|\ell''\|_\infty\left(\|\ell\|_\infty + \|f\|_\infty\right)$. By summing over i from 1 to n, normalizing by $1/n$, taking the integral in Θ w.r.t. $p_0{}^{\boldsymbol{D}}$, inequality (3.15) becomes

$$\begin{aligned} \int_\Theta \left\|f_{\boldsymbol{\theta}} - f\right\|_n^2 p_0{}^{\boldsymbol{D}}(d\boldsymbol{\theta}) \le \left\|f_{\boldsymbol{\theta}^*} - f\right\|_n^2 + C_{f,\ell}\int_{\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \left[\boldsymbol{X}_i^\top(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\right]^2 p_0{}^{\boldsymbol{D}}(d\boldsymbol{\theta}) \\ + \frac{2}{n}\sum_{i=1}^n \left(f_{\boldsymbol{\theta}^*}(x_i) - f(x_i)\right)\ell'\left(\boldsymbol{X}_i^\top\boldsymbol{\theta}^*\right)\boldsymbol{X}_i^\top \int_\Theta (\boldsymbol{\theta}-\boldsymbol{\theta}^*) p_0{}^{\boldsymbol{D}}(d\boldsymbol{\theta}). \end{aligned} \tag{3.16}$$

Note that, the right term of inequality (3.16) corresponds to a sum of three components. In the following, we keep the first component and treat the other two.

Let us first show that the last component vanishes. Indeed, let $\boldsymbol{\theta} \in \Theta_{p_0{}^{\boldsymbol{D}}}$, from (3.14) and the fact that $a \in]0,1]$, we have

$$\begin{aligned} \left\|\boldsymbol{D}^\top\boldsymbol{\theta}\right\|_{a,2}^a = \sum_{l=1}^L \left\|[\boldsymbol{D}^\top\boldsymbol{\theta}]_{b_l}\right\|_2^a &\le \sum_{l=1}^L \left(\left\|[\boldsymbol{D}^\top\boldsymbol{\theta} - \boldsymbol{D}^\top\boldsymbol{\theta}^*]_{b_l}\right\|_2 + \left\|[\boldsymbol{D}^\top\boldsymbol{\theta}^*]_{b_l}\right\|_2\right)^a \\ &\le \left\|\boldsymbol{D}^\top\boldsymbol{\theta} - \boldsymbol{D}^\top\boldsymbol{\theta}^*\right\|_{a,2}^a + \left\|\boldsymbol{D}^\top\boldsymbol{\theta}^*\right\|_{a,2}^a \\ &\le r_L + \left\|\boldsymbol{D}^\top\boldsymbol{\theta}^*\right\|_{a,2}^a \le R. \end{aligned}$$

Then $\boldsymbol{\theta} \in \left\{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \left\|\boldsymbol{D}^\top\boldsymbol{\theta}\right\|_{a,2}^a \le R\right\} = \Theta$. Therefore, we have the embedding

$$\Theta_{p_0{}^{\boldsymbol{D}}} \subseteq \Theta. \tag{3.17}$$

In what follows, we denote $\mathbb{B}_{a,\mathcal{B}}^a(x) = \left\{\boldsymbol{z} \in \mathbb{R}^q : \left\|\boldsymbol{z}\right\|_{a,2}^a \le x\right\}$, $\forall x > 0$ for brevity. By (3.17), property (2.1), Lemma 2.6.3 and symmetry of $\mathbb{B}_{a,\mathcal{B}}^a(r_L) \cap \mathrm{Span}(\boldsymbol{D}^\top)$, we obtain

$$\begin{aligned} &\int_\Theta (\boldsymbol{\theta}-\boldsymbol{\theta}^*) p_0{}^{\boldsymbol{D}}(d\boldsymbol{\theta}) \\ &\propto \int_{\Theta\cap\Theta_{p_0{}^{\boldsymbol{D}}}} (\boldsymbol{\theta}-\boldsymbol{\theta}^*)\prod_{l=1}^L \exp\left(-\alpha^a \left\|[\boldsymbol{D}^\top\boldsymbol{\theta} - \boldsymbol{D}^\top\boldsymbol{\theta}^*]_{b_l}\right\|_2^a\right) g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{\theta} - \boldsymbol{D}^\top\boldsymbol{\theta}^*]_{b_l}\right\|_2\right) d\boldsymbol{\theta} \\ &= \int_{\Theta_{p_0{}^{\boldsymbol{D}}}} (\boldsymbol{\theta}-\boldsymbol{\theta}^*)\prod_{l=1}^L \exp\left(-\alpha^a \left\|[\boldsymbol{D}^\top\boldsymbol{\theta} - \boldsymbol{D}^\top\boldsymbol{\theta}^*]_{b_l}\right\|_2^a\right) g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{\theta} - \boldsymbol{D}^\top\boldsymbol{\theta}^*]_{b_l}\right\|_2\right) d\boldsymbol{\theta} \\ &= \frac{\widetilde{\boldsymbol{D}}}{\sqrt{\det\left(\boldsymbol{D}\boldsymbol{D}^\top\right)}} \int_{\mathbb{B}_{a,\mathcal{B}}^a(r_L)\cap\mathrm{Span}(\boldsymbol{D}^\top)} \boldsymbol{z}\prod_{l=1}^L \exp\left(-\alpha^a \left\|\boldsymbol{z}_{b_l}\right\|_2^a\right) g\left(\left\|\boldsymbol{z}_{b_l}\right\|_2\right) d\boldsymbol{z} = 0, \end{aligned} \tag{3.18}$$

which is the desired claim.

We now bound the second term in the right hand side of (3.16). Define

$$p_0(d\boldsymbol{u}) = \frac{1}{C_L\sqrt{\det\left(\boldsymbol{D}\boldsymbol{D}^\top\right)}}\prod_{l=1}^{L}\exp\left(-\alpha^a\left\|\boldsymbol{u}_{b_l}\right\|_2^a\right)g\left(\left\|\boldsymbol{u}_{b_l}\right\|_2\right)I_{\mathrm{Span}(\boldsymbol{D}^\top)\cap\mathbb{B}^a_{a,\mathcal{B}}(r_L)}(\boldsymbol{u})d\boldsymbol{u}. \tag{3.19}$$

One can see that p_0 coincides with the probability measure ${p_0}^{\boldsymbol{D}}$ on $\mathbb{R}^p$ via a change of variables of type (2.19). So, p_0 is a probability measure on $\mathbb{R}^q$. For any $i,j\in\{1,\dots,L\}$, $i\neq j$, by a change of variables, we get $\int_{\mathbb{R}^q}\boldsymbol{u}_{b_i}\boldsymbol{u}_{b_j}^\top p_0(d\boldsymbol{u}) = -\int_{\mathbb{R}^q}\boldsymbol{u}_{b_i}\boldsymbol{u}_{b_j}^\top p_0(d\boldsymbol{u})$, so

$$\int_{\mathbb{R}^q}\boldsymbol{u}_{b_i}\boldsymbol{u}_{b_j}^\top p_0(d\boldsymbol{u}) = 0. \tag{3.20}$$

For any $j\in\{1,\dots,L\}$, as all groups have the same size, we have

$$\int_{\mathbb{R}^q}\boldsymbol{u}_{b_j}\boldsymbol{u}_{b_j}^\top p_0(d\boldsymbol{u}) = \int_{\mathbb{R}^q}\boldsymbol{u}_{b_1}\boldsymbol{u}_{b_1}^\top p_0(d\boldsymbol{u}). \tag{3.21}$$

We obtain

$$\begin{aligned}
&\int_{\mathbb{R}^p}\frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{X}_i^\top(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\right]^2 {p_0}^{\boldsymbol{D}}(d\boldsymbol{\theta})\\
\text{((2.1) and Lemma 2.6.3)} &= \frac{1}{n}\int_{\mathbb{R}^q}\left[\boldsymbol{X}\widetilde{\boldsymbol{D}}\boldsymbol{u}\right]^\top\boldsymbol{X}\widetilde{\boldsymbol{D}}\boldsymbol{u}p_0(d\boldsymbol{u})\\
&= \frac{1}{n}\int_{\mathbb{R}^q}\mathrm{tr}\left(\boldsymbol{u}^\top\widetilde{\boldsymbol{D}}^\top\boldsymbol{X}^\top\boldsymbol{X}\widetilde{\boldsymbol{D}}\boldsymbol{u}\right)p_0(d\boldsymbol{u})\\
&= \frac{1}{n}\mathrm{tr}\left(\left(\boldsymbol{X}\widetilde{\boldsymbol{D}}\right)^\top\boldsymbol{X}\widetilde{\boldsymbol{D}}\int_{\mathbb{R}^q}\boldsymbol{u}\boldsymbol{u}^\top p_0(d\boldsymbol{u})\right)\\
\text{((3.20) and (3.21))} &= \frac{1}{n}\sum_{l=1}^{L}\mathrm{tr}\left(\left[\left(\boldsymbol{X}\widetilde{\boldsymbol{D}}\right)^\top\boldsymbol{X}\widetilde{\boldsymbol{D}}\right]_{b_l,b_l}\int_{\mathbb{R}^q}\boldsymbol{u}_{b_1}\boldsymbol{u}_{b_1}^\top p_0(d\boldsymbol{u})\right)\\
\text{(Von Neumann's trace inequality)} &\leq \frac{1}{n}\sum_{l=1}^{L}\sum_{j=1}^{K}\sigma_j\left(\left[\left(\boldsymbol{X}\widetilde{\boldsymbol{D}}\right)^\top\boldsymbol{X}\widetilde{\boldsymbol{D}}\right]_{b_l,b_l}\right)\sigma_j\left(\int_{\mathbb{R}^q}\boldsymbol{u}_{b_1}\boldsymbol{u}_{b_1}^\top p_0(d\boldsymbol{u})\right)\\
&\leq \frac{1}{n}\int_{\mathbb{R}^q}\left\|\boldsymbol{u}_{b_1}\right\|_2^2 p_0(d\boldsymbol{u})\sum_{l=1}^{L}\mathrm{tr}\left(\left[\left(\boldsymbol{X}\widetilde{\boldsymbol{D}}\right)^\top\boldsymbol{X}\widetilde{\boldsymbol{D}}\right]_{b_l,b_l}\right)\\
&= \frac{1}{n}\mathrm{tr}\left(\left(\boldsymbol{X}\widetilde{\boldsymbol{D}}\right)^\top\boldsymbol{X}\widetilde{\boldsymbol{D}}\right)\int_{\mathbb{R}^q}\left\|\boldsymbol{u}_{b_1}\right\|_2^2 p_0(d\boldsymbol{u}).
\end{aligned} \tag{3.22}$$

Moreover, by inequality (2.2), Assumption **(H.3)** and Von Neumann's trace inequality, we obtain

$$\frac{\mathrm{tr}\left(\left(\boldsymbol{X}\widetilde{\boldsymbol{D}}\right)^\top\boldsymbol{X}\widetilde{\boldsymbol{D}}\right)}{n} \leq \sum_{j=1}^{p}\sigma_j\left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\right)\sigma_j\left(\widetilde{\boldsymbol{D}}\widetilde{\boldsymbol{D}}^\top\right) \leq \sigma_1\left(\widetilde{\boldsymbol{D}}\widetilde{\boldsymbol{D}}^\top\right)\sum_{j=1}^{p}\sigma_j\left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\right) \leq \frac{p}{\mu}. \tag{3.23}$$

Putting together (3.22) and (3.23), we get the bound

$$C_{f,\ell}\int_{\mathbb{R}^p}\frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{X}_i^\top(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\right]^2 {p_0}^{\boldsymbol{D}}(d\boldsymbol{\theta}) \leq C_{f,\ell}\frac{p}{\mu}\int_{\mathbb{R}^q}\left\|\boldsymbol{u}_{b_1}\right\|_2^2 p_0(d\boldsymbol{u}). \tag{3.24}$$

Thanks to (3.18) and (3.24), inequality (3.16) becomes

$$\int_\Theta\left\|f_{\boldsymbol{\theta}}-f\right\|_n^2 {p_0}^{\boldsymbol{D}}(d\boldsymbol{\theta}) \leq \left\|f_{\boldsymbol{\theta}^*}-f\right\|_n^2 + C_{f,\ell}\frac{p}{\mu}\int_{\mathbb{R}^q}\left\|\boldsymbol{u}_{b_1}\right\|_2^2 p_0(d\boldsymbol{u}). \tag{3.25}$$

Now, inserting (3.25) into Theorem 3.2.2 (with $p = {p_0}^{\boldsymbol{D}}$), we arrive at

$$\mathbb{E}\left[\left\|\widehat{f}_n-f\right\|_n^2\right] \leq \left\|f_{\boldsymbol{\theta}^*}-f\right\|_n^2 + C_{f,\ell}\frac{p}{\mu}\int_{\mathbb{R}^q}\left\|\boldsymbol{u}_{b_1}\right\|_2^2 p_0(d\boldsymbol{u}) + \frac{\beta\,\mathrm{KL}({p_0}^{\boldsymbol{D}},\pi)}{n}. \tag{3.26}$$

To complete the proof, it remains to bound the last two terms in the right hand side of (3.26). This is the goal of the following lemma.

Lemma 3.5.1. *Consider the same framework as the one in Theorem 3.4.1, we have*

$$\int_{\mathbb{R}^q} \|\boldsymbol{u}_{b_1}\|_2^2 p_0(d\boldsymbol{u}) \leq 2K_{1,g}^{\boldsymbol{D}} e^{r_L \alpha^a}, \tag{3.27}$$

and

$$\mathrm{KL}(p_0{}^{\boldsymbol{D}}, \pi) \leq 1 + r_L \alpha^a + \lambda \sum_{l=1}^{L} \log\left\{ h\left(\left\| \left[\boldsymbol{D}^\top \boldsymbol{\theta}^*\right]_{b_l} \right\|_2 \right) \right\} + \alpha^a \|\boldsymbol{D}^\top \boldsymbol{\theta}^*\|_{a,2}^a. \tag{3.28}$$

With $r_L = 3\sqrt{K_{a,g}^{\boldsymbol{D}}} L$, it follows from (3.26) and Lemma 3.5.1 that

$$\begin{aligned}\mathbb{E}\left[\|\widehat{f}_n - f\|_n^2\right] \leq \|f_{\boldsymbol{\theta}^*} - f\|_n^2 + \frac{\beta}{n}\left(1 + 3\sqrt{K_{a,g}^{\boldsymbol{D}}} L \alpha^a + \alpha^a \|\boldsymbol{D}^\top \boldsymbol{\theta}^*\|_{a,2}^a\right) \\ + \frac{\lambda\beta}{n} \sum_{l=1}^{L} \log\left\{ h\left(\left\| \left[\boldsymbol{D}^\top \boldsymbol{\theta}^*\right]_{b_l} \right\|_2 \right) \right\} + \frac{2K_{1,g}^{\boldsymbol{D}} e^{3\sqrt{K_{a,g}^{\boldsymbol{D}}} L \alpha^a} p C_{f,\ell}}{\mu}.\end{aligned}$$

According to (3.14), this completes the proof of Theorem 3.4.1. □

3.5.2 Proof of Lemma 3.5.1

To prove Lemma 3.5.1, we need an intermediate result.

Lemma 3.5.2. *Let $s > L\sqrt{K_{a,g}^{\boldsymbol{D}}}$. The following inequality holds*

$$\frac{1}{T} \int_{\left\{\boldsymbol{u} \in \mathbb{R}^p \,:\, \|\boldsymbol{D}^\top \boldsymbol{u}\|_{a,2}^a > s\right\}} \prod_{l=1}^{L} g\left(\left\| [\boldsymbol{D}^\top \boldsymbol{u}]_{b_l} \right\|_2 \right) d\boldsymbol{u} \leq \frac{L^2 K_{a,g}^{\boldsymbol{D}}}{\left(s - L\sqrt{K_{a,g}^{\boldsymbol{D}}}\right)^2},$$

where $T = \int_{\mathbb{R}^p} \prod_{l=1}^{L} g\left(\left\| [\boldsymbol{D}^\top \boldsymbol{u}]_{b_l} \right\|_2 \right) d\boldsymbol{u}$.

Proof. Let $\boldsymbol{U}$ be a random vector in $\mathbb{R}^p$ with density $\boldsymbol{u} \mapsto \frac{1}{T} \prod_{l=1}^{L} g\left(\left\| [\boldsymbol{D}^\top \boldsymbol{u}]_{b_l} \right\|_2 \right)$, where $T < +\infty$ by Assumption (G.2). By Chebyshev inequality, we have

$$\begin{aligned}&\frac{1}{T} \int_{\left\{\boldsymbol{u} \in \mathbb{R}^p \,:\, \|\boldsymbol{D}^\top \boldsymbol{u}\|_{a,2}^a > s\right\}} \prod_{l=1}^{L} g\left(\left\| [\boldsymbol{D}^\top \boldsymbol{u}]_{b_l} \right\|_2 \right) d\boldsymbol{u} \\ &= \mathbb{P}\left[\sum_{l=1}^{L} \left\| [\boldsymbol{D}^\top \boldsymbol{U}]_{b_l} \right\|_2^a > s \right] \\ &= \mathbb{P}\left[\sum_{l=1}^{L} \left\| [\boldsymbol{D}^\top \boldsymbol{U}]_{b_l} \right\|_2^a - \mathbb{E}\left[\left\| [\boldsymbol{D}^\top \boldsymbol{U}]_{b_l} \right\|_2^a \right] > s - \sum_{l=1}^{L} \mathbb{E}\left[\left\| [\boldsymbol{D}^\top \boldsymbol{U}]_{b_l} \right\|_2^a \right] \right] \\ &\leq \frac{\mathbb{E}\left[\left[\sum_{l=1}^{L} \left\| [\boldsymbol{D}^\top \boldsymbol{U}]_{b_l} \right\|_2^a - \mathbb{E}\left[\left\| [\boldsymbol{D}^\top \boldsymbol{U}]_{b_l} \right\|_2^a \right] \right]^2 \right]}{\left(s - \sum_{l=1}^{L} \mathbb{E}\left[\left\| [\boldsymbol{D}^\top \boldsymbol{U}]_{b_l} \right\|_2^a \right]\right)^2} \\ &= \frac{\mathrm{var}\left(\sum_{l=1}^{L} \left\| [\boldsymbol{D}^\top \boldsymbol{U}]_{b_l} \right\|_2^a \right)}{\left(s - \sum_{l=1}^{L} \mathbb{E}\left[\left\| [\boldsymbol{D}^\top \boldsymbol{U}]_{b_l} \right\|_2^a \right]\right)^2} \\ &\leq \frac{\mathbb{E}\left[\left(\sum_{l=1}^{L} \left\| [\boldsymbol{D}^\top \boldsymbol{U}]_{b_l} \right\|_2^a \right)^2 \right]}{\left(s - \sum_{l=1}^{L} \mathbb{E}\left[\left\| [\boldsymbol{D}^\top \boldsymbol{U}]_{b_l} \right\|_2^a \right]\right)^2}.\end{aligned} \tag{3.29}$$

Next, by Cauchy-Schwartz inequality and Remark 3.3.5, we obtain

$$\mathbb{E}\left[\left(\sum_{l=1}^{L}\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{b_l}\right\|_2^a\right)^2\right] \leq \mathbb{E}\left[L\sum_{l=1}^{L}\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{b_l}\right\|_2^{2a}\right] \leq L^2 K_{a,g}^{\boldsymbol{D}} \tag{3.30}$$

and by Jensen inequality

$$s - \sum_{l=1}^{L}\mathbb{E}\left[\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{b_l}\right\|_2^a\right] \geq s - \sum_{l=1}^{L}\sqrt{\mathbb{E}\left[\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{b_l}\right\|_2^{2a}\right]} \geq s - L\sqrt{K_{a,g}^{\boldsymbol{D}}} > 0. \tag{3.31}$$

Thus, combining (3.29), (3.30) and (3.31), we get

$$\frac{1}{T}\int_{\left\{\boldsymbol{u}\in\mathbb{R}^p\,:\,\left\|\boldsymbol{D}^\top\boldsymbol{u}\right\|_{a,2}^a > s\right\}}\prod_{l=1}^{L} g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{u}]_{b_l}\right\|_2\right)d\boldsymbol{u} \leq \frac{L^2 K_{a,g}^{\boldsymbol{D}}}{\left(s - L\sqrt{K_{a,g}^{\boldsymbol{D}}}\right)^2}.$$

□

We now turn to the proof of Lemma 3.5.1

Proof. Let us begin by the proof of inequality (3.27). We have

$$\begin{aligned}\int_{\mathbb{R}^q}\left\|\boldsymbol{u}_{b_1}\right\|_2^2 p_0(d\boldsymbol{u}) &= \frac{1}{C_L\sqrt{\det\left(\boldsymbol{D}\boldsymbol{D}^\top\right)}}\int_{\mathbb{B}_{a,\mathcal{B}}^a(r_L)\cap\mathrm{Span}(\boldsymbol{D}^\top)}\left\|\boldsymbol{u}_{b_1}\right\|_2^2\prod_{l=1}^{L}e^{-\alpha^a\left\|\boldsymbol{u}_{b_l}\right\|_2^a} g\left(\left\|\boldsymbol{u}_{b_l}\right\|_2\right)d\boldsymbol{u} \\ &\leq \frac{1}{C_L\sqrt{\det\left(\boldsymbol{D}\boldsymbol{D}^\top\right)}}\int_{\mathbb{B}_{a,\mathcal{B}}^a(r_L)\cap\mathrm{Span}(\boldsymbol{D}^\top)}\left\|\boldsymbol{u}_{b_1}\right\|_2^2\prod_{l=1}^{L} g\left(\left\|\boldsymbol{u}_{b_l}\right\|_2\right)d\boldsymbol{u}. \end{aligned} \tag{3.32}$$

In the following, we show inequality (3.27) by bounding the right term of inequality (3.32). By Lemma 2.6.3 and Remark 3.3.5, we get

$$\begin{aligned}\frac{\int_{\mathbb{B}_{a,\mathcal{B}}^a(r_L)\cap\mathrm{Span}(\boldsymbol{D}^\top)}\left\|\boldsymbol{u}_{b_1}\right\|_2^2\prod_{l=1}^{L} g\left(\left\|\boldsymbol{u}_{b_l}\right\|_2\right)d\boldsymbol{u}}{\sqrt{\det\left(\boldsymbol{D}\boldsymbol{D}^\top\right)}\int_{\mathbb{R}^p}\prod_{l=1}^{L} g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{u}]_{b_l}\right\|_2\right)d\boldsymbol{u}} &\leq \frac{\int_{\mathrm{Span}(\boldsymbol{D}^\top)}\left\|\boldsymbol{u}_{b_1}\right\|_2^2\prod_{l=1}^{L} g\left(\left\|\boldsymbol{u}_{b_l}\right\|_2\right)d\boldsymbol{u}}{\sqrt{\det\left(\boldsymbol{D}\boldsymbol{D}^\top\right)}\int_{\mathbb{R}^p}\prod_{l=1}^{L} g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{u}]_{b_l}\right\|_2\right)d\boldsymbol{u}} \\ &= \frac{\int_{\mathbb{R}^p}\left\|[\boldsymbol{D}^\top\boldsymbol{u}]_{b_1}\right\|_2^2\prod_{l=1}^{L} g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{u}]_{b_l}\right\|_2\right)d\boldsymbol{u}}{\int_{\mathbb{R}^p}\prod_{l=1}^{L} g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{u}]_{b_l}\right\|_2\right)d\boldsymbol{u}} \leq K_{1,g}^{\boldsymbol{D}}.\end{aligned}$$

Then

$$\frac{1}{\sqrt{\det\left(\boldsymbol{D}\boldsymbol{D}^\top\right)}}\int_{\mathbb{B}_{a,\mathcal{B}}^a(r_L)\cap\mathrm{Span}(\boldsymbol{D}^\top)}\left\|\boldsymbol{u}_{b_1}\right\|_2^2\prod_{l=1}^{L} g\left(\left\|\boldsymbol{u}_{b_l}\right\|_2\right)d\boldsymbol{u} \leq K_{1,g}^{\boldsymbol{D}}T. \tag{3.33}$$

We now bound ${C_L}^{-1}$. By a change of variables, we obtain

$$\begin{aligned}{C_L}^{-1} &= \left(\int_{\Theta_{p_0^D}}\prod_{l=1}^{L}e^{-\alpha^a\left\|[\boldsymbol{D}^\top\boldsymbol{\theta}-\boldsymbol{D}^\top\boldsymbol{\theta}^*]_{b_l}\right\|_2^a} g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{\theta}-\boldsymbol{D}^\top\boldsymbol{\theta}^*]_{b_l}\right\|_2\right)d\boldsymbol{\theta}\right)^{-1} \\ &= \left(\int_{\left\{\boldsymbol{u}\in\mathbb{R}^p\,:\,\left\|\boldsymbol{D}^\top\boldsymbol{u}\right\|_{a,2}^a\leq r_L\right\}} e^{-\alpha^a\left\|\boldsymbol{D}^\top\boldsymbol{u}\right\|_{a,2}^a}\prod_{l=1}^{L} g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{u}]_{b_l}\right\|_2\right)d\boldsymbol{u}\right)^{-1} \\ &\leq e^{r_L\alpha^a}\left(\int_{\left\{\boldsymbol{u}\in\mathbb{R}^p\,:\,\left\|\boldsymbol{D}^\top\boldsymbol{u}\right\|_{a,2}^a\leq r_L\right\}}\prod_{l=1}^{L} g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{u}]_{b_l}\right\|_2\right)d\boldsymbol{u}\right)^{-1}.\end{aligned}$$

Since $r_L = 3\sqrt{K_{a,g}^{\boldsymbol{D}}}L > \sqrt{K_{a,g}^{\boldsymbol{D}}}L$, Lemma 3.5.2 gives us

$$\begin{aligned}
{C_L}^{-1} &\leq e^{r_L\alpha^a}\left[T\left(1-\frac{1}{T}\int_{\left\{\boldsymbol{u}\in\mathbb{R}^p\,:\,\left\|\boldsymbol{D}^\top\boldsymbol{u}\right\|_{a,2}^a > r_L\right\}}\prod_{l=1}^{L} g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{u}]_{b_l}\right\|_2\right)d\boldsymbol{u}\right)\right]^{-1} \\
&\leq e^{r_L\alpha^a}T^{-1}\left(1-\frac{L^2K_{a,g}^{\boldsymbol{D}}}{(r_L-L\sqrt{K_{a,g}^{\boldsymbol{D}}})^2}\right)^{-1} \\
&= e^{r_L\alpha^a}T^{-1}\left(1-\frac{1}{4}\right)^{-1} \leq 2e^{r_L\alpha^a}T^{-1}. \qquad (3.34)
\end{aligned}$$

Combining (3.33) and (3.34), (3.32) becomes $\int_{\mathbb{R}^q}\left\|\boldsymbol{u}_{b_1}\right\|_2^2 p_0(d\boldsymbol{u}) \leq 2K_{1,g}^{\boldsymbol{D}}e^{r_L\alpha^a}$. That concludes the proof of inequality (3.27) in Lemma 3.5.1.

Next, we prove inequality (3.28). Remind that $\operatorname{supp}(\pi)=\Theta$, $\operatorname{supp}({p_0}^{\boldsymbol{D}})=\Theta_{{p_0}^{\boldsymbol{D}}}$. By (3.17), we get $\Theta_{{p_0}^{\boldsymbol{D}}}\subseteq\Theta$ implying that ${p_0}^{\boldsymbol{D}}$ is absolutely continuous w.r.t. π. So $\mathrm{KL}({p_0}^{\boldsymbol{D}},\pi)<+\infty$ which can be bounded. The bound in (3.28) can be proved as follows. By Lemma 2.6.3, we have

$$\begin{aligned}
\mathrm{KL}({p_0}^{\boldsymbol{D}},\pi) &= \int_{\mathbb{R}^p}\log\left(\frac{{p_0}^{\boldsymbol{D}}(d\boldsymbol{\theta})}{\pi(d\boldsymbol{\theta})}\right){p_0}^{\boldsymbol{D}}(d\boldsymbol{\theta}) \\
&= \int_{\mathbb{R}^p}\log\left(\frac{C_{\alpha,g,R}}{C_L}\frac{\prod_{l=1}^{L}e^{-\alpha^a\left\|[\boldsymbol{D}^\top\boldsymbol{\theta}-\boldsymbol{D}^\top\boldsymbol{\theta}^*]_{b_l}\right\|_2^a}g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{\theta}-\boldsymbol{D}^\top\boldsymbol{\theta}^*]_{b_l}\right\|_2\right)}{\prod_{l=1}^{L}e^{-\alpha^a\left\|[\boldsymbol{D}^\top\boldsymbol{\theta}]_{b_l}\right\|_2^a}g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{\theta}]_{b_l}\right\|_2\right)}\right){p_0}^{\boldsymbol{D}}(d\boldsymbol{\theta}) \\
&= \int_{\mathbb{R}^q}\log\left(\frac{C_{\alpha,g,R}}{C_L}\prod_{l=1}^{L}\frac{e^{\alpha^a\left\|\boldsymbol{t}_{b_l}\right\|_2^a}g\left(\left\|\boldsymbol{t}_{b_l}-\boldsymbol{t}^*_{b_l}\right\|_2\right)}{e^{\alpha^a\left\|\boldsymbol{t}_{b_l}-\boldsymbol{t}^*_{b_l}\right\|_2^a}g\left(\left\|\boldsymbol{t}_{b_l}\right\|_2\right)}\right)p_0(d\boldsymbol{t}) \\
&= \log\left(\frac{C_{\alpha,g,R}}{C_L}\right)+\alpha^a\sum_{l=1}^{L}\int_{\mathbb{R}^q}\left[\left\|\boldsymbol{t}_{b_l}\right\|_2^a-\left\|\boldsymbol{t}_{b_l}-\boldsymbol{t}^*_{b_l}\right\|_2^a\right]p_0(d\boldsymbol{t}) \\
&\quad+\sum_{l=1}^{L}\int_{\mathbb{R}^q}\log\left(\frac{g\left(\left\|\boldsymbol{t}_{b_l}-\boldsymbol{t}^*_{b_l}\right\|_2\right)}{g\left(\left\|\boldsymbol{t}_{b_l}\right\|_2\right)}\right)p_0(d\boldsymbol{t}),
\end{aligned}$$

where p_0 is a probability measure in $\mathbb{R}^q$ defined in (3.19). We know that $\boldsymbol{t}^*=\boldsymbol{D}^\top\boldsymbol{\theta}^*$, according to the fact that $\left\|\boldsymbol{t}_{b_l}\right\|_2^a-\left\|\boldsymbol{t}_{b_l}-\boldsymbol{t}^*_{b_l}\right\|_2^a\leq\left\|\boldsymbol{t}^*_{b_l}\right\|_2^a$ and Assumption (G.4), we get

$$\mathrm{KL}({p_0}^{\boldsymbol{D}},\pi)\leq\log\left(\frac{C_{\alpha,g,R}}{C_L}\right)+\alpha^a\left\|\boldsymbol{D}^\top\boldsymbol{\theta}^*\right\|_{a,2}^a+\lambda\sum_{l=1}^{L}\log\left\{h\left(\left\|\left[\boldsymbol{D}^\top\boldsymbol{\theta}^*\right]_{b_l}\right\|_2\right)\right\}. \qquad (3.35)$$

Now, it remains to bound $\log(C_{\alpha,g,R}/C_L)$. Remind that $C_{\alpha,g,R}$ is the normalization factor of π, and thus

$$C_{\alpha,g,R}=\int_\Theta\prod_{l=1}^{L}\exp\left(-\alpha^a\left\|[\boldsymbol{D}^\top\boldsymbol{\theta}]_{b_l}\right\|_2^a\right)g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{\theta}]_{b_l}\right\|_2\right)d\boldsymbol{\theta}\leq\int_{\mathbb{R}^p}\prod_{l=1}^{L}g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{\theta}]_{b_l}\right\|_2\right)d\boldsymbol{\theta}=T.$$

Combining this with the bound of ${C_L}^{-1}$ in (3.34), we obtain

$$\log\left(\frac{C_{\alpha,g,R}}{C_L}\right)\leq r_L\alpha^a+\log(2)\leq 1+r_L\alpha^a. \qquad (3.36)$$

Inserting (3.36) into (3.35), we get inequality (3.28). This completes the proof. □

Chapter 4

Sharp Oracle Inequalities for Low-complexity Priors

Main contributions of this chapter

- Show in Theorem 4.3.1 and Theorem 4.3.3 that the EWA in (1.12) and the variational/penalized estimator in (1.9) satisfy (deterministic) sharp oracle inequalities for prediction with optimal remainder terms, for a general class of data losses and penalties.
- Etablish oracle inequalities in probability (see Section 4.3.3) for random observations.
- Refine these results for the standard linear regression with Gaussian or sub-Gaussian noise, and a quadratic loss, and specialize them to the popular penalties in the literature (see Section 4.4).
- Discuss minimax optimality of the obtained bounds.

The results in this chapter can be found in [101].

Contents

4.1 Introduction

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

4.1.1 Problem statement

Let $\boldsymbol{y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)^\top \in \mathbb{R}^n$ i.i.d. observations drawn from a general regression problem in high dimension. By the aggregation approach, the regression function f is approximated by $f_{\boldsymbol{\theta}_0}$ (see Section 1.1.1) with

$$\boldsymbol{\theta}_0 \in \operatorname*{Argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathbb{E}\left[F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y})\right].$$

The loss function F is supposed to be smooth and convex. Our goal is to provide general oracle inequalities in prediction for two estimators of $\boldsymbol{\theta}_0$: the penalized estimator (1.9) and exponential weighted aggregation (1.12) in which $J_{\boldsymbol{\lambda}_n} = \lambda_n J$ with $\lambda_n > 0$ is the regularization parameter, and J is a proper closed convex function that promotes some specific notion of simplicity/low-complexity.

4.1.2 Chapter organization

Section 4.2 states our main assumptions on the data loss and the prior penalty. In Section 4.3, we prove our main oracle inequalities, and their versions in probability. We then tackle the case of linear

regression with quadratic data loss in Section 4.4. A key intermediate result in the proof of our main results is established in Section 4.5 with an elegant argument relying on Moreau-Yosida regularization.

4.2 Estimation with low-complexity penalties

The estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ in (1.9) and (1.12) 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.

4.2.1 Data loss

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

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

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

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

(H.2) For any $\overline{\boldsymbol{\theta}} \in \mathbb{R}^p$ and $\boldsymbol{y} \in \mathbb{R}^n$,

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

Recall that by Lemma 2.3.8, the monotone conjugate φ^+ of φ is a proper, closed, convex, strongly coercive and non-decreasing function on $\mathbb{R}_+$ that vanishes at 0. Moreover, $\varphi^{++} = \varphi$. φ^+ is finite-valued on $\mathbb{R}_+$ if φ is strongly coercive, and it vanishes only at 0 under e.g. Lemma 2.3.8(iii).

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

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

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

Example 4.2.1. Consider the case where[1]

$$\varphi(t) = t^q/q, \quad q \in]1, +\infty[,$$

or equivalently

$$\varphi^+(t) = t^{q_*}/q_*, \quad \text{where} \quad 1/q + 1/q_* = 1.$$

For $q = q_* = 2$, **(H.1)** amounts to saying that $F(\cdot, \boldsymbol{y})$ is strongly convex for all $\boldsymbol{y}$. In particular, [7, Proposition 10.13] shows that $F(\boldsymbol{u}, \boldsymbol{y}) = \|\boldsymbol{u} - \boldsymbol{y}\|_2^q/q$ is uniformly convex for $q \in [2, +\infty[$ with modulus $\varphi(t) = C_q t^q/q$, where $C_q > 0$ is a constant that depends solely on q.

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

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

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

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

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

4.2.2 Prior penalty

Recall the main definitions and results from convex analysis that are collected in Section 2.3. Our main assumption on J is the following.

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

By Lemma 2.3.5, 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 2.3.13 and Lemma 2.3.10, the polar gauge $J^\circ \overset{\text{def}}{=} \gamma_{\mathcal{C}^\circ}$ enjoys the same properties as J in **(H.3)**.

4.3 Oracle inequalities for a general loss

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

We recall $T_{\boldsymbol{\theta}}$ and $e_{\boldsymbol{\theta}}$ the model subspace and vector associated to $\boldsymbol{\theta}$ (see Definition 2.4.1). Denote $S_{\boldsymbol{\theta}} = T_{\boldsymbol{\theta}}^{\perp}$. Given two coercive finite-valued gauges $J_1 = \gamma_{\mathcal{C}_1}$ and $J_2 = \gamma_{\mathcal{C}_2}$, and a linear operator $\boldsymbol{A}$, we define $\|\boldsymbol{A}\|_{J_1 \to J_2}$ the *operator bound* as

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

Note that $\|\boldsymbol{A}\|_{J_1 \to J_2}$ is bounded (this follows from Lemma 2.3.5(v)). Furthermore, we have from Lemma 2.3.13 that

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

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

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

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

This generalizes the compatibility factor introduced in [150] for the Lasso (and used in [47]). 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 (1.9); see [67, 145, 147, 144, 146] (see also Chapter 5). One can see in particular that $\Upsilon(T, c)$ is larger than the smallest singular value of $\boldsymbol{X}_T$.

The oracle inequalites will provided in terms of the loss

$$R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \frac{1}{n} D_F\left(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{X}\boldsymbol{\theta}_0\right).$$

where we recall that D_F is the Bregman distance associated to F at $\boldsymbol{X}\boldsymbol{\theta}_0$.

4.3.1 Oracle inequality for $\widehat{\boldsymbol{\theta}}_n^{\text{EWA}}$

We are now ready to establish our first main result: an oracle inequality for the EWA in (1.12).

Theorem 4.3.1. *Consider the EWA $\widehat{\boldsymbol{\theta}}_n^{\text{EWA}}$ in (1.12), where F and J satisfy Assumptions **(H.1)**-**(H.2)** and **(H.3)**. Then, for any $\tau > 1$ such that*

$$\lambda_n \geq \frac{\tau J^\circ\left(-\boldsymbol{X}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\right)}{n},$$

the following holds,

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

Remark 4.3.2.

1. It should be emphasized that Theorem 4.3.1 is actually a deterministic statement for a fixed choice of λ_n. Probabilistic analysis will be required when the result is applied to particular statistical models as we will see later. For this, we will use concentration inequalities in order to provide bounds that hold with high probability over the data.
2. The oracle inequality is sharp. The remainder in it has two terms. The first one encodes the complexity of the model promoted by J. The second one, $p\beta$, captures the influence of the temperature parameter. In particular, taking β sufficiently small of the order $O\left((pn)^{-1}\right)$, this term becomes $O(n^{-1})$.
3. When $\varphi(t) = \nu t^2/2$, i.e. $F(\cdot, \boldsymbol{y})$ is ν-strongly convex, then $\varphi^+(t) = t^2/(2\nu)$, and the reminder term becomes
$$\frac{\lambda_n^2\big(\tau J^\circ(e_{\boldsymbol{\theta}})+1\big)^2\big\| \mathrm{P}_{T_{\boldsymbol{\theta}}} \big\|_{2\to J}^2}{2\tau^2\nu \Upsilon\left(T_{\boldsymbol{\theta}}, \frac{\tau J^\circ(e_{\boldsymbol{\theta}})+1}{\tau-1}\right)^2}. \tag{4.3}$$
If, moreover, ∇F is also κ-Lipschitz continuous, then it can be shown that $R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big)$ is equivalent to a quadratic loss. This means that the oracle inequality in Theorem 4.3.1 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.
4. If J is such that $e_{\boldsymbol{\theta}} \in \partial J(\boldsymbol{\theta}) \subset \mathcal{C}^\circ$ (typically for a strong gauge by (2.5)), 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).
5. A close inspection of the proof of Theorem 4.3.1 reveals that the term $p\beta$ can be improved to the smaller bound
$$p\beta + \Big(V_n(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}) - \mathbb{E}_{\widehat{\mu}_n}\left[V_n(\boldsymbol{\theta})\right]\Big),$$
where the upper-bound is a consequence of Jensen inequality.

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

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

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

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

This holds for any $\boldsymbol{\eta} \in \partial V_n(\boldsymbol{\theta})$, and in particular at the minimal selection $\big(\partial V_n(\boldsymbol{\theta})\big)^0$ (see Section 4.5 for details). It then follows from the pillar result in Proposition 4.5.2[2] that

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

We thus deduce the inequality

$$V_n(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}) - V_n(\overline{\boldsymbol{\theta}}) \leq p\beta - \frac{1}{n}\varphi\big(\big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\overline{\boldsymbol{\theta}}\big\|_2\big), \quad \forall \overline{\boldsymbol{\theta}} \in \mathbb{R}^p. \tag{4.4}$$

[2] In the appendix, we provide a self-contained proof based on a novel Moreau-Yosida regularization argument. In [47, Corollary 1 and 2], an alternative proof is given using an absolute continuity argument since $\widehat{\mu}_n$ is locally Lipschitz, hence a Sobolev function.

By definition of the Bregman divergence, we have

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

By virtue of the duality inequality (2.4), we have

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

Denote $\boldsymbol{\omega} = \widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{\theta}$. By virtue of **(H.3)**, Corollary 2.4.3 and (2.4), we obtain

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

This inequality together with (4.4) (applied with $\overline{\boldsymbol{\theta}} = \boldsymbol{\theta}$) and (4.1) yield

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

where we applied Fenchel-Young inequality (2.3) to get the last bound. Taking the infimum over $\boldsymbol{\theta} \in \mathbb{R}^p$ yields the desired result. □

Stratifiable functions Theorem 4.3.1 has a nice instanciation when $\mathbb{R}^p$ can be partitioned into a collection of subsets $\{\mathcal{M}_i\}_i$ that form a stratification of $\mathbb{R}^p$. That is, $\mathbb{R}^p$ is a finite disjoint union $\cup_i \mathcal{M}_i$ such that the partitioning sets $\mathcal{M}_i$ (called strata) fit nicely together and the stratification is endowed with a partial ordering for the closure operation. For example, it is known that a polyhedral function has a polyhedral stratification, and more generally, semialgebraic functions induce stratifications into finite disjoint unions of manifolds; see, e.g., [42]. Another example is that of partly smooth convex functions thoroughly studied in [145, 147, 144, 146] for various statistical and inverse problems. These functions induce a stratification into strata that are C^2-smooth submanifolds of $\mathbb{R}^p$. In turns out that all popular penalty functions discussed in this chapter are partly smooth (see [144, 146]). Let's denote $\mathcal{M}$ the set of strata associated to J. With this notation at hand, the oracle inequality (4.2) now reads

$$R_n(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0) \le \inf_{\substack{\mathcal{M}\in\mathcal{M} \\ \boldsymbol{\theta}\in\mathcal{M}}} \left(R_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) + \frac{1}{n}\varphi^+\left(\frac{\lambda_n\sqrt{n}\big(\tau J^\circ(e_{\boldsymbol{\theta}}) + 1\big)\|\mathrm{P}_{T_{\boldsymbol{\theta}}}\|_{2\to J}}{\tau\Upsilon\left(T_{\boldsymbol{\theta}}, \frac{\tau J^\circ(e_{\boldsymbol{\theta}})+1}{\tau-1}\right)}\right)\right) + p\beta. \tag{4.5}$$

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

The next result establishes that $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ satisfies a sharp prediction oracle inequality that we will compare to (4.2).

Theorem 4.3.3. *Consider the penalized estimator* $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ *in* (1.9)*, where* F *and* J *satisfy Assumptions* **(H.1)** *and* **(H.3)**. *Then, for any* $\tau > 1$ *such that*

$$\lambda_n \geq \frac{\tau J^\circ\left(-\boldsymbol{X}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\right)}{n},$$

the following holds,

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}, \boldsymbol{\theta}_0\big) \leq \inf_{\boldsymbol{\theta}\in\mathbb{R}^p} \left(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \frac{1}{n}\varphi^+ \left(\frac{\lambda_n\sqrt{n}\big(\tau J^\circ(e_{\boldsymbol{\theta}}) + 1\big)\big\| \mathrm{P}_{T_{\boldsymbol{\theta}}} \big\|_{2\to J}}{\tau \Upsilon\left(T_{\boldsymbol{\theta}}, \frac{\tau J^\circ(e_{\boldsymbol{\theta}})+1}{\tau-1}\right)} \right) \right). \tag{4.6}$$

Proof. The proof follows the same lines as that of Theorem 4.3.1 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 $\boldsymbol{\theta} \in \mathbb{R}^p$

$$V_n(\boldsymbol{\theta}) \geq V_n(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}) + \frac{1}{n}\varphi\big(\big\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}\big\|_2\big). \tag{4.7}$$

Continuing exactly as just after (4.4), replacing $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ with $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ and invoking (4.7) instead of (4.4), we arrive at the claimed result. □

Remark 4.3.4.

1. Observe that the penalized estimator $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ does not require the moment assumption **(H.2)** for (4.6) to hold. The convexity assumption on φ in **(H.1)**, which was important to apply Jensen's inequality in the proof of (4.2), is not needed either to get (4.6).
2. As we remarked for Theorem 4.3.1, Theorem 4.3.3 is also a deterministic statement for a fixed choice of λ_n that holds for any minimizer $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, which is not unique in general. The condition on λ_n is similar to the one in [108] where authors established different guarantees for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$.

One clearly sees that the difference between the prediction performance of $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ lies in the term $p\beta$ (or rather its lower-bound in Remark 4.3.2-5). Thus letting $\beta \to 0$ in (4.2), one recovers the oracle inequality (4.6) of penalized estimators. In particular, for $\beta = O\left((pn)^{-1}\right)$, this is on the order $O(n^{-1})$.

4.3.3 Oracle inequalities in probability

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

(H.4) $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n)$ are independent and identically distributed observations, and

$$F(\boldsymbol{u}, \boldsymbol{y}) = \sum_{i=1}^n f_i(\boldsymbol{u}_i, \boldsymbol{y}_i), \quad f_i : \mathbb{R}\times\mathbb{R} \to \mathbb{R}.$$

Moreover,

(i) $\mathbb{E}\left[|f_i((\boldsymbol{X}\boldsymbol{\theta}_0)_i, \boldsymbol{y}_i)|\right] < +\infty$, $\forall i \in \{1, \ldots, n\}$;

(ii) $|f_i'((\boldsymbol{X}\boldsymbol{\theta}_0)_i, t)| \leq g(t)$, where $\mathbb{E}\left[g(\boldsymbol{y}_i)\right] < +\infty$, $\forall i \in \{1, \ldots, n\}$;

(iii) Bernstein moment condition: $\forall 1 \leq i \leq n$ and all integers $m \geq 2$,

$$\mathbb{E}\left[\left|f_i'((\boldsymbol{X}\boldsymbol{\theta}_0)_i, \boldsymbol{y}_i)\right|^m\right] \leq m!\kappa^{m-2}\sigma_i^2/2$$

for some constants $\kappa > 0$, $\sigma_i > 0$ independent of n.

Observe that under **(H.4)**, and by virtue of Lemma 2.3.13(iv) and [80, Proposition V.3.3.4], we have

$$J^\circ\big(-\boldsymbol{X}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\big) = \sigma_{\mathcal{C}}\big(-\boldsymbol{X}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})\big) = \sup_{\boldsymbol{z}\in\boldsymbol{X}(\mathcal{C})} -\sum_{i=1}^n f_i'((\boldsymbol{X}\boldsymbol{\theta}_0)_i, \boldsymbol{y}_i)\boldsymbol{z}_i. \tag{4.8}$$

Thus, checking the event $\mathcal{E}$ amounts to establishing a deviation inequality for the supremum of an empirical process[3] above its mean under the weak Bernstein moment condition **(H.4)**(iii), which essentially requires that the $f_i'((\boldsymbol{X}\boldsymbol{\theta}_0)_i, \boldsymbol{y}_i)$ have sub-exponential tails. We will first tackle the case where $\mathcal{C}$ is the convex hull of a finite set (i.e. $\mathcal{C}$ is a polytope).

4.3.3.1 Polyhedral penalty

We here suppose that J is a finite-valued gauge of $\mathcal{C} = \overline{\mathrm{conv}}\,(\mathcal{V})$, where $\mathcal{V}$ is finite, i.e. $\mathcal{C}$ is a polytope with vertices [125, Corollary 19.1.1]. Our first oracle inequality in probability is the following.

Proposition 4.3.5. *Consider the estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, where F and $J \overset{def}{=} \gamma_{\mathcal{C}}$ satisfy Assumptions **(H.1)**, **(H.2)**, **(H.3)** and **(H.4)**, and $\mathcal{C}$ is a polytope with vertices $\mathcal{V}$. Suppose that $\mathrm{rank}(\boldsymbol{X}) = n$ and $\max_{\boldsymbol{v}\in\mathcal{V}} \|\boldsymbol{X}\boldsymbol{v}\|_\infty \leq 1$, and take*

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

for some $\tau > 1$ and $\delta > 1$. Then (4.2) and (4.6) hold with probability at least $1 - 2|\mathcal{V}|^{1-\delta}$.

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

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

Moreover, (4.8) specializes to

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

Let $t = \lambda_n n/\tau$. By the union bound and (4.8), we have

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

The random variables $\left(f_i'((\boldsymbol{X}\boldsymbol{\theta}_0)_i, \boldsymbol{y}_i)\boldsymbol{z}_i\right)_i$ are zero-mean independent, and $\forall i$ and $m \geq 2$

$$\mathbb{E}\left[\left|f_i'((\boldsymbol{X}\boldsymbol{\theta}_0)_i, \boldsymbol{y}_i)\boldsymbol{z}_i\right|^m\right] \leq |\boldsymbol{z}_i|^m m!\kappa^{m-2}\sigma_i^2/2 \leq \max_{\boldsymbol{v}\in\mathcal{V}}\|\boldsymbol{X}\boldsymbol{v}\|_\infty^m\, m!\kappa^{m-2}\sigma_i^2/2 \leq m!\kappa^{m-2}\sigma_i^2/2.$$

We are then in position to apply the Bernstein inequality to get

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

where $\sigma^2 = \max_{1\leq i\leq n}\sigma_i^2$. Every t such that

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

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

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

Remark 4.3.6. In the monograph [19, Lemma 14.12], the authors derived an exponential deviation inequality for the supremum of an empirical process with finite $\mathcal{V}$ and possibly unbounded empirical processes under a Bernstein moment condition similar to ours (in fact ours implies theirs). The very last part of our proof can be obtained by applying their result. We detailed it here for the sake of completeness.

Lasso To lighten the notation, let $I_{\boldsymbol{\theta}} = \operatorname{supp}(\boldsymbol{\theta})$. From (2.7), it is easy to see that

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

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

$$|\mathcal{V}| = 2p \quad \text{and} \quad \max_{\boldsymbol{v}\in\mathcal{V}} \|\boldsymbol{X}\boldsymbol{v}\|_2 = \max_{1\leq i\leq p} \|\boldsymbol{X}_i\|_2 .$$

Inserting this into Proposition 4.3.5, we obtain the following corollary.

Corollary 4.3.7. *Consider the estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, where where J is the Lasso penalty and F satisfies Assumptions* **(H.1)**, **(H.2)** *and* **(H.4)**. *Suppose that* $\operatorname{rank}(\boldsymbol{X}) = n$ *and* $\max_i \|\boldsymbol{X}_i\|_\infty \leq 1$, *and take*

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

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

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\big) \leq \inf_{\substack{I\subset\{1,\dots,p\} \\ \boldsymbol{\theta}:\ \operatorname{supp}(\boldsymbol{\theta})=I}} \left(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \frac{1}{n}\varphi^+\left(\frac{\lambda_n\sqrt{n}\,(\tau+1)\,\sqrt{|I|}}{\tau\Upsilon\left(\operatorname{Span}\{\boldsymbol{a}_i\}_{i\in I}, \frac{\tau+1}{\tau-1}\right)}\right)\right) + p\beta, \tag{4.9}$$

and

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}, \boldsymbol{\theta}_0\big) \leq \inf_{\substack{I\subset\{1,\dots,p\} \\ \boldsymbol{\theta}:\ \operatorname{supp}(\boldsymbol{\theta})=I}} \left(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \frac{1}{n}\varphi^+\left(\frac{\lambda_n\sqrt{n}\,(\tau+1)\,\sqrt{|I|}}{\tau\Upsilon\left(\operatorname{Span}\{\boldsymbol{a}_i\}_{i\in I}, \frac{\tau+1}{\tau-1}\right)}\right)\right). \tag{4.10}$$

For $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, we recover a similar scaling for λ_n and the oracle inequality as in [148], though in the latter the oracle inequality is not sharp unlike ours. Note that the above oracle inequality extends readily to the case of analysis/fused Lasso $\left\|\boldsymbol{D}^\top \cdot\right\|_1$ where $\boldsymbol{D}$ is surjective. We leave the details to the interested reader (see also the analysis-group Lasso example in Section 4.4).

Anti-sparsity From Section 2.4.3.4, recall the saturation support $I_{\boldsymbol{\theta}}^{\mathrm{sat}}$ of $\boldsymbol{\theta}$. From (2.14), we get

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

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

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

We have the following oracle inequalities.

Corollary 4.3.8. *Consider the estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, where where J is anti-sparsity penalty* (2.13), *and F satisfies Assumptions* **(H.1)**, **(H.2)** *and* **(H.4)**. *Suppose that* $\operatorname{rank}(\boldsymbol{X}) = n$ *and* $\max_{i,j} |\boldsymbol{X}_{i,j}| \leq 1/p$, *and take*

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

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

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\big) \leq \inf_{\substack{I \subset \{1,\dots,p\} \\ \boldsymbol{\theta}:\ I_{\boldsymbol{\theta}}^{\mathrm{sat}} = I}} \left(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \frac{1}{n} \varphi^+ \left(\frac{\lambda_n \sqrt{n}\,(\tau+1)}{\tau \Upsilon\left(\left\{ \overline{\boldsymbol{\theta}} \ :\ \overline{\boldsymbol{\theta}}_I \in \mathbb{R}\,\mathrm{sgn}(\boldsymbol{\theta}_I) \right\}, \frac{\tau+1}{\tau-1} \right)} \right) \right) + p\beta, \quad (4.11)$$

and

$$R_n\big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}, \boldsymbol{\theta}_0\big) \leq \inf_{\substack{I \subset \{1,\dots,p\} \\ \boldsymbol{\theta}:\ I_{\boldsymbol{\theta}}^{\mathrm{sat}} = I}} \left(R_n\big(\boldsymbol{\theta}, \boldsymbol{\theta}_0\big) + \frac{1}{n} \varphi^+ \left(\frac{\lambda_n \sqrt{n}\,(\tau+1)}{\tau \Upsilon\left(\left\{ \overline{\boldsymbol{\theta}} \ :\ \overline{\boldsymbol{\theta}}_I \in \mathbb{R}\,\mathrm{sgn}(\boldsymbol{\theta}_I) \right\}, \frac{\tau+1}{\tau-1} \right)} \right) \right). \quad (4.12)$$

We are not aware of any result of this kind in the literature. The bound imposed on $\boldsymbol{X}$ is similar to what is generally assumed in the vector quantization literature [102, 133].

4.3.3.2 General penalty

Extending the above reasoning to a general penalty requires a deviation inequality for the supremum of an empirical process in (4.8) under the Bernstein moment condition **(H.4)**(iii), but without the need of uniform boundedness. This can be achieved via generic chaining along a tree using entropy with bracketing; see [151, Theorem 8]. The resulting deviation bound will thus depend on the entropies with bracketing. These quantities capture the complexity of the set $\boldsymbol{X}(\mathcal{C})$ but are intricate to compute in general. This subject deserves further investigation that we leave to a future work.

Remark 4.3.9 (Group Lasso). Using the union bound, we have

$$\mathbb{P}\left[\max_{i \in \{1,\dots,L\}} \left\| \boldsymbol{X}_{b_i}^\top \boldsymbol{\xi} \right\|_2 \geq \lambda_n n/\tau \right] \leq \sum_{i=1}^{L} \mathbb{P}\left[\left\| \boldsymbol{X}_{b_i}^\top \boldsymbol{\xi} \right\|_2 \geq \lambda_n n/\tau \right].$$

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

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

then one can use [8, Theorem 3]. Indeed, assume the nroamlization $\max_i \left\| \boldsymbol{X}_{b_i}^\top \boldsymbol{X}_{b_i} \right\|_{2\to 2} \leq n$, which entails

$$\mathbb{E}\left[\left\| \boldsymbol{X}_{b_i}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y}) \right\|_2 \right] \leq \mathbb{E}\left[\left\| \boldsymbol{X}_{b_i}^\top \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y}) \right\|_2^2 \right]^{1/2} \leq \sigma \sqrt{Kn/2}.$$

It then follows that taking

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

the oracle inequalities (4.17) and (4.18) hold for the group Lasso with probability at least $1 - L^{1-\delta}$. A similar result can be proved for the analysis-group Lasso just as well with a proper normalization assumption on $\boldsymbol{X}$ (see Section 4.4.3.3).

4.4 Oracle inequalities for low-complexity linear regression

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

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_0 + \boldsymbol{\xi}, \quad (4.13)$$

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

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

$$\mathbb{P}\left[|\langle \boldsymbol{\xi}, \boldsymbol{z} \rangle| \geq t\right] \leq 2e^{-t^2/(2\left\|\boldsymbol{z}\right\|_2^2 \sigma^2)}, \quad \forall \boldsymbol{z} \in \mathbb{R}^n. \tag{4.14}$$

In this case, the bounds of Section 4.3.3 can be improved.

4.4.1 General penalty

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

Definition 4.4.1. The Gaussian width of a subset $\mathcal{S} \subset \mathbb{R}^n$ is defined as

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

The concept of Gaussian width has appeared in the literature in different contexts. In particular, it has been used to establish sample complexity bounds to ensure exact recovery (noiseless case) and mean-square estimation stability (noisy case) for low-complexity penalized estimators from Gaussian measurements; see e.g. [127, 37, 139, 152, 146].

The Gaussian width has deep connections to convex geometry and it enjoys many useful properties. It is well-known that it is positively homogeneous, monotonic w.r.t. inclusion, and invariant under orthogonal transformations. Moreover, $w(\overline{\text{conv}}\,(\mathcal{S})) = w(\mathcal{S})$. From Lemma 2.3.10(ii)-(iii), $w(\mathcal{S})$ is a non-negative finite quantity whenever the set $\mathcal{S}$ is bounded and contains the origin.

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

Proposition 4.4.2. *Let the data generated by (4.13) where $\boldsymbol{\xi}$ is a zero-mean sub-Gaussian random vector with parameter σ. Consider the estimators $\widehat{\boldsymbol{\theta}}_n^{\text{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\text{PEN}}$, where F and $J \overset{def}{=} \gamma_{\mathcal{C}}$ satisfy Assumptions* **(H.1)**-**(H.2)** *and* **(H.3)**. *Suppose that*

$$\lambda_n \geq \frac{\tau \sigma c_1 \sqrt{2\log(c_2/\delta)}\, w\left(\boldsymbol{X}(\mathcal{C})\right)}{n},$$

for some $\tau > 1$ and $0 < \delta < \min(c_2, 1)$, where c_1 and c_2 are positive absolute constants. Then with probability at least $1 - \delta$, (4.2) and (4.6) hold with the remainder term given by (4.3) with $\nu = 1$.

The proof requires sophisticated ideas from the theory of generic chaining [136], but we only apply these results. The constants c_1 and c_2 can be traced back to the proof of these results as detailed in [136].

Proof. First, from (4.14), we have the bound

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

i.e. the increment condition [136, (0.4)] is verified. Thus combining (4.8) with the probability bound in [136, page 11], the generic chaining theorem [136, Theorem 1.2.6] and the majorizing measure theorem [136, Theorem 2.1.1], we have

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

□

If the noise is Gaussian, an enhanced version can be proved by invoking Gaussian concentration of Lipschitz functions [93].

Proposition 4.4.3. *Let the data generated by (4.13) with noise $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. Consider the estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, where F and $J \overset{def}{=} \gamma_{\mathcal{C}}$ satisfy Assumptions (H.1)-(H.2) and (H.3). Suppose that*

$$\lambda_n \geq \frac{(1+\delta)\tau\sigma w\left(\boldsymbol{X}(\mathcal{C})\right)}{n},$$

for some $\tau > 1$ and $\delta > 0$. Then with probability at least $1 - \exp\left(-\frac{\delta^2 w(\boldsymbol{X}(\mathcal{C}))^2}{2\|\boldsymbol{X}\|_{J\to 2}^2}\right)$, (4.2) and (4.6) hold with the remainder term given by (4.3) with $\nu = 1$.

Proof. Thanks to sublinearity (see Lemma 2.3.5(i) and Lemma 2.3.13), the function $\boldsymbol{\xi} \mapsto J^\circ(\boldsymbol{X}^\top \boldsymbol{\xi})$ is Lipschitz continuous with Lipschitz constant $\|\boldsymbol{X}^\top\|_{2\to J^\circ} = \|\boldsymbol{X}\|_{J\to 2}$. From (4.8), we also have

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

Observe that $\boldsymbol{X}(\mathcal{C})$ is a convex compact set containing the origin. Setting $\epsilon = \lambda_n n/\tau - \sigma w\left(\boldsymbol{X}(\mathcal{C})\right) \geq \delta\sigma w\left(\boldsymbol{X}(\mathcal{C})\right)$, it follows from (4.8) and the Gaussian concentration of Lipschitz functions [93] that

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

□

Estimating theoretically the Gaussian width of a set[4] is a non-trivial problem that has been extensively studied in the areas of probability in Banach spaces and stochastic processes. There are classical bounds on the Gaussian width (Sudakov's and Dudley's inequalities), but they are difficult to estimate in most cases and neither of these bounds is tight for all sets. When the set is a convex cone (intersected with a sphere), tractable estimates based on polarity arguments were proposed in, e.g., [37].

4.4.2 Polyhedral penalty

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

Proposition 4.4.4. *Let the data generated by (4.13) where $\boldsymbol{\xi}$ is a zero-mean sub-Gaussian random vector with parameter σ. Consider the estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, where F and $J \overset{def}{=} \gamma_{\mathcal{C}}$ satisfy Assumptions (H.1)-(H.2) and (H.3), and moreover $\mathcal{C}$ is a polytope with vertices $\mathcal{V}$. Suppose that*

$$\lambda_n \geq \frac{\tau\sigma\left(\max_{\boldsymbol{v}\in\mathcal{V}} \|\boldsymbol{X}\boldsymbol{v}\|_2\right)\sqrt{2\delta\log(|\mathcal{V}|)}}{n},$$

for some $\tau > 1$ and $\delta > 1$. Then with probability at least $1 - 2|\mathcal{V}|^{1-\delta}$, (4.2) and (4.6) hold with the remainder term given by (4.3) with $\nu = 1$.

In particular, if $\max_{\boldsymbol{v}\in\mathcal{V}} \|\boldsymbol{X}\boldsymbol{v}\|_2 \leq \sqrt{n}$, then one can take

$$\lambda_n \geq \tau\sigma\sqrt{\frac{2\delta\log(|\mathcal{V}|)}{n}}.$$

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

Proof. From (4.8) we have

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

where in the last inequality, we used the fact that a convex function attains its maximum on $\mathcal{C}$ at an extreme point $\mathcal{V}$. Let $\epsilon = \sigma\big(\max_{\boldsymbol{v}\in\mathcal{V}} \|\boldsymbol{X}\boldsymbol{v}\|_2\big)\sqrt{2\delta\log(|\mathcal{V}|)}$. By the union bound, (4.14) and (4.8), we have

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

□

4.4.3 Applications

In this section, we exemplify our oracle inequalities for the penalties described in Section 2.4.3.

4.4.3.1 Lasso

Recall the derivations for the Lasso in Section 4.3.3.1. We obtain the following corollary of Proposition 4.4.4.

Corollary 4.4.5. *Let the data generated by (4.13) where $\boldsymbol{\xi}$ is a zero-mean sub-Gaussian random vector with parameter σ. Assume that $\boldsymbol{X}$ is such that $\max_i \|\boldsymbol{X}_i\|_2 \leq \sqrt{n}$. Consider the estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, where J is the Lasso penalty (2.6) and F satisfies Assumptions* **(H.1)**-**(H.2)**. *Suppose that*

$$\lambda_n \geq \tau\sigma\sqrt{\frac{2\delta\log(2p)}{n}},$$

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

$$\frac{1}{n}\big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\boldsymbol{\theta}_0\big\|_2^2 \leq \inf_{\substack{I\subset\{1,\dots,p\} \\ \boldsymbol{\theta}:\ \mathrm{supp}(\boldsymbol{\theta})=I}} \left(\frac{1}{n}\big\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\big\|_2^2 + \frac{\lambda_n^2\,(\tau+1)^2\,|I|}{\tau^2\Upsilon\left(\mathrm{Span}\{\boldsymbol{a}_i\}_{i\in I}, \frac{\tau+1}{\tau-1}\right)^2}\right) + p\beta, \quad (4.15)$$

and

$$\frac{1}{n}\big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\big\|_2^2 \leq \inf_{\substack{I\subset\{1,\dots,p\} \\ \boldsymbol{\theta}:\ \mathrm{supp}(\boldsymbol{\theta})=I}} \left(\frac{1}{n}\big\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\big\|_2^2 + \frac{\lambda_n^2\,(\tau+1)^2\,|I|}{\tau^2\Upsilon\left(\mathrm{Span}\{\boldsymbol{a}_i\}_{i\in I}, \frac{\tau+1}{\tau-1}\right)^2}\right). \quad (4.16)$$

The remainder term grows as $\frac{|I|\log(p)}{n}$. The oracle inequality (4.16) recovers [47, Theorem 1] in the exactly sparse case, and (4.16) the one in [135, Theorem 4] (see also [88, Theorem 11] and [48, Theorem 2]). It is worth mentioning, however, that [47, Theorem 1] handles the inexactly sparse case while we do not.

4.4.3.2 Group Lasso

Recall the notations in Section 2.4.3.2, and denote $I_{\boldsymbol{\theta}} = \mathrm{supp}_{\mathcal{B}}(\boldsymbol{\theta})$ the set indexing active blocks in $\boldsymbol{\theta}$. From (2.9), we have

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

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

We have the following oracle inequalities as corollaries of Proposition 4.4.2 and Proposition 4.4.3.

Corollary 4.4.6. *Let the data generated by (4.13). 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 (2.8) with L non-overlapping blocks of equal size K. Assume that $\boldsymbol{X}$ is such that $\max_i \left\|\boldsymbol{X}_{b_i}^\top \boldsymbol{X}_{b_i}\right\|_{2\to 2} \leq n$.*

(i) $\boldsymbol{\xi}$ is a zero-mean sub-Gaussian random vector with parameter σ: suppose that

$$\lambda_n \geq 3\tau\sigma c_1 \frac{\sqrt{2\log(c_2/\delta)}\left(\sqrt{K}+\sqrt{2\log(L)}\right)}{\sqrt{n}},$$

for some $\tau > 1$ and $0 < \delta < \min(c_2, 1)$, where c_1 and c_2 are the positive absolute constants in Proposition 4.4.2. Then, with probability at least $1-\delta$, the following holds

$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 \leq \inf_{\substack{I\subset\{1,\dots,L\} \\ \boldsymbol{\theta}:\ \mathrm{supp}_{\mathcal{B}}(\boldsymbol{\theta})=I}} \left(\frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 + \frac{\lambda_n^2 (\tau+1)^2 |I|}{\tau^2 \Upsilon\left(\mathrm{Span}\{a_j\}_{j\in b_i, i\in I}, \frac{\tau+1}{\tau-1}\right)^2} \right) + p\beta, \tag{4.17}$$

and

$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 \leq \inf_{\substack{I\subset\{1,\dots,L\} \\ \boldsymbol{\theta}:\ \mathrm{supp}_{\mathcal{B}}(\boldsymbol{\theta})=I}} \left(\frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 + \frac{\lambda_n^2 (\tau+1)^2 |I|}{\tau^2 \Upsilon\left(\mathrm{Span}\{a_j\}_{j\in b_i, i\in I}, \frac{\tau+1}{\tau-1}\right)^2} \right). \tag{4.18}$$

(ii) $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$: suppose that

$$\lambda_n \geq \tau\sigma \frac{\sqrt{K}+\sqrt{2\delta\log(L)}}{\sqrt{n}},$$

for some $\tau > 1$ and $\delta > 1$. Then, with probability at least $1 - L^{1-\delta}$, (4.17) and (4.18) hold.

The first remainder term is on the order $\frac{|I|\left(\sqrt{K}+\sqrt{2\log(L)}\right)^2}{n}$. This is similar to the scaling that has been provided in the literature for EWA with other group sparsity priors and noises [123] and Chapter 3. Similar rates were given for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ with the group Lasso in [108, 99, 149].

Proof.

(i) This is a consequence of Proposition 4.4.2, for which we need to bound

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

We first have, for any block b_i

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

Furthermore, $\left\|\boldsymbol{X}_{b_i}^\top \cdot\right\|_2$ is Lipschitz continuous with Lipschitz constant $\left\|\boldsymbol{X}_{b_i}\right\|_{2\to 2} \leq \sqrt{n}$. Thus the union bound and Gaussian concentration of Lipschitz functions [93] yield, for any $t > 0$,

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

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

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

(ii) The proof follows the lines of Proposition 4.4.3 where we additionally use the union bound. Indeed,

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

where used the Gaussian concentration of Lipschitz functions [93] in the last inequality.

□

Remark 4.4.7. We observe in passing that another way to prove the oracle inequalities in the sub-Gaussian is to use Dudley's inequality on the sphere in $\mathbb{R}^K$ after applying a union bound on the L blocks. In addition, in the Gaussian case, the (similar) bound $\lambda_n \geq 3\delta\tau\sigma\frac{\sqrt{K}+\sqrt{2\log(L)}}{\sqrt{n}}$ can be obtained by combining Proposition 4.4.3 and the estimate $w(\boldsymbol{X}(\mathcal{C})) \leq 3(\sqrt{Kn} + \sqrt{2n\log(L)})$ in the proof of (i). The corresponding probability of success would be at least $1 - L^{-9(\delta-1)^2}$.

4.4.3.3 Analysis-group Lasso

We now turn to the prior penalty (2.10). Recall the notations in Section 2.4.3.3, and remind $\Lambda_{\boldsymbol{\theta}} = \bigcup_{i\in\mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top\boldsymbol{\theta})} b_i$. We assume that $\boldsymbol{D}$ is a frame of $\mathbb{R}^p$, hence surjective, meaning that there exist $c, d > 0$ such that for any $\boldsymbol{\omega} \in \mathbb{R}^p$

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

This together with (2.11)-(2.12) and Cauchy-Schwarz inequality entail

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

Note, however, that from (2.11), we do not have in general $\left\|\boldsymbol{D}^+\mathrm{P}_{\mathrm{Ker}(\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}^c}^\top)}\boldsymbol{D}e^{\|\|_{1,2}}_{\boldsymbol{D}^\top\boldsymbol{\theta}}\right\|_{\infty,2} \leq 1$.

With exactly the same arguments to those for proving Corollary 4.4.6, replacing $\boldsymbol{X}$ by $\boldsymbol{X}\boldsymbol{D}$, we arrive at the following oracle inequalities.

Corollary 4.4.8. *Let the data generated by (4.13). 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 (2.10) with L blocks of equal size K. Assume that $\boldsymbol{D}$ is a frame, and $\boldsymbol{X}$ is such that $\max_i \left\|\boldsymbol{D}_{b_i}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{D}_{b_i}\right\|_{2\to 2} \leq n$.*

(i) $\boldsymbol{\xi}$ is a zero-mean sub-Gaussian random vector with parameter σ: suppose that

$$\lambda_n \geq 3\tau\sigma c_1 \frac{\sqrt{\log(c_2/\delta)}\left(\sqrt{K}+\sqrt{2\log(L)}\right)}{\sqrt{n}},$$

for some $\tau > 1$ and $0 < \delta < \min(c_2, 1)$, where c_1 and c_2 are the positive absolute constants in Proposition 4.4.2. Then, with probability at least $1-\delta$, the following holds

$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 \leq \inf_{\substack{I\subset\{1,\dots,L\} \\ \boldsymbol{\theta}:\ \mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top\boldsymbol{\theta})=I}} \left(\frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 + \frac{c\lambda_n^2\left(\tau\left\|\boldsymbol{D}^+ \mathrm{P}_{\mathrm{Ker}(\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}^c}^\top)} \boldsymbol{D} e^{\|\|_{1,2}}_{\boldsymbol{D}^\top\boldsymbol{\theta}}\right\|_{\infty,2}+1\right)^2 |I|}{\tau^2 \Upsilon\left(\mathrm{Ker}(\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}^c}^\top), \frac{\tau\left\|\boldsymbol{D}^+ \mathrm{P}_{\mathrm{Ker}(\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}^c}^\top)} \boldsymbol{D} e^{\|\|_{1,2}}_{\boldsymbol{D}^\top\boldsymbol{\theta}}\right\|_{\infty,2}+1}{\tau-1}\right)^2} \right) + p\beta, \tag{4.19}$$

and

$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 \leq \inf_{\substack{I\subset\{1,\dots,L\} \\ \boldsymbol{\theta}:\ \mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top\boldsymbol{\theta})=I}} \left(\frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 + \frac{c\lambda_n^2\left(\tau\left\|\boldsymbol{D}^+ \mathrm{P}_{\mathrm{Ker}(\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}^c}^\top)} \boldsymbol{D} e^{\|\|_{1,2}}_{\boldsymbol{D}^\top\boldsymbol{\theta}}\right\|_{\infty,2}+1\right)^2 |I|}{\tau^2 \Upsilon\left(\mathrm{Ker}(\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}^c}^\top), \frac{\tau\left\|\boldsymbol{D}^+ \mathrm{P}_{\mathrm{Ker}(\boldsymbol{D}_{\Lambda_{\boldsymbol{\theta}}^c}^\top)} \boldsymbol{D} e^{\|\|_{1,2}}_{\boldsymbol{D}^\top\boldsymbol{\theta}}\right\|_{\infty,2}+1}{\tau-1}\right)^2} \right). \tag{4.20}$$

(ii) $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$: suppose that

$$\lambda_n \geq \tau\sigma \frac{\sqrt{K}+\sqrt{2\delta\log(L)}}{\sqrt{n}},$$

for some $\tau > 1$ and $\delta > 1$. Then, with probability at least $1 - L^{1-\delta}$, (4.19) and (4.20) hold.

To the best of our knowledge, this result is new to the literature. The scaling of the remainder term is the same as in Chapter 3 (though with a different prior) and [123] with analysis sparsity priors different from ours (the authors in the latter also assume that $\boldsymbol{D}$ is invertible).

4.4.3.4 Anti-sparsity

Recall the derivations for the ℓ_∞ norm example in Section 4.3.3.1. We have the following oracle inequalities from Proposition 4.4.4.

Corollary 4.4.9. *Let the data generated by (4.13) where $\boldsymbol{\xi}$ is a zero-mean sub-Gaussian random vector with parameter σ. Assume that $\boldsymbol{X}$ is such that $\max_{i,j} |\boldsymbol{X}_{i,j}| \leq 1/p$. Consider the estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, where F satisfies Assumptions (H.1)-(H.2), and J is the anti-sparsity penalty (2.13). Suppose*

that

$$\lambda_n \geq \tau\sigma\sqrt{2\delta\log(2)}\sqrt{\frac{p}{n}},$$

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

$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 \leq \inf_{\substack{I\subset\{1,\dots,p\}\\ \boldsymbol{\theta}:\ I_{\boldsymbol{\theta}}^{\mathrm{sat}}=I}} \left(\frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 + \frac{\lambda_n^2\,(\tau+1)^2}{\tau^2\Upsilon\left(\left\{\overline{\boldsymbol{\theta}}\ :\ \overline{\boldsymbol{\theta}}_I \in \mathbb{R}\,\mathrm{sgn}(\boldsymbol{\theta}_I)\right\}, \frac{\tau+1}{\tau-1}\right)^2}\right) + p\beta, \tag{4.21}$$

and

$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 \leq \inf_{\substack{I\subset\{1,\dots,p\}\\ \boldsymbol{\theta}:\ I_{\boldsymbol{\theta}}^{\mathrm{sat}}=I}} \left(\frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 + \frac{\lambda_n^2\,(\tau+1)^2}{\tau^2\Upsilon\left(\left\{\overline{\boldsymbol{\theta}}\ :\ \overline{\boldsymbol{\theta}}_I \in \mathbb{R}\,\mathrm{sgn}(\boldsymbol{\theta}_I)\right\}, \frac{\tau+1}{\tau-1}\right)^2}\right). \tag{4.22}$$

The first remainder term scales as $\frac{p}{n}$ which reflects that anti-sparsity regularization requires an overdetermined regime to ensure good stability performance. This is in agreement with [145, Theorem 7]. This phenomenon was also observed by [58] who studied sample complexity thresholds for noiseless recovery from random projections of the hypercube.

4.4.3.5 Nuclear norm

We now turn to the nuclear norm case. Recall the notations of Section 2.4.3.5. For matrices $\boldsymbol{\theta} \in \mathbb{R}^{p_1\times p_2}$, a measurement map $\boldsymbol{X}$ takes the form of a linear operator whose ith component is given by the Frobenius scalar product

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

where $\boldsymbol{X}^i$ is a matrix in $\mathbb{R}^{p_1\times p_2}$. We denote $\|\cdot\|_{\mathrm{F}}$ the associated norm. From (2.16), 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, from (2.16), we have

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

To apply Proposition 4.4.2 and Proposition 4.4.3, we need to bound $w(\boldsymbol{X}(\mathcal{C}))$ ($\mathcal{C}$ is the nuclear ball), or equivalently, to bound

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

which is the expectation of the operator norm of a random series with matrix coefficients. Thus using [140, Theorem 4.1.1(4.1.5)] to get this bound, and inserting it into Proposition 4.4.2 and Proposition 4.4.3, 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 4.4.10. *Let the data generated by* (4.13) *with a linear operator* $\boldsymbol{X}: \mathbb{R}^{p_1\times p_2} \to \mathbb{R}^n$. *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* (2.15).

(i) $\boldsymbol{\xi}$ *is a zero-mean sub-Gaussian random vector with parameter* σ*: suppose that*

$$\lambda_n \geq 2\tau\sigma c_1 \sqrt{\frac{\log(c_2/\delta)\log(p_1+p_2)}{n}},$$

for some $\tau > 1$ *and* $0 < \delta < \min(c_2, 1)$*, where* c_1 *and* c_2 *are the positive absolute constants in Proposition 4.4.2. Then, with probability at least* $1-\delta$*, the following holds*

$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 \leq \inf_{\substack{r\in\{1,\dots,\min(p_1,p_2)\} \\ \boldsymbol{\theta}:\ \mathrm{rank}(\boldsymbol{\theta})=r}} \left(\frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 + \frac{2\lambda_n^2 (\tau+1)^2 r}{\tau^2 \Upsilon\left(T_{\boldsymbol{\theta}}, \frac{\tau+1}{\tau-1}\right)^2} \right) + p_1 p_2 \beta, \tag{4.23}$$

and

$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 \leq \inf_{\substack{r\in\{1,\dots,\min(p_1,p_2)\} \\ \boldsymbol{\theta}:\ \mathrm{rank}(\boldsymbol{\theta})=r}} \left(\frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}\boldsymbol{\theta}_0\right\|_2^2 + \frac{2\lambda_n^2 (\tau+1)^2 r}{\tau^2 \Upsilon\left(T_{\boldsymbol{\theta}}, \frac{\tau+1}{\tau-1}\right)^2} \right). \tag{4.24}$$

(ii) $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$*: suppose that*

$$\lambda_n \geq (1+\delta)\tau\sigma \sqrt{\frac{2\log(p_1+p_2)}{n}},$$

for some $\tau > 1$ *and* $\delta > 0$*. Then, with probability at least* $1 - (p_1+p_2)^{-\delta^2}$*, (4.23) and (4.24) hold.*

The set over which the infimum is taken just reminds us that the nuclear norm is partly smooth (see above) relative to the constant rank manifold (which is a Riemannian submanifold of $\mathbb{R}^{p_1\times p_2}$) [53, Theorem 3.19]. The first remainder term now scales as $\frac{r\log(p_1+p_2)}{n}$. In the i.i.d. Gaussian case, we recover the same rate as in [47, Theorem 3] for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and in [88, Theorem 2] for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$.

4.4.4 Discussion of minimax optimality

In this section, we discuss the optimality of the estimators $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ (we remind the reader that the design $\boldsymbol{X}$ is fixed). Recall the discussion on stratification at the end of Section 4.3.1. Let $\mathcal{M}_0 \in \mathcal{M}$ be the stratum active at $\boldsymbol{\theta}_0 \in \mathcal{M}_0$. In this setting, with $\beta = O(1/(pn))$, (4.5) and Proposition 4.4.3 ensure that

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

and

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

with high probability. In particular, for a polyhedral gauge penalty, in which case $\mathcal{M}_0 = T_{\boldsymbol{\theta}_0}$ (see [145]), and under the normalization $\max_{\boldsymbol{v}\in\mathcal{V}} \|\boldsymbol{X}\boldsymbol{v}\|_2 \leq \sqrt{n}$, Proposition 4.4.4 entails

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

and

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

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

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

where ψ_n is the rate, which ideally, should be comparable to the risk bounds above. A standard path to derive such a lower bound is to exhibit a subset of $\mathcal{M}_0$ of well-separated points while controlling its diameter, see [142, Chapter 2] or [104, Section 4.3]. This however must be worked out on a case-by-case basis.

Example 4.4.11. For the Lasso case, $\mathcal{M}_0 = T_{\boldsymbol{\theta}_0}$ is the subspace of vectors whose support is contained in that of $\boldsymbol{\theta}_0$. Let $I = \mathrm{supp}(\boldsymbol{\theta}_0)$ and $s = \|\boldsymbol{\theta}_0\|_0$. Define the set

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

We have $\mathcal{B}_0\subset\mathcal{M}_0$ and $\left\|\boldsymbol{\theta}-\boldsymbol{\theta}'\right\|_0\leq 2s$ for all $(\boldsymbol{\theta},\boldsymbol{\theta}')\in\mathcal{B}_0$. Define $\mathcal{F}_0 \overset{\mathrm{def}}{=} \left\{r\boldsymbol{X}\boldsymbol{\theta} \;:\; \boldsymbol{\theta}\in\mathcal{B}_0\right\}$, for $r>0$ to be specified later. Due to the Varshamov-Gilbert lemma [104, Lemma 4.7], given $a\in]0,1[$, there exists a subset $\mathcal{B}\subset\mathcal{B}_0$ with cardinality $|\mathcal{B}|\geq 2^{\rho s/2}$ such that for two distinct elements $\boldsymbol{X}\boldsymbol{\theta}$ and $\boldsymbol{X}\boldsymbol{\theta}'$ in $\mathcal{F}_0$

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

where

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

Standard results from random matrix theory ensure that $\underline{\kappa}>0$ for a Gaussian design with high probability as long as $n\geq s+C\sqrt{s}$ [139] for some positive absolute constant C.

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

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

We are now in position to apply [142, Theorem 2.5] to conclude that there exists $\eta\in]0,1[$ (that depends on a) such that

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

This lower bound together with Corollary 4.4.5 show that $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ (with $\beta=O(1/(pn))$) and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ are nearly minimax (up to a logarithmic factor) over $\mathcal{M}_0$.

One can generalize this reasoning to get a minimax lower bound over the larger class of s-sparse vectors, i.e.

$$\bigcup\left\{V=\mathrm{Span}\{(\boldsymbol{a}_j)_{1\leq j\leq p}\} \;:\; \dim(V)=s\right\},$$

which is a finite union of subspaces that contains $\mathcal{M}_0$. Let $(a,b)\in]0,1[^2$ such that $1\leq s\leq abp$ and $a(-1+b-\log(b))\geq\log(2)$ [5], $c\in]0,1/8[$. Then combining [142, Theorem 2.5] and [104, Lemma 4.6

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

and Lemma 4.10], we have for $\eta \overset{\text{def}}{=} \frac{1}{1+(ab)^{\rho s/2}} \left(1 - 2c - \sqrt{\frac{2c}{-\rho \log(ab)}}\right) \in]0,1[$

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

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

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

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

One can see that the difference between this lower bound and the one on $\mathcal{M}_0$ lies in the $\log(p/s)$ factor, which basically derives from the control over the union of subspaces. The minimax prediction risk (in expectation) over the ℓ_0-ball were studied in [122, 116, 153, 160, 155], where similar lower bounds were obtained.

Example 4.4.12. For the group Lasso with L groups of equal size K, $\mathcal{M}_0$ is the subspace group sparse vectors whose group support is included in that of $\boldsymbol{\theta}_0$. Let s be the number of non-zero (active) groups in $\boldsymbol{\theta}_0$. Following exactly the same reasoning as for the Lasso, one can show that the risk lower bound in probability scales as $C\sigma^2 sK/n$, which together with Corollary 4.4.6, shows that $\widehat{\boldsymbol{\theta}}_n^{\text{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\text{PEN}}$ are nearly minimax (up again to a logarithmic factor) over $\mathcal{M}_0$. One can also derive the lower bound $C\sigma^2 s(K + \log(L/s))/n$ over the set of s-block sparse vectors. Such minimax lower bound is comparable to the one in [99].

Example 4.4.13. Let us consider the ℓ_∞-penalty. Denote the saturation support of $\boldsymbol{\theta}_0$ as I^{sat} and recall the subspace $T_{\boldsymbol{\theta}_0}$ form (2.14). Thus, $\mathcal{M}_0 = T_{\boldsymbol{\theta}_0}$ is the subspace of vectors which are collinear to $\operatorname{sgn}(\boldsymbol{\theta}_0)$ on I^{sat} and free on its complement. Observe that $\dim(\mathcal{M}_0) = p - s + 1$, where $s = |I^{\text{sat}}|$. Define the set

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

By construction, $\mathcal{B}_0 \subset \mathcal{M}_0$, and $\left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_0 \leq 2(p-s)$ for all $(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \mathcal{B}_0$. Thus following the same arguments as for the Lasso example (using again Varshamov-Gilbert lemma and [142, Theorem 2.5]), we conclude that there exists $\eta \in]0,1[$ (that depends on a) such that

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

where the restricted isometry constants are defined similarly to the Lasso but with respect to the model subspace $\mathcal{M}_0$ of the ℓ_∞ norm. Again, for a Gaussian design, $\underline{\kappa} > 0$ with high probability as long as $n \geq (p-s+1) + C\sqrt{p-s+1}$ [139].

The obtained minimax lower bound is consistent with the sample complexity thresholds derived in [58] for noiseless recovery from random projections of the hypercube. For a saturation support size small compared to p, the bound of Corollary 4.4.9 comes close to the minimax lower bound.

Example 4.4.14. Let $r = \operatorname{rank}(\boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0 \in \mathbb{R}^{p_1 \times p_2}$, and $p = \max(p_1, p_2)$. For the nuclear norm, $\mathcal{M}_0$ is the manifold of rank-r matrices. Thus arguing as in [88, Theorem 5] (who use the Varshamov-Gilbert lemma [104] to find the covering set), one can show that the minimax risk lower bound over $\mathcal{M}_0$ is $C\sigma^2 r/n$. In view of Corollary 4.4.10, we deduce that $\widehat{\boldsymbol{\theta}}_n^{\text{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\text{PEN}}$ are nearly minimax over the constant rank manifolds.

4.5 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 $\left(\mathcal{C}\right)^0$ 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

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

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

Lemma 4.5.1. *Let f be a finite-valued and convex function. Then*

(i) $\left(^{\gamma}f(\boldsymbol{\theta})\right)_{\gamma>0}$ *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) $\forall \boldsymbol{\theta} \in \mathbb{R}^p$, $\nabla\, ^{\gamma}f(\boldsymbol{\theta}) \to \left(\partial f(\boldsymbol{\theta})\right)^0$ *and* $\left\|\nabla\, ^{\gamma}f(\boldsymbol{\theta})\right\|_2 \nearrow \left\|\left(\partial f(\boldsymbol{\theta})\right)^0\right\|_2$ *as* $\gamma \searrow 0$.

Proof. (i) [7, Proposition 12.32]. (ii) [7, Proposition 12.29]. (iii) by assumption, f is subdifferentiable everywhere and its subdifferential is a maximal monotone operator with domain $\mathbb{R}^p$, and the result follows from [7, Corollary 23.46(i)]. □

We are now equipped to prove the following important result[6].

Proposition 4.5.2. *Let the density $\widehat{\mu}_n$ in (1.12), where*

(a) F satisfies Assumptions **(H.1)**-**(H.2)**;

(b) J is a finite-valued lower-bounded convex function, and $\exists R > 0$ and $\rho \geq 0$, such that $\forall \boldsymbol{\theta} \in \mathbb{R}^p$, $\left\|\left(\partial J(\boldsymbol{\theta})\right)^0\right\|_2 \leq R \left\|\boldsymbol{\theta}\right\|_2^{\rho}$;

(c) and V_n is coercive.

Then, $\forall \overline{\boldsymbol{\theta}} \in \mathbb{R}^p$,

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

This result covers of course the situation where J fulfills **(H.3)**. In this case, since $\partial J(\boldsymbol{\theta}) \subset \mathcal{C}^\circ$ by Corollary 2.4.3(i), we have $\rho = 0$ and $R = \operatorname{diam}(\mathcal{C}^\circ)$, the diameter of the convex compact set $\mathcal{C}^\circ$ containing the origin. It can be shown that, when $F(\cdot, \boldsymbol{y})$ is strongly coercive, the coercivity assumption (c) can be equivalently stated as $J_\infty(\boldsymbol{\theta}) > 0$, $\forall \boldsymbol{\theta} \in \ker(\boldsymbol{X}) \setminus \{0\}$, where J_∞ is the recession/asymptotic function of J; see e.g. [126].

Proof. Let

$$V_n^{\gamma}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \frac{1}{n} F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) + \lambda_n\, ^{\gamma}J(\boldsymbol{\theta})$$

and define

$$\mu_n^{\gamma}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \exp\left(-V_n^{\gamma}(\boldsymbol{\theta})/\beta\right)/Z,$$

where $0 < Z < +\infty$ is the normalizing constant of the density $\widehat{\mu}_n$. Assumption **(H.1)** and Lemma 4.5.1(ii)-(iii) tell us that $V_n^{\gamma} \in C^1(\mathbb{R}^p)$ and $\nabla V_n^{\gamma}(\boldsymbol{\theta}) \to \left(\partial V_n(\boldsymbol{\theta})\right)^0$ as $\gamma \to 0$. Thus

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

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

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

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

From coercivity of V_n, the objective in the min is also coercive in $(\boldsymbol{\theta}, \overline{\boldsymbol{\theta}})$ by [126, Exercise 3.29(b)]. It then follows from [126, Theorem 3.31] that V_n^γ is also coercive. In turn, [126, Theorem 11.8(c) and 3.26(a)] allow to assert that for some $a \in]0, +\infty[$, $\exists b \in]-\infty, +\infty[$ such that for all $\gamma > 0$ and $\boldsymbol{\theta} \in \mathbb{R}^p$

$$\mu_n^\gamma(\boldsymbol{\theta}) \leq \exp\left(-a \|\boldsymbol{\theta}\|_2 - b\right) / Z. \tag{4.25}$$

Lemma 4.5.1 (iii) and Assumption (b) on J entail that for any $\boldsymbol{\theta} \in \mathbb{R}^p$,

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

Altogether, we have

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

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

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

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

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

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

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

where we used coercivity of V_n^γ (see (4.25)) 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 4.5.1(i) that $\mu_n^\gamma \to \widehat{\mu}_n$ as $\gamma \to 0$. Thus, arguing again as in (4.25), we can apply the dominated convergence theorem to conclude that

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

This concludes the proof. □

Chapter 5

Estimation Bounds with Low-complexity Priors

Main contributions of this chapter

- Develop bounds guaranteeing that both $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ in (1.12) and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ in (1.9) stably estimate $\boldsymbol{\theta}_0$ from the noisy measurements $\boldsymbol{y}$ (see Theorem 5.3.6).
- Provide sample complexity bounds (see Section 5.4.2) that guarantee that our consistency bounds hold with high probability in the case of Gaussian design.
- Exemplify our bounds on several penalties routinely used in the literature (see Section 5.4.3).

Contents

5.1 Introduction

In this chapter, we again consider a high-dimensional statistical estimation problem in which the number of parameters is comparable or larger than the sample size. We provide a unified framework for establishing consistency bounds the parameter estimates for exponential weighted aggregation and penalized estimators with a general class of data losses and log-concave priors. In the case of high-dimensional regression with a Gaussian design, we provide sample complexity bounds that guarantee that our consistency bounds hold with high probability. These results are applied to several instances including the Lasso, the group Lasso, their analysis-type counterparts, the ℓ_∞ and the nuclear norm penalties. We also discuss extension beyond Gaussian design.

5.1.1 Problem statement

Let $\boldsymbol{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ be i.i.d. observations drawn from a general regression problem in high dimension. By the aggregation approach, the regression function f is approximated by $f_{\boldsymbol{\theta}_0}$ (see Section 1.1.1) with

$$\boldsymbol{\theta}_0 \in \operatorname*{Argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathbb{E}\left[F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y})\right].$$

The loss function F is supposed to be smooth and convex. Our goal is to provide bounds guaranteeing that both the penalized estimator (1.9) and exponential weighted aggregation (1.12) stably estimate $\boldsymbol{\theta}_0$ from the noisy measurements $\boldsymbol{y}$. Here, we consider $J_{\boldsymbol{\lambda}_n} = \lambda_n J$ where $\lambda_n > 0$ is the regularization parameter and J is a proper closed convex function that promotes some specific notion of simplicity/low-complexity.

5.1.2 Chapter organization

In Section 5.2, we state our main assumptions on the data loss and the prior penalty, discuss key properties and some examples which are popular in the literature. In Section 5.3, we prove our estimation bounds. We will apply them to the penalty examples for the Gaussian design in Section 5.4. We also discuss extension beyond Gaussian designs.

5.2 Estimation with log-concave priors

The estimator $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ in (1.12) 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.

5.2.1 Data loss

The class of loss functions F that we consider obey the following assumptions (which are the same as in Chapter 4):

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

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

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

(H.2) For any $\overline{\boldsymbol{\theta}} \in \mathbb{R}^p$ and $\boldsymbol{y} \in \mathbb{R}^n$,

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

See Section 4.2.1 for a further discussion on this class of data loss

5.2.2 Prior penalty

Throughout, we assume the following on J.

(H.3) $J : \mathbb{R}^p \to \mathbb{R}$ is a lower-bounded convex function.

This assumption implies that J is finite-valued, locally Lipschitz continuous on $\mathrm{dom}(J) = \mathbb{R}^p$, and for any point $\boldsymbol{\theta} \in \mathbb{R}^p$, $\partial J(\boldsymbol{\theta})$ a non-empty compact set. The class of penalties J we consider in this chapter is then much larger than the one of gauges in Chapter 4.

Closure properties The set of functions satisfying **(H.3)** is closed under addition[1], smooth perturbation and pre-composition by a linear operator. Consequently, more intricate regularizers can be built starting from simple penalties. Moreover, all the geometrical objects defined in for these regularizers can be deduced from those of the simple building penalties; see [145, Proposition 8, Proposition 10 and Corollary 2]. In turn, our unified analysis in Section 5.3 will apply to them just as well.

5.3 Main results

In the rest of the chapter, denote $T_{\boldsymbol{\theta}}$ and $e_{\boldsymbol{\theta}}$ respectively the model subspace and vector associated to $\boldsymbol{\theta}$, and let $S_{\boldsymbol{\theta}} = T_{\boldsymbol{\theta}}^\top$. Let $f_{\boldsymbol{\theta}} \in \mathrm{ri}(\partial J(\boldsymbol{\theta}))$, $J_{f_{\boldsymbol{\theta}}}^\circ$ and $J_{f_{\boldsymbol{\theta}}}$ as given in Definition 2.3.14.

Definition 5.3.1. For a vector $\boldsymbol{\theta} \in \mathbb{R}^p$, the set of *dual certificates* for a vector $\boldsymbol{\theta} \in \mathbb{R}^p$ is defined as

$$\mathcal{D}_{\boldsymbol{\theta}} = \mathrm{Span}(\boldsymbol{X}^\top) \cap \partial J(\boldsymbol{\theta}). \tag{5.1}$$

The so-called *source* or *range* condition is verified if and only if $\mathcal{D}_{\boldsymbol{\theta}} \neq \emptyset$. The set of *non-degenerate* dual certificates is

$$\breve{\mathcal{D}}_{\boldsymbol{\theta}} = \mathrm{Span}(\boldsymbol{X}^\top) \cap \mathrm{ri}(\partial J(\boldsymbol{\theta})). \tag{5.2}$$

Observe that $\breve{\mathcal{D}}_{\boldsymbol{\theta}} = \mathrm{ri}(\mathcal{D}_{\boldsymbol{\theta}})$ whenever the source condition is verified.

In the following, we will used the shorthand notation $\boldsymbol{\zeta} \stackrel{\mathrm{def}}{=} \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y})$.

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

5.3.1 Prediction and Bregman divergence bounds

We start by providing some preliminary key bounds.

Lemma 5.3.2. *Assume that* **(H.1)**, **(H.2)** *and* **(H.3)** *hold. Suppose that* $\mathcal{D}_{\boldsymbol{\theta}_0} \neq \emptyset$. *Consider* $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ *and* $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ *with* $\lambda_n = c\|\boldsymbol{\zeta}\|_2/n$ *for some positive constant* c. *Then,*

$$
\begin{aligned}
D_J^{\boldsymbol{\eta}_0}\left(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\right) &\leq \frac{np\beta + 2\varphi^+\big(\|\boldsymbol{\zeta}\|_2(1+c\|\boldsymbol{\alpha}_0\|_2)/2\big)}{c\|\boldsymbol{\zeta}\|_2}, \\
\varphi\Big(\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2\Big) &\leq np\beta + \varphi^+\big(\|\boldsymbol{\zeta}\|_2(1+c\|\boldsymbol{\alpha}_0\|_2)\big), \\
D_J^{\boldsymbol{\eta}_0}\left(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}, \boldsymbol{\theta}_0\right) &\leq \frac{2}{c\|\boldsymbol{\zeta}\|_2}\varphi^+\big(\|\boldsymbol{\zeta}\|_2(1+c\|\boldsymbol{\alpha}_0\|_2)/2\big), \\
\varphi\Big(\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2\Big) &\leq \varphi^+\big(\|\boldsymbol{\zeta}\|_2(1+c\|\boldsymbol{\alpha}_0\|_2)\big).
\end{aligned}
$$

where $\boldsymbol{\alpha}_0$ *is such that* $\boldsymbol{\eta}_0 = \boldsymbol{X}^\top\boldsymbol{\alpha}_0 \in \mathcal{D}_{\boldsymbol{\theta}_0}$. *In particular, taking* $\beta = \frac{C}{np}\varphi^+\big(\|\boldsymbol{\zeta}\|_2/2\big)$ *for some positive constant* C, *we have*

$$
\begin{aligned}
D_J^{\boldsymbol{\eta}_0}\left(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}, \boldsymbol{\theta}_0\right) &\leq \frac{C+2}{c\|\boldsymbol{\zeta}\|_2}\varphi^+\big(\|\boldsymbol{\zeta}\|_2(1+c\|\boldsymbol{\alpha}_0\|_2)/2\big), \\
\varphi\Big(\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2\Big) &\leq (C+1)\varphi^+\big(\|\boldsymbol{\zeta}\|_2(1+c\|\boldsymbol{\alpha}_0\|_2)\big).
\end{aligned}
$$

Proof. Combining the fact that $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ is a global minimizer of V_n with Assumption **(H.1)**, we have

$$
\begin{aligned}
&\frac{1}{n}\varphi\Big(\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2\Big) + \frac{1}{n}\Big\langle \nabla F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y}), \boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\Big\rangle \\
&\leq \frac{1}{n}F(\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}, \boldsymbol{y}) - \frac{1}{n}F(\boldsymbol{X}\boldsymbol{\theta}_0, \boldsymbol{y}) \\
&\leq -\lambda_n(J(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}) - J(\boldsymbol{\theta}_0)) - \frac{1}{n}\varphi\Big(\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2\Big) \\
&= -\lambda_n D_J^{\boldsymbol{\eta}_0}\left(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}, \boldsymbol{\theta}_0\right) - \lambda_n\Big\langle \boldsymbol{\alpha}_0, \boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\Big\rangle - \frac{1}{n}\varphi\Big(\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2\Big). \qquad (5.3)
\end{aligned}
$$

We then deduce from Cauchy-Schwarz and Fenchel-Young inequalities that

$$
\begin{aligned}
\lambda_n D_J^{\boldsymbol{\eta}_0}\left(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}, \boldsymbol{\theta}_0\right) &\leq -\Big\langle \frac{1}{n}\boldsymbol{\zeta} + \lambda_n\boldsymbol{\alpha}_0, \boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\Big\rangle - \frac{2}{n}\varphi\Big(\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2\Big) \\
&\leq \big(\frac{1}{n}\|\boldsymbol{\zeta}\|_2 + \lambda_n\|\boldsymbol{\alpha}_0\|_2\big)\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2 - \frac{2}{n}\varphi\Big(\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2\Big) \\
&\leq \lambda_n(1/c + \|\boldsymbol{\alpha}_0\|_2)\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2 - \frac{2}{n}\varphi\Big(\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2\Big) \\
&\leq \frac{2}{n}\varphi^+\big(n\lambda_n(1/c + \|\boldsymbol{\alpha}_0\|_2)/2\big).
\end{aligned}
$$

Departing again from (5.3) and using non-negativity of the Bregman divergence, we obtain

$$
\begin{aligned}
\frac{1}{n}\varphi\Big(\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2\Big) &\leq (1/c + \|\boldsymbol{\alpha}_0\|_2)\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2\lambda_n - \frac{1}{n}\varphi\Big(\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2\Big) \\
&\leq \frac{1}{n}\varphi^+\big(n\lambda_n(1/c + \|\boldsymbol{\alpha}_0\|_2)\big).
\end{aligned}
$$

Let us now turn to $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$. From the proof of the inequality (4.4), we have

$$
V_n(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}) - V(\boldsymbol{\theta}_0) \leq p\beta - \frac{1}{n}\varphi\Big(\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2\Big).
$$

We thus infer, using again Assumption **(H.1)**, that

$$\frac{1}{n}\varphi\Big(\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}-\boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2\Big)\le p\beta-\lambda_n D_J^{\boldsymbol{\eta}_0}\Big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}},\boldsymbol{\theta}_0\Big)-\Big\langle\frac{1}{n}\boldsymbol{\zeta}+\lambda_n\boldsymbol{\alpha}_0,\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}-\boldsymbol{X}\boldsymbol{\theta}_0\Big\rangle - \frac{1}{n}\varphi\Big(\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}-\boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2\Big).$$

The Cauchy-Schwarz and Fenchel-Young inequalities then yield

$$\begin{aligned}\lambda_n D_J^{\boldsymbol{\eta}_0}\Big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}},\boldsymbol{\theta}_0\Big)&\le p\beta-\Big\langle\frac{1}{n}\boldsymbol{\zeta}+\lambda_n\boldsymbol{\alpha}_0,\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}-\boldsymbol{X}\boldsymbol{\theta}_0\Big\rangle-\frac{2}{n}\varphi\Big(\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}-\boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2\Big)\\&\le p\beta+\frac{2}{n}\varphi^+\big(n\lambda_n(1/c+\|\boldsymbol{\alpha}_0\|_2)/2\big)\\&=\frac{1}{n}\Big(C\varphi^+\big(n\lambda_n/(2c)\big)+2\varphi^+\big(n\lambda_n(1/c+\|\boldsymbol{\alpha}_0\|_2)/2\big)\Big)\\&\le\frac{1}{n}(C+2)\varphi^+\big(n\lambda_n(1/c+\|\boldsymbol{\alpha}_0\|_2)/2\big)\end{aligned}$$

where, in the last inequality, we used that φ^+ is non-decreasing on $\mathbb{R}_+$. On the other hand,

$$\begin{aligned}\frac{1}{n}\varphi\Big(\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}-\boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2\Big)&\le p\beta-\Big\langle\frac{1}{n}\boldsymbol{\zeta}+\lambda_n\boldsymbol{\alpha}_0,\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}-\boldsymbol{X}\boldsymbol{\theta}_0\Big\rangle-\frac{1}{n}\varphi\Big(\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}-\boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2\Big)\\&\le p\beta+\frac{1}{n}\varphi^+\big(n\lambda_n(1/c+\|\boldsymbol{\alpha}_0\|_2)\big)\\&=\frac{1}{n}\Big(C\varphi^+\big(n\lambda_n/(2c)\big)+\varphi^+\big(n\lambda_n(1/c+\|\boldsymbol{\alpha}_0\|_2)\big)\Big)\\&\le\frac{1}{n}(C+1)\varphi^+\big(n\lambda_n(1/c+\|\boldsymbol{\alpha}_0\|_2)\big).\end{aligned}$$

□

With the prescribed choice of β, the bounds for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ are of the same order.

Remark 5.3.3. For the penalized estimator, the obtained bounds are generalizations beyond the strongly convex case of those in [67, 143, 146].

Example 5.3.4. Consider the case where $\varphi : t\in\mathbb{R}_+\mapsto t^q/q$, $q\in]1,+\infty[$, in which case $\varphi^+(t)=t^{q_*}/q_*$ where $1/q+1/q_*=1$. With the choice $\beta=C2^{\frac{q}{1-q}}\frac{(\|\boldsymbol{\zeta}\|_2)^{\frac{q}{q-1}}}{np}$, and straightforward algebraic manipulations, the bounds of Lemma 5.3.2 specialize to

$$\begin{aligned}D_J^{\boldsymbol{\eta}_0}\Big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}},\boldsymbol{\theta}_0\Big)&\le\frac{q-1}{cq}(C+2)2^{\frac{q}{1-q}}(1+c\|\boldsymbol{\alpha}_0\|_2)^{\frac{q}{q-1}}\|\boldsymbol{\zeta}\|_2^{\frac{1}{q-1}},\\\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}-\boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2&\le(q-1)^{1/q}(C+1)^{1/q}2^{\frac{1}{1-q}}(1+c\|\boldsymbol{\alpha}_0\|_2)^{\frac{1}{q-1}}\|\boldsymbol{\zeta}\|_2^{\frac{1}{q-1}},\\D_J^{\boldsymbol{\eta}_0}\Big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}},\boldsymbol{\theta}_0\Big)&\le\frac{q-1}{cq}2^{\frac{1}{1-q}}(1+c\|\boldsymbol{\alpha}_0\|_2)^{\frac{q}{q-1}}\|\boldsymbol{\zeta}\|_2^{\frac{1}{q-1}},\\\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}-\boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2&\le(q-1)^{1/q}2^{\frac{1}{(1-q)}}(1+c\|\boldsymbol{\alpha}_0\|_2)^{\frac{1}{q-1}}\|\boldsymbol{\zeta}\|_2^{\frac{1}{q-1}}.\end{aligned}$$

In particular, for $q\ge2$, as is the case when $F(\boldsymbol{u},\boldsymbol{y})=\|\boldsymbol{u}-\boldsymbol{y}\|_2^q/q$ (see Example 4.2.1), we have the normalized estimates

$$\begin{aligned}n^{-1/2}D_J^{\boldsymbol{\eta}_0}\Big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}},\boldsymbol{\theta}_0\Big)&\le\frac{q-1}{cq}(C+2)2^{\frac{q}{1-q}}(1+c\|\boldsymbol{\alpha}_0\|_2)^{\frac{q}{q-1}}\Big\|n^{-1/2}\boldsymbol{\zeta}\Big\|_2^{\frac{1}{q-1}}n^{-\frac{q-2}{2(q-1)}},\\n^{-1/2}\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}-\boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2&\le(q-1)^{1/q}(C+1)^{1/q}2^{\frac{1}{1-q}}(1+c\|\boldsymbol{\alpha}_0\|_2)^{\frac{1}{q-1}}\Big\|n^{-1/2}\boldsymbol{\zeta}\Big\|_2^{\frac{1}{q-1}}n^{-\frac{q-2}{2(q-1)}},\\n^{-1/2}D_J^{\boldsymbol{\eta}_0}\Big(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}},\boldsymbol{\theta}_0\Big)&\le\frac{q-1}{cq}2^{\frac{1}{1-q}}(1+c\|\boldsymbol{\alpha}_0\|_2)^{\frac{q}{q-1}}\Big\|n^{-1/2}\boldsymbol{\zeta}\Big\|_2^{\frac{1}{q-1}}n^{-\frac{q-2}{2(q-1)}},\\n^{-1/2}\Big\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}-\boldsymbol{X}\boldsymbol{\theta}_0\Big\|_2&\le(q-1)^{1/q}2^{\frac{1}{(1-q)}}(1+c\|\boldsymbol{\alpha}_0\|_2)^{\frac{1}{q-1}}\Big\|n^{-1/2}\boldsymbol{\zeta}\Big\|_2^{\frac{1}{q-1}}n^{-\frac{q-2}{2(q-1)}}.\end{aligned}$$

with the choice $\beta = C2^{\frac{q}{1-q}} \frac{\|n^{-1/2}\boldsymbol{\zeta}\|_2^{\frac{q}{q-1}}}{n^{\frac{q-2}{2(q-1)}}p}$.

The following lemma is a key towards establishing our estimation bound.

Lemma 5.3.5. *Assume that* **(H.3)** *holds. Then, for any* $\boldsymbol{\eta}_0 \in \mathrm{ri}(\partial J(\boldsymbol{\theta}_0))$ *and* $\boldsymbol{\theta} \in \mathbb{R}^p$

$$\left\|\mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right\|_2 \leq \|\,\mathrm{P}_{S_{\boldsymbol{\theta}_0}}\,\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2} \frac{D_J^{\boldsymbol{\eta}_0}(\boldsymbol{\theta},\boldsymbol{\theta}_0)}{1-J^\circ_{f_{\boldsymbol{\theta}_0}}\big(\mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\boldsymbol{\eta}_0-f_{\boldsymbol{\theta}_0})\big)}.$$

Proof. Let $\boldsymbol{\eta} \in \partial J(\boldsymbol{\theta}_0)$ at which Theorem 2.4.2(ii) holds for $\boldsymbol{\omega} = \boldsymbol{\theta}-\boldsymbol{\theta}_0$. Thus

$$\begin{aligned}
D_J^{\boldsymbol{\eta}_0}(\boldsymbol{\theta},\boldsymbol{\theta}_0) &\geq D_J^{\boldsymbol{\eta}_0}(\boldsymbol{\theta},\boldsymbol{\theta}_0) - D_J^{\boldsymbol{\eta}}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\\
&= \langle \boldsymbol{\eta}-\boldsymbol{\eta}_0, \boldsymbol{\theta}-\boldsymbol{\theta}_0\rangle\\
&= \Big\langle \mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\boldsymbol{\eta}-f_{\boldsymbol{\theta}_0}) - \mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\boldsymbol{\eta}_0-f_{\boldsymbol{\theta}_0}), \boldsymbol{\theta}-\boldsymbol{\theta}_0\Big\rangle\\
\text{(Theorem 2.4.2(ii))} &= J_{f_{\boldsymbol{\theta}_0}}(\mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)) - \Big\langle \mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\boldsymbol{\eta}_0-f_{\boldsymbol{\theta}_0}), \mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\Big\rangle\\
&\geq J_{f_{\boldsymbol{\theta}_0}}(\mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0))\big(1-J^\circ_{f_{\boldsymbol{\theta}_0}}(\boldsymbol{\eta}_{S_{\boldsymbol{\theta}_0}} - \mathrm{P}_{S_{\boldsymbol{\theta}_0}} f_{\boldsymbol{\theta}_0})\big),
\end{aligned}$$

where in the last inequality, we used the duality inequality (2.4) on $\mathrm{dom}(J^\circ_{f_{\boldsymbol{\theta}_0}}) \times \mathrm{dom}(J_{f_{\boldsymbol{\theta}_0}})$, where, by Lemma 2.3.15, $\mathrm{dom}(J^\circ_{f_{\boldsymbol{\theta}_0}}) = S_{\boldsymbol{\theta}_0}$ and $\mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) \in \mathrm{dom}(J_{f_{\boldsymbol{\theta}_0}})$. In view of the definition of $\boldsymbol{\eta}_0$ and the last assertion of Theorem 2.4.2(i), the denominator never vanishes. Moreover, since $J_{f_{\boldsymbol{\theta}_0}}$ is positively homogeneous and coercive on $S_{\boldsymbol{\theta}_0}$, it follows that for all $\boldsymbol{\theta} \in \mathbb{R}^p$

$$\left\|\mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right\|_2 \leq \|\,\mathrm{P}_{S_{\boldsymbol{\theta}_0}}\,\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2} J_{f_{\boldsymbol{\theta}_0}}(\mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)).$$

This concludes the proof. □

5.3.2 Bounds on the parameter estimates

We are now ready to state our main estimation bounds.

Theorem 5.3.6. *Assume that* **(H.1)**, **(H.2)** *and* **(H.3)** *hold, with* φ *strictly increasing. Suppose that*

$$\breve{\mathcal{D}}_{\boldsymbol{\theta}_0} \neq \emptyset \quad \text{and} \quad \ker(\boldsymbol{X}) \cap T_{\boldsymbol{\theta}_0} = \{0\}. \tag{5.4}$$

Let $\boldsymbol{\eta}_0 = \boldsymbol{X}^\top \boldsymbol{\alpha}_0 \in \breve{\mathcal{D}}_{\boldsymbol{\theta}_0}$. *Consider* $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ *and* $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ *with* $\lambda_n = c\|\boldsymbol{\zeta}\|_2/n$ *and* $\beta = \frac{C}{np}\varphi^+(\|\boldsymbol{\zeta}\|_2/2)$ *for some positive constants* c *and* C. *Then,*

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{\theta}_0\right\|_2 &\leq \|\boldsymbol{X}^+_{T_{\boldsymbol{\theta}_0}}\|_{2\to 2}\varphi^{-1}\big((C+1)\varphi^+(\|\boldsymbol{\zeta}\|_2(1+c\|\boldsymbol{\alpha}_0\|_2))\big)\\
&\quad + \frac{(C+2)\big(\|\,\mathrm{P}_{S_{\boldsymbol{\theta}_0}}\,\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2} + \|\boldsymbol{X}^+_{T_{\boldsymbol{\theta}_0}}\|_{2\to 2}\|\boldsymbol{X}_{S_{\boldsymbol{\theta}_0}}\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2}\big)}{c\|\boldsymbol{\zeta}\|_2\big(1-J^\circ_{f_{\boldsymbol{\theta}_0}}\big(\mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\boldsymbol{\eta}_0-f_{\boldsymbol{\theta}_0})\big)\big)}\varphi^+(\|\boldsymbol{\zeta}\|_2(1+c\|\boldsymbol{\alpha}_0\|_2)/2),\\
\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{\theta}_0\right\|_2 &\leq \|\boldsymbol{X}^+_{T_{\boldsymbol{\theta}_0}}\|_{2\to 2}\varphi^{-1}\big(\varphi^+(\|\boldsymbol{\zeta}\|_2(1+c\|\boldsymbol{\alpha}_0\|_2))\big)\\
&\quad + \frac{2\big(\|\,\mathrm{P}_{S_{\boldsymbol{\theta}_0}}\,\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2} + \|\boldsymbol{X}^+_{T_{\boldsymbol{\theta}_0}}\|_{2\to 2}\|\boldsymbol{X}_{S_{\boldsymbol{\theta}_0}}\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2}\big)}{c\|\boldsymbol{\zeta}\|_2\big(1-J^\circ_{f_{\boldsymbol{\theta}_0}}\big(\mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\boldsymbol{\eta}_0-f_{\boldsymbol{\theta}_0})\big)\big)}\varphi^+(\|\boldsymbol{\zeta}\|_2(1+c\|\boldsymbol{\alpha}_0\|_2)/2).
\end{aligned}$$

Proof. We give the proof for $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$, that of $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ follows exactly the same lines. In view of (5.4), we have $\boldsymbol{X}^+_{T_{\boldsymbol{\theta}_0}} = (\boldsymbol{X}^*_{T_{\boldsymbol{\theta}_0}}\boldsymbol{X}_{T_{\boldsymbol{\theta}_0}})^+\boldsymbol{X}^*_{T_{\boldsymbol{\theta}_0}}$. Thus, by the triangle inequality, Lemma 5.3.2 and Lemma 5.3.5,

we get

$$
\begin{aligned}
&\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}-\boldsymbol{\theta}_0\right\|_2\\
&\le \left\|\mathrm{P}_{T_{\boldsymbol{\theta}_0}}(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}-\boldsymbol{\theta}_0)\right\|_2+\left\|\mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}-\boldsymbol{\theta}_0)\right\|_2\\
&= \left\|\boldsymbol{X}_{T_{\boldsymbol{\theta}_0}}^{+}\boldsymbol{X}_{T_{\boldsymbol{\theta}_0}}(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}-\boldsymbol{\theta}_0)\right\|_2+\left\|\mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}-\boldsymbol{\theta}_0)\right\|_2\\
&\le \left\|\boldsymbol{X}_{T_{\boldsymbol{\theta}_0}}^{+}\right\|_{2\to 2}\left\|\boldsymbol{X}_{T_{\boldsymbol{\theta}_0}}(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}-\boldsymbol{\theta}_0)\right\|_2+\left\|\,\mathrm{P}_{S_{\boldsymbol{\theta}_0}}\right\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2}J_{f_{\boldsymbol{\theta}_0}}\left(\mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}-\boldsymbol{\theta}_0)\right)\\
&\le \left\|\boldsymbol{X}_{T_{\boldsymbol{\theta}_0}}^{+}\right\|_{2\to 2}\left\|\boldsymbol{X}(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}-\boldsymbol{\theta}_0)\right\|_2+\left(\left\|\,\mathrm{P}_{S_{\boldsymbol{\theta}_0}}\right\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2}+\left\|\boldsymbol{X}_{T_{\boldsymbol{\theta}_0}}^{+}\right\|_{2\to 2}\left\|\boldsymbol{X}_{S_{\boldsymbol{\theta}_0}}\right\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2}\right)J_{f_{\boldsymbol{\theta}_0}}\left(\mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}-\boldsymbol{\theta}_0)\right)\\
&\le \left\|\boldsymbol{X}_{T_{\boldsymbol{\theta}_0}}^{+}\right\|_{2\to 2}\varphi^{-1}\big(\varphi^{+}\big(\|\boldsymbol{\zeta}\|_2(1+c\|\boldsymbol{\alpha}_0\|_2)\big)\big)\\
&\quad+\frac{2\left(\left\|\,\mathrm{P}_{S_{\boldsymbol{\theta}_0}}\right\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2}+\left\|\boldsymbol{X}_{T_{\boldsymbol{\theta}_0}}^{+}\right\|_{2\to 2}\left\|\boldsymbol{X}_{S_{\boldsymbol{\theta}_0}}\right\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2}\right)}{c\|\boldsymbol{\zeta}\|_2\big(1-J^{\circ}_{f_{\boldsymbol{\theta}_0}}\big(\mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\boldsymbol{\eta}_0-f_{\boldsymbol{\theta}_0})\big)\big)}\varphi^{+}\big(\|\boldsymbol{\zeta}\|_2(1+c\|\boldsymbol{\alpha}_0\|_2)/2\big).
\end{aligned}
$$

□

Example 5.3.7. Consider again the case where $\varphi : t \in \mathbb{R}_+ \mapsto t^q/q$, $q \in]1,+\infty[$. Then $\beta = C2^{\frac{q}{1-q}}\frac{\left(\|\boldsymbol{\zeta}\|_2\right)^{\frac{q}{q-1}}}{np}$, and the bounds of Theorem 5.3.6 read

$$
\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}-\boldsymbol{\theta}_0\right\|_2 \le \Bigg[&\left\|\boldsymbol{X}_{T_{\boldsymbol{\theta}_0}}^{+}\right\|_{2\to 2}(q-1)^{1/q}(C+1)^{1/q}2^{\frac{1}{(1-q)}}(1+c\|\boldsymbol{\alpha}_0\|_2)^{\frac{1}{q-1}}\\
&+\frac{\frac{q-1}{cq}(C+2)^{\frac{q}{1-q}}\left(\left\|\,\mathrm{P}_{S_{\boldsymbol{\theta}_0}}\right\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2}+\left\|\boldsymbol{X}_{T_{\boldsymbol{\theta}_0}}^{+}\right\|_{2\to 2}\left\|\boldsymbol{X}_{S_{\boldsymbol{\theta}_0}}\right\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2}\right)(1+c\|\boldsymbol{\alpha}_0\|_2)^{\frac{q}{q-1}}}{\big(1-J^{\circ}_{f_{\boldsymbol{\theta}_0}}\big(\mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\boldsymbol{\eta}_0-f_{\boldsymbol{\theta}_0})\big)\big)}\Bigg]\|\boldsymbol{\zeta}\|_2^{\frac{1}{q-1}},\\
\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}-\boldsymbol{\theta}_0\right\|_2 \le \Bigg[&\left\|\boldsymbol{X}_{T_{\boldsymbol{\theta}_0}}^{+}\right\|_{2\to 2}(q-1)^{1/q}2^{\frac{1}{(1-q)}}(1+c\|\boldsymbol{\alpha}_0\|_2)^{\frac{1}{q-1}}\\
&+\frac{\frac{q-1}{cq}2^{\frac{1}{1-q}}\left(\left\|\,\mathrm{P}_{S_{\boldsymbol{\theta}_0}}\right\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2}+\left\|\boldsymbol{X}_{T_{\boldsymbol{\theta}_0}}^{+}\right\|_{2\to 2}\left\|\boldsymbol{X}_{S_{\boldsymbol{\theta}_0}}\right\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2}\right)(1+c\|\boldsymbol{\alpha}_0\|_2)^{\frac{q}{q-1}}}{\big(1-J^{\circ}_{f_{\boldsymbol{\theta}_0}}\big(\mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\boldsymbol{\eta}_0-f_{\boldsymbol{\theta}_0})\big)\big)}\Bigg]\|\boldsymbol{\zeta}\|_2^{\frac{1}{q-1}}.
\end{aligned}
$$

In particular, for $F(\boldsymbol{u},\boldsymbol{y}) = \frac{1}{q}\|\boldsymbol{y}-\boldsymbol{u}\|_2^q/q$, $q \in [2,+\infty[$, Assumptions **(H.1)**-**(H.2)** are fulfilled with $\varphi(t) = C_q t^q/q$, $C_q > 0$; see Example 4.2.1. In addition, we have $\|\nabla F(\boldsymbol{X}\boldsymbol{\theta}_0,\boldsymbol{y})\|_2 = \|\boldsymbol{\xi}\|_2^{q-1}$, $\boldsymbol{\xi} = \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}_0$, and the above bounds now read

$$
\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}-\boldsymbol{\theta}_0\right\|_2 \le \Bigg[&\left\|\boldsymbol{X}_{T_{\boldsymbol{\theta}_0}}^{+}\right\|_{2\to 2}(q-1)^{1/q}(C+1)^{1/q}2^{\frac{1}{(1-q)}}(1+c\|\boldsymbol{\alpha}_0\|_2)^{\frac{1}{q-1}}\\
&+\frac{\frac{q-1}{cq}(C+2)^{\frac{q}{1-q}}\left(\left\|\,\mathrm{P}_{S_{\boldsymbol{\theta}_0}}\right\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2}+\left\|\boldsymbol{X}_{T_{\boldsymbol{\theta}_0}}^{+}\right\|_{2\to 2}\left\|\boldsymbol{X}_{S_{\boldsymbol{\theta}_0}}\right\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2}\right)(1+c\|\boldsymbol{\alpha}_0\|_2)^{\frac{q}{q-1}}}{\big(1-J^{\circ}_{f_{\boldsymbol{\theta}_0}}\big(\mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\boldsymbol{\eta}_0-f_{\boldsymbol{\theta}_0})\big)\big)}\Bigg]\|\boldsymbol{\xi}\|_2,\\
\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}-\boldsymbol{\theta}_0\right\|_2 \le \Bigg[&\left\|\boldsymbol{X}_{T_{\boldsymbol{\theta}_0}}^{+}\right\|_{2\to 2}(q-1)^{1/q}2^{\frac{1}{(1-q)}}(1+c\|\boldsymbol{\alpha}_0\|_2)^{\frac{1}{q-1}}\\
&+\frac{\frac{q-1}{cq}2^{\frac{1}{1-q}}\left(\left\|\,\mathrm{P}_{S_{\boldsymbol{\theta}_0}}\right\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2}+\left\|\boldsymbol{X}_{T_{\boldsymbol{\theta}_0}}^{+}\right\|_{2\to 2}\left\|\boldsymbol{X}_{S_{\boldsymbol{\theta}_0}}\right\|_{J_{f_{\boldsymbol{\theta}_0}}\to 2}\right)(1+c\|\boldsymbol{\alpha}_0\|_2)^{\frac{q}{q-1}}}{\big(1-J^{\circ}_{f_{\boldsymbol{\theta}_0}}\big(\mathrm{P}_{S_{\boldsymbol{\theta}_0}}(\boldsymbol{\eta}_0-f_{\boldsymbol{\theta}_0})\big)\big)}\Bigg]\|\boldsymbol{\xi}\|_2.
\end{aligned}
$$

with the choice $\lambda_n = c\frac{\|\boldsymbol{\xi}\|_2^{q-1}}{n}$ and $\beta = C2^{\frac{q}{1-q}}\frac{\|\boldsymbol{\xi}\|_2^q}{np}$. In plain words, this bound tells us that the distance of $\boldsymbol{\theta}_0$ to $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ or to any minimizer $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ is within a factor of the noise level. This justi-

fies the terminology "linear convergence rate" widely used in the inverse problem literature; see the monograph [131].

5.4 Bounds with a random design

The non-degenerate source condition (first part of (5.4)) is an abstract condition, which is not trivial to check in practice. In fact, exhibiting a valid non-degenerate certificate is not obvious for a general design $\boldsymbol{X}$. Our aim now is to answer this question when the design is drawn from the standard Gaussian ensemble, i.e. $\boldsymbol{X}_{i,j} \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$[2]. This will allow us to derive sample complexity bounds, i.e. lower bounds on the number of observations n, which will ensure that (5.4) holds with overwhelming probability on the design.

Throughout this section, to lighten the notation, we drop the dependence on $\boldsymbol{\theta}_0$ of T, S, e and f, i.e. we denote $T \overset{\text{def}}{=} T_{\boldsymbol{\theta}_0}$ and similarly for the other quantities. We will also denote $d \overset{\text{def}}{=} \dim(T)$.

5.4.1 Minimal norm certificate

Definition 5.4.1 (Linearized pre-certificate). Assume that

$$\ker(\boldsymbol{X}) \cap T = \{0\}.$$

The "linearized pre-certificate" at $\boldsymbol{\theta}_0$ is

$$\boldsymbol{\eta}_{\mathsf{F}} \overset{\text{def}}{=} \boldsymbol{X}^\top \operatorname*{Argmin}_{\boldsymbol{X}^\top \boldsymbol{\alpha} \in \operatorname{aff}(\partial J(\boldsymbol{\theta}_0))} \|\boldsymbol{\alpha}\|_2.$$

The subscript "F" is a tribute to J.-J. Fuchs who first introduced this certificate in the context of stable support recovery in linear regression by solving the classical Lasso problem [72]. It can be easily shown, by definition of the model subspace T, that that $\boldsymbol{\eta}_{\mathsf{F}}$ can be equivalently expressed in closed form as

$$\boldsymbol{\eta}_{\mathsf{F}} = \boldsymbol{X}^\top \boldsymbol{X}_T^{+,\top} e,$$

whence the name "linearized pre-certificate".

Remark 5.4.2. It can be shown, see e.g. [147, Proposition 1], that if $\boldsymbol{\eta}_{\mathsf{F}} \in \breve{\mathcal{D}}_{\boldsymbol{\theta}_0}$, then $\boldsymbol{\eta}_{\mathsf{F}}$ is actually the minimal ℓ_2-norm dual certificate, i.e.

$$\boldsymbol{\eta}_{\mathsf{F}} = \boldsymbol{X}^\top \operatorname*{Argmin}_{\boldsymbol{X}^\top \boldsymbol{\alpha} \in \partial J(\boldsymbol{\theta}_0)} \|\boldsymbol{\alpha}\|_2.$$

This certificate plays a pivotal role in the compressed sensing literature.

5.4.2 Bounds on the number of measurements

We now investigate under which condition we can ensure that $\boldsymbol{\eta}_{\mathsf{F}} \in \breve{\mathcal{D}}_{\boldsymbol{\theta}_0}$ with high probability, or equivalently, from Theorem 2.4.2(i), that

$$J_f^\circ\big(\mathrm{P}_S(\boldsymbol{\eta}_{\mathsf{F}} - f)\big) < 1.$$

Our approach is inspired by that of [23]. The key ingredient is the fact that, owing to the isotropy of the Gaussian ensemble, $\boldsymbol{\alpha}_{\mathsf{F}} \overset{\text{def}}{=} \boldsymbol{X}_T^{+,\top} e$ and $\boldsymbol{X}_S^\top$ are independent, no matter what T is. Thus, for some $\tau > 0$ and $\nu \in]0,1]$

$$\mathbb{P}\left(J_f^\circ\big(\mathrm{P}_S(\boldsymbol{\eta}_{\mathsf{F}} - f)\big) \geq \nu\right) \leq \mathbb{P}\left(J_f^\circ\big(\mathrm{P}_S(\boldsymbol{\eta}_{\mathsf{F}} - f)\big) \geq \nu \middle| \|\boldsymbol{\alpha}_{\mathsf{F}}\|_2 \leq \tau\right) + \mathbb{P}\left(\|\boldsymbol{\alpha}_{\mathsf{F}}\|_2 \geq \tau\right). \tag{5.5}$$

[2] Another normalization most used in the the compressive sensing literature is to take the entries with variance $1/n$. This normalization implies that the columns of $\boldsymbol{X}$ are unit-normed with high probability. Our results can be adapted easily for such a normalization.

The first term in this inequality will be bounded on a case-by-case basis, and uses the fact that conditionally on $\boldsymbol{\alpha}_{\mathsf{F}}$, the entries of $\boldsymbol{\eta}_F = \boldsymbol{X}^\top \boldsymbol{\alpha}_{\mathsf{F}}$ are i.i.d. $\mathcal{N}(0, \|\boldsymbol{\alpha}_{\mathsf{F}}\|_2^2)$. For the second term, observe that as long as $n \geq d$, $\boldsymbol{X}_T$ is injective. Thus

$$\|\boldsymbol{\alpha}_{\mathsf{F}}\|_2^2 = \left\langle e, \left(\boldsymbol{X}_T^\top \boldsymbol{X}_T\right)^{-1} e \right\rangle.$$

$\left(\boldsymbol{X}_T^\top \boldsymbol{X}_T\right)^{-1}$ is an inverse Wishart matrix with n degrees of freedom. To estimate the deviation of this quadratic form, we use classical results on inverse χ^2 random variables with $n - d + 1$ degrees of freedom and we get the tail bound

$$\mathbb{P}\left(\|\boldsymbol{\alpha}_{\mathsf{F}}\|_2 \geq \tau\right) \leq e^{-\frac{t^2}{4(n-d+1)}} \tag{5.6}$$

for $\tau = \frac{\|e\|_2}{\sqrt{n-d+1}-t}$ and $t > 0$.

We now turn to bounding the first term in (5.5) on a case-by-case basis.

5.4.2.1 Finite gauge of a polytope

We here suppose that J is a finite gauge of a polytope $\mathcal{C}$. We use the shorthand notation $\mathcal{V}_S$ for the vertices of $\mathrm{P}_S(\mathcal{C})$. We can assert the following.

Proposition 5.4.3. *Assume that f is chosen such that $\kappa \overset{def}{=} \nu + \inf_{\boldsymbol{v}\in\mathcal{V}_S} \langle \boldsymbol{v}, f\rangle > 0$. Let $a \overset{def}{=} \kappa^{-2}\|e\|_2^2 \max_{\boldsymbol{v}\in\mathcal{V}_S} \|\boldsymbol{v}\|_2^2$. If $\boldsymbol{X}$ is drawn from the standard Gaussian ensemble with*

$$n \geq 2\delta a \log(|\mathcal{V}_S|/2) + d, \quad \text{for some } \delta > 1,$$

then

$$J_f^\circ\big(\mathrm{P}_S(\boldsymbol{\eta}_{\mathsf{F}} - f)\big) < \nu \in]0, 1],$$

with probability at least $1 - 2|\mathcal{V}_S|/2^{-\left(\sqrt{\frac{\delta}{2a}+\delta-1}-\sqrt{\frac{\delta}{2a}}\right)^2}$.

In turn, with high probability, (5.4) and the parameter estimation bounds of Theorem 5.3.6 hold with $\boldsymbol{\alpha}_0 = \boldsymbol{\alpha}_{\mathsf{F}}$ and $\boldsymbol{\eta}_0 = \boldsymbol{\eta}_{\mathsf{F}} \in \breve{\mathcal{D}}_{\boldsymbol{\theta}_0}$.

Proof. From [145, Proposition 5(iii)], and the fact that a gauge is non-negative, we have,

$$J_f^\circ(\boldsymbol{\eta}_S) = \inf_{\tau\geq 0} \max(J^\circ(\tau f + \boldsymbol{\eta}_S), \tau) \leq J^\circ(\boldsymbol{\eta}_S) = \sigma_{\mathcal{C}}(\boldsymbol{\eta}_S) = \max_{\boldsymbol{v}\in\mathcal{C}} \langle \boldsymbol{\eta}_S, \boldsymbol{v}\rangle = \max_{\boldsymbol{v}\in\mathcal{V}_S} \langle \boldsymbol{\eta}, \boldsymbol{v}\rangle.$$

Thus, using a union bound and classical tail bounds of the Gaussian distribution, we get

$$\begin{aligned}
\mathbb{P}\left(J_f^\circ(\mathrm{P}_S(\boldsymbol{\eta}_{\mathsf{F}} - f)) \geq \nu \middle| \|\boldsymbol{\alpha}_{\mathsf{F}}\|_2 \leq \tau\right) &\leq \mathbb{P}\left(\max_{\boldsymbol{v}\in\mathcal{V}_S} \langle \boldsymbol{\eta}_{\mathsf{F}}, \boldsymbol{v}\rangle \geq \kappa \middle| \|\boldsymbol{\alpha}_{\mathsf{F}}\|_2 \leq \tau\right) \\
&\leq |\mathcal{V}_S| \max_{\boldsymbol{v}\in\mathcal{V}_S} \mathbb{P}\left(\langle \boldsymbol{\eta}_{\mathsf{F}}, \boldsymbol{v}\rangle \geq \kappa \middle| \|\boldsymbol{\alpha}_{\mathsf{F}}\|_2 \leq \tau\right) \\
&\leq |\mathcal{V}_S|/2 \max_{\boldsymbol{v}\in\mathcal{V}_S} e^{-\frac{\kappa^2}{2\tau^2\|\boldsymbol{v}\|_2^2}} \\
&\leq |\mathcal{V}_S|/2e^{-\frac{\kappa^2}{2\tau^2 \max_{\boldsymbol{v}\in\mathcal{V}_S}\|\boldsymbol{v}\|_2^2}}.
\end{aligned}$$

Set $q = n - d + 1$, which satisfies $q \geq 1$ under the restricted injectivity assumption. With the choice of τ devised above, we get

$$\mathbb{P}\left(J_f^\circ(\mathrm{P}_S(\boldsymbol{\eta}_{\mathsf{F}} - f)) \geq 1 \middle| \|\boldsymbol{\alpha}_{\mathsf{F}}\|_2 \leq \tau\right) \leq |\mathcal{V}_S|/2e^{-\frac{q-t}{2a}}. \tag{5.7}$$

Equating the arguments of the exponentials in (5.6)-(5.7), and solving

$$\frac{t^2}{4q} + \frac{t}{2a} - \left(\frac{q}{2a} - \log\left(\frac{|\mathcal{V}_S|}{2}\right)\right) = 0$$

for t to get equal probabilities, we obtain

$$t = \frac{q}{a}\left(\sqrt{1 + 2a\left(1 - \frac{2a\log\left(\frac{|\mathcal{V}_S|}{2}\right)}{q}\right)} - 1\right).$$

Setting

$$\delta = \frac{q}{2a\log\left(\frac{|\mathcal{V}_S|}{2}\right)},$$

we get under the bound on n that $\delta > 1$, and

$$t = 2\delta\log\left(\frac{|\mathcal{V}_S|}{2}\right)\left(\sqrt{1 + 2a\frac{\delta - 1}{\delta}} - 1\right).$$

Inserting t in one of the probability terms, and after basic algebraic rearrangements, we get the claimed probability of success. □

Let us now exemplify Proposition 5.4.3.

Example 5.4.4 (Lasso). Denote $I = \operatorname{supp}(\boldsymbol{\theta}_0)$. From (2.7), we have $\mathrm{P}_S f = 0$ hence $\kappa = \nu$, $\|e\|_2^2 = |I| = s$. Moreover, $\mathcal{V}$ is the set of unit-norm one-sparse vectors. Thus

$$|\mathcal{V}_S| = 2(p - s) \leq 2p, \quad d = s \quad \text{and} \quad \max_{\boldsymbol{v}\in\mathcal{V}_S} \|\boldsymbol{v}\|_2 = 1.$$

Taking

$$n \geq 2\delta\nu^{-2}s\log(p) + s, \quad \text{for some } \delta > 1,$$

we have

$$\|(\boldsymbol{\eta}_{\mathsf{F}})_{I^c}\|_\infty < \nu$$

with probability at least $1 - 2p^{-\left(\sqrt{\frac{\delta\nu^2}{2s} + \delta - 1} - \sqrt{\frac{\delta\nu^2}{2s}}\right)^2}$. The bound on n coincides with that of [23, Theorem 1.1].

Example 5.4.5 (Anti-sparsity). Recall the discussion of Section 2.4.3.4. Denote I^{sat} the saturation of $\boldsymbol{\theta}_0$ and set $s = |I^{\mathrm{sat}}|$. We assume $s \geq 2$ since $s = 1$ is trivial. The ℓ_∞ norm is the gauge of the hypercube whose vertices are $\mathcal{V} = \{\pm 1\}^p$. Thus

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

Moreover,

$$S = \operatorname{Span}\left((\boldsymbol{b}_i)_{i\in I^{\mathrm{sat}}}\right), \quad \dim(S) = s - 1 \quad \text{and} \quad d = p - s + 1,$$

where the j-th entry of $\boldsymbol{b}_i$ is

$$\begin{cases} 1 - \operatorname{sgn}((\boldsymbol{\theta}_0)_j)/s & \text{if } j = i, \\ -\operatorname{sgn}((\boldsymbol{\theta}_0)_j)/s & \text{if } i \neq j \in I^{\mathrm{sat}}, \\ 0 & \text{if } j \notin I^{\mathrm{sat}}. \end{cases}$$

The projection of the hypercube on S is a $(s-1)$-dimensional zonotope with at most s generators collinear to $(\boldsymbol{b}_i)_{i\in I^{\mathrm{sat}}}$. A classical exact upper bound on the number of vertices of a zonotope gives, see e.g. [65],

$$|\mathcal{V}_S| \leq 2\sum_{i=0}^{s-2}\binom{s-1}{i} = 2\left(\sum_{i=0}^{s-1}\binom{s-1}{i} - 1\right) = 2(2^{s-1} - 1) \leq 2^s,$$

where we used the binomial theorem.

On the other hand, we have $\mathrm{P}_S f = 0$ hence $\kappa = \nu$, and $\|e\|_2^2 = 1/s$. Moreover, a (crude) bound yields

$$\max_{\boldsymbol{v} \in \mathcal{V}_S} \|\boldsymbol{v}\|_2 \leq \sqrt{p}.$$

Thus, taking

$$n \geq (1 + 2\delta\nu^{-2}\log(2))p - s + 1, \quad \text{for some } \delta > 1,$$

(5.4) holds with probability at least $1 - 2\, 2^{-(s-1)\left(\sqrt{\frac{\delta\nu^2 s}{2p} + \delta - 1} - \sqrt{\frac{\delta\nu^2 s}{2p}}\right)^2}$. This is similar to the sample bound in [145, Theorem 7], and that by [58] who studied sample complexity thresholds for noiseless recovery from random projections of the hypercube. Though the proof argument of the former was tied to the structure of the subdifferential gauge J_f° associated to the ℓ_∞ norm.

5.4.2.2 Analysis-group Lasso

Consider the analysis-group Lasso penalty, see Section 2.4.3.3, with L blocks of equal size K. Let $I = \mathrm{supp}_{\mathcal{B}}(\boldsymbol{D}^\top \boldsymbol{\theta}_0)$, $s = |I|$ the number of active blocks in $\boldsymbol{D}^\top\boldsymbol{\theta}_0$, and $\Lambda = \bigcup_{i \in I} b_i$, and Λ^c its complement. We have

$$S = \mathrm{Span}(\boldsymbol{D}_{\Lambda^c}).$$

Moreover, [145, Proposition 10(iii)] tells us that

$$f = \boldsymbol{D} e^{\|\|_{1,2}}_{\boldsymbol{D}^\top\boldsymbol{\theta}_0} \quad \text{and}$$
$$J_f^\circ(\boldsymbol{\eta}_S) = \inf_{\boldsymbol{\omega} \in \mathrm{Ker}(\boldsymbol{D}_{\Lambda^c})} \max_{i \in I^c} \|(\boldsymbol{D}_{\Lambda^c}^+ \boldsymbol{\eta}_S + \boldsymbol{\omega})_{b_i}\|_2 \leq \max_{i \in I^c} \|(\boldsymbol{D}_{\Lambda^c}^+ \boldsymbol{\eta}_S)_{b_i}\|_2 = \max_{i \in I^c} \|(\boldsymbol{D}_{\Lambda^c}^+ \boldsymbol{\eta})_{b_i}\|_2,$$

where we used properties of the Moore-Penrose pseudo-inverse that yield

$$\boldsymbol{D}_{\Lambda^c}^+ \mathrm{P}_S = \boldsymbol{D}_{\Lambda^c}^+.$$

Denote $\boldsymbol{A}^i \overset{\text{def}}{=} \boldsymbol{D}_{\Lambda^c}^{+,*} \mathrm{P}_{b_i} \boldsymbol{D}_{\Lambda^c}^+$, which is a symmetric semidefinite positive matrix, and $r \overset{\text{def}}{=} \max_{i \in I^c} \mathrm{rank}(\boldsymbol{A}^i)$.

Proposition 5.4.6. *Assume that f is such that $\kappa \overset{def}{=} \nu - \max_{i \in I^c} \|f\|_{\boldsymbol{A}^i} > 0$. If $\boldsymbol{X}$ is drawn from the standard Gaussian ensemble with*

$$n \geq (1+\delta)\frac{\|\boldsymbol{D}\|_{2\to 2}^2 \|\boldsymbol{D}_{\Lambda^c}^+\|_{2\to 2}^2}{\kappa^2} s \left(K + 4\max\left(1, \sqrt{\frac{r}{4\log(L)}}\right)\log(L)\right) + \dim(\mathrm{Ker}(\boldsymbol{D}_{\Lambda^c}^\top)),$$

for some $\delta > 0$, then

$$J_f^\circ\big(\mathrm{P}_S(\boldsymbol{\eta}_{\mathsf{F}} - f)\big) < \nu \in]0, 1],$$

with probability at least

$$1 - \left(L^{-\delta/2} + L^{-\frac{\delta^2 \|\boldsymbol{D}\|_{2\to 2}^2 \|\boldsymbol{D}_{\Lambda^c}^+\|_{2\to 2}^2}{\kappa^2(4\delta+4)}}\right).$$

Thus, with high probability, (5.4) and the parameter estimation bounds of Theorem 5.3.6 hold with $\boldsymbol{\alpha}_0 = \boldsymbol{\alpha}_{\mathsf{F}}$ and $\boldsymbol{\eta}_0 = \boldsymbol{\eta}_{\mathsf{F}} \in \breve{\mathcal{D}}_{\boldsymbol{\theta}_0}$.

Proof. With a union bound, we have

$$\begin{aligned}
\mathbb{P}\left(J_f^\circ(\mathrm{P}_S(\boldsymbol{\eta}_{\mathsf{F}} - f)) \geq \nu \middle| \|\boldsymbol{\alpha}_{\mathsf{F}}\|_2 \leq \tau\right) &\leq \mathbb{P}\left(\max_{i \in I^c} \|(\boldsymbol{D}_{\Lambda^c}^+(\boldsymbol{\eta} - f))_{b_i}\|_2 \geq \nu \middle| \|\boldsymbol{\alpha}_{\mathsf{F}}\|_2 \leq \tau\right) \\
&\leq L \max_{i \in I^c} \mathbb{P}\left(\|\boldsymbol{\eta}\|_{\boldsymbol{A}^i} \geq \kappa \middle| \|\boldsymbol{\alpha}_{\mathsf{F}}\|_2 \leq \tau\right) \\
&\leq L \max_{i \in I^c} \mathbb{P}\left(\|\boldsymbol{Z}\|_{\boldsymbol{A}^i} \geq \kappa/\tau\right),
\end{aligned}$$

where $\boldsymbol{Z}$ is a centered standard Gaussian vector. We then need to bound the quadratic form $\|\boldsymbol{Z}\|_{\boldsymbol{A}^i}$, i.e. a Gaussian chaos of order 2. From [16, Example 2.12], we get

$$\begin{aligned}&\mathbb{P}\left(\|\boldsymbol{Z}\|_{\boldsymbol{A}^i}^2 - \mathbb{E}\left[\|\boldsymbol{Z}\|_{\boldsymbol{A}^i}^2\right] \geq \kappa^2/\tau^2 - \mathbb{E}\left[\|\boldsymbol{Z}\|_{\boldsymbol{A}^i}^2\right]\right)\\ &\leq \exp\left(-\min\left(\frac{\left(\kappa^2/\tau^2 - \mathbb{E}\left[\|\boldsymbol{Z}\|_{\boldsymbol{A}^i}^2\right]\right)^2}{4\|\boldsymbol{A}^i\|_{\mathrm{F}}^2}, \frac{\kappa^2/\tau^2 - \mathbb{E}\left[\|\boldsymbol{Z}\|_{\boldsymbol{A}^i}^2\right]}{4\|\boldsymbol{A}^i\|_{2\to 2}}\right)\right)\end{aligned}$$

with the proviso that $\kappa^2/\tau^2 \geq \mathbb{E}\left[\|\boldsymbol{Z}\|_{\boldsymbol{A}^i}^2\right]$. By the Von Neumann's trace inequality, we have

$$\mathbb{E}\left[\|\boldsymbol{Z}\|_{\boldsymbol{A}^i}^2\right] = \mathrm{tr}(\boldsymbol{A}^i) = \mathrm{tr}(\mathrm{P}_{b_i}\boldsymbol{D}^+_{\Lambda^c}\boldsymbol{D}^{+,*}_{\Lambda^c}) \leq K\|\boldsymbol{D}^+_{\Lambda^c}\|_{2\to 2}^2.$$

Furthermore

$$\|\boldsymbol{A}^i\|_{2\to 2} \leq \|\boldsymbol{D}^+_{\Lambda^c}\|_{2\to 2}^2 \quad \text{and} \quad \|\boldsymbol{A}^i\|_{\mathrm{F}}^2 \leq r\|\boldsymbol{A}^i\|_{2\to 2}^2 \leq r\|\boldsymbol{D}^+_{\Lambda^c}\|_{2\to 2}^4.$$

Altogether, this leads to

$$\mathbb{P}\left(\|\boldsymbol{Z}\|_{\boldsymbol{A}^i} \geq \kappa/\tau\right) \leq \exp\left(-\frac{1}{4}\min\left(r^{-1}\left(\kappa^2(\tau\|\boldsymbol{D}^+_{\Lambda^c}\|_{2\to 2})^{-2} - K\right)^2, \kappa^2(\tau\|\boldsymbol{D}^+_{\Lambda^c}\|_{2\to 2})^{-2} - K\right)\right)$$

provided that $\kappa^2/\tau^2 \geq K\|\boldsymbol{D}^+_{\Lambda^c}\|_{2\to 2}^2$. This is true with

$$\tau = \sqrt{\frac{1}{n-d+1-t}}\,\|e\|_2 \leq \sqrt{\frac{s}{n-d+1-t}}\|\boldsymbol{D}\|_{2\to 2},$$

the choice made for n and

$$t = \frac{\delta}{2}\frac{\|\boldsymbol{D}\|_{2\to 2}^2\|\boldsymbol{D}^+_{\Lambda^c}\|_{2\to 2}^2}{\kappa^2}s\left(K + 4\max\left(1, \sqrt{\frac{r}{4\log(L)}}\right)\log(L)\right).$$

Substituting appropriately, we arrive at

$$\begin{aligned}&\mathbb{P}\left(J_f^\circ(\mathrm{P}_S(\boldsymbol{\eta}_{\mathsf{F}} - f)) \geq \nu \middle| \|\boldsymbol{\alpha}_{\mathsf{F}}\|_2 \leq \tau\right)\\ &\leq L\exp\left(-\min\left(r^{-1}\left(2(1+\delta/2)\max\left(1, \sqrt{\frac{r}{4\log(L)}}\right)\log(L)\right)^2, (1+\delta/2)\max\left(1, \sqrt{\frac{r}{4\log(L)}}\right)\log(L)\right)\right)\\ &\leq L\exp\left(-\min\left((1+\delta/2)^2\log(L), (1+\delta/2)\log(L)\right)\right)\\ &\leq L\exp\left(-(1+\delta/2)\log(L)\right) \leq L^{-\delta/2},\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}\left(\|\boldsymbol{\alpha}_{\mathsf{F}}\|_2 \geq \tau\right) &\leq e^{-\frac{t^2}{4(n-d+1)}}\\ &\leq \exp\left(-\frac{\delta^2\|\boldsymbol{D}\|_{2\to 2}^2\|\boldsymbol{D}^+_{\Lambda^c}\|_{2\to 2}^2 s\left(K + 4\max\left(1, \sqrt{\frac{r}{4\log(L)}}\right)\log(L)\right)}{\kappa^2(16\delta+16)}\right)\\ &\leq \exp\left(-\frac{\delta^2\|\boldsymbol{D}\|_{2\to 2}^2\|\boldsymbol{D}^+_{\Lambda^c}\|_{2\to 2}^2\log(L)}{\kappa^2(4\delta+4)}\right)\\ &= L^{-\frac{\delta^2\|\boldsymbol{D}\|_{2\to 2}^2\|\boldsymbol{D}^+_{\Lambda^c}\|_{2\to 2}^2}{\kappa^2(4\delta+4)}}.\end{aligned}$$

Summing these the last two bounds gives the desired probability. □

When $\boldsymbol{D} = \mathbf{Id}$, i.e. the group Lasso, we have the following corollary.

Corollary 5.4.7. *If $\boldsymbol{X}$ is drawn from the standard Gaussian ensemble with*

$$n \geq (1+\delta)\nu^{-2} s \left(\sqrt{K} + 2\sqrt{\log(L)}\right)^2 + sK,$$

for some $\delta > 0$, then

$$\max_{i \in I^c} \|(\boldsymbol{\eta}_{\mathsf{F}})_{b_i}\|_2 < \nu$$

with probability at least

$$1 - \left(L^{-\delta/2} + L^{-\frac{\delta^2}{\nu^2(4\delta+4)}}\right).$$

We recover a sample bound similar to [23, Theorem 3.1].

Proof. First, observe that for the group Lasso, it can be straightforwardly checked that

$$\dim(\mathrm{Ker}(\boldsymbol{D}_{\Lambda^c}^\top) = |\Lambda| = sK, \ \|\boldsymbol{D}\|_{2\to 2} = \|\boldsymbol{D}_{\Lambda^c}^+\|_{2\to 2} = 1, \ \kappa = \nu \text{ and } r = K.$$

Under the sample lower-bound on n of the corollary, that of Proposition 5.4.6 is in force. We conclude applying the latter. □

5.4.2.3 Nuclear norm

We now turn to the case where J is the nuclear norm. Recall the notations of Section 2.4.3.5. For matrices $\boldsymbol{\theta} \in \mathbb{R}^{p_1 \times p_2}$, a Gaussian measurement map $\boldsymbol{X}$ takes the form of a linear operator whose i-th component is given by the Frobenius scalar product

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

where each matrix $\boldsymbol{X}^i \in \mathbb{R}^{p_1 \times p_2}$ is drawn from the standard Gaussian ensemble. For the nuclear norm, we have

$$\mathrm{P}_S(\boldsymbol{\theta}) = \boldsymbol{U}^\perp \boldsymbol{\theta} \boldsymbol{V}^\perp, \ \mathrm{P}_S(f) = \mathbf{0}, \text{ and } J_f^\circ\big(\mathrm{P}_S(\boldsymbol{\eta} - f)\big)(\boldsymbol{\eta}) = \left\|\boldsymbol{U}^\perp \boldsymbol{\eta} \boldsymbol{V}^\perp\right\|_{2\to 2},$$

where $\boldsymbol{\theta}_0 = \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{\sigma}(\boldsymbol{\theta}_0))\boldsymbol{V}^\top$ is a reduced rank-r SVD decomposition of $\boldsymbol{\theta}_0$, $\boldsymbol{U}^\perp = \mathbf{Id} - \boldsymbol{U}\boldsymbol{U}^\top$, $\boldsymbol{V}^\perp = \mathbf{Id} - \boldsymbol{V}\boldsymbol{V}^\top$, and $\mathbf{Id}$ is the identity operator on the space of $p_1 \times p_2$ matrices (should not be confused with the identity matrix).

We get the following results, whose proof is a slight modification of that of [23, Theorem 1.2].

Proposition 5.4.8. *Let $\boldsymbol{\theta}_0 \in \mathbb{R}^{p_1 \times p_2}$ be a rank-r matrix. If the Gaussian measurement map $\boldsymbol{X}$ is drawn with*

$$n \geq \delta r((1+2\nu^{-2})(p_1+p_2) - (1+4\nu^{-2})r), \quad \text{for some } \delta > 1,$$

then with probability at least $1 - 2e^{-\nu^2(\delta-1)\max(p_1,p_2)/8}$

$$\left\|\boldsymbol{U}^\perp \boldsymbol{\eta}_{\mathsf{F}} \boldsymbol{V}^\perp\right\|_{2\to 2} < \nu.$$

Observe that the sample bound ensures $n \geq d = r(p_1 + p_2 - r)$.

5.4.3 Bounds on the parameter estimates

Here, we are interested in how the bounds of Theorem 5.3.6 scale for each of the examples studied in the previous section.

In the above section, we have shown that for $\nu \in]0,1]$,

$$\frac{1}{1 - J_f^\circ\big(\mathrm{P}_S(\boldsymbol{\eta}_{\mathsf{F}} - f)\big)} \leq \frac{1}{1-\nu}$$

with high probability. To complete our analysis, we need to bound $\|\boldsymbol{X}_T^+\|_{2\to 2}$, $\|\mathrm{P}_S\|_{J_f \to 2}$, and $\|\boldsymbol{X}_S\|_{J_f \to 2}$. This will again be treated for each case separately.

5.4.3.1 Lasso

Denote $I = \mathrm{supp}(\boldsymbol{\theta}_0)$ and $s = |I|$. From (2.7), we have $\mathrm{P}_S f = 0$, $\|e\|_2^2 = |I| = s$, and $J_f = \|\cdot\|_1$. Thus

$$\|\mathrm{P}_S\|_{1\to 2} = \sup_{\boldsymbol{\theta}} \frac{\|\boldsymbol{\theta}_{I^c}\|_2}{\|\boldsymbol{\theta}_{I^c}\|_1} = 1.$$

In addition, we have for any $\boldsymbol{\theta} \in \mathbb{R}^p$

$$\frac{\|\sum_{i\in I^c} \boldsymbol{X}_i \boldsymbol{\theta}_i\|_2}{\|\boldsymbol{\theta}_{I^c}\|_1} \leq \max_{i\in I^c} \|\boldsymbol{X}_i\|_2 \frac{\sum_{i\in I^c} |\boldsymbol{\theta}_i|}{\|\boldsymbol{\theta}_{I^c}\|_1} = \max_{i\in I^c} \|\boldsymbol{X}_i\|_2,$$

whence we get the upper bound

$$\|\boldsymbol{X}_{I^c}\|_{1\to 2} \leq \max_{i\in I^c} \|\boldsymbol{X}_i\|_2.$$

By a union bound and Gaussian concentration of Lipschitz functions, it is immediate to show that

$$\max_{i\in I^c} \|\boldsymbol{X}_i\|_2 \leq \sqrt{n} + \sqrt{2t\log(p)}, \text{ for some } t > 1,$$

with probability at least $1 - p^{1-t}$.

Let us turn to bounding $\|\boldsymbol{X}_I^+\|_{2\to 2}$. Assume that $p \geq 2$. Arguing as in the Davidson-Szarek concentration inequality for the extreme singular values for Gaussian random matrices [54] (see also [139, Proposition 3.3]), we have for any $\epsilon > 0$,

$$\mathbb{P}\left(\|\boldsymbol{X}_I^+\|_{2\to 2} \leq 1/\epsilon\right) = \mathbb{P}\left(\boldsymbol{\sigma}_{\min}(\boldsymbol{X}_I) \geq \epsilon\right) \geq 1 - e^{-(\sqrt{n}-\sqrt{s}-\epsilon)^2/2}$$

provided that $n > (\sqrt{d}+\epsilon)^2$. This condition is in force for n obeying the bound in Example 5.4.4 and $\epsilon = \sqrt{n} - \sqrt{2s\log(p)} > 0$. We thus deduce that

$$\|\boldsymbol{X}_I^+\|_{2\to 2} \leq \frac{n^{-1/2}}{1 - \sqrt{\frac{2s\log(p)}{n}}} \leq \frac{1}{1-\delta^{-1/2}} n^{-1/2}$$

and

$$\|\boldsymbol{\alpha}_{\mathsf{F}}\|_2 \leq \|\boldsymbol{X}_I^+\|_{2\to 2} \|\mathrm{sgn}((\boldsymbol{\theta}_0)_I)\|_2 \leq \frac{1}{1-\delta^{-1/2}} \sqrt{\frac{s}{n}}$$

with probability exceeding $1 - e^{-(\sqrt{2\log(p)}-1)^2/2}$. Thus, choosing $t = \delta$, with probability at least $1 - (p^{1-\delta} + e^{-(\sqrt{2\log(p)}-1)^2/2})$ the following holds

$$\|\boldsymbol{X}_I^+\|_{2\to 2} \|\boldsymbol{X}_{I^c}\|_{1\to 2} \leq \frac{1}{1-\delta^{-1/2}} \left(1 + \sqrt{\frac{2\delta\log(p)}{n}}\right) \leq \frac{1}{1-\delta^{-1/2}} (1 + s^{-1/2}) \leq \frac{2}{1-\delta^{-1/2}}.$$

To sum up, we have proved the following.

Proposition 5.4.9. *Consider $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ with the Lasso penalty under the conditions of Theorem 5.3.6. If $\boldsymbol{X}$ is drawn from the standard Gaussian ensemble with*

$$n \geq 2\delta\nu^{-2} s\log(p) + s, \quad \text{for some } \delta > 1 \text{ and } \nu \in]0, 1],$$

then with probability at least

$$1 - \left(2p^{-\left(\sqrt{\frac{\delta\nu^2}{2s}+\delta-1}-\sqrt{\frac{\delta\nu^2}{2s}}\right)^2} + p^{1-\delta} + e^{-(\sqrt{2\log(p)}-1)^2/2}\right)$$

the following holds:

$$\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{\theta}_0\right\|_2 \leq \frac{1}{1-\delta^{-1/2}} n^{-1/2} \varphi^{-1}\left((C+1)\varphi^+\left(\|\boldsymbol{\zeta}\|_2\left(1 + c\frac{1}{1-\delta^{-1/2}}\sqrt{\frac{s}{n}}\right)\right)\right) + \frac{(C+2)\left(1+\frac{2}{1-\delta^{-1/2}}\right)}{c\|\boldsymbol{\zeta}\|_2(1-\nu)}\varphi^+\left(\|\boldsymbol{\zeta}\|_2\left(1 + c\frac{1}{1-\delta^{-1/2}}\sqrt{\frac{s}{n}}\right)/2\right),$$

$$\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{\theta}_0\right\|_2 \leq \frac{1}{1-\delta^{-1/2}} n^{-1/2} \varphi^{-1}\left(\varphi^+\left(\|\boldsymbol{\zeta}\|_2\left(1 + c\frac{1}{1-\delta^{-1/2}}\sqrt{\frac{s}{n}}\right)\right)\right) + \frac{2\left(1+\frac{2}{1-\delta^{-1/2}}\right)}{c\|\boldsymbol{\zeta}\|_2(1-\nu)}\varphi^+\left(\|\boldsymbol{\zeta}\|_2\left(1 + c\frac{1}{1-\delta^{-1/2}}\sqrt{\frac{s}{n}}\right)/2\right).$$

In the setting of Example 5.3.7, these bounds read

$$\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{\theta}_0\right\|_2 \leq \left[\frac{1}{1-\delta^{-1/2}} n^{-1/2}(q-1)^{1/q}(C+1)^{1/q} 2^{\frac{1}{(1-q)}} + \frac{\frac{q-1}{cq}(C+2)^{\frac{q}{1-q}}\left(1+\frac{2}{1-\delta^{-1/2}}\right)\left(1+c\frac{1}{1-\delta^{-1/2}}\sqrt{\frac{s}{n}}\right)}{1-\nu}\right]\left(\left(1 + c\frac{1}{1-\delta^{-1/2}}\sqrt{\frac{s}{n}}\right)\|\boldsymbol{\zeta}\|_2\right)^{\frac{1}{q-1}},$$

$$\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{\theta}_0\right\|_2 \leq \left[\frac{1}{1-\delta^{-1/2}} n^{-1/2}(q-1)^{1/q} 2^{\frac{1}{(1-q)}} + \frac{\frac{q-1}{cq}2^{\frac{q}{1-q}}\left(1+\frac{2}{1-\delta^{-1/2}}\right)\left(1+c\frac{1}{1-\delta^{-1/2}}\sqrt{\frac{s}{n}}\right)}{1-\nu}\right]\left(\left(1 + c\frac{1}{1-\delta^{-1/2}}\sqrt{\frac{s}{n}}\right)\|\boldsymbol{\zeta}\|_2\right)^{\frac{1}{q-1}}.$$

For $q = 2$, this scaling is consistent with what is encountered in the compressed sensing literature.

5.4.3.2 Anti-sparsity

Let I^{sat} be the saturation of $\boldsymbol{\theta}_0$ and $s = |I^{\mathrm{sat}}|$. Recall that $(e)_{I^{\mathrm{sat}}} = \mathrm{sgn}((\boldsymbol{\theta}_0)_I^{\mathrm{sat}})/s$ and 0 otherwise. Since $\mathrm{P}_S f = 0$, $J_f = \|\cdot\|_\infty$ on S; see [145, Proposition 5(iv)]. Moreover, S is contained in the linear subspace of vectors supported on I^{sat}. In turn

$$\|\mathrm{P}_S\|_{\infty\to 2} = \sup_{\boldsymbol{\theta}\in S}\frac{\|\boldsymbol{\theta}\|_2}{\|\boldsymbol{\theta}\|_\infty} \leq \sup_{\boldsymbol{\theta}}\frac{\|\boldsymbol{\theta}_{I^{\mathrm{sat}}}\|_2}{\|\boldsymbol{\theta}_{I^{\mathrm{sat}}}\|_\infty} = \sqrt{s}.$$

Similarly, we have

$$\|\boldsymbol{X}_S\|_{J_f\to 2} = \sup_{\boldsymbol{\theta}}\frac{\|\boldsymbol{X}_{I^{\mathrm{sat}}}\boldsymbol{\theta}_S\|_2}{\|\boldsymbol{\theta}_S\|_\infty} \leq \|\boldsymbol{X}_{I^{\mathrm{sat}}}\|_{2\to 2}\|\mathrm{P}_S\|_{\infty\to 2} \leq \sqrt{s}\|\boldsymbol{X}_{I^{\mathrm{sat}}}\|_{2\to 2}.$$

In view of the bound on n in Example 5.4.5, we then apply the concentration inequality for the largest singular value of the Gaussian matrix $\boldsymbol{X}_{I^{\mathrm{sat}}}$ to show that

$$\|\boldsymbol{X}_{I^{\mathrm{sat}}}\|_{2\to 2} \leq \delta(\sqrt{n} + \sqrt{s}), \quad \delta > 1,$$

with probability at least $1 - p^{1-\delta}$. Following the same steps as for the Lasso (see previous section), we also have

$$\|\boldsymbol{X}_T^+\|_{2\to 2} \leq C_\delta n^{-1/2}$$

and

$$\|\boldsymbol{\alpha}_{\mathsf{F}}\|_2 \leq \|\boldsymbol{X}_T^+\|_{2\to 2} s^{-1/2} \leq C_\delta\sqrt{\frac{1}{sn}}$$

with high probability for some constant C_δ that depends on δ. We then have

$$\|\boldsymbol{X}_T^+\|_{2\to 2}\|\boldsymbol{X}_S\|_{J_f\to 2} \leq C_\delta\delta\left(1 + \sqrt{\frac{s}{n}}\right)\sqrt{s}$$

with large probability. Altogether, we have the following estimation bounds.

Proposition 5.4.10. *Consider* $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ *and* $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ *with the* ℓ_∞ *penalty under the conditions of Theorem 5.3.6. If* $\boldsymbol{X}$ *is drawn from the standard Gaussian ensemble with*

$$n \geq (1 + 2\delta\nu^{-2}\log(2))p - s + 1, \quad \text{for some } \delta > 1 \text{ and } \nu \in]0,1],$$

then with large probability the following holds:

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{\theta}_0\right\|_2 &\leq C_\delta n^{-1/2}\varphi^{-1}\left((C+1)\varphi^+\left(\|\boldsymbol{\zeta}\|_2\left(1 + cC_\delta\sqrt{\frac{1}{sn}}\right)\right)\right) \\
&\quad + \sqrt{s}\frac{2(C+2)\left(1 + \delta C_\delta\left(1 + \sqrt{\frac{s}{n}}\right)\right)}{c\|\boldsymbol{\zeta}\|_2(1-\nu)}\varphi^+\left(\|\boldsymbol{\zeta}\|_2\left(1 + cC_\delta\sqrt{\frac{1}{sn}}\right)/2\right), \\
\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{\theta}_0\right\|_2 &\leq C_\delta n^{-1/2}\varphi^{-1}\left(\varphi^+\left(\|\boldsymbol{\zeta}\|_2\left(1 + cC_\delta\sqrt{\frac{1}{sn}}\right)\right)\right) \\
&\quad + \sqrt{s}\frac{4\left(1 + \delta C_\delta\left(1 + \sqrt{\frac{s}{n}}\right)\right)}{c\|\boldsymbol{\zeta}\|_2(1-\nu)}\varphi^+\left(\|\boldsymbol{\zeta}\|_2\left(1 + cC_\delta\sqrt{\frac{1}{sn}}\right)/2\right).
\end{aligned}$$

In the context of Example 5.3.7, these bounds specialize to

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{\theta}_0\right\|_2 &\leq \left[C_\delta n^{-1/2}(q-1)^{1/q}(C+1)^{1/q}2^{\frac{1}{(1-q)}}\right. \\
&\quad \left. + \sqrt{s}\frac{\frac{q-1}{cq}(C+2)^{\frac{q}{1-q}}2(C+2)\left(1 + \delta C_\delta\left(1 + \sqrt{\frac{s}{n}}\right)\right)\left(1 + cC_\delta\sqrt{\frac{1}{sn}}\right)}{1-\nu}\right] \\
&\quad \left(\left(1 + cC_\delta\sqrt{\frac{1}{sn}}\right)\|\boldsymbol{\zeta}\|_2\right)^{\frac{1}{q-1}}, \\
\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{\theta}_0\right\|_2 &\leq \left[C_\delta n^{-1/2}(q-1)^{1/q}2^{\frac{1}{(1-q)}}\right. \\
&\quad \left. + \sqrt{s}\frac{\frac{q-1}{cq}(C+2)^{\frac{q}{1-q}}4\left(1 + \delta C_\delta\left(1 + \sqrt{\frac{s}{n}}\right)\right)\left(1 + cC_\delta\sqrt{\frac{1}{sn}}\right)}{1-\nu}\right]\left(\left(1 + cC_\delta\sqrt{\frac{1}{sn}}\right)\|\boldsymbol{\zeta}\|_2\right)^{\frac{1}{q-1}}.
\end{aligned}$$

The dominant term is the one that scales at least as $O(\sqrt{s})$, which is also consistent with the bounds we got for the prediction oracle inequalities in the previous chapter. This reflects the anticipated intuitive idea that the larger the saturation support, the more difficult the estimation.

5.4.3.3 Group Lasso

We here discuss the group Lasso for the sake of simplicity, but a similar bound can be derived for the analysis version. Recall the group Lasso penalty (Section 2.4.3.2) with L blocks of equal size K. Let $I = \mathrm{supp}_{\mathcal{B}}(\boldsymbol{\theta}_0)$, $s = |I|$ the number of active blocks in $\boldsymbol{\theta}_0$, and $\Lambda = \bigcup_{i\in I} b_i$, and Λ^c its complement. By the triangle inequality, we can upper-bound

$$\frac{\|\boldsymbol{\theta}_{\Lambda^c}\|_2}{\|\boldsymbol{\theta}_{\Lambda^c}\|_{1,2}} \leq \frac{\sum_{i\in I^c}\|\boldsymbol{\theta}_{b_i}\|_2}{\sum_{i\in I^c}\|\boldsymbol{\theta}_{b_i}\|_2} = 1,$$

and thus

$$\|\,\mathrm{P}_S\,\|_{(1,2)\to 2} = \sup_{\boldsymbol{\theta}}\frac{\|\boldsymbol{\theta}_{\Lambda^c}\|_2}{\|\boldsymbol{\theta}_{\Lambda^c}\|_{1,2}} = 1.$$

Moreover, for any $\boldsymbol{\theta} \in \mathbb{R}^p$

$$\frac{\left\|\sum_{i\in I^c} \boldsymbol{X}_{b_i}\boldsymbol{\theta}_{b_i}\right\|_2}{\sum_{i\in I^c}\left\|\boldsymbol{\theta}_{b_i}\right\|_2} \leq \frac{\sum_{i\in I^c}\left\|\boldsymbol{X}_{b_i}\right\|_{2\to 2}\boldsymbol{\theta}_{b_i}}{\sum_{i\in I^c}\left\|\boldsymbol{\theta}_{b_i}\right\|_2} \leq \max_{i\in I^c}\left\|\boldsymbol{X}_{b_i}\right\|_{2\to 2}.$$

This yields

$$\left\|\boldsymbol{X}_{\Lambda^c}\right\|_{(1,2)\to 2} \leq \max_{i\in I^c}\left\|\boldsymbol{X}_{b_i}\right\|_{2\to 2}.$$

Under the sample bound on n in Corollary 5.4.7, we can invoke the concentration of the largest singular value of $\left\|\boldsymbol{X}_{b_i}\right\|_{2\to 2}$ and a union bound to get

$$\max_{i\in I^c}\left\|\boldsymbol{X}_{b_i}\right\|_{2\to 2} \leq \sqrt{n}+\sqrt{K}+\sqrt{2t\log(L)}, \text{ for some } t>1,$$

with probability at least $1-L^{1-t}$.

To bound $\left\|\boldsymbol{X}_\Lambda^+\right\|_{2\to 2}$, we use now concentration of the smallest singular value of $\boldsymbol{X}_\Lambda$, and we obtain

$$\left\|\boldsymbol{X}_\Lambda^+\right\|_{2\to 2} \leq \frac{1}{\sqrt{n}-\sqrt{sK}-2\sqrt{s\log(L)}} \leq \frac{n^{-1/2}}{1-(1+\delta)^{-1/2}}$$

with probability larger than $1-L^{-2}$, where we used the sample bound of Corollary 5.4.7 in the last inequality. It then follows that

$$\left\|\boldsymbol{\alpha}_{\mathsf{F}}\right\|_2 \leq \left\|\boldsymbol{X}_\Lambda^+\right\|_{2\to 2}\sqrt{s} \leq \frac{1}{1-(1+\delta)^{-1/2}}\sqrt{\frac{s}{n}}.$$

Choosing $t=\sqrt{1+\delta}$, $\delta>0$, and using again the sample bound of Corollary 5.4.7, we have

$$\left\|\boldsymbol{X}_\Lambda^+\right\|_{2\to 2}\left\|\boldsymbol{X}_{\Lambda^c}\right\|_{(1,2)\to 2} \leq \frac{2}{1-(1+\delta)^{-1/2}}$$

with probability exceeding $1-(L^{1-\sqrt{1+\delta}}+L^{-2})$. Combining this discussion with Theorem 5.3.6 and Corollary 5.4.7, we have proved the following.

Proposition 5.4.11. *Consider $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$ with the group Lasso penalty under the conditions of Theorem 5.3.6. If $\boldsymbol{X}$ is drawn from the standard Gaussian ensemble with*

$$n \geq (1+\delta)\nu^{-2}s\left(\sqrt{K}+2\sqrt{\log(L)}\right)^2 + sK, \text{ for some } \delta>0,$$

then with probability at least

$$1-\left(L^{-\delta/2}+L^{-\frac{\delta^2}{\nu^2(4\delta+4)}}+L^{1-\sqrt{1+\delta}}+L^{-2}\right)$$

the following holds:

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}-\boldsymbol{\theta}_0\right\|_2 &\leq \frac{1}{1-(1+\delta)^{-1/2}}n^{-1/2}\varphi^{-1}\left((C+1)\varphi^+\left(\|\boldsymbol{\zeta}\|_2\left(1+c\frac{1}{1-(1+\delta)^{-1/2}}\sqrt{\frac{s}{n}}\right)\right)\right)\\
&\quad+\frac{(C+2)\left(1+\frac{2}{1-(1+\delta)^{-1/2}}\right)}{c\|\boldsymbol{\zeta}\|_2(1-\nu)}\varphi^+\left(\|\boldsymbol{\zeta}\|_2\left(1+c\frac{1}{1-(1+\delta)^{-1/2}}\sqrt{\frac{s}{n}}\right)/2\right),\\
\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}-\boldsymbol{\theta}_0\right\|_2 &\leq \frac{1}{1-(1+\delta)^{-1/2}}n^{-1/2}\varphi^{-1}\left(\varphi^+\left(\|\boldsymbol{\zeta}\|_2\left(1+c\frac{1}{1-(1+\delta)^{-1/2}}\sqrt{\frac{s}{n}}\right)\right)\right)\\
&\quad+\frac{2\left(1+\frac{2}{1-(1+\delta)^{-1/2}}\right)}{c\|\boldsymbol{\zeta}\|_2(1-\nu)}\varphi^+\left(\|\boldsymbol{\zeta}\|_2\left(1+c\frac{1}{1-(1+\delta)^{-1/2}}\sqrt{\frac{s}{n}}\right)/2\right).
\end{aligned}$$

For the data loss of Example 5.3.7, these bounds become

$$\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{\theta}_0\right\|_2 \leq \left[\frac{1}{1-(1+\delta)^{-1/2}} n^{-1/2} (q-1)^{1/q} (C+1)^{1/q} 2^{\frac{1}{(1-q)}} + \frac{\frac{q-1}{cq}(C+2)^{\frac{q}{1-q}} \left(1+\frac{2}{1-(1+\delta)^{-1/2}}\right)\left(1+c\frac{1}{1-(1+\delta)^{-1/2}}\sqrt{\frac{s}{n}}\right)}{1-\nu}\right] \left(\left(1+c\frac{1}{1-(1+\delta)^{-1/2}}\sqrt{\frac{s}{n}}\right)\|\boldsymbol{\zeta}\|_2\right)^{\frac{1}{q-1}},$$

$$\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{\theta}_0\right\|_2 \leq \left[\frac{1}{1-(1+\delta)^{-1/2}} n^{-1/2} (q-1)^{1/q} 2^{\frac{1}{(1-q)}} + \frac{\frac{q-1}{cq}2^{\frac{q}{1-q}} \left(1+\frac{2}{1-(1+\delta)^{-1/2}}\right)\left(1+c\frac{1}{1-(1+\delta)^{-1/2}}\sqrt{\frac{s}{n}}\right)}{1-\nu}\right] \left(\left(1+c\frac{1}{1-(1+\delta)^{-1/2}}\sqrt{\frac{s}{n}}\right)\|\boldsymbol{\zeta}\|_2\right)^{\frac{1}{q-1}}.$$

5.4.3.4 Nuclear norm

We follow the notation of Section 5.4.2.3. We have

$$\|\mathrm{P}_S\|_{*\to F} = \sup_{\boldsymbol{\theta}\in S} \frac{\|\boldsymbol{\theta}\|_{\mathrm{F}}}{\|\boldsymbol{\theta}\|_*} = \sup_{\boldsymbol{\theta}\in S} \frac{\|\boldsymbol{\sigma}(\boldsymbol{\theta})\|_2}{\|\boldsymbol{\sigma}(\boldsymbol{\theta})\|_1} = 1.$$

Moreover,

$$\|\boldsymbol{X}\,\mathrm{P}_S\|_{*\to 2} = \|\mathrm{P}_S\,\boldsymbol{X}^*\|_{2\to S_\infty} = \sup_{\boldsymbol{u}\in\mathbb{R}^n} \frac{\|\sum_{i=1}^n \boldsymbol{u}_i\,\mathrm{P}_S(\boldsymbol{X}^i)\|_{2\to 2}}{\|\boldsymbol{u}\|_2}.$$

We now argue that $\mathrm{P}_S(\boldsymbol{X}^i) = \boldsymbol{U}^\perp \boldsymbol{X}^i \boldsymbol{V}^\perp$, and the $\boldsymbol{X}^i$'s are independent $p_1 \times p_2$ random matrix with i.i.d. standard Gaussian entries. It follows that $\boldsymbol{Z} \stackrel{\mathrm{def}}{=} \sum_{i=1}^n \boldsymbol{u}_i\,\mathrm{P}_S(\boldsymbol{X}^i)$ is identically distributed to a rotation of an $(p_1-r)\times(p_2-r)$ Gaussian random matrix whose entries are i.i.d. $\mathcal{N}(0, n\|\boldsymbol{u}\|_2^2)$. Applying the concentration inequality for the largest singular value of $\boldsymbol{Z}$, we get

$$\frac{\|\sum_{i=1}^n \boldsymbol{u}_i\,\mathrm{P}_S(\boldsymbol{X}^i)\|_{2\to 2}}{\|\boldsymbol{u}\|_2} \leq \epsilon\sqrt{n}$$

with probability larger than $1 - e^{-\frac{(\epsilon-\sqrt{p_1-r}-\sqrt{p_2-r})^2}{2}}$. Set $\epsilon = 2t\sqrt{p-r}$ for some $t > 1$, where $p = \max(p_1, p_2)$. Since $\sqrt{a}+\sqrt{b} \leq \sqrt{2(a+b)}$, and owing to the sample bound of Proposition 5.4.8, we can ensure that

$$\|\boldsymbol{X}\,\mathrm{P}_S\|_{*\to 2} \leq 2t\sqrt{n(p-r)},$$

with high probability. One could define a basis for T and write $\boldsymbol{X}_T$ as an $n \times d$ dimensional standard Gaussian matrix. Consequently, and using exactly the same argument as for the previous penalties, one can show that

$$\|\boldsymbol{X}_T^+\|_{F\to F} \leq C_\delta n^{-1/2},$$

and

$$\|\boldsymbol{\alpha}_{\mathsf{F}}\|_F \leq C_\delta\sqrt{\frac{r}{n}},$$

with large probability, where C_δ is a positive constant that depends on $\delta > 1$ in Proposition 5.4.8. Consequently, taking $t = \delta$, we get

$$\|\boldsymbol{X}_T^+\|_{F\to F}\|\boldsymbol{X}\,\mathrm{P}_S\|_{*\to 2} \leq 2C_\delta\delta\sqrt{p-r}.$$

Putting together the above bounds, we have the following claim.

Proposition 5.4.12. *Let $\boldsymbol{\theta}_0 \in \mathbb{R}^{p_1 \times p_2}$ be a rank-r matrix. If the Gaussian measurement map $\boldsymbol{X}$ is drawn with*

$$n \geq \delta r((1+2\nu^{-2})(p_1+p_2) - (1+4\nu^{-2})r), \quad \text{for some } \delta > 1,$$

then with large probability, the following holds:

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{\theta}_0\right\|_{\mathrm{F}} &\leq C_\delta n^{-1/2} \varphi^{-1}\left((C+1)\varphi^+\left(\|\boldsymbol{\zeta}\|_2\left(1+cC_\delta\sqrt{\frac{r}{n}}\right)\right)\right) \\
&\quad + \frac{(C+2)\left(1+2C_\delta\delta\sqrt{p-r}\right)}{c\|\boldsymbol{\zeta}\|_2(1-\nu)}\varphi^+\left(\|\boldsymbol{\zeta}\|_2\left(1+cC_\delta\sqrt{\frac{r}{n}}\right)/2\right), \\
\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{\theta}_0\right\|_{\mathrm{F}} &\leq C_\delta n^{-1/2} \varphi^{-1}\left(\varphi^+\left(\|\boldsymbol{\zeta}\|_2\left(1+cC_\delta\sqrt{\frac{r}{n}}\right)\right)\right) \\
&\quad + \frac{2\left(1+2C_\delta\delta\sqrt{p-r}\right)}{c\|\boldsymbol{\zeta}\|_2(1-\nu)}\varphi^+\left(\|\boldsymbol{\zeta}\|_2\left(1+cC_\delta\sqrt{\frac{r}{n}}\right)/2\right).
\end{aligned}$$

Revisiting Example 5.3.7 with these bounds, we obtain

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} - \boldsymbol{\theta}_0\right\|_2 &\leq \left[C_\delta n^{-1/2}(q-1)^{1/q}(C+1)^{1/q}2^{\frac{1}{(1-q)}}\right. \\
&\quad \left. + \frac{\frac{q-1}{cq}(C+2)^{\frac{q}{1-q}}\left(1+2cC_\delta\delta\sqrt{p-r}\right)\left(1+cC_\delta\sqrt{\frac{r}{n}}\right)}{1-\nu}\right]\left(\left(1+cC_\delta\sqrt{\frac{r}{n}}\right)\|\boldsymbol{\zeta}\|_2\right)^{\frac{1}{q-1}}, \\
\left\|\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} - \boldsymbol{\theta}_0\right\|_2 &\leq \left[C_\delta n^{-1/2}(q-1)^{1/q}2^{\frac{1}{(1-q)}}\right. \\
&\quad \left. + \frac{\frac{q-1}{cq}2^{\frac{q}{1-q}}\left(1+2cC_\delta\delta\sqrt{p-r}\right)\left(1+cC_\delta\sqrt{\frac{r}{n}}\right)}{1-\nu}\right]\left(\left(1+cC_\delta\sqrt{\frac{r}{n}}\right)\|\boldsymbol{\zeta}\|_2\right)^{\frac{1}{q-1}}.
\end{aligned}$$

5.4.4 Beyond Gaussian design

One may wonder whether the above results can be extended beyond Gaussian designs. For instance, adapting the arguments of [23], the results for polyhedral regularization and the (analysis) group Lasso can be extended to matrices whose entries are i.i.d. sub-Gaussian (see [23]). The situation is however much more intricate for the nuclear norm.

Another approach, which still to be investigated, is to go beyond the dual certificate $\boldsymbol{\eta}_{\mathsf{F}}$. The reasoning chain we are thinking of, which is inspired by [143] (see also [146]), is to investigate the connections between existence of a dual certificate satisfying (5.4), and injectivity of $\boldsymbol{X}$ on the so-called descent cone of J at $\boldsymbol{\theta}_0$. As advocated in [37], the latter property can be characterized very sharply via the Gaussian width of the descent cone of J (restricted to a sphere). The Gaussian width is closely related to another geometric quantity called the statistical dimension in conic integral geometry, which has been investigated in the context of recovery from Gaussian measurements in [1], and later extended in [139] for sub-Gaussian designs and even beyond.

Part II

Algorithms

Chapter 6

EWA for Non-smooth Priors through Langevin Diffusion and Proximal Splitting

Main contributions of this chapter

- Propose two algorithms based on forward-backward proximal splitting to sample from non-smooth distributions which are not necessarily differentiable nor log-concave. Their perfomances are verified by theoretical consistency guarantees.
- Apply these algorithms to compute the EWA with several popular penalties in the literature.

The results in this chapter can be found in [100].

Contents

6.1 Introduction

In this chapter, we propose proximal splitting-type algorithms for sampling from distributions whose densities are not necessarily smooth nor log-concave. Our approach brings together tools from, on the one hand, variational analysis and non-smooth optimization, and on the other hand, stochastic differential equations (SDE), and in particular the Langevin diffusion. We establish in particular consistency guarantees of our algorithms seen as discretization schemes in this context. These algorithms are then applied to compute the exponentially weighted aggregates for regression problems involving non-smooth priors encouraging some notion of simplicity/complexity.

6.1.1 Problem statement

We consider the EWA (1.12) with $\Theta = \mathbb{R}^p$ and $J_{\boldsymbol{\lambda}_n} = \frac{1}{n} J_{\boldsymbol{\lambda}}$ where $J_{\boldsymbol{\lambda}} : \mathbb{R}^p \to \mathbb{R}$ is the regularizing penalty promoting some specific notion of simplicity/low-complexity which depends on a vector of parameters $\boldsymbol{\lambda}$. Hence, the aggregators are defined via the probability density function

$$\widehat{\mu}_n(\boldsymbol{\theta}) = \frac{\exp\left(-V(\boldsymbol{\theta})/(n\beta)\right)}{\int_{\mathbb{R}^p} \exp\left(-V(\boldsymbol{\omega})/(n\beta)\right) d\boldsymbol{\omega}},$$

where $V(\boldsymbol{\theta}) \stackrel{\text{def}}{=} F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) + J_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$. Without loss of generality, we consider $n\beta$ as β.

In this chapter, we focus on the computation of EWA using the Langevin diffusion (see Section 1.1.5) with a large family of priors $\widehat{\mu}_n$ which are not necessarily differentiable nor log-concave.

6.1.2 Chapter organization

Consider the SDE (1.13) with Moreau-Yosida regularized version of μ under mild assumptions of the latter. Well-posedness of this SDE and consistency guarantees for its discrete approximations are proven in Section 6.2. Section 6.3 provides a large class of functions, namely prox-regular functions, for which the previous theoretical analysis applies. From this analysis, two algorithms are derived in Section 6.4 and applied in Section 6.5 to compute the EWA with several penalties.

6.2 Langevin diffusion with Moreau-Yosida regularization

We aim to enlarge the family of μ covered by [52, 111, 62, 63] by relaxing some underlying conditions. Especially, μ is not necessarily differentiable nor log-concave. Namely, denote $\widetilde{C^{1,+}}(\mathbb{R}^d)$ the class of differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$ whose gradient is locally Lipschitz continuous and there exists $K > 0$ such that

$$\langle \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle \leq K(1 + \|\boldsymbol{x}\|_2^2), \quad \forall \boldsymbol{x} \in \mathbb{R}^d,$$

our target distributions μ is defined as

$$\mu(\boldsymbol{\theta}) \propto \exp\Big(-\Big(L(\boldsymbol{\theta}) + H \circ \boldsymbol{D}^\top(\boldsymbol{\theta})\Big)\Big), \tag{6.1}$$

where $\boldsymbol{D} \in \mathbb{R}^{p\times q}$, $H : \mathbb{R}^q \to \mathbb{R}$ and $L \in \widetilde{C^{1,+}}(\mathbb{R}^p)$. To avoid trivialities, from now on, we assume that $\operatorname{Argmin}(H) \neq \emptyset$. In our framework, let $\boldsymbol{M}$ be a symmetric definite positive matrix, we impose the following assumptions on H.

(**H.1**) $H \in \mathcal{J}(\mathbb{R}^q)$.

(**H.2**) $\operatorname{prox}_{\gamma H}^{\boldsymbol{M}}$ is single-valued.

(**H.3**) $\operatorname{prox}_{\gamma H}^{\boldsymbol{M}}$ is locally Lipschitz continuous.

(**H.4**) There exists $C > 0$ such that

$$\Big\langle \boldsymbol{D}^\top\boldsymbol{\theta}, \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{D}^\top\boldsymbol{\theta}) \Big\rangle_{\boldsymbol{M}} \leq C(1 + \|\boldsymbol{\theta}\|_2^2), \quad \forall \boldsymbol{\theta} \in \mathbb{R}^p.$$

Let us define the following SDE with the Moreau-Yosida regularized version of H

$$\begin{aligned} d\boldsymbol{L}(t) &= \boldsymbol{\psi}(\boldsymbol{L}(t))dt + d\boldsymbol{W}(t), \ t > 0, \\ \text{where} \quad \boldsymbol{\psi} : \boldsymbol{\theta} \in \mathbb{R}^p &\mapsto -\frac{1}{2}\nabla\big(L + ({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top\big)(\boldsymbol{\theta}), \end{aligned} \tag{6.2}$$

$\boldsymbol{\psi}$ is the drift coefficient.

Recall that Assumptions (**H.1**) and (**H.2**) are mild assumptions required to establish key properties of Moreau-Yosida regularization stated in Lemmas 2.5.5 and 2.5.4. Lemma 2.5.5 allows us to compute $\nabla\, {}^{\boldsymbol{M},\gamma}H$ by exploiting its the relation between $\nabla\, {}^{\boldsymbol{M},\gamma}H$ and $\operatorname{prox}_{\gamma H}^{\boldsymbol{M}}$.

To guarantee well-posedness (existence and uniqueness) and discretization consistency of the SDE (6.2), we will also need Assumptions (**H.3**) and (**H.4**).

6.2.1 Well-posedness

We start with the following characterization of the drift $\boldsymbol{\psi}$.

Proposition 6.2.1. *Assume that Assumptions (**H.1**), (**H.2**), (**H.3**) and (**H.4**) hold. Then,*

$$\langle \boldsymbol{\psi}(\boldsymbol{\theta}), \boldsymbol{\theta} \rangle \leq K(1 + \|\boldsymbol{\theta}\|_2^2), \text{ for some } K > 0, \tag{6.3}$$

and

$$\boldsymbol{\psi} \text{ is locally Lipschitz continuous.} \tag{6.4}$$

Proof. In view of Lemma 2.5.5, the drift term reads

$$\boldsymbol{\psi}(\boldsymbol{\theta}) = -\frac{1}{2}\nabla(L + ({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top)(\boldsymbol{\theta}) = -\frac{1}{2}\nabla L(\boldsymbol{\theta}) - \frac{1}{2\gamma}\boldsymbol{D}\boldsymbol{M}\boldsymbol{D}^\top\boldsymbol{\theta} + \frac{1}{2\gamma}\boldsymbol{D}\boldsymbol{M}\operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{D}^\top\boldsymbol{\theta}).$$

Since $L \in \widetilde{C^{1,+}}(\mathbb{R}^p)$ and Assumption (H.4) holds, there exist $K_1 > 0$ and $K_2 > 0$ such that

$$\begin{aligned}\langle \boldsymbol{\psi}(\boldsymbol{\theta}), \boldsymbol{\theta} \rangle &= -\frac{1}{2}\langle \nabla L(\boldsymbol{\theta}), \boldsymbol{\theta} \rangle - \frac{1}{2\gamma} \left\| \boldsymbol{D}^\top \boldsymbol{\theta} \right\|_{\boldsymbol{M}}^2 + \frac{1}{2} \left\langle \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{D}^\top \boldsymbol{\theta}), \boldsymbol{D}^\top \boldsymbol{\theta} \right\rangle_{\boldsymbol{M}} \\ &\leq K_1(1 + \left\| \boldsymbol{\theta} \right\|_2^2) + \frac{\|\boldsymbol{D}\|_{2\to 2}{}^2 \|\boldsymbol{M}\|_{2\to 2} \left\| \boldsymbol{\theta} \right\|_2^2}{2\gamma} + K_2(1 + \left\| \boldsymbol{\theta} \right\|_2^2) \\ &\leq K(1 + \left\| \boldsymbol{\theta} \right\|_2^2),\end{aligned}$$

where $K \geq K_1 + K_2 + \frac{\|\boldsymbol{D}\|_{2\to 2}{}^2 \|\boldsymbol{M}\|_{2\to 2}}{2\gamma}$. Moreover, under Assumption (H.3), $({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top$ is locally Lipschitz continuous by virtue of Lemma 2.5.5, which applies thanks to Assumptions (H.1) and (H.2). Clearly $({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top \in \widetilde{C^{1,+}}(\mathbb{R}^p)$. Since $\widetilde{C^{1,+}}(\mathbb{R}^p)$ is closed under addition, we conclude the proof. □

The following proposition guarantees the well-posedness of the SDE (6.2).

Proposition 6.2.2. *Assume that (6.3) and (6.4) hold. Then, for every initial point $\boldsymbol{L}(0)$ such that $\mathbb{E}\left[\left\| \boldsymbol{L}(0) \right\|_2^2\right] < +\infty$,*

(i) there exists a unique solution to the SDE (6.2) which is strongly Markovian, and the diffusion is non-explosive, i.e. $\mathbb{E}\left[\left\| \boldsymbol{L}(t) \right\|_2^2\right] < +\infty$ for all $t > 0$,

(ii) $\boldsymbol{L}$ admits an (unique) invariant measure having the density μ_γ defined as

$$\mu_\gamma(\boldsymbol{\theta}) = \frac{\exp\left(-\left(L(\boldsymbol{\theta}) + ({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top(\boldsymbol{\theta})\right)\right)}{Z_\gamma}, \tag{6.5}$$

where

$$Z_\gamma = \int_{\mathbb{R}^p} \exp\left(-\left(L(\boldsymbol{\theta}') + ({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top(\boldsymbol{\theta}')\right)\right) d\boldsymbol{\theta}'.$$

Proof. Claim (i) follows by combining Proposition 6.2.1 and [158, Theorem 3.6, Chapter II]. Claim (ii) is a consequence of Proposition 6.2.1 and [124, Theorem 2.1]. □

The following proposition answers the natural question on the behaviour of $\mu_\gamma - \mu$ as a function of γ.

Proposition 6.2.3. *Assume that Assumption (H.1) holds. Then, μ_γ converges to μ in total variation as $\gamma \to 0$.*

Proof. With some abuse of notation, we denote with the same symbol the measure and its density with respect to the Lebesgue measure. Thus

$$\left\| \mu_\gamma - \mu \right\|_{\mathrm{TV}} = \int_{\mathbb{R}^M} |\mu_\gamma(\boldsymbol{\theta}) - \mu(\boldsymbol{\theta})| \, d\boldsymbol{\theta},$$

where

$$\mu_\gamma(\boldsymbol{\theta}) = \exp\left(-\left(L(\boldsymbol{\theta}) + ({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top(\boldsymbol{\theta})\right)\right) / Z_\gamma,$$

$$\mu(\boldsymbol{\theta}) = \exp\left(-\left(L(\boldsymbol{\theta}) + H \circ \boldsymbol{D}^\top(\boldsymbol{\theta})\right)\right) / Z,$$

and

$$Z = \int_{\mathbb{R}^M} \exp\left(-(L(\boldsymbol{\theta}') + H \circ \boldsymbol{D}^\top(\boldsymbol{\theta}'))\right) d\boldsymbol{\theta}'.$$

In view of Lemma 2.5.4(ii), applying the monotone convergence theorem (c.f. Theorem 2.1.9), we conclude that $Z_\gamma \to Z$ when $\gamma \to 0$. This together with Lemma 2.5.4(ii) again yield that μ_γ converges to μ pointwise. We conclude using Scheffé(-Riesz) theorem (c.f. Theorem 2.1.11). □

6.2.2 Discretization

6.2.2.1 Approach 1

Inserting the identities of Lemma 2.5.5 into (6.2), we get the SDE

$$d\boldsymbol{L}(t) = -\frac{1}{2}\left(\nabla L + \gamma^{-1}\boldsymbol{D}\boldsymbol{M}\left(\mathbf{I}_q - \mathrm{prox}_{\gamma H}^{\boldsymbol{M}}\right)\circ \boldsymbol{D}^\top\right)(\boldsymbol{L}(t))dt + d\boldsymbol{W}(t),\ \boldsymbol{L}(0) = \boldsymbol{l}_0,\ t > 0. \tag{6.6}$$

Consider now the forward Euler discretization of (6.6) with step-size $\delta > 0$, which can be rearranged as

$$\boldsymbol{L}_{k+1} = \boldsymbol{L}_k - \frac{\delta}{2}\nabla L(\boldsymbol{L}_k) - \frac{\delta}{2\gamma}\boldsymbol{D}\boldsymbol{M}\left(\boldsymbol{D}^\top\boldsymbol{L}_k - \mathrm{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{D}^\top\boldsymbol{L}_k)\right) + \sqrt{\delta}\boldsymbol{Z}_k,\ t > 0,\ \boldsymbol{L}_0 = \boldsymbol{l}_0. \tag{6.7}$$

Note that by Lemma 2.5.5, and without the stochastic term $\sqrt{\delta}\boldsymbol{Z}_k$, (6.7) amounts to a relaxed form of gradient descent on L and the Moreau envelope of H in the metric $\boldsymbol{M}$ with step-size δ.

From (6.7), an Euler approximate solution is defined as

$$\boldsymbol{L}^\delta(t) \overset{\text{def}}{=} \boldsymbol{L}_0 - \frac{1}{2}\int_0^t \left(\nabla L(\overline{\boldsymbol{L}}(s)) - \gamma^{-1}\boldsymbol{D}\boldsymbol{M}\left(\boldsymbol{D}^\top\overline{\boldsymbol{L}}(s) - \mathrm{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{D}^\top\overline{\boldsymbol{L}}(s))\right)\right)ds + \int_0^t d\boldsymbol{W}(s),$$

where $\overline{\boldsymbol{L}}(t) = \boldsymbol{L}_k$ for $t \in [k\delta, (k+1)\delta[$. Observe that $\boldsymbol{L}^\delta(k\delta) = \overline{\boldsymbol{L}}(k\delta) = \boldsymbol{L}_k$, hence $\boldsymbol{L}^\delta(t)$ and $\overline{\boldsymbol{L}}(t)$ are continuous-time extensions to the discrete-time chain $\{\boldsymbol{L}_k\}_k$.

Mean square convergence of the pathwise approximation (6.7) and of its first-order moment can be established as follows.

Theorem 6.2.4. *Assume that (6.3) and (6.4) hold and* $\mathbb{E}\left[\left\|\boldsymbol{L}(0)\right\|_2^r\right] < +\infty$ *for any* $r \geq 2$. *Then*

$$\left\|\mathbb{E}\left[\boldsymbol{L}^\delta(T)\right] - \mathbb{E}\left[\boldsymbol{L}(T)\right]\right\|_2 \leq \mathbb{E}\left[\sup_{0\leq t\leq T}\left\|\boldsymbol{L}^\delta(t) - \boldsymbol{L}(t)\right\|_2\right] \underset{\delta\to 0}{\longrightarrow} 0. \tag{6.8}$$

The convergence rate is of order $\delta^{1/2}$ *when* $\mathrm{prox}_{\gamma H}^{\boldsymbol{M}}$ *is globally Lipschitz continuous.*

Proof. Owing to Proposition 6.2.1 and [158, Theorem 4.1, Chapter II], we get that the r-th moments of $\boldsymbol{L}(t)$ are bounded for any $r \geq 2$ and $t \geq 0$. A similar reasoning also entails that the r-th moments of the continuous-time extension $\boldsymbol{L}^\delta$ are also bounded. Moreover, according to Proposition 6.2.1, the drift $\boldsymbol{\psi}$ is locally Lipschitz continuous. The claim then follows from [79, Theorem 2.2] and Jensen's inequality. In the globally Lipschitz continuous case, we get the claimed rate by putting together Lemma 2.1.8, Jensen's inequality and [158, Theorem 7.3, Chapter II] or [86, Theorem 10.2.2 and Remark 10.2.3]. □

6.2.2.2 Approach 2

Assume now that the metric also depends on $\gamma \in]0, \gamma_0[$ with $\gamma_0 > 0$, and we emphasize this by denoting it $\boldsymbol{M}_\gamma$ such that

(i) $\boldsymbol{M}_\gamma$ is symmetric positive definite for any $\gamma \in]0, \gamma_0[$,

(ii) for any $\boldsymbol{\theta} \in \mathbb{R}^q$, $\gamma \to \|\boldsymbol{\theta}\|_{\boldsymbol{M}_\gamma}$ is a decreasing mapping on $]0, \gamma_0[$,

(iii) $\boldsymbol{M}_\gamma \underset{\gamma\to 0}{\to} \mathbf{I}_q$ (such a choice is motivated by the scheme described in Section 6.4.1).

One can consider an alternative version of the SDE (6.2), i.e.

$$d\boldsymbol{L}(t) = -\frac{1}{2}\nabla\left(\left(L + ({}^{\boldsymbol{M}_\gamma,\gamma}H)\circ\boldsymbol{D}^\top\right)\circ {\boldsymbol{M}_\gamma}^{-1/2}\right)(\boldsymbol{L}(t))dt + {\boldsymbol{M}_\gamma}^{1/2}d\boldsymbol{W}(t),\ t > 0. \tag{6.9}$$

Denote the drift coefficient of (6.9) by $\boldsymbol{\phi}$, we get that

$$\langle\boldsymbol{\phi}(\boldsymbol{\theta}), \boldsymbol{\theta}\rangle = \langle\boldsymbol{\psi}(\boldsymbol{u}), \boldsymbol{u}\rangle,$$

where $\boldsymbol{u} = {\boldsymbol{M}_\gamma}^{-1/2}\boldsymbol{\theta}$. Therefore, it is easily seen that $\boldsymbol{\phi}$ also satisfies (6.3) and (6.4) under Assumptions (**H.1**), (**H.2**), (**H.3**) and (**H.4**). In turn, Proposition 6.2.2 applies to (6.9) the diffusion $\boldsymbol{L}$ is unique, non explosive and admits an unique invariant measure μ_γ having density

$$\boldsymbol{\theta} \mapsto \exp\left(-\left(L + ({}^{\boldsymbol{M}_\gamma,\gamma}H) \circ \boldsymbol{D}^\top\right) \circ {\boldsymbol{M}_\gamma}^{-1/2}(\boldsymbol{\theta})\right)/Z_\gamma$$

where

$$Z_\gamma = \sqrt{\det(\boldsymbol{M}_\gamma)} \int_{\mathbb{R}^p} \exp\left(-\left(L + ({}^{\boldsymbol{M}_\gamma,\gamma}H) \circ \boldsymbol{D}^\top\right)(\boldsymbol{u})\right) d\boldsymbol{u}.$$

Since $\det(\boldsymbol{M}_\gamma) \underset{\gamma\to 0}{\to} 1$, applying the reasoning in the proof of Proposition 6.2.3, we also deduce that μ_γ converges to μ in total variation as $\gamma \to 0$.

By the change of variable $\boldsymbol{U}(t) = {\boldsymbol{M}_\gamma}^{-1/2}\boldsymbol{L}(t)$, we get the following SDE

$$d\boldsymbol{U}(t) = -\frac{1}{2}{\boldsymbol{M}_\gamma}^{-1}\nabla\left(L + ({}^{\boldsymbol{M}_\gamma,\gamma}H) \circ \boldsymbol{D}^\top\right)(\boldsymbol{U}(t))dt + d\boldsymbol{W}(t),\ t > 0. \tag{6.10}$$

In an analogous way to (6.7), the forward Euler discretization of (6.10) has a deterministic part which is a relaxed gradient descent in the metric ${\boldsymbol{M}_\gamma}^{-1}$. In turn, mean square convergence of the Euler discretizations of both (6.9) and (6.10) and of their first-order moments can be established exactly in the same way as in Theorem 6.2.4. We omit the details here for the sake of brevity.

6.3 Prox-regular penalties

Let us consider a prox-regular function (see Section 2.5.4) satisfying Assumption (**H.1**). Owing to the following lemma, such type of functions also fullfills Assumptions (**H.2**) and (**H.3**).

Lemma 6.3.1. *Let $\boldsymbol{M} \in \mathbb{R}^{p\times p}$ symmetric positive definite and γ small enough, assume that $H : \mathbb{R}^p \to \mathbb{R}$ is prox-regular and satisfies Assumption (**H.1**). Then $\operatorname{prox}_{\gamma H}^{\boldsymbol{M}}$ is single-valued and locally Lipschitz continuous.*

Proof. The proof of Lemma 6.3.1 is based on the one of [126, Proposition 13.37] and generalizes to the proximal mapping in metric $\boldsymbol{M}$ for any $\boldsymbol{M} \in \mathbb{R}^{p\times p}$ symmetric positive definite.

Without loss of generality, we prove the claim on a neighbourhood of $\overline{\boldsymbol{x}}$ where H is lsc. Let $\overline{\boldsymbol{x}} \in \mathbb{R}^p$, $\overline{\boldsymbol{v}} \in \partial H(\overline{\boldsymbol{x}})$, since H is prox-regular at $\overline{\boldsymbol{x}}$ for $\overline{\boldsymbol{v}}$ and H is prox-bounded, owing to [10, Lemma 4.1], there exist $\epsilon > 0$ and $\lambda_0 > 0$ such that

$$\begin{aligned} H(\boldsymbol{x}') &> H(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{x}' - \boldsymbol{x}\rangle - \frac{1}{2\lambda_0}\|\boldsymbol{x}' - \boldsymbol{x}\|_2^2 \\ &> H(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{x}' - \boldsymbol{x}\rangle - \frac{1}{2\lambda_0\sigma_{\min}(\boldsymbol{M})}\|\boldsymbol{x}' - \boldsymbol{x}\|_{\boldsymbol{M}}^2, \end{aligned} \tag{6.11}$$

for any $\boldsymbol{x}' \neq \boldsymbol{x}$ and $(\boldsymbol{x}, \boldsymbol{v}) \in \operatorname{gph} \mathrm{T}^H_{\epsilon,\overline{\boldsymbol{x}},\overline{\boldsymbol{v}}}$. Let $\gamma_0 = \lambda_0\sigma_{\min}(\boldsymbol{M})$, $\gamma \in]0, \gamma_0[$ and $\boldsymbol{u} = \boldsymbol{x} + \gamma\boldsymbol{M}^{-1}\boldsymbol{v}$, (6.11) becomes

$$H(\boldsymbol{x}') + \frac{1}{2\gamma}\|\boldsymbol{x}' - \boldsymbol{u}\|_{\boldsymbol{M}}^2 > H(\boldsymbol{x}) + \frac{1}{2\gamma}\|\boldsymbol{x} - \boldsymbol{u}\|_{\boldsymbol{M}}^2.$$

Therefore, $\operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{u}) = \boldsymbol{x}$ where $(\boldsymbol{x}, \boldsymbol{v}) \in \operatorname{gph} \mathrm{T}^H_{\epsilon,\overline{\boldsymbol{x}},\overline{\boldsymbol{v}}}$. That yields

$$\operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\overline{\boldsymbol{x}} + \gamma\boldsymbol{M}^{-1}\overline{\boldsymbol{v}}) = \overline{\boldsymbol{x}}.$$

Since H is lsc, proper and prox-bounded, from [126, Theorem 1.17(c)] (see also [126, Theorem 1.25]), we have

$$\boldsymbol{x} \in \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{u}), \boldsymbol{u} \to \overline{\boldsymbol{x}} + \gamma\boldsymbol{M}^{-1}\overline{\boldsymbol{v}} \implies \begin{cases} \boldsymbol{x} \to \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\overline{\boldsymbol{x}} + \gamma\boldsymbol{M}^{-1}\overline{\boldsymbol{v}}) = \overline{\boldsymbol{x}}, \\ H(\boldsymbol{x}) = {}^{\boldsymbol{M},\gamma}H(\boldsymbol{u}) - \frac{1}{2\gamma}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 \to H(\overline{\boldsymbol{x}}). \end{cases} \tag{6.12}$$

For any $\boldsymbol{x} \in \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{u})$, by Fermat rules we get

$$\boldsymbol{v} = \frac{\boldsymbol{M}}{\gamma}(\boldsymbol{u} - \boldsymbol{x}) \in \partial H(\boldsymbol{x}). \tag{6.13}$$

For any $\gamma \in]0, \gamma_0[$, owing to (6.12) and (6.13), there exists $\mathcal{N}_{\gamma,\overline{\boldsymbol{x}},\overline{\boldsymbol{v}}}$ a neighbourhood of $\overline{\boldsymbol{x}} + \gamma \boldsymbol{M}^{-1}\overline{\boldsymbol{v}}$ such that for any $\boldsymbol{u} \in \mathcal{N}_{\gamma,\overline{\boldsymbol{x}},\overline{\boldsymbol{v}}}$, $\left\|\boldsymbol{x} - \overline{\boldsymbol{x}}\right\|_2 \leq \epsilon$, $\left\|H(\boldsymbol{x}) - H(\overline{\boldsymbol{x}})\right\|_2 \leq \epsilon$ and $\left\|\boldsymbol{v} - \overline{\boldsymbol{v}}\right\|_2 \leq \epsilon$. We get then

$$\operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{u}) = \boldsymbol{x} \implies \boldsymbol{v} = \frac{\boldsymbol{M}}{\gamma}(\boldsymbol{u} - \boldsymbol{x}) \in \mathrm{T}^{H}_{\epsilon,\overline{\boldsymbol{x}},\overline{\boldsymbol{v}}}(\boldsymbol{x}).$$

So that

$$\operatorname{prox}_{\gamma H}^{\boldsymbol{M}} = (\boldsymbol{M} + \gamma \mathrm{T}^{H}_{\epsilon,\overline{\boldsymbol{x}},\overline{\boldsymbol{v}}})^{-1} \circ \boldsymbol{M} = (\boldsymbol{M} + \delta^{-1} S)^{-1} \circ (\gamma\delta)^{-1}\boldsymbol{M},$$

where $\delta = 1/\gamma - 1/\gamma_0$,

$$S = \mathrm{T}^{H}_{\epsilon,\overline{\boldsymbol{x}},\overline{\boldsymbol{v}}} + \frac{\boldsymbol{M}}{\gamma_0}.$$

From (6.11), S is maximal monotone, the latter operator is well defined as a single-valued operator (see [6, Proposition 3.22 (ii)(d)]). Let $\boldsymbol{p} = \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{x})$ and $\boldsymbol{p}' = \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{x}')$. It then follows that

$$\boldsymbol{M}\boldsymbol{x} - \gamma\delta\boldsymbol{M}\boldsymbol{p} \in \gamma S(\boldsymbol{p}) \quad \text{and} \quad \boldsymbol{M}\boldsymbol{x}' - \gamma\delta\boldsymbol{M}\boldsymbol{p}' \in \gamma S(\boldsymbol{p}'),$$

and monotonicity of S yields

$$\left\langle \boldsymbol{p}' - \boldsymbol{p}, \boldsymbol{M}(\boldsymbol{x}' - \boldsymbol{x}) \right\rangle \geq \gamma\delta \left\|\boldsymbol{p}' - \boldsymbol{p}\right\|_{\boldsymbol{M}}^2 \geq \gamma\delta\sigma_{\min}(\boldsymbol{M}) \left\|\boldsymbol{p}' - \boldsymbol{p}\right\|_2^2.$$

Using Cauchy-Schwarz's inequality, we obtain

$$\left\|\boldsymbol{p}' - \boldsymbol{p}\right\|_2 \leq K \left\|\boldsymbol{x}' - \boldsymbol{x}\right\|_2,$$

where

$$K^{-1} = \frac{\gamma\delta\sigma_{\min}(\boldsymbol{M})}{\|\boldsymbol{M}\|_{2\to 2}} = \frac{(1 - \gamma/\gamma_0)\sigma_{\min}(\boldsymbol{M})}{\|\boldsymbol{M}\|_{2\to 2}}.$$

Let us note that when γ decreases, inequality (6.11) can hold for a larger ϵ, which enlarges $\mathcal{N}_{\gamma,\overline{\boldsymbol{x}},\overline{\boldsymbol{v}}}$, and thus, $\overline{\boldsymbol{x}} + \gamma\boldsymbol{M}^{-1}\overline{\boldsymbol{v}}$ concentrates around $\overline{\boldsymbol{x}}$ for any $\overline{\boldsymbol{v}}$. Thus, when γ is small enough, there exists a neighbourhood of $\overline{\boldsymbol{x}}$ that is included in $\mathcal{N}_{\gamma,\overline{\boldsymbol{x}},\overline{\boldsymbol{v}}}$ for any $\overline{\boldsymbol{v}} \in \partial H(\overline{\boldsymbol{x}})$. This concludes the proof. □

Lower-C^2 (or semi-convex) functions, see Example 2.5.9-(ii), satisfy the global counterpart of Lemma 2.5.10-(ii). For a lower-C^2 penalty H satisfying Assumption **(H.1)**, the following lemma shows that $\operatorname{prox}_{\gamma H}^{\boldsymbol{M}}$ is globally Lipschitz continuous with a proper choice of γ which in turn implies directly Assumption **(H.4)** according to Lemma 2.1.8.

Lemma 6.3.2. *Assume that H is lower-C^2 (with constant r) satisfying Assumption **(H.1)** and $\gamma \in]0, r\sigma_{\min}(\boldsymbol{M})[$, $\operatorname{prox}_{\gamma H}^{\boldsymbol{M}}$ is single-valued and Lipschitz continuous with constant $\frac{\|\boldsymbol{M}\|_{2\to 2}}{\sigma_{\min}(\boldsymbol{M})}\left(1 - \frac{\gamma}{r\sigma_{\min}(\boldsymbol{M})}\right)^{-1}$.*

Proof. From [126, Example 12.28(b)], ∂H is hypomonotone of modulus $\frac{1}{r}$. In turn

$$S = \partial H + \frac{1}{\gamma_0}\boldsymbol{M} = \partial\left(H + \frac{1}{2\gamma_0}\left\|\cdot\right\|_{\boldsymbol{M}}^2\right)$$

is monotone with $\gamma_0 = r\sigma_{\min}(\boldsymbol{M})$, or equivalently that $H + \frac{1}{2\gamma_0}\left\|\cdot\right\|_{\boldsymbol{M}}^2$ is convex [126, Example 12.28(b)]. Let $\delta = \frac{1}{\gamma} - \frac{1}{\gamma_0}$ and

$$W(\boldsymbol{w}, \boldsymbol{\theta}) = H(\boldsymbol{w}) + \frac{r'}{2}\left\|\boldsymbol{w} - \boldsymbol{\theta}\right\|_{\boldsymbol{M}}^2.$$

Thus

$$H(\boldsymbol{w}) + \frac{1}{2\gamma}\left\|\boldsymbol{w} - \boldsymbol{\theta}\right\|_{\boldsymbol{M}}^2 = W(\boldsymbol{w}, \boldsymbol{\theta}) + \frac{\delta}{2}\left\|\boldsymbol{w} - \boldsymbol{\theta}\right\|_{\boldsymbol{M}}^2.$$

$W(\cdot, \boldsymbol{\theta})$ is a convex function on $\mathbb{R}^p$ and $\delta > 0$ as $\gamma < \gamma_0$. Altogether, this entails that $W(\cdot, \boldsymbol{\theta}) + \frac{\delta}{2} \|\cdot - \boldsymbol{\theta}\|_{\boldsymbol{M}}^2$ is strongly convex uniformly in $\boldsymbol{\theta}$ and γ complying with $\gamma < \gamma_0$. It then follows that $\mathrm{prox}_{\gamma H}^{\boldsymbol{M}}$ is single-valued. We have

$$\boldsymbol{M} + \gamma \partial H = \gamma \left(\delta \boldsymbol{M} + S\right) = \gamma\delta \left(\boldsymbol{M} + \delta^{-1} S\right).$$

By Fermat's rule, we then get

$$\mathrm{prox}_{\gamma H}^{\boldsymbol{M}} = \left(\boldsymbol{M} + \gamma \partial H\right)^{-1} \circ \boldsymbol{M} = \left(\boldsymbol{M} + \delta^{-1} S\right)^{-1} \circ (\gamma\delta)^{-1} \boldsymbol{M},$$

and the latter operator is well-defined as a single-valued operator since S is maximal monotone; see [6, Proposition 3.22 (ii)(d)]. Let $\boldsymbol{p} = \mathrm{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{\theta})$ and $\boldsymbol{p}' = \mathrm{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{\theta}')$. It then follows that

$$\boldsymbol{M}\boldsymbol{\theta} - \gamma\delta \boldsymbol{M}\boldsymbol{p} \in \gamma S(\boldsymbol{p}) \text{ and } \boldsymbol{M}\boldsymbol{\theta}' - \gamma\delta \boldsymbol{M}\boldsymbol{p}' \in \gamma S(\boldsymbol{p}'),$$

and monotonicity of S yields

$$\left\langle \boldsymbol{p}' - \boldsymbol{p}, \boldsymbol{M}(\boldsymbol{\theta}' - \boldsymbol{\theta}) \right\rangle \geq \gamma\delta \left\|\boldsymbol{p}' - \boldsymbol{p}\right\|_{\boldsymbol{M}}^2 \geq \gamma\delta\sigma_{\min}(\boldsymbol{M}) \left\|\boldsymbol{p}' - \boldsymbol{p}\right\|_2^2.$$

Using Cauchy-Schwartz inequality, we then obtain

$$\left\|\boldsymbol{p}' - \boldsymbol{p}\right\|_2 \leq \kappa \left\|\boldsymbol{\theta}' - \boldsymbol{\theta}\right\|_2,$$

where

$$\kappa^{-1} = \frac{\gamma\delta\sigma_{\min}(\boldsymbol{M})}{\|\boldsymbol{M}\|_{2\to 2}} = \frac{\sigma_{\min}(\boldsymbol{M})}{\|\boldsymbol{M}\|_{2\to 2}} \left(1 - \frac{\gamma}{\gamma_0}\right) = \frac{\sigma_{\min}(\boldsymbol{M})}{\|\boldsymbol{M}\|_{2\to 2}} \left(1 - \frac{\gamma}{r\sigma_{\min}(\boldsymbol{M})}\right).$$

That concludes the proof of Lemma 6.3.2. □

Remark 6.3.3. As a consequence of the these results, when $\mathrm{prox}_{\gamma H}^{\boldsymbol{M}}$ is globally Lipschitz continuous, the optimal convergence rate in (6.8) is of order $\delta^{1/2}$ in view of Theorem 6.2.4.

6.4 Forward-Backward type LMC algorithms

Let us now deal with our main goal: computing the EWA in (1.12) by sampling from $\widehat{\mu}_n$. Remind that

$$\widehat{\mu}_n(\boldsymbol{\theta}) \propto \exp\left(-\frac{F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) + J_{\boldsymbol{\lambda}}(\boldsymbol{\theta})}{\beta}\right),$$

where $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a general loss and $J_{\boldsymbol{\lambda}} : \mathbb{R}^p \to \mathbb{R}$ is the penalty. Assume that $F(\boldsymbol{X}\cdot, \boldsymbol{y}) \in \widetilde{C^{1,+}}(\mathbb{R}^p)$ and the penalty takes the form $J_{\boldsymbol{\lambda}} = W_{\boldsymbol{\lambda}} \circ \boldsymbol{D}^\top$. Let us impose the following assumptions on $W_{\boldsymbol{\lambda}}$.

(**H.1'**) $W_{\boldsymbol{\lambda}} \in \mathcal{J}(\mathbb{R}^q)$.

(**H.2'**) $\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}}$ is single-valued.

(**H.3'**) $\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}}$ is locally Lipschitz continuous.

To lighten notation, we will write $F_\beta \stackrel{\text{def}}{=} F(\boldsymbol{X}\cdot, \boldsymbol{y})/\beta$. This section aims to describe our Forward-Backward type Langevin Monte-Carlo (LMC) algorithms to implement (1.12). These algorithms are based on wise specializations of the results reported in Section 6.2.

6.4.1 Forward-backward LMC (FBLMC)

In (6.1), we set $\boldsymbol{D} = \mathbf{I}_p$ (hence $J_{\boldsymbol{\lambda}} = W_{\boldsymbol{\lambda}}$), $L \equiv 0$, and $H = F_\beta + J_{\boldsymbol{\lambda}}/\beta$, where F is a quadratic loss, i.e. $F_\beta(\boldsymbol{\theta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2/\beta$. Observe that H satisfies Assumption (**H.1**) by Assumption (**H.1'**). To check Assumptions (**H.2**), (**H.3**) and (**H.4**), we need to design a metric in which $\mathrm{prox}_{\gamma H}^{\boldsymbol{M}}$ is expressed as a function of $\mathrm{prox}_{\gamma J_{\boldsymbol{\lambda}}/\beta}$. This idea is formalized in the following lemma.

Lemma 6.4.1. *Assume that* $0 < \gamma < \beta/(2\|\boldsymbol{X}\|_{2\to 2}{}^2)$ *and Assumption* (**H.1'**) *holds. Define* $\boldsymbol{M}_\gamma \overset{def}{=} \mathbf{I}_p - (2\gamma/\beta)\boldsymbol{X}^\top\boldsymbol{X}$*, which is symmetric positive definite. Then*

$$\mathrm{prox}_{\gamma H}^{\boldsymbol{M}_\gamma} = \mathrm{prox}_{\gamma J_{\boldsymbol{\lambda}}/\beta} \circ (\mathbf{I}_p - \gamma\nabla F_\beta)\,. \tag{6.14}$$

Proof. We have

$$\begin{aligned}\mathrm{prox}_{\gamma H}^{\boldsymbol{M}_\gamma}(\boldsymbol{\theta}) &= \underset{\boldsymbol{w}\in\mathbb{R}^p}{\mathrm{Argmin}}\ \frac{1}{2\gamma}\|\boldsymbol{w}-\boldsymbol{\theta}\|_{\boldsymbol{M}_\gamma}^2 + H(\boldsymbol{w}) \\ &= \underset{\boldsymbol{w}\in\mathbb{R}^p}{\mathrm{Argmin}}\ \frac{1}{2}\|\boldsymbol{w}-\boldsymbol{\theta}\|_2^2 - \frac{\gamma}{\beta}\|\boldsymbol{X}(\boldsymbol{w}-\boldsymbol{\theta})\|_2^2 + \frac{\gamma}{\beta}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2 + \frac{\gamma}{\beta}J_{\boldsymbol{\lambda}}(\boldsymbol{w}).\end{aligned}$$

By the Pythagoras relation, we then get

$$\begin{aligned}\mathrm{prox}_{\gamma H}^{\boldsymbol{M}_\gamma}(\boldsymbol{\theta}) &= \underset{\boldsymbol{w}\in\mathbb{R}^p}{\mathrm{Argmin}}\ \frac{1}{2}\|\boldsymbol{w}-\boldsymbol{\theta}\|_2^2 + \frac{\gamma}{\beta}\left(\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|_2^2 - \langle\boldsymbol{X}(\boldsymbol{\theta}-\boldsymbol{w}), \boldsymbol{X}\boldsymbol{\theta}-\boldsymbol{y}\rangle\right) + \frac{\gamma}{\beta}J_{\boldsymbol{\lambda}}(\boldsymbol{w}) \\ &= \underset{\boldsymbol{w}\in\mathbb{R}^p}{\mathrm{Argmin}}\ \frac{1}{2}\|\boldsymbol{w}-\boldsymbol{\theta}\|_2^2 - \frac{\gamma}{\beta}\left\langle\boldsymbol{w}-\boldsymbol{\theta}, \boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})\right\rangle + \frac{\gamma}{\beta}J_{\boldsymbol{\lambda}}(\boldsymbol{w}) \\ &= \underset{\boldsymbol{w}\in\mathbb{R}^p}{\mathrm{Argmin}}\ \frac{1}{2}\left\|\boldsymbol{w}-\left(\boldsymbol{\theta}-\frac{2\gamma}{\beta}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\theta}-\boldsymbol{y})\right)\right\|_2^2 + \frac{\gamma}{\beta}J_{\boldsymbol{\lambda}}(\boldsymbol{w}) \\ &= \mathrm{prox}_{\gamma J_{\boldsymbol{\lambda}}/\beta}\left(\boldsymbol{\theta}-\gamma\nabla F(\boldsymbol{\theta})\right).\end{aligned}$$

We conclude the proof of Lemma 6.4.1. □

In view of Lemma 6.14, Assumptions (**H.2'**) and (**H.3'**), it is immediate to check that Assumptions (**H.2**) and (**H.3**) are satisfied.

It remains now to verify Assumption (**H.4**) which is fulfilled by imposing the following assumption on $W_{\boldsymbol{\lambda}}$ (or $J_{\boldsymbol{\lambda}}$).

(**H.4'-FB**) There exists $C'_{\mathrm{FB}} > 0$ such that

$$\left\langle \mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta} \circ (\mathbf{I}_p - \gamma\nabla F_\beta)(\boldsymbol{\theta}), \boldsymbol{\theta}\right\rangle_{\boldsymbol{M}_\gamma} \le C'_{\mathrm{FB}}(1 + \|\boldsymbol{\theta}\|_2^2), \quad \forall\boldsymbol{\theta}\in\mathbb{R}^p.$$

By Lemma 2.1.8, a sufficient condition for Assumption (**H.4'-FB**) to hold is that the proximal mapping of $W_{\boldsymbol{\lambda}}$ is Lipschitz continuous.

From Lemmas 2.5.5 and 6.4.1, we get

$$\nabla^{\boldsymbol{M}_\gamma,\gamma}H = \gamma^{-1}\boldsymbol{M}_\gamma\left(\mathbf{I}_p - \mathrm{prox}_{\gamma H}^{\boldsymbol{M}_\gamma}\right) = \gamma^{-1}\boldsymbol{M}_\gamma\left(\mathbf{I}_p - \mathrm{prox}_{\gamma J_{\boldsymbol{\lambda}}/\beta}(\mathbf{I}_p - \gamma\nabla F_\beta)\right).$$

With this expression at hand, the forward Euler discretization of the SDE (6.2), specialized to the current case, reads

$$\boldsymbol{L}_{k+1} = \boldsymbol{L}_k - \frac{\delta}{2\gamma}\boldsymbol{M}_\gamma\left(\boldsymbol{L}_k - \mathrm{prox}_{\gamma J_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{L}_k - \gamma\nabla F_\beta(\boldsymbol{L}_k))\right) + \sqrt{\delta}\boldsymbol{Z}_k,\ t>0,\ \boldsymbol{L}_0 = \boldsymbol{l}_0. \tag{6.15}$$

Similarly, the forward Euler discretization of the SDE (6.10) is given by

$$\boldsymbol{U}_{k+1} = (1-\frac{\delta}{2\gamma})\boldsymbol{U}_k + \frac{\delta}{2\gamma}\mathrm{prox}_{\gamma J_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{U}_k - \gamma\nabla F_\beta(\boldsymbol{U}_k)) + \sqrt{\delta}\boldsymbol{Z}_k,\ t>0,\ \boldsymbol{U}_0 = \boldsymbol{l}_0. \tag{6.16}$$

The familiar reader may have recognized that the deterministic part of (6.16) is nothing but the relaxed form of the so-called Forward-Backward proximal splitting algorithm [7]. This terminology reflects that there is a forward Euler discretization on F_β and a Euler backward discretization on $J_{\boldsymbol{\lambda}}$.

6.4.2 Semi-Forward-Backward LMC (Semi-FBLMC)

The main limitation of (6.15) is that the proximal mapping of $J_{\boldsymbol{\lambda}}$ must be easy to compute. This may not be true even if the proximal mapping of $W_{\boldsymbol{\lambda}}$ is accessible as, for for example, when $\boldsymbol{D}$ is not a tight frame [7]. Our goal is to circumvent this difficulty.

Toward this goal, in (6.1), consider now $L = F_\beta$, $H = W_{\boldsymbol{\lambda}}/\beta$ and $\boldsymbol{M} = \mathbf{I}_q$. Owing to Assumptions (**H.1'**), (**H.2'**) and (**H.3'**), one can check that Assumptions (**H.1**), (**H.2**) and (**H.3**) are fulfilled. Assumption (**H.4**) is verified by imposing the following on $W_{\boldsymbol{\lambda}}$.

(**H.4'-SFB**) There exists $C'_{\mathrm{SFB}} > 0$ such that

$$\left\langle \mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{u}), \boldsymbol{u} \right\rangle \leq C'_{\mathrm{SFB}}(1 + \left\|\boldsymbol{u}\right\|_2^2), \quad \forall \boldsymbol{u} \in \mathbb{R}^q.$$

From Lemma 2.5.5, we obtain

$$\nabla \left(({}^{\gamma}H) \circ \boldsymbol{D}^\top \right)(\boldsymbol{\theta}) = \gamma^{-1}\boldsymbol{D}(\boldsymbol{D}^\top\boldsymbol{\theta} - \mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{D}^\top\boldsymbol{\theta})).$$

Thus, the forward Euler discretization of SDE (6.2) now reads

$$\boldsymbol{L}_{k+1} = \boldsymbol{L}_k - \frac{\delta}{2}\nabla F_\beta(\boldsymbol{L}_k) - \frac{\delta}{2\gamma}\boldsymbol{D}\left(\boldsymbol{D}^\top\boldsymbol{L}_k - \mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{D}^\top\boldsymbol{L}_k)\right) + \sqrt{\delta}\boldsymbol{Z}_k, \; t > 0, \; \boldsymbol{L}_0 = \boldsymbol{l}_0. \tag{6.17}$$

In the case where $\boldsymbol{D} = \mathbf{I}_p$, F_β and $W_{\boldsymbol{\lambda}}$ are convex, we recover the scheme studied in [63].

6.5 Applications to penalties in statistics

In this section, we exemplify our LMC sampling algorithms for some penalties, some of which cover those studied in the previous chapters. Our goal is by no means to be exhaustive, but rather to be illustrative and show the versatility of our framework. For each penalty, we aim at checking that Assumptions (**H.1'**), (**H.2'**), (**H.3'**), (**H.4'-FB**) and (**H.4'-SFB**) hold, and to compute $\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}$. In turn, this allows to apply our algorithms (6.16) and (6.17) to compute EWA with such penalties.

6.5.1 Analysis-group-separable penalties

We first focus on a class of penalties where $J_{\boldsymbol{\lambda}}$ is analysis-group-separable, i.e.

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = W_{\boldsymbol{\lambda}}(\boldsymbol{D}^\top\boldsymbol{\theta}), \quad \text{where} \quad W_{\boldsymbol{\lambda}}(\boldsymbol{u}) = \textstyle\sum_{l=1}^L w_{\boldsymbol{\lambda}}\left(\left\|\boldsymbol{u}_{b_l}\right\|_2\right), \tag{6.18}$$

for $w_{\boldsymbol{\lambda}} : \mathbb{R}_+ \to \mathbb{R}$, and some uniform partition $(b_l)_{l\in\{1,\dots,L\}}$ of $\{1,\dots,q\}$, i.e. $\cup_{l=1}^L b_l = \{1,\dots,q\}$ and $b_l \cap b_{l'}$, $\forall l \neq l'$.

Remark 6.5.1. It is worth mentioning that separability of $W_{\boldsymbol{\lambda}}$ does not entail that of $J_{\boldsymbol{\lambda}}$. In fact, overlapping groups can be easily taken intro account as any overlapping-group penalty can be written as the composition of $W_{\boldsymbol{\lambda}}$ with a linear operator, say $\boldsymbol{B}$, such that $\boldsymbol{B}^\top\boldsymbol{B}$ is diagonal, and $\boldsymbol{B}$ acts as a group extractor, see [113, 39] (see also Example 3.3.2) .

A first consequence of separability is that $\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}$ is also separable under the following mild assumptions on $w_{\boldsymbol{\lambda}}$.

(**W.1**) $w_{\boldsymbol{\lambda}}$ is bounded from below on $]0, +\infty[$.

(**W.2**) $w_{\boldsymbol{\lambda}}$ is non-decreasing on $]0, +\infty[$.

Lemma 6.5.2. *Assume that Assumptions (**W.1**) and (**W.2**) hold, and $w_{\boldsymbol{\lambda}}$ is continuous on $]0, +\infty[$. For any $\boldsymbol{u} \in \mathbb{R}^q$ and $\gamma > 0$, we have*

$$\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{u}) = \begin{pmatrix} \mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}\left(\left\|\boldsymbol{u}_{b_1}\right\|_2\right) \frac{\boldsymbol{u}_{b_1}}{\left\|\boldsymbol{u}_{b_1}\right\|_2} \\ \vdots \\ \mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}\left(\left\|\boldsymbol{u}_{b_L}\right\|_2\right) \frac{\boldsymbol{u}_{b_L}}{\left\|\boldsymbol{u}_{b_L}\right\|_2} \end{pmatrix}.$$

Proof. This is a probably known result, for which we provide a simple proof. Since $W_{\boldsymbol{\lambda}}$ is separable and $w_{\boldsymbol{\lambda}}$ is continuous and lower-bounded, we have

$$\min_{\boldsymbol{w}\in\mathbb{R}^q} \frac{1}{2}\|\boldsymbol{w}-\boldsymbol{u}\|_2^2 + \frac{\gamma}{\beta}W_{\boldsymbol{\lambda}}(\boldsymbol{w}) = \sum_{l=1}^{L} \min_{\boldsymbol{v}\in\mathbb{R}^K} \frac{1}{2}\|\boldsymbol{v}-\boldsymbol{u}_{b_l}\|_2^2 + \frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}\left(\|\boldsymbol{v}\|_2\right),$$

and thus, $\forall l \in \{1,\dots,L\}$,

$$\left[\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{u})\right]_{b_l} = \operatorname*{Argmin}_{\boldsymbol{v}\in\mathbb{R}^K} \frac{1}{2}\|\boldsymbol{v}-\boldsymbol{u}_{b_l}\|_2^2 + \frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}\left(\|\boldsymbol{v}\|_2\right). \tag{6.19}$$

If $\boldsymbol{u}_{b_l}=0$, then as $w_{\boldsymbol{\lambda}}$ is an increasing function, $\left[\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{u})\right]_{b_l}=0$. For $\boldsymbol{u}_{b_l}\neq 0$, by isotropy of problem (6.19), we can write

$$\min_{\boldsymbol{v}\in\mathbb{R}^K} \frac{1}{2}\|\boldsymbol{v}-\boldsymbol{u}_{b_l}\|_2^2 + \frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}\left(\|\boldsymbol{v}\|_2\right) = \min_{t\geq 0} \frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}(t) + \left(\min_{\|\boldsymbol{v}\|_2=t} \frac{1}{2}\|\boldsymbol{v}-\boldsymbol{u}_{b_l}\|_2^2\right). \tag{6.20}$$

The inner minimization problem amounts to solving for the orthogonal projector on the ℓ_2 sphere in $\mathbb{R}^K$ of radius t, which is obviously $\boldsymbol{v} = t\frac{\boldsymbol{u}_{b_l}}{\|\boldsymbol{u}_{b_l}\|_2}$ since $\boldsymbol{u}_{b_l}\neq 0$. Inserting this into (6.20) and rearranging the terms, (6.19) becomes

$$\left[\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{u})\right]_{b_l} = \frac{\boldsymbol{u}_{b_l}}{\|\boldsymbol{u}_{b_l}\|_2} \operatorname*{Argmin}_{t\geq 0} \frac{1}{2}\left(t-\|\boldsymbol{u}_{b_l}\|_2\right)^2 + \frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}(t) = \frac{\boldsymbol{u}_{b_l}}{\|\boldsymbol{u}_{b_l}\|_2}\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(\|\boldsymbol{u}_{b_l}\|_2),$$

where we used even-symmetry of $w_{\boldsymbol{\lambda}}$. □

Our aim is now to design a family of penalties that will allow to establish Assumptions **(H.1')**, **(H.2')**, **(H.3')**, **(H.4'-FB)** and **(H.4'-SFB)**, while involving a form of shrinkage which is ubiquitous in low-complexity regularization. Inspired by the work of [3], we make the following assumptions on $w_{\boldsymbol{\lambda}}$.

(W.3) $w_{\boldsymbol{\lambda}}$ is continuously differentiable on $]0,+\infty[$ and the problem

$$\min_{t\in[0,+\infty[} \{t + \frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}'(t)\}$$

has a unique solution at 0 for a given γ.

Under these assumptions, $\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}$ has a indeed convenient shrinkage-type form.

Lemma 6.5.3 ([3, Theorem 1]). *Assume that Assumptions* **(W.2)** *and* **(W.3)** *hold for some $\gamma>0$. Then, $\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}$ are the single-valued continuous mappings, and satisfy, for $t\in[0,+\infty[$,*

$$\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t) = \begin{cases} 0 & \text{if } t \leq \frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}'(0^+), \\ t - \frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}'\left(\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t)\right) & \text{if } t > \frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}'(0^+). \end{cases} \tag{6.21}$$

Let us turn to check our Assumptions. Assumptions **(H.1')**, **(H.2')** and **(H.3')** are fulfilled thanks to Assumptions **(W.1)**, **(W.2)** and **(W.3)**. It remains to check Assumptions **(H.4'-FB)** and **(H.4'-SFB)**. This is the subject of the following lemma.

Lemma 6.5.4. *Assume that Assumptions* **(W.2)** *and* **(W.3)** *hold for some $\gamma>0$, then Assumptions* **(H.4'-FB)** *and* **(H.4'-SFB)** *also hold.*

Proof. Before proceeding, let us discuss about the term $\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}$. In view of Assumption **(W.2)**, $w_{\boldsymbol{\lambda}}'/\beta$ is positive on $]0,+\infty[$. According to Lemma 6.5.3 we get that, for any $t\geq 0$, $\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t)=0$ if $t\leq\frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}'(0)$ and $\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t) = t - \frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}'(\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t)) \leq t$ otherwise. Hence for any $t\geq 0$,

$$0 \leq \mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t) \leq t, \quad \forall t \geq 0. \tag{6.22}$$

Set $\boldsymbol{u} = \boldsymbol{D}^\top \boldsymbol{\theta}$, from Lemma 6.5.2 and (6.22), we get that

$$\left\langle \mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{u}), \boldsymbol{u} \right\rangle = \sum_{l=1}^{L} \left\langle [\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{u})]_{b_l}, \boldsymbol{u}_{b_l} \right\rangle = \sum_{l=1}^{L} \frac{\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}\left(\|\boldsymbol{u}_{b_l}\|_2\right)}{\|\boldsymbol{u}_{b_l}\|_2} \|\boldsymbol{u}_{b_l}\|_2^2 \leq \|\boldsymbol{u}\|_2^2.$$

According to the fact that $\|\boldsymbol{u}\|_2^2 = \|\boldsymbol{D}^\top \boldsymbol{\theta}\|_2^2 \leq \|\boldsymbol{D}\|_{2\to 2}{}^2 \|\boldsymbol{\theta}\|_2^2$, Assumption **(H.4'-SFB)** holds.

Set $\boldsymbol{v} = 2\gamma \boldsymbol{X}^\top \boldsymbol{y}/\beta$ and $\boldsymbol{t}_{\boldsymbol{\theta}} = \boldsymbol{\theta} - \gamma \nabla F_\beta(\boldsymbol{\theta}) = \boldsymbol{M}_\gamma \boldsymbol{\theta} + \boldsymbol{v}$, by Young's inequality, we obtain that

$$\left\langle \mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{t}_{\boldsymbol{\theta}}), \boldsymbol{\theta} \right\rangle_{\boldsymbol{M}_\gamma} = \left\langle \boldsymbol{M}_\gamma \mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{t}_{\boldsymbol{\theta}}), \boldsymbol{\theta} \right\rangle \leq \frac{1}{2} \|\boldsymbol{M}_\gamma\|_{2\to 2}{}^2 \left\|\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{t}_{\boldsymbol{\theta}})\right\|_2^2 + \frac{1}{2}\|\boldsymbol{\theta}\|_2^2.$$

Moreover, owing to Lemma 6.5.2 and (6.22), we get that

$$\begin{aligned}
\left\|\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{t}_{\boldsymbol{\theta}})\right\|_2^2 = \left\|\sum_{l=1}^{L} \frac{\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}\left(\left\|[\boldsymbol{t}_{\boldsymbol{\theta}}]_{b_l}\right\|_2\right)}{\left\|[\boldsymbol{t}_{\boldsymbol{\theta}}]_{b_l}\right\|_2} [\boldsymbol{t}_{\boldsymbol{\theta}}]_{b_l}\right\|_2^2 &\leq \left(\sum_{l=1}^{L} \left|\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}\left(\left\|[\boldsymbol{t}_{\boldsymbol{\theta}}]_{b_l}\right\|_2\right)\right|\right)^2 \\
&\leq \left(\sum_{l=1}^{L} \left\|[\boldsymbol{t}_{\boldsymbol{\theta}}]_{b_l}\right\|_2\right)^2 \\
&\leq L\left\|\boldsymbol{t}_{\boldsymbol{\theta}}\right\|_2^2 \\
&\leq 2L\left(\|\boldsymbol{M}_\gamma\|_{2\to 2}{}^2 \|\boldsymbol{\theta}\|_2^2 + \|\boldsymbol{v}\|_2^2\right).
\end{aligned}$$

Thus, Assumption **(H.4'-FB)** holds and we conclude the proof of Lemma 6.5.4. □

We now discuss some popular penalties $w_{\boldsymbol{\lambda}}$ that satisfy Assumptions **(W.1)**, **(W.2)** and **(W.3)** for some $\gamma > 0$.

6.5.2 Examples

ℓ_1 penalty Take $w_{\boldsymbol{\lambda}} : t \in \mathbb{R}_+ \mapsto \lambda t$. This entails the analysis-group Lasso penalty

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \lambda \sum_{l=1}^{L} \left\|[\boldsymbol{D}^\top \boldsymbol{\theta}]_{b_l}\right\|_2.$$

Clearly, $w_{\boldsymbol{\lambda}}$ is a continuous positive convex function which verifies Assumptions **(W.1)**, **(W.2)** and **(W.3)** for any $\gamma > 0$, and its proximal mapping corresponds to soft-thresholding, i.e.

$$\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t) = (t - \gamma\lambda/\beta)_+, \quad \forall t \geq 0.$$

FIRM penalty The FIRM penalty is given by [73]

$$w_{\boldsymbol{\lambda}}(t) = \begin{cases} \lambda\left(t - \frac{t^2}{2\mu}\right) & \text{if } 0 \leq t \leq \mu, \\ \frac{\lambda\mu}{2} & \text{if } t > \mu. \end{cases} \tag{6.23}$$

which entails the corresponding analysis-group FIRM penalty $J_{\boldsymbol{\lambda}}$. Since ${w_{\boldsymbol{\lambda}}}'(t) = \lambda\left(1 - \frac{t}{\mu}\right)_+ \geq 0$, $w_{\boldsymbol{\lambda}}$ is non-decreasing and bounded from below by $w_{\boldsymbol{\lambda}}(0) = 0$ on $]0, +\infty[$. Thus, $w_{\boldsymbol{\lambda}}$ satisfies Assumptions **(W.1)** and **(W.2)**. Assumption **(W.3)** also holds for any $\gamma < \beta\mu/\lambda$. The operator $\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}$ can be constructed from [157, Definition II.3]. Its formula is defined as

$$\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t) = \begin{cases} 0 & \text{if } 0 \leq t \leq \alpha, \\ \frac{\mu}{\mu - \alpha}(t - \alpha) & \text{if } \alpha < t \leq \mu, \\ t & \text{if } t > \mu, \end{cases} \tag{6.24}$$

where $\alpha = \gamma\lambda/\beta$. The formula (6.24) can also be found using Lemma 6.5.3. Observe that the FIRM shrinkage (6.24) interpolates between hard- (see [157, Definition II.2]) and soft-thresholding. In particular, (6.24) coincides with soft-thresholding when $\mu \to +\infty$.

SCAD penalty The SCAD penalty, proposed in [68] is parameterized by $\boldsymbol{\lambda} = (\lambda, a) \in]0, +\infty[\times]2, +\infty[$ as

$$w_{\boldsymbol{\lambda}}(t) = \begin{cases} \lambda t & \text{if } 0 \leq t \leq \lambda, \\ -\frac{t^2 - 2a\lambda t + \lambda^2}{2(a-1)} & \text{if } \lambda < t \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } t > a\lambda, \end{cases} \tag{6.25}$$

The following lemma establishes the validity of $w_{\boldsymbol{\lambda}}$ and computes $\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}$.

Lemma 6.5.5. *Let $w_{\boldsymbol{\lambda}}$ defined in (6.25), and $\kappa = \gamma/\beta$. For any $\gamma < (a-1)\beta$,*

(i) $w_{\boldsymbol{\lambda}}$ satisfies Assumptions (W.1), (W.2) and (W.3),

(ii) The proximal mapping of the SCAD penalty is given by the shrinkage

$$\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t) = \begin{cases} (t - \kappa\lambda)_+ & \text{if } 0 \leq t \leq (\kappa+1)\lambda, \\ \frac{(a-1)t - ka\lambda}{a-1-\kappa} & \text{if } (\kappa+1)\lambda < t \leq a\lambda, \\ t & \text{if } t > a\lambda. \end{cases} \tag{6.26}$$

Proof.

(i) Observe that $w_{\boldsymbol{\lambda}}$ is continuously differentiable on $]0, +\infty[$ with

$$w_{\boldsymbol{\lambda}}{}'(t) = \kappa\lambda \left(I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right) \geq 0,$$

$w_{\boldsymbol{\lambda}}$ is then non decreasing and bounded from below by $w_{\boldsymbol{\lambda}}(0) = 0$ on $]0, +\infty[$. Thus, $w_{\boldsymbol{\lambda}}$ satisfies Assumptions (W.1) and (W.2). Let us check Assumption (W.3). Let $u(t) = t + \kappa w_{\boldsymbol{\lambda}}{}'(t)$, we obtain that

- $u(0) = \kappa\lambda$,
- if $0 < t \leq \lambda$, $u(t) = t + \kappa\lambda > \kappa\lambda$,
- if $\lambda < t \leq a\lambda$, since $a - 1 > \kappa > 0$, $u(t) = t + \frac{\kappa(a\lambda - t)}{a-1} = \kappa\lambda + \frac{a-1-\kappa}{a-1} t + \frac{\kappa\lambda}{a-1} > \kappa\lambda$,
- if $t > a\lambda$, since $a - 1 > \kappa$, $u(t) = t > a\lambda > \kappa\lambda$.

Thus, $t = 0$ is the unique mimimum in $[0, +\infty[$ of $t + p'_\lambda(t)$. In other words, $w_{\boldsymbol{\lambda}}$ satisfies Assumption (W.3).

(ii) For the sake of simplified notation, we denote $p = \mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t)$. Owing to Lemma 6.5.3, we obtain that

$$p = \begin{cases} 0 & \text{if } t \leq \kappa\lambda, \\ t - \kappa\lambda \left(I(p \leq \lambda) + \frac{(a\lambda - p)_+}{(a-1)\lambda} I(p > \lambda) \right) & \text{otherwise.} \end{cases} \tag{6.27}$$

From (6.27), we get the following assertions when $t > \kappa\lambda$,

- if $p \leq \lambda$, $p = t - \kappa\lambda$, and $t = p + \kappa\lambda \leq (\kappa+1)\lambda$,
- if $\lambda < p \leq a\lambda$, $p = t - \kappa(a\lambda - p)/(a-1)$ implies that $p = \frac{(a-1)t - \kappa a\lambda}{a-1-\kappa}$. Since $\lambda < p \leq a\lambda$, $\kappa < a - 1$ and $a > 2$, we also get that

$$(1+\kappa)\lambda < t = \frac{a-1-\kappa}{a-1} p + \frac{\kappa a\lambda}{a-1} \leq a\lambda,$$

- if $p > a\lambda$, $p = t$, and $t > a\lambda$.

That concludes the proof of (ii), Lemma 6.5.5.

□

Since $a > 2$, one can set $\kappa = 1$. In this case, (6.26) specializes to [68, Equation (2.8)].

ℓ_∞ penalty The ℓ_∞ norm penalty is convex and continuous but is not separable, unlike the previous ones. It is a suitable prior to promote flat vectors, and has found applications in several fields [84, 102, 133]. It entails the following penalty $W_{\boldsymbol{\lambda}}$:

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = W_{\boldsymbol{\lambda}}(\boldsymbol{D}^\top \boldsymbol{\theta}) \quad \text{where} \quad W_{\boldsymbol{\lambda}}(\boldsymbol{u}) = \lambda \max_{l \in \{1,\dots,L\}} \left\{ \|[\boldsymbol{u}]_{b_l}\|_2 \right\}, \tag{6.28}$$

where $\boldsymbol{\lambda} = \lambda > 0$. Since $W_{\boldsymbol{\lambda}}$ is not separable, Lemma 6.5.2 is not applicable. Nevertheless, the proximal mapping of $W_{\boldsymbol{\lambda}}$ can still be obtained easily from the projector on the ℓ_1 unit ball, i.e.

$$\operatorname{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{u}) = \boldsymbol{u} - \mathrm{P}_{\left\{\boldsymbol{x} \,:\, \sum_l \left\|\boldsymbol{x}_{b_l}\right\|_2 \leq \frac{\beta}{\lambda\gamma}\right\}}(\boldsymbol{u}). \tag{6.29}$$

This projector can be obtained from [66, Proposition 2] (see also references therein). One can see that Assumptions **(H.1')**, **(H.2')** and **(H.3')** hold. We report the verification of Assumptions **(H.4'-FB)** and **(H.4'-SFB)** in the proof of the following lemma.

Lemma 6.5.6. *Let $W_{\boldsymbol{\lambda}}$ in (6.28). Then Assumptions **(H.4'-FB)** and **(H.4'-SFB)** hold.*

Proof. Set $\boldsymbol{u} = \boldsymbol{D}^\top \boldsymbol{\theta}$, $\alpha = \gamma\lambda/\beta$ and $\boldsymbol{p}_{\boldsymbol{u}} = \mathrm{P}_{\left\{\boldsymbol{x} \,:\, \alpha \sum_l \left\|\boldsymbol{x}_{b_l}\right\|_2 \leq 1\right\}}(\boldsymbol{u})$. Owing to (6.29) and Young's inequality, we obtain that

$$\left\langle \boldsymbol{u}, \operatorname{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{u}) \right\rangle = \langle \boldsymbol{u}, \boldsymbol{u} - \boldsymbol{p}_{\boldsymbol{u}} \rangle \leq \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{u}\|_2 \|\boldsymbol{p}_{\boldsymbol{u}}\|_2 \leq \frac{3}{2}\|\boldsymbol{u}\|_2^2 + \frac{1}{2}\|\boldsymbol{p}_{\boldsymbol{u}}\|_2^2 \leq \frac{3}{2}\|\boldsymbol{u}\|_2^2 + \frac{1}{2\alpha^2}.$$

According to the fact that $\|\boldsymbol{u}\|_2^2 = \|\boldsymbol{D}^\top \boldsymbol{\theta}\|_2^2 \leq \|\boldsymbol{D}\|_{2\to 2}{}^2 \|\boldsymbol{\theta}\|_2^2$, Assumption **(H.4'-SFB)** holds.

Set $\boldsymbol{v} = 2\gamma \boldsymbol{X}^\top \boldsymbol{y}/\beta$, $\boldsymbol{t}_{\boldsymbol{\theta}} = \boldsymbol{\theta} - \gamma \nabla F_\beta(\boldsymbol{\theta}) = \boldsymbol{M}_\gamma \boldsymbol{\theta} + \boldsymbol{v}$ and $\boldsymbol{p}_{\boldsymbol{t}_{\boldsymbol{\theta}}} = \mathrm{P}_{\left\{\boldsymbol{x} \,:\, \alpha \sum_l \left\|\boldsymbol{x}_{b_l}\right\|_2 \leq 1\right\}}(\boldsymbol{t}_{\boldsymbol{\theta}})$. By Young's inequality, we obtain that

$$\left\langle \operatorname{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{t}_{\boldsymbol{\theta}}), \boldsymbol{\theta} \right\rangle_{\boldsymbol{M}_\gamma} = \left\langle \boldsymbol{M}_\gamma \operatorname{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{t}_{\boldsymbol{\theta}}), \boldsymbol{\theta} \right\rangle \leq \frac{1}{2}\|\boldsymbol{M}_\gamma\|_{2\to 2}{}^2 \left\|\operatorname{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{t}_{\boldsymbol{\theta}})\right\|_2^2 + \frac{1}{2}\|\boldsymbol{\theta}\|_2^2.$$

Moreover, owing to (6.29), we get that

$$\left\|\operatorname{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{t}_{\boldsymbol{\theta}})\right\|_2^2 = \left\|\boldsymbol{t}_{\boldsymbol{\theta}} - \boldsymbol{p}_{\boldsymbol{t}_{\boldsymbol{\theta}}}\right\|_2^2 \leq 2\|\boldsymbol{t}_{\boldsymbol{\theta}}\|_2^2 + 2\|\boldsymbol{p}_{\boldsymbol{t}_{\boldsymbol{\theta}}}\|_2^2 \leq 4\|\boldsymbol{M}_\gamma\|_{2\to 2}{}^2 \|\boldsymbol{\theta}\|_2^2 + \left(4\|\boldsymbol{v}\|_2^2 + \frac{2}{\alpha^2}\right).$$

Thus, Assumption **(H.4'-FB)** holds and we conclude the proof of Lemma 6.5.6. □

The proposed prior in Chapter 3 Consider the prior π in (3.7) with g is given in Example 3.3.10 and $H = +\infty$. Then,

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = W_{\boldsymbol{\lambda}}(\boldsymbol{D}^\top \boldsymbol{\theta}), \quad \text{where} \quad W_{\boldsymbol{\lambda}}(\boldsymbol{u}) = \sum_{l=1}^{L} w_{\boldsymbol{\lambda}}(\boldsymbol{u}),$$

with

$$w_{\boldsymbol{\lambda}}(x) = \beta \left(\alpha^a x^a + c \log(\tau^b + x^b) \right), \tag{6.30}$$

$w_{\boldsymbol{\lambda}}$ is parameterized by $\boldsymbol{\lambda} = (a, b, c, \alpha, \tau) \in]0,1] \times]0,1] \times]\frac{2+K}{b}, +\infty[\times \mathbb{R}_+ \times \mathbb{R}^{+,*}$. Set $a = 1$ and $b = 1$, the following lemma checks the validity of Assumptions **(W.1)**, **(W.2)** and **(W.3)**.

Lemma 6.5.7. *Let $w_{\boldsymbol{\lambda}}$ given by (6.30). On $]0, +\infty[$, the function $w_{\boldsymbol{\lambda}}$ is bounded from below, non-decreasing and continuously differentiable. Fix $a = 1$ and $b = 1$, the problem $\min_{t \in [0,+\infty[} \{t + \frac{\gamma}{\beta} w_{\boldsymbol{\lambda}}{}'(t)\}$ has a unique solution at 0 for a given $0 < \gamma \leq \tau^2/c$.*

Proof. $w_{\boldsymbol{\lambda}}$ is clearly nondecreasing, bounded from below by $w_{\boldsymbol{\lambda}}(0)$, and continously differentiable (in fact even C^∞) on $]0, +\infty[$. Let us set $a = 1$ and $b = 1$, and

$$u(x) = x + \frac{\gamma}{\beta} w_{\boldsymbol{\lambda}}{}'(x) = x + \frac{\gamma c}{\tau + x} + \gamma\alpha.$$

One can see that u admits a local maximum at $x_0 = -\sqrt{\gamma c} - \tau \notin [0, +\infty[$ and a local minimum at $x_1 = \sqrt{\gamma c} - \tau$. Thus, the problem $\min_{t \in [0,+\infty[} u(t)$ has a unique solution at 0 when $x_1 \leq 0$ equivalent to $\gamma \leq \tau^2/c$. □

In view of Lemma 6.5.4, the EWA with the prior (3.7) is computable by FBLMC and Semi-FBLMC algorithms the proximal mapping $\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}$ is computed through Lemmas 6.5.2 and 6.5.3.

Chapter 7

Numerical Results

Main contributions of this chapter

- Collect the numerical experiments to illustrate and validate
  - the performance of the EWA proposed in Chapter 3,
  - the algorithms proposed in Chapter 6.

The results in this chapter can be found in [100] and [64].

Contents

7.1 Introduction

7.1.1 Problem statement

We consider a linear regression problem

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_0 + \boldsymbol{\xi}, \tag{7.1}$$

where $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. The noise level σ is chosen according to the simulated $\boldsymbol{\theta}_0$ through the signal-to-noise ratio SNR, i.e. $\sigma = n^{-1/2} \left\|\boldsymbol{X}\boldsymbol{\theta}_0\right\|_2 / 10^{\mathrm{SNR}/10}$. In our experiments, we take $\mathrm{SNR} = 5$.

The goal is estimating $\boldsymbol{\theta}_0$ by computing the EWA via the priors proposed in Section 6.5. Three types of problems are considered: compressed sensing, inpainting and deconvolution whose regression function described in what follows.

Compressed sensing In this case $\boldsymbol{X}$ is drawn from a random ensemble. In our experiments, $\boldsymbol{X}$ is drawn uniformly at random from the Rademacher ensemble, i.e. its entries are i.i.d. Rademacher random variables. We also set $n = 9p/16$.

Inpainting In this case, $\boldsymbol{X}$ acts as a masking operator. Let $\mathcal{M} \subset \{1, \ldots, p\}$ be the set indexing masked pixels. Thus

$$\boldsymbol{X} \operatorname{vec}(\boldsymbol{\theta}_0) = \left(\operatorname{vec}(\boldsymbol{\theta})_{j \in \{1,\ldots,p\} \setminus \mathcal{M}}\right).$$

In our numerical experiments, we mask out 20% of the pixels, and thus $n = p - \lfloor 20\% p \rfloor$ where $\lfloor p \rfloor$ the stands of integer part of p. The masked positions are chosen randomly from the uniform distribution.

Deconvolution In this case $\boldsymbol{X}$ is the convolution operator with a Gaussian kernel with periodic boundary conditions, such that $\boldsymbol{X}$ is diagonalized in the discrete Fourier basis. In this experiment, the standard deviation of the kernel is set to 1.

Two types of processing experiments are considered.

Image processing experiments Let $\boldsymbol{\theta}_0$ is a 2-D image which is a matrix in $\mathbb{R}^{128 \times 128}$. The model (7.1) becomes

$$\boldsymbol{y} = \boldsymbol{X} \operatorname{vec}(\boldsymbol{\theta}_0) + \boldsymbol{\xi}. \tag{7.2}$$

Note that $p = 128^2$.

Assuming that the targeted image is piecewise smooth, a popular prior is the so-called isotropic total variation [128]. To cas this into our framework, define $\boldsymbol{D}_c : \mathbb{R}^{\sqrt{p} \times \sqrt{p}} \to \mathbb{R}^{\sqrt{p} \times \sqrt{p}}$ and $\boldsymbol{D}_r : \mathbb{R}^{\sqrt{p} \times \sqrt{p}} \to$

$\mathbb{R}^{\sqrt{p}\times\sqrt{p}}$ the finite difference operators along, respectively, the columns and rows of an image, with Neumann boundary conditions. We define $\boldsymbol{D}_{\mathrm{TV}}$ as

$$\boldsymbol{D}_{\mathrm{TV}}\colon \boldsymbol{\theta} \in \mathbb{R}^{\sqrt{p}\times\sqrt{p}} \mapsto \mathrm{vec}\left((\mathrm{vec}(\boldsymbol{D}_r(\boldsymbol{\theta})), \mathrm{vec}(\boldsymbol{D}_c(\boldsymbol{\theta})))^\top\right)^\top \in \mathbb{R}^{2p}.$$

With this notation, our prior penalty $J_{\boldsymbol{\lambda}}$ reads

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \sum_{l=1}^{p} w_{\boldsymbol{\lambda}}\left(\sqrt{\mathrm{vec}(\boldsymbol{D}_r(\boldsymbol{\theta}))_l^2 + \mathrm{vec}(\boldsymbol{D}_c(\boldsymbol{\theta}))_l^2}\right) = W_{\boldsymbol{\lambda}}(\boldsymbol{D}_{\mathrm{TV}}\boldsymbol{\theta}), \tag{7.3}$$

which clearly has the form (6.18) with p blocks of equal size 2.

For image processing experiments, the EWA will be computed by Semi-FBLMC Algorithm.

Signal processing experiments Here we consider reconstructing a 1D signal, with $p = 128$, from compressed sensing measurements using the EWA with ℓ_∞ penalty and the one proposed in Chapter 3.

(i) For EWA-ℓ_∞ estimator: The signal is piecewise flat whose coordinates are valued in $\{-1, 1\}$. Here, $\boldsymbol{X}$ is drawn from a Rademacher ensemble with $n > p$ [1].

(ii) For EWA proposed in Chapter 3: The signal is either individual sparse or group sparse. The non-zero entries of $\boldsymbol{\theta}_0$ are equal to 1.

Here, we consider $\boldsymbol{D} = \mathbf{I}_p$.

For signal processing experiments, the EWA will be computed by FBLMC Algorithm.

7.1.2 Chapter organization

The performance of the EWA proposed in Chapter 3 is illustrated by the numerical experiments in Section 7.2. The forward-backward LMC type algorithms are validated in Section 7.3 for EWA with ℓ_1, SCAD, FIRM, ℓ_∞ penalties.

7.2 Numerical results on EWA for analysis-group sparsity

7.2.1 Signal processing experiments

We set $\boldsymbol{D} = \mathbf{I}_p$, which corresponds to the classical group sparsity. The design matrix is drawn uniformly at random from the Rademacher ensemble, i.e. its entries are i.i.d. variates valued in $\{-1, 1\}$ with equal probabilities. The non-zero entries of $\boldsymbol{\theta}_0$ are equal to 1 and we denote $s = \|\boldsymbol{\theta}_0\|_0$ the sparsity level of $\boldsymbol{\theta}_0$. Two types of sparsity behavior are considered: individual sparsity where $K_{\boldsymbol{\theta}_0} = 1$; group structured sparsity with $K_{\boldsymbol{\theta}_0} = 4$. Besides, the positions of the non-zero/active entries (for $K_{\boldsymbol{\theta}_0} = 1$) or groups (for $K_{\boldsymbol{\theta}_0} = 4$) are chosen randomly uniformly on $\{1, \ldots, p\}$.

The experiments are performed by fixing $p = 128$, and taking $s = 4, 8, \ldots, p$, $n = 8, 16, \ldots, p$, step-size $\delta = 4\sigma^2/(np)$ and integration time $T = 3500$. The parameters in the prior are chosen to minimize the remainder term in the oracle inequality (3.13). For each (s, n), and each value of $K_{\boldsymbol{\theta}_0}$, $N_{\mathrm{rep}} = 20$ instances of the problem suite $(\boldsymbol{X}, \boldsymbol{\theta}_0, \boldsymbol{Y})$ are generated, and EWA is applied with a chosen K and the other parameters as detailed above. The estimation quality/success is then assessed by

$$\pi_{s,n} = \frac{1}{N_{\mathrm{rep}}} \sum_{j=1}^{N_{\mathrm{rep}}} I\left(\left\|\widehat{\boldsymbol{\theta}}_n^{(j,s,n)} - \boldsymbol{\theta}_0^{(j,s,n)}\right\|_n \leq \epsilon\right), \tag{7.4}$$

where $\epsilon > 0$ (we choose $\epsilon = 0.4$) and $\widehat{\boldsymbol{\theta}}_n^{(j,s,n)}$ (resp. $\boldsymbol{\theta}_0^{(j,s,n)}$) corresponds to $\widehat{\boldsymbol{\theta}}_n$ (resp. $\boldsymbol{\theta}_0$) in the j-th replication of (s, n).

s/p and n/p are respectively normalized measures of sparsity and problem indeterminacy. We get a two-dimensional phase space $(s/p, n/p) \in [0, 1]^2$ describing the difficulty of a problem instance, i.e.

[1] The overdetermined regime is known to yield good performance for the ℓ_∞ penalty [145].

problems are easier as one moves up (more measurements) and to the left (sparser $\boldsymbol{\theta}_0$). Phase diagrams plotting $\pi_{s,n}$ in (7.4) as a function $(s/p, n/p)$ were widely advocated by Donoho and co-authors for ℓ_1 minimization [59]. Such diagrams often have an interesting two-phase structure (as displayed in Figures 7.1(a)-(d), brighter color indicate better success), with phases separated by a specific curve, called phase transition curve. Thus, a good estimator is intended to have a large bright area which indicates its good performance at a wider range of (s, n).

Figure 7.1(a) (resp. (b)) shows the phase diagrams when $K_{\boldsymbol{\theta}_0} = 1$ and $K = 1$ (resp. $K = 4$) in EWA. In this case, the phase transition curve for $K = 1$, the correct group size, is slightly better that with $K = 4$. The situation reverses for Figures 7.1(c)-(d) where $K_{\boldsymbol{\theta}_0} = 4$, and one observes that the success area is significantly better using $K = 4$ than $K = 1$. This is expected as it reveals better performance of EWA when used with the choice $K = K_{\boldsymbol{\theta}_0}$. This is also confirmed by visual inspection of Figures 7.1(c')-(d'), where we plotted instances of recovered vectors $\widehat{\boldsymbol{\theta}}_n^{(j,s,n)}$ when $(s, K_{\boldsymbol{\theta}_0}) \in \{4, 8\} \times \{1, 4\}$ and $n/p = 1/2$. EWA was again applied with $K = 1$ and $K = 4$ in each case. Large spurious entries appear outside the true support when the group size is not correctly chosen, though the impact is less important for $K = 1$.

It is worth observing that $s/p = \|\boldsymbol{\theta}_0\|_{0,2} K_{\boldsymbol{\theta}_0}/p$. As far as the expected phase transition curve is concerned, one has from Corollary 3.4.6 that it is expected to occur for

$$n/p = C_\epsilon \|\boldsymbol{\theta}_0\|_{0,2} K_{\boldsymbol{\theta}_0}/p \left(\log(p/K_{\boldsymbol{\theta}_0})/K_{\boldsymbol{\theta}_0}\right) = C_\epsilon s/p \left(\log(p/K_{\boldsymbol{\theta}_0})/K_{\boldsymbol{\theta}_0}\right)$$

for some constant $C_\epsilon > 0$ depending on ϵ. That is, the phase transition curve is linear (p and $K_{\boldsymbol{\theta}_0}$ are fixed for each diagram), which is confirmed by visual inspection of Figures 7.1(a)-(d), where the overlayed blue line is the fitted linear phase transition curve.

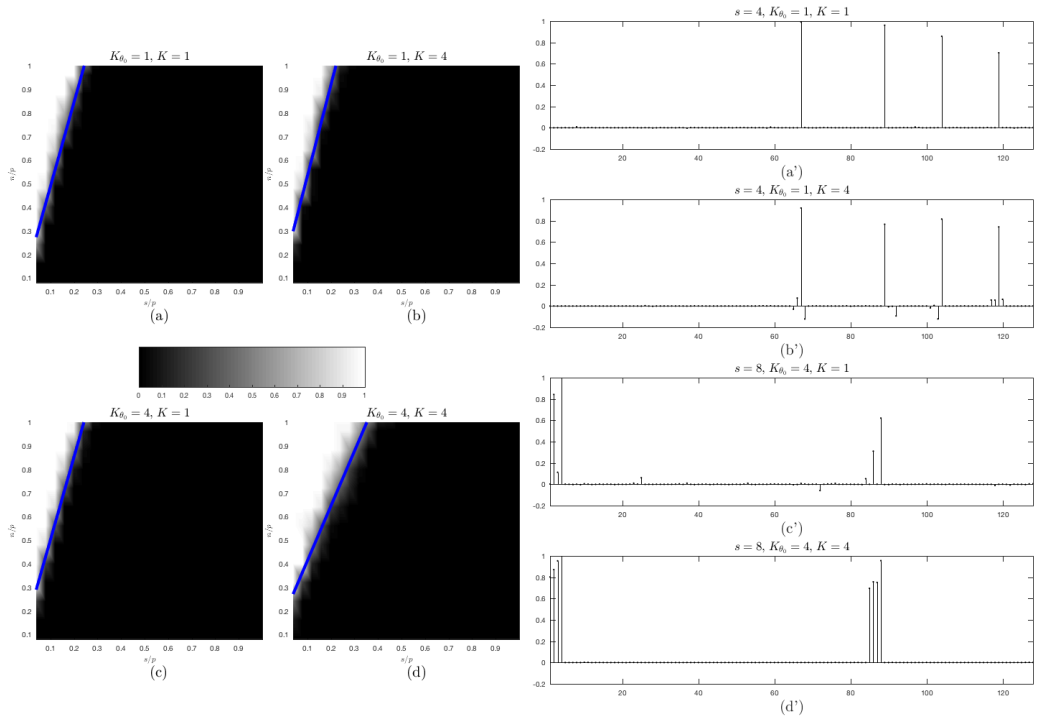

Figure 7.1: (a)-(d): Phase diagrams of EWA for $\boldsymbol{D} = \mathbf{I}_p$, the color bar ranges from dark ($\pi_{s,n} = 0$) to bright ($\pi_{s,n} = 1$). The blue line is the fitted phase transition curve. (a')-(d'): Examples of vectors $\widehat{\boldsymbol{\theta}}_n^{(j,s,n)}$ recovered by EWA with $n/p = 1/2$, two sparsity levels $s = 4$ and $s = 8$ and two group sizes $K_{\boldsymbol{\theta}_0} = 1$ and $K_{\boldsymbol{\theta}_0} = 4$.

7.2.2 Image processing experiments

In the second numerical experiment, $\boldsymbol{\theta}_0$ is a 2-D image which is a matrix in $\mathbb{R}^{160\times 160}$ (a close-up of the known Shepp-Logan phantom, see Figure 7.2(a)). Thus $\mathrm{vec}(\boldsymbol{\theta}_0)$ is vector in $\mathbb{R}^p$ with $p = 160^2$. Our goal is to recover $\boldsymbol{\theta}_0$ in the compressed sensing problem.

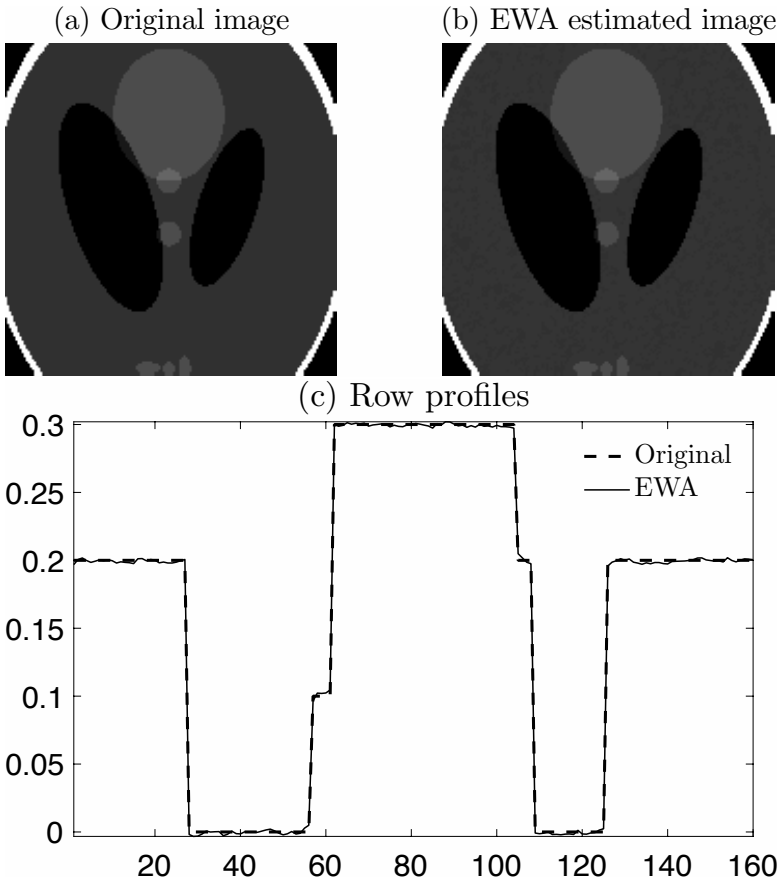

Figure 7.2: (a): Original close-up of Shepp-Logan phantom image. (b): Image recovered by EWA with $\delta = 2 \cdot 10^{-8}$ and $T = 10^4$. (c) Profiles of a row extracted from each image.

The results are depicted in Figure 7.2. In this experiment, the number of observations is $n = 9p/16 = 14400$, and we have $\|\boldsymbol{D}_{\mathrm{TV}}(\theta_0)\|_{0,2} = 1376 \ll n$. A notable property of the EWA estimate is that it does not suffer from the stair-casing effect, unlike total variation minimization.

7.3 Numerical results on Forward-Backward LMC type algorithms

7.3.1 Image processing experiments

Let $\boldsymbol{\theta}_0$ is a 2-D image which is a matrix in $\mathbb{R}^{128\times 128}$. Thus $\mathrm{vec}(\boldsymbol{\theta}_0)$ is vector in $\mathbb{R}^p$ with $p = 128^2$. The goal is estimating $\boldsymbol{\theta}_0$ by computing the EWA with $w_{\boldsymbol{\lambda}}$ as the ℓ_1, SCAD and FIRM penalties. Three types of problems are considered: compressed sensing, inpainting and deconvolution. The corresponding estimators are denoted respectively EWA-ℓ_1, EWA-SCAD and EWA-FIRM. Because of the presence of the analysis operator $\boldsymbol{D}_{\mathrm{TV}}$, which is not unitary, we applied Semi-FBLMC scheme (6.17) to compute EWA with $\beta = 1/(pn)$ (see Remark 4.3.2) , $\gamma = \beta$, and $\delta = \{5\beta/10^3, 5\beta/10^2, 5\beta/10^6\}$ respectively associated to inpainting, deconvolution and compressed sensing problems. The results are depicted in Figure 7.4.

7.3.2 Signal processing experiments

Here we consider reconstructing a piecewise flat 1D signal from compressed sensing measurements using EWA whose coordinates are valued in $\{-1, 1\}$. We set $F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) = \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2$, $J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \left\|\boldsymbol{\theta}\right\|_\infty$, i.e. $\boldsymbol{D} = \mathbf{I}_p$ and the size of groups is 1. We can then use the FBLMC scheme (6.16), where we choose $\beta = 1/(pn)$ (see Remark 4.3.2), $\gamma = \beta$, and $\delta = 5/10^2$. The results are shown in Figure 7.3.

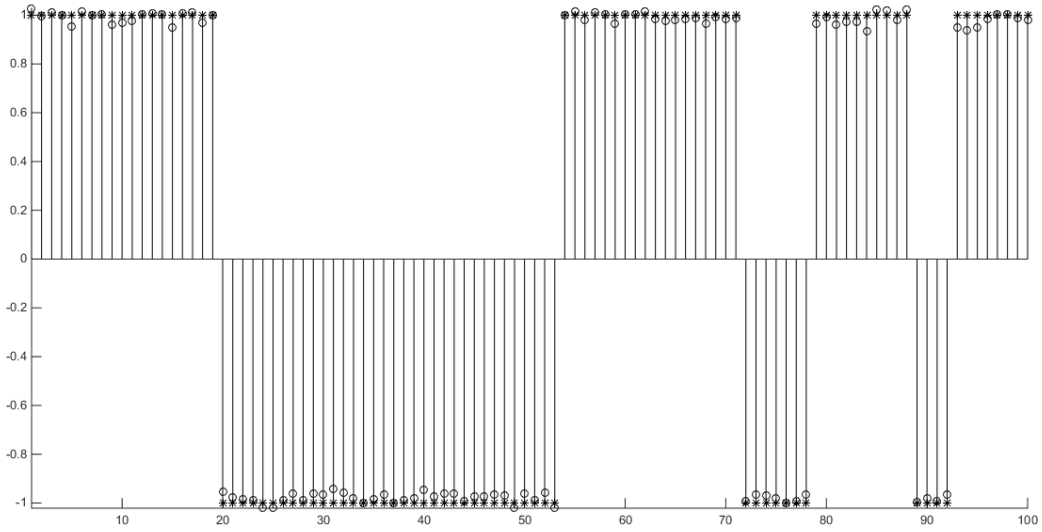

Figure 7.3: Compressed sensing with EWA using the ℓ_∞ penalty. $'*'$ is the original signal and $'\circ'$ is the the estimated one.

7.4 Reproducible research

Following the philosophy of reproducible research, all the code implementing our EWA algorithms and reproducing the experiments of this manuscript are made publicly available for download on:

(i) For experiments in Section 7.2: `https://github.com/luuduytung/GroupAnalyseEWAToolbox`

(ii) For experiments in Section 7.3: `https://github.com/luuduytung/LMCToolbox`
analysis-group

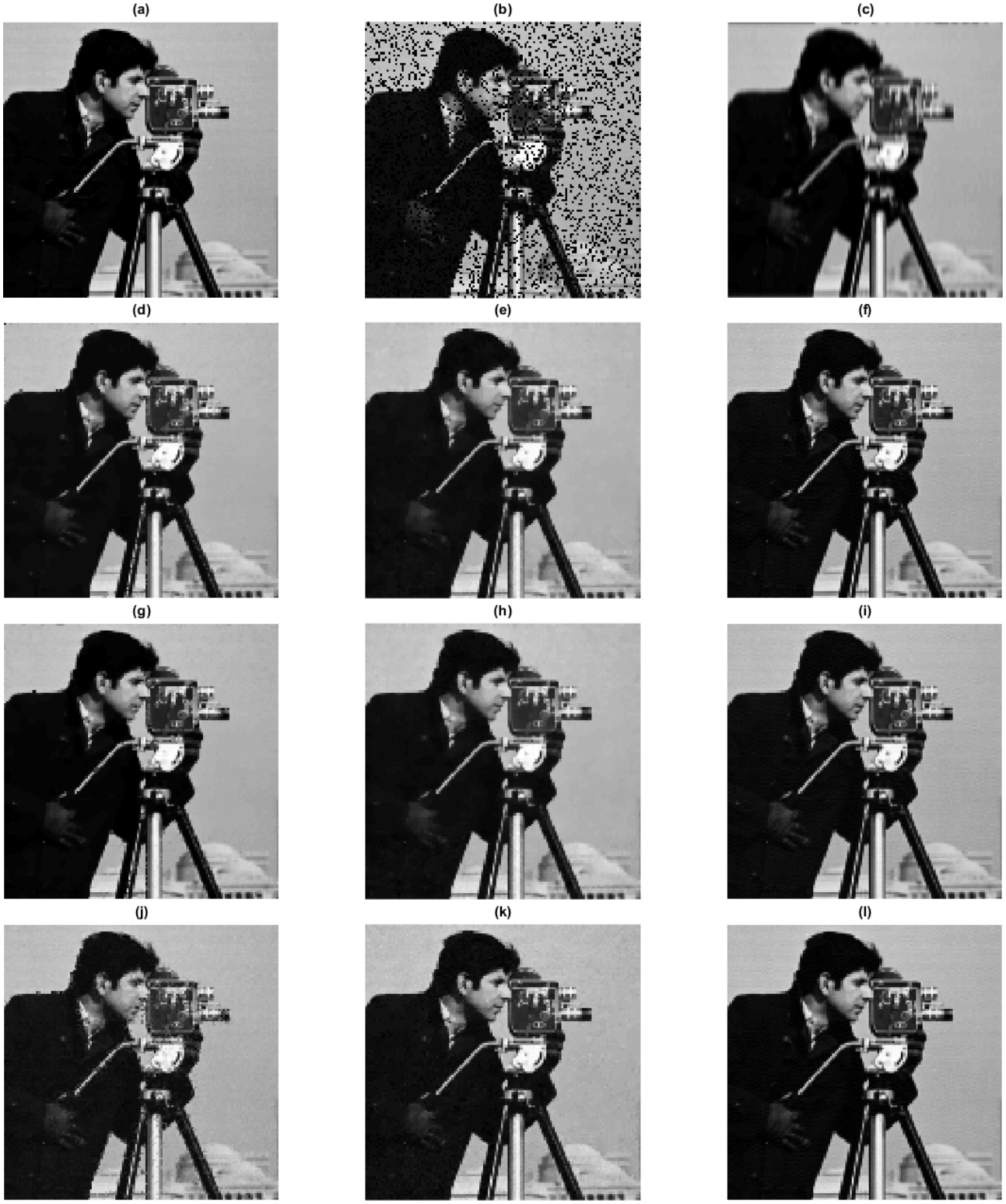

Figure 7.4: (a): Original image. (b,c) Observed masked and blurry images. (d, e, f): EWA-ℓ_1 estimated images from masked image, compressed sensing measurements, and blurry image. (g, h, i): EWA-FIRM estimated images from masked image, compressed sensing measurements, and blurry image. (j, k, l): EWA-SCAD estimated images from masked image, compressed sensing measurements, and blurry image.

Chapter 8

Conclusions and Perspectives

This manuscript provides new results on penalized estimators and EWA around two main themes: theoretical performance guarantees, namely prediction oracle inequalities (Chapter 3 and 4) and estimation bounds which are the heart of Chapter 5; new MC algorithms to sample from structured non-necessarily smooth nor log-concave densities based on Langevin diffusion and proximal splitting, together with their consistency guarantees (Chapter 6). Our results were also supported by several numerical experiments (Chapter 7).

Many of our results are unifying with an unprecedented level of generality, by highlighting the role of geometrical quantities tied to the involved low-complexity regularizing penalties. Consequently, they cover many popular priors in the literature as corollaries. Our algorithmic results provide an insightful and theoretically-grounded support to proximal splitting-based sampling algorithms that were recently applied in the literature by some authors, but whose justification and guarantees were until now open problems. This is in particular true for the FBLMC algorithm.

From this work, many conclusions and take-away messages can be drawn.

Take away messages:

(i) Our family of priors proposed in Chapter 3 offers flexibility thanks to the tuning parameters. This allows us to generalize the results in [52] to the analysis-group sparsity. This also optimizes the performance of EWA by tuning these parameters for the particular dataset at hand.

(ii) From the unified analysis for EWA and the penalized estimator, we constructed their theoretical guaranties for a general family of loss functions (beyond the quadratic loss) and penalties, and developed them in a random context. Namely, we established the prediction oracle inequalities (resp. estimation bounds) in probability when the observations (resp. the design matrix) are random. Moreover, we refined them by assuming that the random part is drawn from a sub-Gaussian or Gaussian distribution, and specialized them to several instances in the literature including the Lasso, the group Lasso, their analysis type counterparts, the ℓ_∞ and the nuclear norm penalties.

(iii) Two algorithms were proposed: FBLMC and Semi-FBLMC. These algorithms exploit the composite structure of the distribution and do not require smoothness nor log-concavity. However, in sone situations, the computation of the implicit step in FBLMC may become expensive, in which case, the Semi-FBLMC is clearly a preferable option.

(iv) The numerical experiments illustrated the performance of the EWA with several instances of priors, and in several numerical problems. In particular, by the phase diagram, we validated the oracle inequality guaranteeing the performance of the EWA proposed in Chapter 3. An important remark is that, in the image processing experiments with TV prior, the EWA estimate does not suffer from the stair-casing effect, unlike total variation minimization.

Perspectives The research program investigated in this work has many open questions that are yet to be answered.

(i) **Oracle inequalities under milder assumptions:** In Chapter 4, we considered a finite valued and positively homogeneous penalty J. In the random context, the observations $\boldsymbol{y}$ were assumed to be i.i.d. and the loss function $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is decomposed into a n-independent sum of functions $f_i : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. Then, it would be interesting to generalize those results to J beyond the 1-homogeneous case, and $\boldsymbol{y}$ and F beyond independency.

(ii) **Estimation bounds under milder assumptions:** In Chapter 5, we considered the case where the design matrix has components drawn from the standard Gaussian. The extension for structured design matrices (e.g. partial Fourier) is a direction for further research.

(iii) **Unified analysis of minimax lower-bounds:** Our sharp prediction oracle inequalities were shown to be close to the minimax lower-bounds. Minimax lower-bounds for estimation clearly need to be worked out in a general setting.

(iv) **Model selection properties for EWA:** In their work on total variation denoising (i.e. $\boldsymbol{X} = \mathbf{I}_n$), the authors in [98] has shown that EWA does not suffer the stair-casing effect. Investigating this for arbitrary design and beyond TV regularization is clearly an important perspective.

(v) **Convergence rate guarantees for FBLMC and variants in the non-convex case:** In Chapter 6, we proved the consistency of the FBLMC and Semi-FBLMC algorithms. However, convergence rates to the stationary distribution is still an open problem. In the convex case, we can show exponential convergence for (Semi-)FBLMC with a weighted average by capitalizing on the results of [63, 91]. However, exploring the convergence rate for our algorithms in the non-convex case requires new arguments that are still to be developed. We believe that this is an important direction for further research.

List of Publications

Preprints

(3) T. D. Luu, J. Fadili, C. Chesneau, *PAC-Bayesian risk bounds for group-analysis sparse regression by exponential weighting*, submitted to Electronic Journal of Statistics (EJS), *hal-01367742*.

(4) T. D. Luu, J. Fadili, C. Chesneau, *Sharp oracle inequalities for low-complexity priors*, submitted to Annals of the Institute of Statistical Mathematics (AISM), *arXiv:1702.03166*.

(5) T. D. Luu, J. Fadili, C. Chesneau, *Sampling from non-smooth distribution through Langevin diffusion*, submitted to Computational Statistics and Data Analysis (CSDA), *arXiv:1412.7392*.

Conference Proceedings

(9) T. Luu, J. Fadili et C. Chesneau *Agrégation à poids exponentiels: Algorithmes d'échantillonnage*, Colloque sur le Traitement du Signal et des Images (GRETSI), 2017.

(10) T. Luu, J. Fadili et C. Chesneau, *Agrégation à poids exponentiels et estimation penalisée : Inégalités oracles*, Colloque sur le Traitement du Signal et des Images (GRETSI), 2017.

(11) T. Luu, J. Fadili et C. Chesneau *Agrégation à poids exponentiels: Algorithmes d'échantillonnage*, Colloque ORASIS, 2017.

List of Notations

Sets

$\mathbb{R}$: Set of real numbers.
$\mathbb{R}_+$: Set of non-negative real numbers.
$\overline{\mathbb{R}}$: $\mathbb{R} \cup \{+\infty\}$.
$\mathbb{R}^d$: d-dimensional real Euclidean space.
$\mathbb{R}^{d\times r}$: Set of $d \times r$ real matrices.
$\{a, \ldots, b\}$: Set of integers x such that $a \leq x \leq b$.
$\mathcal{H}$: Dictionary of aggregation.
$\mathcal{F}_\Theta$: EWA dictionary.

Functions

$\mathrm{dom}(f)$: Effective domain of a function f.
$f_{\boldsymbol{\theta}}$: Linear combination of elements in the dictionary $\mathcal{H}$ with the weights in $\boldsymbol{\theta}$.
f^*: Legendre-Fenchel conjugate of a function f.
f^+: Monotone conjugate of a function f.
f°: Polar of a function f.
${}^\gamma f$: Moreau envelope of a function f of order $\gamma > 0$.
${}^{\boldsymbol{M},\gamma} f$: Moreau envelope of a function f of order $\gamma > 0$ in the metric $\boldsymbol{M}$.
$\sigma_{\mathcal{C}}$: Support function of a set $\mathcal{C}$.
$\gamma_{\mathcal{C}}$: Gauge (Minkowski functional) of a set $\mathcal{C}$.
$J^\circ_{f_{\boldsymbol{x}}}$: Subdifferential gauge associated to $f_{\boldsymbol{x}} \in \mathrm{ri}(\partial J(\boldsymbol{x}))$.
∇f: Gradient of a function f.
∇g: Gradient w.r.t. the first variable of a bivariate function g.
$\partial^F f$: Fréchet subdifferential of a function f.
∂f: Subdifferential of a function f.
$D_f^{\boldsymbol{\eta}}(\cdot, \boldsymbol{x})$: Bregman divergence associated to a convex function f at $\boldsymbol{x}$ w.r.t. $\boldsymbol{\eta} \in \partial f(\boldsymbol{x})$.
$\mathrm{T}^f_{\epsilon,\overline{\boldsymbol{x}},\overline{\boldsymbol{v}}}(\cdot)$: f-attentive ϵ-localization (with $\epsilon > 0$) of ∂f around $(\overline{\boldsymbol{x}}, \overline{\boldsymbol{v}})$.
$C^k(\mathbb{R}^d)$: Class of functions $f : \mathbb{R}^d \to \mathbb{R}$ such that its first k derivatives are all exist and are continuous (with k is a non-negative integer).
$C^\infty(\mathbb{R}^d)$: Class of functions $f : \mathbb{R}^d \to \mathbb{R}$ such that its k-th derivative exists and continuous for any non-negative integer numbers k.
$\widetilde{C^{1,+}}(\mathbb{R}^d)$: Class of differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$ whose gradient is locally Lipschitz continuous and there exists $K > 0$, $\langle \boldsymbol{x}, \nabla f(\boldsymbol{x})\rangle \leq K(1 + \left\|\boldsymbol{x}\right\|_2^2)$, $\forall \boldsymbol{x} \in \mathbb{R}^d$.
$\mathcal{J}(\mathbb{R}^d)$: Class of functions which are proper, lsc and bounded from below.
$\mathcal{G}(\mathbb{R}^d)$: Class of finite-valued gauges.
$\Gamma_0(\mathbb{R}^d)$: Class of proper lsc convex functions.

Operators

$\mathrm{gph}(S)$: Graph of an operator S.

sgn: Sign operators.
$(\cdot)_+$: Positive part of a real number.
$\lfloor \cdot \rfloor$: Stands of integer part.
Γ: Gamma function.
$\mathrm{prox}_{\gamma f}$: Moreau proximal mapping of a function f.
$\mathrm{prox}_{\gamma f}^{\boldsymbol{M}}$: Moreau proximal mapping of a function f in the metric $\boldsymbol{M}$.
$\mathrm{P}_{\mathcal{C}}$: Orthogonal projector on a set $\mathcal{C}$.
Id: Identity operator on the space of $p_1 \times p_2$ matrices (should not be confused with the identity matrix).

Operations on sets

$|\mathcal{C}|$: Cardinality of a set $\mathcal{C}$.
$\mathcal{C}^c$: Complement of a set $\mathcal{C}$.
$\mathcal{C}^\circ$: Polar of a set $\mathcal{C}$.
$\iota_{\mathcal{C}}$: Indicator operator of a set $\mathcal{C}$.
$I_{\mathcal{C}}$: Characteristic operator of a set $\mathcal{C}$.
$\mathrm{bd}(\mathcal{C})$: Boundary of a set $\mathcal{C}$.
$\mathrm{Span}(\mathcal{C})$: Smallest linear manifold containing the set $\mathcal{C}$.
$\mathrm{conv}\,(\mathcal{C})$: Convex hull of a set $\mathcal{C}$.
$\overline{\mathrm{conv}}\,(\mathcal{C})$: Closure of the convex hull of a set $\mathcal{C}$.
$\mathrm{aff}(\mathcal{C})$: Affine hull of a set $\mathcal{C}$.
$\mathrm{par}(\mathcal{C})$: Parallel subspace of a set $\mathcal{C}$.
$\mathrm{ri}(\mathcal{C})$: Relative interior of a set $\mathcal{C}$.

Notations in linear algebra

Bold uppercase letters: Matrices in an Euclidean space.
Bold lowercase letters: Vectors in an Euclidean space.
$\mathbf{I}_d$: Identity matrix on $\mathbb{R}^d$.
$\ker(\boldsymbol{M})$: Kernel of a matrix $\boldsymbol{M}$.
$\mathrm{Span}(\boldsymbol{M})$: Image of a matrix $\boldsymbol{M}$.
$\mathrm{rank}(\boldsymbol{M})$: Rank of a matrix $\boldsymbol{M}$.
$\mathrm{tr}(\boldsymbol{M})$: Trace of a square matrix $\boldsymbol{M}$.
$\det(\boldsymbol{M})$: Determinant of a square matrix $\boldsymbol{M}$.
$\boldsymbol{\sigma}(\boldsymbol{M})$: Vector of singular values of a matrix $\boldsymbol{M}$ in non-increasing order.
$\sigma_{\min}(\boldsymbol{M})$: The smallest singular values of a matrix $\boldsymbol{M}$.
$\boldsymbol{M}^\top$: Transpose of a matrix $\boldsymbol{M}$.
$\mathrm{vec}(\boldsymbol{M})$: Vectorization opertator of a matrix $\boldsymbol{M}$.
$\mathrm{supp}(\boldsymbol{x})$: Support of a vector $\boldsymbol{x}$.
$\boldsymbol{x}_I$: Subvector whose entries are those of the vector $\boldsymbol{x}$ indexed by a index set I.
$\boldsymbol{M}_I$: Submatrix whose columns are those of the matrix $\boldsymbol{M}$ indexed by a index set I.
$\boldsymbol{M}_{I,J}$: Submatrix whose columns and rows are those of the matrix $\boldsymbol{M}$ indexed by index sets I and J.
$\widetilde{\boldsymbol{M}}$: Canonical dual frame associated to a frame $\boldsymbol{M}$.
$\boldsymbol{M}^+$: Moore-Penrose pseudo inverse of a matrix $\boldsymbol{M}$.
$\boldsymbol{X}$: Design matrix.
$\mathrm{diag}(\boldsymbol{x})$: Diagonal matrix whose diagonal entries are the components of $\boldsymbol{x}$.
$\langle \boldsymbol{x}, \boldsymbol{y} \rangle$: Scalar product of two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.

$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{M}}$: Scalar product of two vectors $\boldsymbol{x},\ \boldsymbol{y} \in \mathbb{R}^d$ associated to the metric $\boldsymbol{M}$.
$\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F$: Frobenius scalar product of two matrices $\boldsymbol{A},\ \boldsymbol{B} \in \mathbb{R}^{d \times d}$.
$\|\boldsymbol{x}\|_{\boldsymbol{M}}$: norm of a vector $\boldsymbol{x}$ associated to the metric $\boldsymbol{M}$.
$\|\boldsymbol{x}\|_p$: ℓ_p norm (or pseudo-norm or semi-norm) of a vector $\boldsymbol{x}$ for $p \in [0, +\infty]$.
$\|\boldsymbol{x}\|_{p,2}$: $\ell_{p,2}$-group norms (or pseudo-norm or semi-norm) of a vector $\boldsymbol{x}$ for $p \in [0, +\infty]$.
$\|\boldsymbol{M}\|_{S_p}$: Schatten p-norm (or pseudo-norm or semi-norm) of a matrix $\boldsymbol{M}$ for $p \in [0, +\infty]$.
$\|\boldsymbol{M}\|_{J_1 \to J_2}$: Operator bound of a matrix $\boldsymbol{M}$ for $J_1, J_2 \in \mathcal{G}(\mathbb{R}^d)$.
$\|\boldsymbol{M}\|_*$: Nuclear norm of a matrix $\boldsymbol{M}$.
$\|\boldsymbol{M}\|_F$: Frobenius norm of a matrix $\boldsymbol{M}$.
$\|\boldsymbol{M}\|_{2 \to 2}$: Operator (Spectral) norm of a matrix $\boldsymbol{M}$.
$\dim(T)$: Dimension of a subspace T.
$T^\perp$: Orthogonal subspace of a subspace T.
$\boldsymbol{x}_T$: Orthogonal projection of a vector $\boldsymbol{x}$ on the subspace T.
$\boldsymbol{M}_T$: $\boldsymbol{M}\, \mathrm{P}_T$ for a matrix $\boldsymbol{M}$ and subspace T.
$\boldsymbol{A}^*$: Adjoint of a linear operator $\boldsymbol{A}$.

Geometrical decomposability

$e^J_{\boldsymbol{x}}$: $\mathrm{P}_{\mathrm{aff}(\partial J(\boldsymbol{x}))}(0)$ for $J \in \Gamma_0(\mathbb{R}^d)$.
$S^J_{\boldsymbol{x}}$: $\mathrm{par}(\partial J(\boldsymbol{x}))$ for $J \in \Gamma_0(\mathbb{R}^d)$.
$T^J_{\boldsymbol{x}}$: ${S^J_{\boldsymbol{x}}}^\perp$ for $J \in \Gamma_0(\mathbb{R}^d)$ (Model subspace of a vector $\boldsymbol{x}$ associated to the function $J \in \mathcal{G}(\mathbb{R}^d)$).

Regression problem

$\mathcal{X}$: Design space.
$\mathcal{Y}$: Observation space.
$\{x_i\}_{i \in \{1,\dots,n\}}$: Designs.
$\{y_i\}_{i \in \{1,\dots,n\}}$: Observations.
$\{\xi_i\}_{i \in \{1,\dots,n\}}$: Noises.
$\boldsymbol{x}$: Design vector.
$\boldsymbol{y}$: Observation vector.
$\boldsymbol{\xi}$: Noise vector.
f: Regression function.
$\boldsymbol{g}$: $(g(x_1), \dots, g(x_n))^\top$ for a function $g : \mathcal{X} \to \mathbb{R}$.
$\|g\|_n$: $\sqrt{\sum_{i=1}^n g^2(x_i)/n}$ for a function $g : \mathcal{X} \to \mathbb{R}$.

List of Acronyms

EWA: Exponential Weighted Aggregation.
PEN: Penalization.
SURE: Stein's Unbiased Risk Estimate
SOI: Sparse Oracle Inequality.
lsc: Lower Semicontinuous.
i.i.d.: Independent and identically distributed.
SDE: Stochastic Differential Equation.
LMC: Langevin Monte-Carlo.
FBLMC: Forward-backward Langevin Monte-Carlo.
Semi-FBLMC: Semi-Forward-backward Langevin Monte-Carlo.
LASSO: Least Absolute Shrinkage and Selection Operator.
SCAD: Smoothly Clipped Absolute Deviation.
TV: Total Variation.

List of Figures

Bibliography

[1] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA (IMAIAI)*, 3, 2013.

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

[3] A. Antoniadis and J. Fan. Regularization of Wavelet Approximations. *Journal of the American Statistical Association*, 96:939–967, 2001.

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

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

[6] H. H. Bauschke, J. M. Borwein, and P. L. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on Control and Optimization*, 42(2):596–636, 2003.

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

[8] P. Bellec. Concentration of quadratic forms under a bernstein moment assumption. Technical report, Ecole Polytechnique, 2014.

[9] P. Bellec. Optimal bounds for aggregation of affine estimators. Working Papers 2015-06, Centre de Recherche en Economie et Statistique, 2015.

[10] F. Bernard and L. Thibault. Prox-regular functions in hilbert spaces. *Journal of Mathematical Analysis and Applications*, 303(1):1 – 14, 2005.

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

[12] 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.

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

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

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

[16] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

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

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

[19] 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.

[20] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 08 2007.

[21] E. Candès and D. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Communications on Pure and Applied Mathematics*, 57(2):219–266, 2004.

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

[23] E. Candes and B. Recht. Simple bounds for low-complexity model reconstruction. *Arxiv preprint arXiv:1106.1474*, 2011.

[24] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[25] E. J. Candès. Ridgelets: Estimating with Ridge Functions. *Annals of Statistics*, 31, 1999. 1561–1599.

[26] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31(1):59–73, 2011.

[27] 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.

[28] E. J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[29] E. J. Candès and Y. Plan. A probabilistic and RIPless theory of compressed sensing. *Information Theory, IEEE Transactions on*, 57(11):7235–7254, 2011.

[30] E. J. Candès and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on*, 57(4):2342–2359, 2011.

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

[32] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[33] 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.

[34] E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52(12):5406–5425, 2006.

[35] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 56(5):2053–2080, 2010.

[36] L. Chaari, J.-Y. Tourneret, C. Chaux, and H. Batatia. A hamiltonian monte carlo method for non-smooth energy sampling. Technical Report arXiv:1401.3988, , 2014.

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

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

[39] 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.

[40] C. Chesneau, M. Fadili, and J.-L. Starck. Stein block thresholding for image denoising. *Applied and Computational Harmonic Analysis*, 28(1):67–88, 2010.

[41] R. Coifman and D. Donoho. Translation invariant de-noising. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*. Springer-Verlag, 1995. 125–150.

[42] M. Coste. An introduction to semialgebraic geometry. Technical report, Institut de Recherche Mathematiques de Rennes, October 2002.

[43] D. Dai, P. Rigollet, and T. Zhang. Deviation optimal learning using greedy q-aggregation. *Ann. Statist.*, 40(3):1878–1905, 06 2012.

[44] 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.

[45] 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.

[46] A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Ann. Statist.*, 40, 2016.

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

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

[49] A. S. Dalalyan and J. Salmon. Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.*, 40(4):2327–2355, 08 2012.

[50] 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.

[51] A. S. Dalalyan and A. B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 18(3):914–944, 08 2012.

[52] 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.

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

[54] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and banach spaces. In B. Johnson and J. Lindenstrauss, editors, *Handbook on the Geometry of Banach spaces*, pages 317–366. Elsevier Scientific, 2001.

[55] C.-A. Deledalle, S. Vaiter, G. Peyré, and M. J. Fadili. Stein unbiased gradient estimator of the risk (SUGAR) for multiple parameter selection. *SIAM J. Imaging Sciences*, 7(4):2448–2487, 2014.

[56] 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.

[57] D. Donoho and I. Johnstone. Adapting to Unknown Smoothness Via Wavelet Shrinkage. *JASA*, 90(432):1200–1224, 1995.

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

[59] D. L. Donoho. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Phil. Trans. Royal Soc. A*, 367(1906):4273–4293, 2009.

[60] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[61] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia. *Journal of the Royal Statistical Society, Ser. B*, pages 371–394, 1995.

[62] A. Durmus and E. Moulines. Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm. Preprint hal-01176132, July 2015.

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

[64] 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.

[65] H. EF. he number of partitions of a set of n points in k dimensions induced by hyperplanes. *Proc. Edinburgh Math. Soc.*, 15:285–289, 1967.

[66] J. Fadili and G. Peyré. Total variation projection with first order schemes. *IEEE Transactions on Image Processing*, 20(3):657–669, 2011.

[67] 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.

[68] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties, 2001.

[69] K. Fang, S. Kotz, and K. Ng. *Symmetric multivariate and related distributions.* Monographs on statistics and applied probability. Chapman and Hall, 1990.

[70] 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.

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

[72] J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Transactions on Information Theory*, 50(6):1341–1344, 2004.

[73] H.-Y. Gao and A. Bruce. Waveshrink with firm shrinkage. *Statist. Sinica*, 7:855– 874, 1997.

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

[75] I. S. Gradshteyn and I. M. Ryshik. *Table of Integrals, Series, and Products.* Academic Press, London, 4th edition, 1965.

[76] M. Grasmair. Linear convergence rates for Tikhonov regularization with positively homogeneous functionals. *Inverse Problems*, 27(7):075014, 2011.

[77] M. Grasmair, O. Scherzer, and M. Haltmeier. Necessary and sufficient conditions for linear convergence of l1-regularization. *Communications on Pure and Applied Mathematics*, 64(2):161–182, 2011.

[78] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *Information Theory, IEEE Transactions on*, 57(3):1548–1566, 2011.

[79] D. Higham, X. Mao, and A. Stuart. Strong convergence of euler-type methods for nonlinear stochastic differential equations. *SIAM J. Numer. Anal.*, 40(3):1041–1063, 2003.

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

[81] J. Huang, J. L. Horowitz, and F. Wei. Variable selection in nonparametric additive models. *Ann. Statist.*, 38(4):2282–2313, 08 2010.

[82] N. Ikeda and S. Watanabe. *Stochastic differential equations and diffusion processes.* NH, 1989.

[83] 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.

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

[85] J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Inf. Comput.*, 132(1):1–63, Jan. 1997.

[86] P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations.* Stochastic Modelling and Applied Probability. Springer, 1995.

[87] V. Koltchinskii. Oracle inequalities in empirical risk minimization and sparse recovery problems. In *Lectures from the 38th Probability Summer School held in Saint-Flour*, volume 2033 of *Lecture Notes in Mathematics*. Springer, 2008.

[88] 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.

[89] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines, 2008.

[90] N. Kusolitsch. Why the theorem of scheffé should be rather called a theorem of riesz. *Periodica Mathematica Hungarica*, 61(1):225–229, 2010.

[91] D. Lamberton and G. Pages. Recursive computation of the invariant distribution of a diffusion. *Bernoulli*, 8(3):367–405, 2002.

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

[93] 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.

[94] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, Berlin, May 1991.

[95] G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006.

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

[97] D. Lorenz. Convergence rates and source conditions for Tikhonov regularization with sparsity constraints. *Journal of Inverse and Ill-Posed Problems*, 16(5):463–478, 2008.

[98] C. Louchet. *Variational and Bayesian models for image denoising : from total variation towards non-local means*. Theses, Université René Descartes - Paris V, Dec. 2008.

[99] 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.

[100] T. D. Luu, J. Fadili, and C. Chesneau. Sampling from non-smooth distribution through Langevin diffusion. Technical report, hal-01492056, May 2017.

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

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

[103] J. Maly. Lectures on change of variables in integral. Preprint 305, Department of Mathematics, University of Helsinki, 2001.

[104] P. Massart. *Concentration inequalities and model selection*. Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003. Springer Verlag, 2007.

[105] L. Meier, S. V. D. Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 2008.

[106] L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *Ann. Statist.*, 37(6B):3779–3821, 12 2009.

[107] L. Montuelle and E. Le Pennec. PAC-Bayesian aggregation of linear estimators. *ArXiv e-prints*, Oct. 2014.

[108] 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.

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

[110] 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.

[111] M. Pereyra. Proximal markov chain monte carlo algorithms. *Statistics and Computing*, 26(4):745–760, 2016.

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

[113] G. Peyré, J. Fadili, and C. Chesneau. Group sparsity with overlapping partition functions. In *EUSIPCO*, Barcelona, Spain, Aug. 2011.

[114] R. A. Poliquin and R. T. Rockafellar. *Prox-regular functions in variational analysis*, 1996.

[115] R. A. Poliquin, R. T. Rockafellar, and L. Thibault. Local differentiability of distance functions. *Transactions of the American mathematical Society*, 352:5231–5249, 2000.

[116] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over *ell*$_q$ -balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, Oct 2011.

[117] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

[118] 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.

[119] P. Rigollet. *Oracle inequalities, aggregation and adaptation.* Theses, Université Pierre et Marie Curie - Paris VI, Nov. 2006.

[120] P. Rigollet. Kullback?leibler aggregation and misspecified generalized linear models. *Ann. Statist.*, 40(2):639–665, 04 2012.

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

[122] P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771, 04 2011.

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

[124] G. O. Roberts and R. L. Tweedie. Exponential Convergence of Langevin Distributions and Their Discrete Approximations. *Bernoulli*, 2(4):341–363, 1996.

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

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

[127] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

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

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

[130] H. Scheffe. A useful convergence theorem for probability distributions. *Ann. Math. Statist.*, 18(3):434–438, 09 1947.

[131] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen. *Variational methods in imaging*, volume 167. Springer, 2009.

[132] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 03 1978.

[133] 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.

[134] W. Su and E. J. Candès. Slope is adaptive to unknown sparsity and asymptotically minimax. *Annals of Statistics*, 44(3):1038–1068, 2015.

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

[136] M. Talagrand. *The generic chaining. Upper and lower bounds of stochastic processes.* Springer-Verlag, Berlin, 2005.

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

[138] 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.

[139] J. Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*. Birkhäuser, 2014.

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

[141] A. B. Tsybakov. Optimal rates of aggregation. In *COLT*, volume 2777 of *Lecture Notes in Computer Science*, pages 303–313. Springer, 2003.

[142] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 1st edition, 2008.

[143] S. Vaiter. *Low Complexity Regularization of Inverse Problems*. Theses, Université Paris Dauphine - Paris IX, July 2014.

[144] 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*, 69(4):791–832, 2017.

[145] 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)*, 4(3):230–287, 2015.

[146] 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.

[147] S. Vaiter, G. Peyré, and M. J. Fadili. Model consistency of partly smooth regularizers. *IEEE Transactions on Information Theory*, 2017. in press.

[148] S. van de Geer. High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36:614–645, 2008.

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

[150] 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.

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

[152] R. Vershynin. Estimation in high dimensions: a geometric perspective. In *Sampling Theory, a Renaissance*. Birkhäuser, 2014.

[153] N. Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6(0):38–90, 2012.

[154] 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.

[155] Z. Wang, S. Paterlini, F. Gao, and Y. Yang. Adaptive minimax regression estimation over sparse lq-hulls. *J. Mach. Learn. Res.*, 15(1):1675–1711, Jan. 2014.

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

[157] J. Woodworth and R. Chartrand. Compressed sensing recovery via nonconvex shrinkage penalties. *CoRR*, abs/1504.02923, 2015.

[158] M. Xuerong. *Stochastic differential equations and applications*. Woodhead Publishing, 2007.

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

[160] F. Ye and C.-H. Zhang. Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *J. Mach. Learn. Res.*, 11:3519–3540, Dec. 2010.

[161] 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.