PAC-Bayesian risk bounds for group-analysis sparse regression by exponential weighting

Duy Tung Luu, Jalal M. Fadili, Christophe Chesneau

PAC-Bayesian risk bounds for group-analysis sparse regression by exponential weighting

Tung Duy Luu, Jalal Fadili

Normandie Univ, ENSICAEN, CNRS, GREYC, France

Christophe Chesneau

Normandie Univ, UNICAEN, CNRS, LMNO, France

Abstract

In this paper, we consider a high-dimensional non-parametric regression model with fixed design and i.i.d. random errors. We propose an estimator by exponential weighted aggregation (EWA) with a group-analysis sparsity promoting prior on the weights. We prove that our estimator satisfies a sharp group-analysis sparse oracle inequality with a small remainder term ensuring its good theoretical performances. We also propose a forward-backward proximal Langevin Monte-Carlo algorithm to sample from the target distribution (which is not smooth nor log-concave) and derive its convergence guarantees. In turn, this allows us to implement our estimator and validate it on some numerical experiments.

Key words: High-dimensional regression, exponential weighted aggregation, sparse learning, group-analysis sparsity, sparse oracle inequality, frame, forward-backward Langevin Monte-Carlo

AMS subject classifications. 62G07 62G20

1. Introduction

1.1. Problem statement

Let us briefly present our statistical context. Assume that the given data (x_i, Y_i), $i = 1, \ldots, n$, is generated according to the non-parametric regression model

$$Y_i = f(x_i) + \xi_i, \quad i \in \{1, \ldots, n\}, \tag{1.1}$$

where $x_1, \ldots, x_n$ are deterministic in an arbitrary set $\mathcal{X}$, $f : \mathcal{X} \to \mathbb{R}$ is the unknown regression function and $(\xi_1, \ldots, \xi_n)$ are random errors. (1.1) is equivalently written in vector form $\boldsymbol{Y} = \boldsymbol{f} + \boldsymbol{\xi}$. Assume that there exists a dictionary $\mathcal{H} = \left\{f_j : \mathcal{X} \to \mathbb{R},\ j \in \{1, \ldots, p\}\right\}$ such that f is well approximated by a linear combination of elements in $\mathcal{H}$, i.e., there exists $\widetilde{\boldsymbol{\theta}} = (\widetilde{\theta}_1, \ldots, \widetilde{\theta}_p)^\top \in \mathbb{R}^p$ such that $f_{\widetilde{\boldsymbol{\theta}}} = \sum_{j=1}^p \widetilde{\theta}_j f_j$ is a suitable approximation of f. The f_j's are known and may be either fixed atoms in a basis or pre-estimators. In the context of high dimension, the cardinality of $\mathcal{H}$ is much larger than the sample size (i.e., $p \gg n$). Thus, in such a setting, the classical least-squares to estimate $\widetilde{\boldsymbol{\theta}}$ is obviously not applicable.

The idea of aggregating elements in a dictionary has been introduced in machine learning to combine different techniques (see [39, 56]) with some procedures such as bagging [7], boosting [31, 54] and random forests [1, 4–6, 8, 32]. In the recent years, there has been a flurry of research on the use of the concept of sparsity in various areas

Email addresses: duy-tung.luu@ensicaen.fr (Tung Duy Luu), Jalal.Fadili@greyc.ensicaen.fr (Jalal Fadili), christophe.chesneau@unicaen.fr (Christophe Chesneau)

including statistics and machine learning in high dimension. The idea is that even if the cardinality of $\mathcal{H}$ is very large, the number of effective elements in the dictionary is much smaller than the sample size. Namely, the number of non-zero components of $\widetilde{\boldsymbol{\theta}}$ is assumed to be much smaller than n. This makes it possible to build an estimate $f_{\widetilde{\boldsymbol{\theta}}}$ with good provable performance guarantees under appropriate conditions on the dictionary and noise.

1.2. Overview of previous work

1.2.1. Oracle inequalities

This type of guarantees dates back, for instance, to the work [25–27] on orthogonal wavelet thresholding estimators. Oracle inequalities (according to the terminology introduced in, e.g., [26]), which are at the heart of our work, quantify the quality of an estimator compared to the best possible one among a collection of estimators. Formally, let $g : \mathcal{X} \to \mathbb{R}$, and denote

$$\|g\|_n = \sqrt{\tfrac{1}{n}\textstyle\sum_{j=1}^n g^2(x_j)}.$$

The performance of an estimator $\widehat{f}$ is measured by its averaged squared error, i.e.,

$$R(\widehat{f}) = \|\widehat{f} - f\|_n^2.$$

We aim to find an estimator $\widehat{f}$ that mimics as much as possible the performance of the best model of aggregation in a given class Θ (in the probabilistic sense). This idea is expressed in the following type of inequalities:

$$\mathbb{E}\left\{\|\widehat{f} - f\|_n^2\right\} \le C \inf_{\boldsymbol{\theta}\in\Theta}\left[\|f_{\boldsymbol{\theta}} - f\|_n^2 + \Delta_{n,p}(\boldsymbol{\theta})\right], \tag{1.2}$$

where $C \ge 1$ and the remainder term $\Delta_{n,p}(\boldsymbol{\theta})$ depends on the performance of the estimator, the complexity of $\boldsymbol{\theta}$, the dimension p and the sample size n. Such type of inequality is called balanced oracle inequality. Inequalities of type (1.2) are well adapted under the sparsity scenario. Namely, the complexity of $\boldsymbol{\theta}$ in the remaining 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).

An estimator with good oracle properties would correspond to C is close to 1 (ideally, $C = 1$, in which case the inequality is said "sharp"), $\Delta_{n,p}(\boldsymbol{\theta})$ is small even if $n \ll p$ and decreases rapidly to 0 as $n \to +\infty$. Besides, the choice of Θ is crucial: on the one hand, a non suitable choice can lead a large bias term in (1.2). 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.

In the literature, there are mainly two approaches to provide aggregated estimators in high dimension under the sparsity assumption: Penalization and Exponential Weighted Aggregation (EWA). Given $\boldsymbol{Y} = \boldsymbol{y}$, The penalization approach considers the minimization problem

$$\min_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{y} - f_{\boldsymbol{\theta}}\|_n^2 + \mathrm{pen}(\boldsymbol{\theta}),$$

where $\mathrm{pen} : \mathbb{R}^p \to \mathbb{R}^+$ is a sparsity promoting penalty function, see, e.g., [9]. Our work focuses on EWA approach that we briefly describe now.

1.2.2. Exponential Weighted Aggregation (EWA)

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

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

be a given collection ($\mathcal{F}_\Lambda$ is called dictionary of aggregation) where $\boldsymbol{\lambda} \to f_{\boldsymbol{\lambda}}(x)$ is measurable $\forall x \in \mathcal{X}$. The functions $f_{\boldsymbol{\lambda}}$ may be deterministic or random. The aggregators depend on the nature of $f_{\boldsymbol{\lambda}}$ if the latter is random. Otherwise, the aggregators are defined via the probability measure

$$\mu_n(d\boldsymbol{\lambda}) = \frac{\exp\left(-n\|\boldsymbol{Y} - f_{\boldsymbol{\lambda}}\|_n^2/\beta\right)\pi(d\boldsymbol{\lambda})}{\int_\Lambda \exp\left(-n\|\boldsymbol{Y} - f_{\boldsymbol{\omega}}\|_n^2/\beta\right)\pi(d\boldsymbol{\omega})}, \tag{1.3}$$

where $\beta > 0$ called temperature parameter and π called prior which is a probability measure on Λ. Then, we define the EWA aggregate by

$$\widehat{f_n}(x) = \int_\Lambda f_\lambda(x)\mu_n(d\lambda). \tag{1.4}$$

This idea was initially proposed in [35, 39, 56] with a uniform prior on a finite set Λ. Observe that (1.4) can be interpreted as the Bayesian posterior conditional mean in the regression model but only when the noise is Gaussian noise $\mathcal{N}(0, \sigma^2\mathbf{I}_n)$ and provided that $\beta = 2\sigma^2$ and the prior density is π.

In the literature, the works in [18, 19, 21, 23] consider deterministic dictionaries. These papers proposed several PAC-Bayesian[1] type of oracle inequalities under different assumptions. Especially, the assumptions in [23] depend only on the noise and turns out to be fulfilled for a large class of noise. This serves to construct, for a suitable prior and dictionary, a SOI with a remainder term of order $O(\|\boldsymbol{\theta}\|_0 \ln(p)/n)$, which scales linearly with the sparsity level and increases in p only logarithmically.

The random dictionary case are tackled in [45]. The initial idea is to obtain two independent samples from the initial sample by randomization or sample splitting (see [37, 49, 58]). 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 [38]. More recent work such as [20] generalizes the idea where the pre-estimators are affine and the priors are continuous.

A shortcoming of EWA is its suboptimality in deviation. In particular, the work in [16, Section 2] has shown that the EWA leads a suboptimal remainder term for oracle inequalities in probability.

1.2.3. Generalization of sparsity assumption

Analysis sparsity Let $q \geq p$ and $\boldsymbol{D}^\top \in \mathbb{R}^{q\times p}$ be a linear analysis operator. The analysis sparsity assumption means that $\left\|\boldsymbol{D}^\top\widetilde{\boldsymbol{\theta}}\right\|_0 \ll n$. A typical example is total variation [53] where the operator $\boldsymbol{D}^\top$ corresponds to "finite differences" (i.e., $(\boldsymbol{D}^\top\boldsymbol{\theta})_1 = \theta_1$, $(\boldsymbol{D}^\top\boldsymbol{\theta})_j = \theta_j - \theta_{j-1}$, $\forall j \geq 2$). Another example is the fused Lasso [55] where $\boldsymbol{D}^\top$ is a positive combination of the identity and finite differences.

Group sparsity Group sparsity corresponds to saying that the aggregator $\boldsymbol{\theta}$ is block sparse, see Section 4.1 for formal details. Group sparsity is at the heart of the group Lasso and related methods [34, 36, 42, 43, 48, 59]. In the EWA context, the group sparsity prior is considered in [50] as an application of the aggregation of orthogonal projectors.

1.3. Contributions

Our main contributions are summarized as follows:

- We propose an EWA estimator, with a deterministic dictionary, under a group-analysis sparsity prior (see Section 1.2.3). More precisely, we assume that $\boldsymbol{D}$ is a frame, and thus is not necessarily invertible unlike previous work. In a finite space, that equivalents to the fact that $\boldsymbol{D}\boldsymbol{D}^\top$ is invertible. In addition, our prior class (see (4.2)) is much more general than previously proposed ones [23] which recovered as very special cases. It also allows more flexibility to enhance the performance of EWA. This prior class is parameterized through a function $g : \mathbb{R}^+ \to \mathbb{R}^+$ that satisfies mild conditions (see Assumptions **(G.1)**-**(G.4)**).
- We establish a group-analysis SOI where the remainder term depends on the number of active groups in $\boldsymbol{D}^\top\boldsymbol{\theta}$ and on the function g (see Theorem 5.1)
- For an appropriate choice of g which is well-adapted to the group-analysis sparsity scenario, we exhibit a group-analysis SOI where, as expected, the remainder term scales as $O\left(\left\|\boldsymbol{D}^\top\boldsymbol{\theta}\right\|_{0,\mathcal{G}} \ln(L)/n\right)$, where $\left\|\boldsymbol{D}^\top\boldsymbol{\theta}\right\|_{0,\mathcal{G}}$ is the number of active groups in $\boldsymbol{D}^\top\boldsymbol{\theta}$, and L is the total number of groups (see Corollary 5.2). This rate coincides with the classical one $O(\|\boldsymbol{\theta}\|_0 \ln(p)/n)$ under the sparsity scenario, i.e., $\boldsymbol{D} = \mathbf{I}_p$ and $L = p$.

[1] PAC stands for Probably Approximately Correct. A PAC-bound is probably correct as it not a deterministic guarantee allowing a small probability that the estimator does not behave well. It is also approximately correct as it is tolerant to an inexact performance of the estimator.

- We also propose a forward-backward proximal Langevin Monte-Carlo (LMC) algorithm to sample from the target distribution (6.3) (which is not smooth nor log-concave), and establish several of its properties in a general setting. In turn, this allows us to efficiently implement our EWA estimator with the proposed prior. We validate this algorithm on some numerical examples.

1.4. Paper organization

Necessary notations and some preliminaries are first introduced in Section 2. Section 3 reminds the PAC-Bayesian type oracle inequalities proposed in [23] which are a classical starting point in literature for EWA in the deterministic case. In Section 4, we describe our EWA procedure after specifying the aggregation dictionary and our prior family. In Section 5, we establish our main results, namely group-analysis SOI. Section 6 is devoted to the forward-backward proximal LMC algorithm that implements EWA, and the numerical experiments on several numerical settings are described in Section 7. The proofs of all results are collected in Section 9.

2. Notations and Preliminaries

Before proceeding, let us introduce some notations and definitions.

Vectors and matrices For a d-dimensional Euclidean space $\mathbb{R}^d$, we endow it with its usual inner product $\langle\cdot,\cdot\rangle$ and associated norm $\|\cdot\|_2$. $\mathbf{I}_d$ is the identity matrix on $\mathbb{R}^d$. For $q>0$ and $\boldsymbol{x}\in\mathbb{R}^d$, we also define $\|\boldsymbol{x}\|_q=\left(\Sigma_{j=1}^d\left|x_j\right|^q\right)^{1/q}$ with the usual adaptation $\|\boldsymbol{x}\|_\infty=\max_{j\in\{1,\dots,d\}}\left|x_j\right|$. It is the ℓ_q norm for $q\geq 1$, and quasi-norm for $q\in]0,1[$. $\|\boldsymbol{x}\|_0$ is the ℓ_0 pseudo-norm which counts the number of non-zero elements in $\boldsymbol{x}$. Let $\boldsymbol{M}\in\mathbb{R}^{d\times d}$ symmetric positive definite, we denote $\langle\cdot,\cdot\rangle_{\boldsymbol{M}}=\langle\cdot,\boldsymbol{M}\cdot\rangle$ and $\|\cdot\|_{\boldsymbol{M}}$ its associated norm. Of course, $\left\|\cdot\right\|_{\boldsymbol{M}}$ and $\left\|\cdot\right\|_2$ are equivalent.
For a matrix $\boldsymbol{M}\in\mathbb{R}^{d\times r}$, we set $\boldsymbol{\sigma}(\boldsymbol{M})=(\sigma_1(\boldsymbol{M}),\dots,\sigma_r(\boldsymbol{M}))^\top\in\mathbb{R}^r$ be the vector of singular values of $\boldsymbol{M}$ in non-increasing order. Note that, when $\boldsymbol{M}$ is symmetric semi-definite positive, $\boldsymbol{\sigma}(\boldsymbol{M})$ is also the ordered vector or positive eigenvalues of $\boldsymbol{M}$. We denote $\left|\!\left|\!\left|\boldsymbol{M}\right|\!\right|\!\right|$ the spectral norm of $\boldsymbol{M}$.
For $\mathcal{I}\subset\{1,\dots,d\}$, $\boldsymbol{x}\in\mathbb{R}^d$ and $\boldsymbol{M}\in\mathbb{R}^{d\times d}$, $\boldsymbol{x}_{\mathcal{I}}$ is the subvector whose entries are those of $\boldsymbol{x}$ restricted to the indices in $\mathcal{I}$, and $\boldsymbol{M}_{\mathcal{I}}$ the submatrix whose rows and columns are those of $\boldsymbol{M}$ indexed by $\mathcal{I}$.
Let us denote $\mathrm{vec}:\mathbb{R}^{d\times d}\to\mathbb{R}^{d^2}$ the vectorization operator of a matrix. For any matrix $\boldsymbol{M}$, $\boldsymbol{M}^\top$ denotes its transpose. For any square matrix $\boldsymbol{M}$, $\det(\boldsymbol{M})$ is its determinant.

Definition 2.1 (Frame)**.** *A matrix $\boldsymbol{M}\in\mathbb{R}^{d\times r}$ is a frame if there exist two constants ν and κ with $\nu\geq\kappa>0$, called frame bounds, such that the generalized Parseval relation is satisfied, i.e., $\kappa\|\boldsymbol{x}\|_2^2\leq\left\|\boldsymbol{M}^\top\boldsymbol{x}\right\|_2^2\leq\nu\|\boldsymbol{x}\|_2^2$, $\forall\boldsymbol{x}\in\mathbb{R}^d$.*

By the Courant-Fischer theorem, Definition 2.1 is equivalent to the fact that κ (resp. ν) is a lower (resp. upper) bound of the eigenvalues of $\boldsymbol{M}\boldsymbol{M}^\top$. Moreover, since $\kappa>0$, we have that $\boldsymbol{M}\boldsymbol{M}^\top$ is bijective and $\boldsymbol{M}^\top$ is injective. The frame is said tight when $\kappa=\nu$. Typical examples of (tight) frames that have been used in statistics are translation invariant wavelets [15], ridgelets [11] and curvelets [10] (example of groups and what they represent for wavelets/ridgelets/curvelets in applications are discussed in [14]). 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

$$\frac{1}{\kappa}\geq\sigma_1\left(\widetilde{\boldsymbol{M}}^\top\widetilde{\boldsymbol{M}}\right)\geq\ldots\geq\sigma_d\left(\widetilde{\boldsymbol{M}}^\top\widetilde{\boldsymbol{M}}\right)\geq\frac{1}{\nu}. \tag{2.2}$$

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

Functions For a function $f : \mathbb{R}^d \to \mathbb{R}$, we define $\boldsymbol{f} = (f(x_1), \dots, f(x_d))^\top$ its d-sample vector form, $\left\|f\right\|_d = \left(\frac{1}{d}\Sigma_{j=1}^d f^2(x_j)\right)^{1/2}$ and $\left\|f\right\|_\infty = \sup_{\boldsymbol{x}\in\mathbb{R}^d} \left|f(\boldsymbol{x})\right|$.
For a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, its effective domain is $\mathrm{dom}(f) = \{\boldsymbol{x} \in \mathbb{R}^d \ : \ f(\boldsymbol{x}) < +\infty\}$ and f is proper if $f(\boldsymbol{x}) > -\infty$ for all $\boldsymbol{x}$ and $\mathrm{dom}(f) \neq \emptyset$ as is the case when it is finite-valued. A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$ is lower semi continuous (lsc) at $\boldsymbol{x}_0$ if $\liminf_{\boldsymbol{x}\to\boldsymbol{x}_0} f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0)$. For a differentiable function f, ∇f is its (Euclidean) gradient.

For a set Ω, I_Ω is its characteristic function, i.e., 1 if the argument is in Ω and 0 otherwise. Define $\mathrm{sgn} : \mathbb{R} \mapsto \{-1, 1\}$ be the sign function, i.e., $\mathrm{sgn}(x) = 1$ when $x \geq 0$, and $\mathrm{sgn}(x) = -1$ otherwise. Let $x \in \mathbb{R}$, define $\lfloor x \rfloor$ be the stands of integer part of x. Recall the Gamma function $\Gamma :]0, +\infty] \to]0, +\infty]$, $\Gamma : t \mapsto \int_0^\infty x^{t-1} \exp(-x) dx$.

Definition 2.2 (Proximal mapping and Moreau envelope)**.** *Let $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ be 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*

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

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

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

Useful integration formulas The following lemmas contain useful formula used throughout the paper.

Lemma 2.1 ([33, 3.251.11])**.** *Let $p, \gamma, \nu, \eta > 0$. If $\gamma/\nu < \eta + 1$ we have*

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

otherwise this integral is not definite.

Lemma 2.2 (Cartesian to spherical coordinates [30])**.** *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 then have*

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

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

The following lemma provides an efficient change of variables formula, which will be a key tool in the proof of our general group-analysis SOI (see Theorem 5.1).

Lemma 2.3. *Let $\Theta \subseteq \mathbb{R}^d$ be a measurable set. For any $\boldsymbol{M} \in \mathbb{R}^{r\times d}$ corresponding to the analysis operator of a frame of $\mathbb{R}^d$, let $u : \mathbb{R}^r \to \mathbb{R}$ such that the mapping $\boldsymbol{x} \mapsto u(\boldsymbol{M}\boldsymbol{x})$ is measurable on Θ. We have*

$$\int_\Theta u(\boldsymbol{M}\boldsymbol{x}) d\boldsymbol{x} = \frac{1}{\sqrt{\det(\boldsymbol{M}\boldsymbol{M}^\top)}} \int_{\boldsymbol{M}\Theta} u(\boldsymbol{v}) d\boldsymbol{v} \tag{2.3}$$

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

Though quite natural, proving Lemma 2.3 rigorously requires nontrivial arguments from geometric measure theory; see Section 9.1.

3. PAC-Bayesian type oracle inequalities

This section recalls a PAC-Bayesian type oracle inequality which holds for the EWA procedure of type (1.3)-(1.4) with any deterministic aggregation dictionary, any prior and a large class of noises. Such type of oracle inequalities was introduced in [23] for i.i.d. noise. In the present paper, we adapt it to the non i.i.d. case. Indeed, let us start with the two following assumptions.

(**P.1**) The noise vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$ 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 $\mathbb{E}\{\boldsymbol{\zeta}|\boldsymbol{\xi}'\} = 0$.

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

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

where $\mathrm{supp}(\boldsymbol{\xi}')$ is the support of the distribution of $\boldsymbol{\xi}'$.

Assumption (**P.2**) is based on [23, Assumption N], and can be shown to be fulfilled for a large class of noise.

Proposition 3.1. *Assume that $\boldsymbol{\xi}$ has zero mean. Assumption (**P.2**) is fulfilled with $t_0 = \infty$ when*

- *$\boldsymbol{\xi}$ is a Gaussian random vector with covariance matrix $\boldsymbol{\Sigma}$, with $t_0 = \infty$ and $v(\boldsymbol{a}) \equiv |||\boldsymbol{\Sigma}|||$;*
- *$\boldsymbol{\xi}$ is a Laplace random vector with covariance $\boldsymbol{\Sigma}$, with $t_0 < \sqrt{2/|||\boldsymbol{\Sigma}|||}$ and $v(\boldsymbol{a}) \equiv |||\boldsymbol{\Sigma}|||/(1 - t_0^2 |||\boldsymbol{\Sigma}|||^2/2)$;*
- *$\boldsymbol{\xi}$ is a bounded symmetric random vector, i.e., $\Pr\{|\xi_i| \leq B_i\}$ for some $\boldsymbol{B} \in \mathbb{R}^n$, with $t_0 = \infty$ and $v(\boldsymbol{a}) = \|\boldsymbol{a}\|_2 \leq \|\boldsymbol{B}\|_2$.*

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

$$\sup_{(\lambda, \lambda') \in \Lambda^2} \left\| \boldsymbol{f}_\lambda - \boldsymbol{f}_{\lambda'} \right\|_2 \leq H. \tag{3.1}$$

Note that (3.1) 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. We are now ready to state the PAC-Bayesian type oracle inequalities.

Theorem 3.1. *Let Assumptions (**P.1**) and (**P.2**) be satisfied with some function v and let (3.1) holds. Then for any prior π, any probability measure p on Λ 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*

$$\mathbb{E}\left\{\left\|\widehat{f}_n - f\right\|_n^2\right\} \leq \int_\Lambda \|f - f_\lambda\|_n^2 p(d\boldsymbol{\lambda}) + \frac{\beta \, \mathrm{KL}(p, \pi)}{n}, \tag{3.2}$$

where $\widehat{f}_n$ is the aggregate defined in (1.4) and $\mathrm{KL}(p, \pi) = \int_\Lambda \ln\left(p(d\boldsymbol{\lambda})/\pi(d\boldsymbol{\lambda})\right) p(d\boldsymbol{\lambda})$ is the Kullback-Leibler divergence.

The proof of Theorem 3.1 is a mild adaptation of the original one in [23, Section 2], where we used directly Assumption (**P.2**)-3 in the vector $\boldsymbol{\zeta}$ instead of splitting it into $\zeta_{i, i \in \{1, \ldots, n\}}$ (that are no longer i.i.d.).

Related work The work of [19] proposed three types of oracle inequalities which are similar to (3.2) under different assumptions. The first type (see [19, Theorem 1]) holds under a restrictive condition on the noise. The second (see [19, Theorem 2]) involves conditions depending on the noise and also on the dictionary. The last (see [19, 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 3.1 (with Assumption (**P.2**)) is a good trade-off between these types of oracle inequalities.

Moreover, there exist some related forms of (3.2) in different frameworks. For example, when $\xi_i \sim \mathcal{N}(0, \sigma_i^2)$, $i = 1, \dots, n$, the following aggregate was proposed in [20]:

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

where $\widehat{f}_\lambda$, $\boldsymbol{\lambda} \in \Lambda$ are affine estimators satisfying some conditions imposed in [20, Theorem 1] which yield the definition of $\widehat{r}_\lambda$, $\boldsymbol{\lambda} \in \Lambda$. This aggregate satisfies oracle inequalities defined therein which are the counterparts of (3.2) 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 works in [22] 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.

4. EWA estimator

4.1. Group-analysis sparsity

We now describe formally what is intended by group-analysis 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 $\{\mathcal{G}_\ell\}_{1 \leq \ell \leq L}$ such that

$$\mathcal{G} = \bigcup_{\ell=1}^{L} \mathcal{G}_\ell = \{1, \dots, q\} \quad \text{and} \quad \mathcal{G}_\ell \cap \mathcal{G}_k = \emptyset, \ \ \forall l \neq k.$$

For the sake of simplicity, and without loss of generality, the groups are assumed to have the same size $\operatorname{Card} \mathcal{G}_\ell = G \geq 1$ and the total number of blocks L is supposed to be an integer. A vector $\boldsymbol{v} \in \mathbb{R}^q$ can be divided into L vectors $\boldsymbol{v}_{\mathcal{G}_\ell} \in \mathbb{R}^G$ which are the restrictions of $\boldsymbol{v}$ to the coordinates indexed by $\mathcal{G}_\ell$. We define

$$\|\boldsymbol{v}\|_{s,\mathcal{G}} = \left(\sum_{\ell=1}^{L} \|\boldsymbol{v}_{\mathcal{G}_\ell}\|_2^s\right)^{1/s}, \quad s > 0,$$

which is a norm for $s \geq 1$, with $\|\boldsymbol{v}\|_{\infty,\mathcal{G}} = \max_{\ell \in \{1,\dots,L\}} \|\boldsymbol{v}_{\mathcal{G}_\ell}\|_2$. It is a quasi-norm for $s \in]0, 1[$. $\|\boldsymbol{v}\|_{0,\mathcal{G}}$ counts the number of active (i.e., non-zero) groups in $\boldsymbol{v}$. With these notations, the group-analysis sparsity assumption is formalized as follows.

(**H.1**) Let $\boldsymbol{D} \in \mathbb{R}^{p \times q}$, there exists $\widetilde{\boldsymbol{\theta}} \in \mathbb{R}^p$ such that $f_{\widetilde{\boldsymbol{\theta}}} = f$ and $\|\boldsymbol{D}^\top \widetilde{\boldsymbol{\theta}}\|_{0,\mathcal{G}} \ll n$.

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

(**H.2**) $\boldsymbol{D}$ is a frame with frame bounds $\nu \geq \kappa > 0$.

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

Example 4.1 (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 analysis operator is the isotropic total variation called D_{TV} (see [53]). Namely, let $D_c : \mathbb{R}^{\sqrt{p} \times \sqrt{p}} \to \mathbb{R}^{\sqrt{p} \times \sqrt{p}}$ and $D_r : \mathbb{R}^{\sqrt{p} \times \sqrt{p}} \to \mathbb{R}^{\sqrt{p} \times \sqrt{p}}$ the finite difference operators along, respectively, the columns and rows of an image, with Neumann boundary conditions. We define D_{TV} as*

$$D_{\mathrm{TV}} : \boldsymbol{\theta} \in \mathbb{R}^{\sqrt{p} \times \sqrt{p}} \mapsto \operatorname{vec}\left((\operatorname{vec}(D_r(\boldsymbol{\theta})), \operatorname{vec}(D_c(\boldsymbol{\theta})))^\top\right)^\top \in \mathbb{R}^{2p}.$$

*By vectorizing $\widetilde{\boldsymbol{\theta}}$ in Assumption (**H.1**), $\boldsymbol{D}^\top$ can be considered as the matrix version of the linear operator D_{TV}, called $\boldsymbol{D}_{\mathrm{TV}}$. Here, $\boldsymbol{D}_{\mathrm{TV}} = [\boldsymbol{D}_r^\top \ \boldsymbol{D}_c^\top]^\top$ where $\boldsymbol{D}_r \in \mathbb{R}^{p\times p}$ (resp. $\boldsymbol{D}_c \in \mathbb{R}^{p\times p}$) is the matrix counterpart of D_r (resp. D_c). With Neumann boundary conditions, $\boldsymbol{D}_r$ and $\boldsymbol{D}_c$ are bijective implying injectivity of $\boldsymbol{D}_{\mathrm{TV}}$. Thus, $\boldsymbol{D}$ is a frame in view of Courant-Fisher theorem.*
The isotropic total variation prior on $\boldsymbol{\theta}$ promotes sparsity of $\mathrm{vec}\left(\left\|[D_{\mathrm{TV}}(\boldsymbol{\theta}_0)]_{i,j}\right\|_2\right)_{1\leq i,j\leq\sqrt{p}}$. By defining the set of groups by $\mathcal{G} = \bigcup_{(i,j)\in\{1,\dots,s_{p_0}\}^2}\{(i,j,1),(i,j,2)\}$, one immediately realizes that measuring sparsity of the above vectorized form is equivalent to group sparsity of $D_{\mathrm{TV}}(\boldsymbol{\theta}_0)$ with groups of size 2 along the third dimension.

Example 4.2 (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 [13, 47]). In this framework, $\boldsymbol{D}\boldsymbol{D}^\top = \mathrm{diag}(\boldsymbol{B}_1^\top\boldsymbol{B}_1,\dots,\boldsymbol{B}_L^\top\boldsymbol{B}_L)$ where $\boldsymbol{B}_\ell \in \mathbb{R}^{q_\ell\times p}$, for $\ell\in\{1,\dots,L\}$, is a countable collection of localization operators, and then $q = \sum_{\ell=1}^L q_\ell$. Since the localization operators are injective, $\boldsymbol{D}\boldsymbol{D}^\top$ is bijective. 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 group-analysis sparsity. We specify them below.

4.2. Choice of dictionary

Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ where $\boldsymbol{X}_{i,j} = f_j(x_i)$, where we recall that $f_j \in \mathcal{H}$ is a known function (atom) in the deterministic dictionary $\mathcal{H}$.

(**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:

$$\mathcal{F}_\Theta = \left\{f_{\boldsymbol{\theta}} = \mathscr{L}\left(\textstyle\sum_{j=1}^p \theta_j f_j\right) : \boldsymbol{\theta}\in\Theta = \{\boldsymbol{\theta}\in\mathbb{R}^p \ : \ \left\|\boldsymbol{D}^\top\boldsymbol{\theta}\right\|_{a,\mathcal{G}}^a \leq R\}\right\}, \tag{4.1}$$

where $a\in]0,1]$, $R\in]0,+\infty]$ and $\mathscr{L}:\mathbb{R}\to\mathbb{R}$ is twice continuously differentiable and known function that depends on the regression problem: e.g., $\mathscr{L}(x) = e^x$ for the exponential regression, $\mathscr{L}(x) = e^x/(e^x+1)$ for the logistic regression and $\mathscr{L}(x) = x$ for the linear regression. This dictionary of aggregation is similar to the one proposed in [21–23]. However, the set of indices is modified to adapt to the group-analysis sparsity and the exponent a is varied in $]0,1]$ instead of a fixed $a = 1$. The bound H in (3.1) for $\mathcal{F}_\Theta$ in (4.1) is established in the following result.

Proposition 4.1. *Let $\mathcal{F}_\Theta = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta}\in\Theta\}$ defined in (4.1) with some $R > 0$, $a\in]0,1]$ and $\mathscr{L}:\mathbb{R}\to\mathbb{R}$ twice continuously differentiable. Let Assumption (**H.2**) holds for some $\kappa > 0$. We get that*

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

where $\mathcal{B} = \{\boldsymbol{x}\in\mathbb{R}^n \ : \ \|\boldsymbol{x}\|_2 \leq \left\|\boldsymbol{X}\right\|R^{1/a}\kappa^{-1/2}\}$ and $\boldsymbol{\mathscr{L}} : \boldsymbol{x}\in\mathbb{R}^n \to (\mathscr{L}(x_1),\dots,\mathscr{L}(x_n))$.

From Proposition 4.1, one can choose $H = 2\max_{\boldsymbol{x}\in\mathcal{B}}\left\|\boldsymbol{\mathscr{L}}(\boldsymbol{x})\right\|_2$.

Remark 4.1. *By choosing $H = 2\max_{\boldsymbol{x}\in\mathcal{B}}\left\|\boldsymbol{\mathscr{L}}(\boldsymbol{x})\right\|_2$, H depends on $\boldsymbol{X}$ and then on n under Assumption (**H.3**). So $\beta \geq \max(4\|v\|_\infty, 2H/t_0)$ also depends on n. In this case, $\boldsymbol{\xi}$ must satisfy Assumption (**P.2**) with $t_0 = \infty$. In view of Proposition 3.1, we can consider $\boldsymbol{\xi}$ as a Gaussian or a bounded symmetric noise.*

4.3. Choice of prior

4.3.1. Main assumptions

Recall that the goal is to find a prior leading an oracle inequality with a small remainder term while promoting group-analysis sparsity. In order to promote sparsity, it is well-known that the prior is usually expected to be symmetric and sharply peaked around its mode (the origin) while having tails heavier than merely exponential [44]. This is the rationale behind our general prior which takes the form

$$\pi(d\boldsymbol{\theta}) = \frac{1}{C_{\alpha,g,R}}\prod_{\ell=1}^L \exp\left(-\alpha^a\left\|[\boldsymbol{D}^\top\boldsymbol{\theta}]_{\mathcal{G}_\ell}\right\|_2^a\right)g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{\theta}]_{\mathcal{G}_\ell}\right\|_2\right)I_\Theta(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{4.2}$$

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

(**G.1**) Boundedness: $g : \mathbb{R}^+ \to \mathbb{R}^+$ is a bounded function such that $g \not\equiv 0$, and $\boldsymbol{\theta} \mapsto g\Big(\big\|[\boldsymbol{D}^\top \boldsymbol{\theta}]_{\mathcal{G}_\ell}\big\|_2\Big)$ is measurable on $\mathbb{R}^p$, $\forall \ell \in \{1, \dots, L\}$.

(**G.2**) Integrability: $\int_{\mathbb{R}^p} \prod_{\ell=1}^L g\Big(\big\|[\boldsymbol{D}^\top \boldsymbol{u}]_{\mathcal{G}_\ell}\big\|_2\Big) d\boldsymbol{u} < \infty$.

(**G.3**) Moment condition: $\int_{\mathbb{R}^p} \big\|[\boldsymbol{D}^\top \boldsymbol{u}]_{\mathcal{G}_\ell}\big\|_2^2 \prod_{k=1}^L g\Big(\big\|[\boldsymbol{D}^\top \boldsymbol{u}]_{\mathcal{G}_k}\big\|_2\Big) d\boldsymbol{u} < \infty$, $\forall \ell \in \{1, \dots, L\}$.

(**G.4**) Growth condition: there exist $\lambda \geq 0$ and $h : \mathbb{R}^+ \to \mathbb{R}^+$ such that $g(\|\boldsymbol{t} - \boldsymbol{t}^*\|_2)/g(\|\boldsymbol{t}\|_2) \leq h(\|\boldsymbol{t}^*\|_2)^\lambda$, $\forall (\boldsymbol{t}, \boldsymbol{t}^*) \in \mathbb{R}^G \times \mathbb{R}^G$.

The exponential part can be viewed as a generalized Gaussian on the group analysis coefficients with shape parameter a and scale parameter α^{-1}. The choice of $a \in]0, 1]$ favors sparsity. The choice of α will be made clear when discussing the remainder terms in our group-analysis SOI (see for instance Remark 5.1). The role of the function g is (at least) twofold. The first is precisely to capture heavier tails than exponential. Second, it allows more flexibility to adjust to the group sparsity scenario and optimize the performance of EWA (see remarks hereafter for further discussion and examples).

4.3.2. Discussion of the assumptions

Assumption (**G.4**) controls the growth of the function g. The growth function h will impact the remainder term in the main group-analysis SOI stated in Theorem 5.1, and more precisely the term $\Omega^{\boldsymbol{D}}_{\mu_n,n,L,\lambda}(\boldsymbol{\theta})$ therein.

Assumptions (**G.1**)-(**G.3**) play a prominent role in controlling the key constant $K^{\boldsymbol{D}}_{a,g} > 0$ that is involved in the construction of our general group-analysis SOI in Theorem 5.1. The following remark formalizes the existence of this constant and relates it to the integrability and moment conditions on g.

Remark 4.2. *Let $G \geq 1$ and $\boldsymbol{D} \in \mathbb{R}^{p \times q}$ satisfying Assumption (**H.2**). For any function g satisfying Assumptions (**G.1**)-(**G.3**), and any $a \in]0, 1]$, there exists $K^{\boldsymbol{D}}_{a,g} \in]0, \infty[$ such that, $\forall \ell \in \{1, \dots, L\}$,*

$$\frac{\int_{\mathbb{R}^p} \big\|[\boldsymbol{D}^\top \boldsymbol{u}]_{\mathcal{G}_\ell}\big\|_2^{2a} \prod_{k=1}^L g\Big(\big\|[\boldsymbol{D}^\top \boldsymbol{u}]_{\mathcal{G}_k}\big\|_2\Big) d\boldsymbol{u}}{\int_{\mathbb{R}^p} \prod_{k=1}^L g\Big(\big\|[\boldsymbol{D}^\top \boldsymbol{v}]_{\mathcal{G}_k}\big\|_2\Big) d\boldsymbol{v}} \leq K^{\boldsymbol{D}}_{a,g}. \tag{4.3}$$

Proof. From Assumption (**G.3**) and the fact that $g \not\equiv 0$, one can show that (4.3) holds for $a = 1$. Moreover, since $g \not\equiv 0$ and g satisfies Assumption (**G.2**), it holds that

$$\boldsymbol{u} \mapsto \frac{\prod_{\ell=1}^L g\Big(\big\|[\boldsymbol{D}^\top \boldsymbol{u}]_{\mathcal{G}_\ell}\big\|_2\Big) d\boldsymbol{u}}{\int_{\mathbb{R}^p} \prod_{k=1}^L g\Big(\big\|[\boldsymbol{D}^\top \boldsymbol{v}]_{\mathcal{G}_k}\big\|_2\Big) d\boldsymbol{v}}$$

is a probability measure. Therefore, (4.3) holds for any a in $]0, 1]$ by Hölder's inequality. □

It is then legitimate to wonder whether there exists a simpler condition on g which implies Assumptions (**G.2**) and (**G.3**). The answer is affirmative as we now state.

Lemma 4.1. *Let $G \geq 1$ and $\boldsymbol{D} \in \mathbb{R}^{p \times q}$ satisfying Assumption (**H.2**). Suppose that g satisfies Assumption (**G.1**) and*

$$\int_0^\infty z^{G+1} g(z) dz < \infty. \tag{4.4}$$

*Then Assumptions (**G.2**) and (**G.3**) are in force.*

When the operator $\boldsymbol{D}$ is invertible, $K^{\boldsymbol{D}}_{a,g}$ takes an even simpler and explicit form, and moreover, (4.4) is necessary for g to obey Assumption (**G.3**). This is the subject of Remark 4.3 and Remark 4.4.

Remark 4.3. *Let $G \geq 1$ and $\boldsymbol{D} \in \mathbb{R}^{p \times p}$ be invertible. For any function g satisfying Assumptions (**G.1**)-(**G.3**), and for any $a \in]0, 1]$, one can choose $K^{\boldsymbol{D}}_{a,g}$ in (4.3) as*

$$K^{\boldsymbol{D}}_{a,g} = \frac{\int_0^\infty x^{G-1+2a} g(x) dx}{\int_0^\infty z^{G-1} g(z) dz}. \tag{4.5}$$

Proof. The proof follows by combining Lemmas 2.3 and 2.2, i.e.,

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

□

Remark 4.4. *When $\boldsymbol{D}$ is invertible, if g does not satisfy (4.4) then g cannot fulfill Assumption (G.3). Consequently, Assumption (G.3) and condition (4.4) are equivalent in the invertible case.*

Proof. By Lemmas 2.3 and 2.2, we get

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

□

4.3.3. Examples of g

Let us now discuss some choices of g.

Example 4.3. *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 [23] for the individual sparsity scenario, i.e., with $\boldsymbol{D} = \mathbf{I}_p$, $G = 1$ and $a = 1$.

Example 4.4. *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 $c > (2 + G)/b$. 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.

5. Group-analysis sparse oracle inequality

Once a suitable dictionary and prior are chosen according to the above, the EWA is performed via (1.3)-(1.4). Our goal now is to provide a theoretical guarantee for the aggregates by constructing a group-analysis SOI. First of all, based on PAC-Bayesian type oracle inequalities in Section 3, we establish our first main result: a group-analysis SOI for the dictionary (4.1) and the prior (4.2) with a function g obeying Assumptions (G.1)-(G.4).

Theorem 5.1 (General group-analysis sparse oracle inequality)**.** *Let $G \geq 1$, $\boldsymbol{X}$ satisfying Assumption (H.3), and $\boldsymbol{D}$ satisfying Assumption (H.2) with $\kappa > 0$. Let Assumptions (P.1) and (P.2) be satisfied with some function v, (3.1) holds and $\beta \geq \max(4\|v\|_\infty, 2H/t_0)$. For some $a \in]0,1]$, take the dictionary (4.1) and the prior (4.2) with g satisfying Assumptions (G.1)-(G.4). Let $K^{\boldsymbol{D}}_{a,g}$, as defined in (4.3), and assume that $R > 3\sqrt{K^{\boldsymbol{D}}_{a,g}}L$. Then the following group-analysis SOI holds,*

$$\mathbb{E}\left\{\left\|\widehat{f_n} - f\right\|_n^2\right\} \leq \inf_{\Theta^{\boldsymbol{D}}_{\mu_n,L,R}} \left(\left\|f_{\boldsymbol{\theta}} - f\right\|_n^2 + \Phi^{\boldsymbol{D}}_{\mu_n,n,L}(\boldsymbol{\theta}) + \Omega^{\boldsymbol{D}}_{\mu_n,n,L,\lambda}(\boldsymbol{\theta})\right) + \Psi^{\boldsymbol{D}}_{\mu_n,L,p}, \tag{5.1}$$

with

$$\begin{cases} \Theta^{\boldsymbol{D}}_{\mu_n,L,R} = \{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \left\|\boldsymbol{D}^\top\boldsymbol{\theta}\right\|^a_{a,\mathcal{G}} \leq R - 3\sqrt{K^{\boldsymbol{D}}_{a,g}}L\} \\ \Phi^{\boldsymbol{D}}_{\mu_n,n,L}(\boldsymbol{\theta}) = \dfrac{\beta}{n}\left(1 + 3\sqrt{K^{\boldsymbol{D}}_{a,g}}L\alpha^a + \alpha^a\left\|\boldsymbol{D}^\top\boldsymbol{\theta}\right\|^a_{a,\mathcal{G}}\right) \\ \Omega^{\boldsymbol{D}}_{\mu_n,n,L,\lambda}(\boldsymbol{\theta}) = \dfrac{\lambda\beta}{n}\sum_{\ell=1}^L \ln h\left(\left\|\left[\boldsymbol{D}^\top\boldsymbol{\theta}\right]_{\mathcal{G}_\ell}\right\|_2\right) \\ \Psi^{\boldsymbol{D}}_{\mu_n,L,p} = 2\kappa^{-1}K^{\boldsymbol{D}}_{1,g}\exp\left(3\sqrt{K^{\boldsymbol{D}}_{a,g}}L\alpha^a\right)pC_{f,\mathscr{L}}, \end{cases}$$

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

Before proceeding, we pause to make a few important remarks.

Remark 5.1. *The group-analysis SOI (5.1) is sharp. It depends on several parameters as discussed below.*

(i) The parameter R appears in the dictionary. Namely, the EWA estimator $\widehat{f_n}$ mimics the best aggregate $f_{\boldsymbol{\theta}}$ for all possible weights belonging to $\left\{\boldsymbol{\theta} \in \mathbb{R}^p : \left\|\boldsymbol{D}^\top\boldsymbol{\theta}\right\|^a_{a,\mathcal{G}} \leq R - 3\sqrt{K^{\boldsymbol{D}}_{a,g}}L\right\}$. Then R must be sufficiently large to cover the "good" model $f_{\bar{\boldsymbol{\theta}}}$ 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 choice we can make to reduce the rate of $\Phi^{\boldsymbol{D}}_{\mu_n,n,L}(\boldsymbol{\theta})$ as $\left\|\boldsymbol{D}^\top\boldsymbol{\theta}\right\|^a_{a,\mathcal{G}} \leq R$.

(ii) The parameter α is used to cancel the effect of $L\sqrt{K^{\boldsymbol{D}}_{a,g}}$ in the remainder terms. By choosing $\alpha \leq (3K\sqrt{K^{\boldsymbol{D}}_{a,g}})^{-1/a}$, $\Phi^{\boldsymbol{D}}_{\mu_n,n,L}(\boldsymbol{\theta}) \leq \beta n^{-1}(1 + 3\sqrt{K^{\boldsymbol{D}}_{a,g}}\alpha^a + \alpha^a R) \sim n^{-1}$ and $\Psi^{\boldsymbol{D}}_{\mu_n,L,p} \leq \frac{2eC_{f,\mathscr{L}}}{\kappa}K^{\boldsymbol{D}}_{1,g}p$.

(iii) The parameter $K^{\boldsymbol{D}}_{a,g}$ and the function h depend on the choice of g. They respectively control the rate of $\Psi^{\boldsymbol{D}}_{\mu_n,L,p}$ and $\Omega^{\boldsymbol{D}}_{\mu_n,n,L,\lambda}(\boldsymbol{\theta})$.

In what follows, let us state the consequences of Theorem 5.1 with the choices of g in Example 4.3 and 4.4. Especially, we will discuss the rate of $\Omega^{\boldsymbol{D}}_{\mu_n,n,L,\lambda}(\boldsymbol{\theta})$ and $\Psi^{\boldsymbol{D}}_{\mu_n,L,p}$.

We first consider the prior (4.2) in Example 4.3, under the individual sparsity scenario ($\boldsymbol{D} = \mathbf{I}_p$, $G = 1$) and the choice $a = 1$ (i.e., $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \left\|\boldsymbol{\theta}\right\|_1 \leq R\}$). This is the setting considered in [23]. We obtain the following SOI as a corollary of our main result.

Corollary 5.1. *Let $\boldsymbol{X}$ satisfying Assumption (H.3), $\boldsymbol{D} = \mathbf{I}_p$ and fix $G = 1$. Suppose that Assumptions (P.1) and (P.2) hold with some function v, (3.1) holds and $\beta \geq \max(4\|v\|_\infty, 2H/t_0)$. Fix $a = 1$, take the dictionary (4.1) and the prior (4.2) with g defined in Example 4.3 and $\alpha \leq 1/(3p\tau)$. Assume that $R > 3p\tau$. Choosing $\tau^2 \sim 1/(pn)$ and $R \sim p\tau$, SOI (5.1) holds with $\Theta^{\boldsymbol{D}}_{\mu_n,L,R} = \{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \|\boldsymbol{\theta}\|_1 \leq R - 3p\tau\}$,*

$$\Omega^{\boldsymbol{D}}_{\mu_n,n,L,\lambda}(\boldsymbol{\theta}) \sim \frac{\|\boldsymbol{\theta}\|_0 \ln(p)}{n}, \quad \text{and} \quad \Psi^{\boldsymbol{D}}_{\mu_n,L,p} \sim \frac{1}{n}.$$

The order of $\Omega^{\boldsymbol{D}}_{\mu_n,n,L,\lambda}(\boldsymbol{\theta})$ is the classical rate under the sparsity scenario. This scaling is similar as in [23] with the same prior. However, the following remark shows that this prior is not adapted in the group-analysis case for any group size strictly larger than 1.

Remark 5.2. *Suppose that $G \geq 2$, and let $\gamma = G + 2$, $\nu = 2$ and $\eta = 1$. We have $\gamma/\nu \geq \eta + 1$, and thus Lemma 2.1 yields $\int_0^\infty x^{G+1}(\tau^2 + x^2)^{-2}dx$ is not definite. Consequently, condition (4.4) is not fulfilled with g defined in Example 4.3 when $G \geq 2$.*

According to Remark 4.4, Remark 5.2 implies that Assumption **(G.3)** is not fulfilled for g in Example 4.3 when the group size $G \geq 2$ and $\boldsymbol{D}$ invertible. Thus one cannot construct a group-analysis SOI from Theorem 5.1 to guarantee the quality of the corresponding estimator. Overcoming this limitation was yet another motivation behind the choice of g in Example 4.4, which turns out to work well under the group-analysis sparsity scenario. In a nutshell, an aggregate with g in Example 4.4 exhibits the group-analysis SOI defined in the following corollary with any $G \geq 1$, any $\boldsymbol{D} \in \mathbb{R}^{p\times q}$ satisfying Assumption **(H.2)** and any $a \in]0, 1]$.

Corollary 5.2. *Let $\boldsymbol{X}$ satisfying Assumption **(H.3)**, $G \geq 1$ and $\boldsymbol{D}$ satisfying Assumption **(H.2)** with $\kappa > 0$. Let Assumptions **(P.1)** and **(P.2)** be satisfied with some function v, (3.1) holds and $\beta \geq \max(4\|v\|_\infty, 2H/t_0)$. Take the dictionary (4.1) and the prior (4.2) with $a \in]0, 1]$, $\alpha \geq 0$ and g defined in Example 4.4. We get that g satisfies Assumptions **(G.1)**-**(G.4)**. Then, let $K^{\boldsymbol{D}}_{a,g}$ as defined in (4.3), and assume that $R > 3\sqrt{K^{\boldsymbol{D}}_{a,g}}L$. Then the group-analysis SOI (5.1) holds, with $\lambda = c$ and $h(x) = 1 + (x/\tau)^b$.*

To get an explicit control of the remainder term, it is instructive to have a closed-form of $K^{\boldsymbol{D}}_{a,g}$. This can be done for instance when $\boldsymbol{D}$ is invertible, see (4.5). The obtained group-analysis SOI is stated as follows.

Corollary 5.3. *Consider the same framework as Corollary 5.2 with $\boldsymbol{D}$ invertible. For $a \in]0, 1]$, let $\widetilde{K}^{\boldsymbol{D}}_{a,g} = \frac{\Gamma((2a+G)/b)\Gamma(c-(2a+G)/b)}{\Gamma(G/b)\Gamma(c-G/b)}$, and set $\alpha \leq 1/\left(3\tau^a\sqrt{\widetilde{K}^{\boldsymbol{D}}_{a,g}}L\right)^{1/a}$. Choosing $\tau^2 \sim 1/(pn)$ and $R \sim L\tau^a$, the group-analysis SOI (5.1) holds with $\Theta^{\boldsymbol{D}}_{\mu_n,L,R} = \{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \left\|\boldsymbol{D}^\top\boldsymbol{\theta}\right\|^a_{a,\mathcal{G}} \leq R - 3\tau^a\sqrt{\widetilde{K}^{\boldsymbol{D}}_{a,g}}L\}$,*

$$\Omega^{\boldsymbol{D}}_{\mu_n,n,L,\lambda}(\boldsymbol{\theta}) \sim \frac{\left\|\boldsymbol{D}^\top\boldsymbol{\theta}\right\|_{0,\mathcal{G}}\ln(L)}{n}, \quad \textit{and} \quad \Psi^{\boldsymbol{D}}_{\mu_n,L,p} \sim \frac{1}{n}.$$

By Assumption **(H.1)**, $\left\|\boldsymbol{D}^\top\boldsymbol{\theta}\right\|_{0,\mathcal{G}}$ is small when $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$ (with R must be sufficiently large to cover $\widetilde{\boldsymbol{\theta}}$). Thus, $\left\|\boldsymbol{D}^\top\widetilde{\boldsymbol{\theta}}\right\|_{0,\mathcal{G}}\ln(L)$ is small compared to n. Under the sparsity scenario, the order of $\Omega^{\boldsymbol{D}}_{\mu_n,n,L,\lambda}(\boldsymbol{\theta})$ becomes $O(\|\boldsymbol{\theta}\|_0\ln(p)/n)$ which is the same rate as the aggregate with g in Example 4.3.

6. Forward-Backward proximal LMC algorithm

The goal of this section is to implement our EWA estimator with the probability measure (1.3) via a novel forward-backward proximal Monte-Carlo algorithm based on the Langevin diffusion (coined FB-LMC).

Let us consider a linear regression problem where $\mathscr{L}(x) = x$, and thus[2]

$$\widehat{f_n} = \boldsymbol{X}\widehat{\boldsymbol{\theta}}_n,$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the design matrix (see Section 4.2). Recall the EWA estimator

$$\widehat{\boldsymbol{\theta}}_n = \int_{\mathbb{R}^p} \boldsymbol{\theta}\mu_n(\boldsymbol{\theta})d\boldsymbol{\theta},$$

with the measure

$$\mu_n(\boldsymbol{\theta}) \propto \exp\left(\frac{-\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2}{\beta}\right)\pi(\boldsymbol{\theta}). \tag{6.1}$$

[2]Generalization to non-linear link functions is possible from a practical point of view. However, in this case, the convergence guarantees of the discrete Langevin-based MCMC sampling scheme would be much more intricate even for a simple exponential link function. One of the main difficulties lies in that the data loss $\boldsymbol{\theta} \mapsto \left\|\boldsymbol{y} - \mathscr{L}(\boldsymbol{X}\boldsymbol{\theta})\right\|_2^2$ is not necessarily convex which prevents us from showing that the Langevin diffusion is geometrically ergodic. This is an open question that we leave to a future work.

Computing $\widehat{\boldsymbol{\theta}}_n$ corresponds to an integration problem which becomes very involved to solve analytically or even numerically in high-dimension. A classical alternative is to approximate it via a Markov chain Monte-Carlo (MCMC) method which consists in sampling from μ_n by constructing an appropriate Markov chain whose stationary distribution is μ_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 [23], we here use the Langevin diffusion process; see [51].

6.1. The 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)

$$d\boldsymbol{L}(t) = \frac{1}{2}\boldsymbol{\psi}(\boldsymbol{L}(t))dt + d\boldsymbol{W}(t),\ t > 0,\ \boldsymbol{L}(0) = \boldsymbol{l}_0, \tag{6.2}$$

where $\boldsymbol{\psi} = \nabla \ln(\mu)$, μ is everywhere non-zero and a 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 (6.2) has a unique strong solution and, moreover, $\boldsymbol{L}(t)$ has a stationary distribution with density precisely μ [51, Theorem 2.1]. $\boldsymbol{L}(t)$ is therefore interesting for sampling from μ. In particular, this opens the door to approximating integrals $\int_{\mathbb{R}^p} f(\boldsymbol{\theta})\mu(\boldsymbol{\theta})d\boldsymbol{\theta}$ by the average value of a Langevin diffusion, i.e., $\frac{1}{T}\int_0^T f(\boldsymbol{L}(t))dt$ for a large enough T. Under additional assumptions on μ and f in a proper functional class, the expected squared error of the approximation can be controlled [57].

Forward Euler discretization In practice, in simulating the diffusion sample path, we cannot follow exactly the dynamic defined by the SDE (6.2). Instead, we must discretize it. A popular discretization is given by the forward (Euler) scheme, which reads

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

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

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

It is then tempting to approximate $\widehat{\boldsymbol{\theta}}_n$ by applying this discretization strategy to the Langevin diffusion with μ_n in (6.1) as the target density. However, quantitative consistency guarantees of this discretization require μ (hence $\boldsymbol{\psi}$) to be sufficiently smooth, which limits their applicability in our context. To cope with this difficulty, a few works have recently proposed to replace $\ln(\mu)$ with a smoothed version (typically involving the Moreau-Yosida regularization/envelope, see Definition 2.2) [28, 29, 46]. In [29, 46] 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 [46][3], $-\ln(\mu)$ is replaced with its Moreau envelope, while in [29], it is assumed that $-\ln(\mu) = F + G$, F is convex Lipschitz continuously differentiable, and G 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 and thus cannot be applied to our prior. Proximal steps within MCMC methods have been recently proposed for some simple (convex) signal processing problems [12], though without any guarantees.

6.2. Forward-Backward proximal Langevin MC algorithms

Consider the prior π in (4.2) with g is given in Example 4.4 and $H = +\infty$. Then, μ_n is neither differentiable nor log-concave. To overcome these difficulties, we will exploit the structure of μ_n and some arguments from variational

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

analysis [52]. For ease of notation, in the following, we denote with the same symbol the measure and its density with respect to the Lebesgue measure. Thus μ_n reads

$$\mu_n(\boldsymbol{\theta}) \propto \exp\left(-(V(\boldsymbol{\theta}))\right) \tag{6.3}$$

where $V \overset{\text{def}}{=} F_\beta + W_\lambda \circ \boldsymbol{D}^\top$, $F_\beta(\boldsymbol{\theta}) = \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2/\beta$ and $W_\lambda(\boldsymbol{u}) = \sum_{\ell=1}^{L} w_\lambda\Big(\left\|\boldsymbol{u}_{\mathcal{G}_\ell}\right\|_2\Big)$ with

$$w_\lambda : x \in [0, +\infty[\;\mapsto \alpha^a x^a + c\ln(\tau^b + x^b), \tag{6.4}$$

and w_λ is parameterized by $\boldsymbol{\lambda} = (a, b, c, \alpha, \tau) \in]0,1]\times]0,1]\times]2 + G/b, +\infty[\times\mathbb{R}^+ \times \mathbb{R}^{+,*}$.

We start by collecting some important properties on the function w_λ and its proximal operator, as well as their implications. We denote $V_\gamma = F_\beta + ({}^\gamma W_\lambda) \circ \boldsymbol{D}^\top$.

Lemma 6.1. *The function w_λ in (6.4) is bounded from below, increasing and continuously differentiable on $]0, +\infty[$. Fix $a = b = 1$. Then, for any $\gamma \in]0, \tau^2/c[$, the following holds,*

(i) $\operatorname{prox}_{\gamma w_\lambda}$ *is single-valued on* $[0, +\infty[$ *and is given by*

$$\operatorname{prox}_{\gamma w_\lambda}(x) = \begin{cases} 0 & \text{if } x \leq \gamma w_\lambda'(0^+), \\ t - \gamma w_\lambda'\Big(\operatorname{prox}_{\gamma w_\lambda}(x)\Big) & \text{if } x > \gamma w_\lambda'(0^+). \end{cases}$$

(ii) For any $x \geq 0$, $0 \leq \operatorname{prox}_{\gamma w_\lambda}(x) \leq x$.

(iii) $\operatorname{prox}_{\gamma W_\lambda}(\boldsymbol{u}) = \left(\operatorname{prox}_{\gamma w_\lambda}\Big(\left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2\Big)\frac{\boldsymbol{u}_{\mathcal{G}_1}^\top}{\|\boldsymbol{u}_{\mathcal{G}_1}\|_2}, \ldots, \operatorname{prox}_{\gamma w_\lambda}\Big(\left\|\boldsymbol{u}_{\mathcal{G}_L}\right\|_2\Big)\frac{\boldsymbol{u}_{\mathcal{G}_L}^\top}{\|\boldsymbol{u}_{\mathcal{G}_L}\|_2}\right)^\top$.

(iv) $\operatorname{prox}_{\gamma W_\lambda \circ \boldsymbol{D}^\top}$ *is Lipschitz continuous.*

(v) $\nabla({}^\gamma W_\lambda \circ \boldsymbol{D}^\top) = \gamma^{-1}\boldsymbol{D} \circ (\mathbf{I}_q - \operatorname{prox}_{\gamma W_\lambda}) \circ \boldsymbol{D}^\top$. *It is a uniformly bounded and Lipschitz continuous operator and ∇V_γ is a Lipschitz continuous mapping.*

(vi) Assume that $\gamma \in]0, \min(\tau^2/c, \beta/(2\|\boldsymbol{X}\|^2))[$. Let $\boldsymbol{M}_\gamma \overset{def}{=} \mathbf{I}_p - (2\gamma/\beta)\boldsymbol{X}^\top\boldsymbol{X}$ which is a symmetric positive definite matrix. Then, $\nabla^{\boldsymbol{M}_\gamma,\gamma} V = \gamma^{-1}\boldsymbol{M}_\gamma\Big(\mathbf{I}_p - \operatorname{prox}_{\gamma W_\lambda \circ \boldsymbol{D}^\top}(\mathbf{I}_p - \gamma\nabla F_\beta)\Big)$ which is a Lipschitz continuous mapping.

We now describe two Langevin MC (LMC) sampling algorithms, originally proposed in [40], that are based on forward-backward proximal splitting and establish their guarantees for the penalty w_λ. In the rest of the section, we will fix $a = b = 1$.

Semi-Forward-Backward LMC (Semi-FBLMC) Assume that $\gamma \in]0, \tau^2/c[$. Define the following SDE with the Moreau-Yosida regularized version of W_λ

$$d\boldsymbol{L}(t) = -\frac{1}{2}\nabla V_\gamma(\boldsymbol{L}(t))dt + d\boldsymbol{W}(t), \; t > 0. \tag{6.5}$$

Inserting Lemma 6.1(v) into (6.5), the Euler discretization of (6.5) reads

$$\boldsymbol{L}_{k+1} = \boldsymbol{L}_k - \frac{\delta}{2}\nabla F_\beta(\boldsymbol{L}_k) - \frac{\delta}{2\gamma}\boldsymbol{D}\Big(\boldsymbol{D}^\top\boldsymbol{L}_k - \operatorname{prox}_{\gamma W_\lambda}(\boldsymbol{D}^\top\boldsymbol{L}_k)\Big) + \sqrt{\delta}\boldsymbol{Z}_k, \; t > 0, \; \boldsymbol{L}_0 = \boldsymbol{l}_0. \tag{6.6}$$

Forward-Backward LMC (FBLMC) Assume that $\gamma \in]0, \min(\tau^2/c, \beta/(2\|\boldsymbol{X}\|^2))[$. One can consider an alternative version of the SDE (6.5) with the Moreau-Yosida regularized version of V in the metric $\boldsymbol{M}_\gamma$ (see Lemma 6.1(vi)), i.e.,

$$d\boldsymbol{L}(t) = -\frac{1}{2}\nabla\Big({}^{\boldsymbol{M}_\gamma,\gamma}V \circ \boldsymbol{M}_\gamma^{-1/2}\Big)(\boldsymbol{L}(t))dt + \boldsymbol{M}_\gamma^{1/2}d\boldsymbol{W}(t), \; t > 0. \tag{6.7}$$

By the change of variable $\boldsymbol{U}(t) = \boldsymbol{M}_\gamma^{-1/2}\boldsymbol{L}(t)$ we get the following SDE

$$d\boldsymbol{U}(t) = -\frac{1}{2}\boldsymbol{M}_\gamma^{-1}\nabla^{\boldsymbol{M}_\gamma,\gamma}V(\boldsymbol{U}(t))dt + d\boldsymbol{W}(t),\ t > 0. \tag{6.8}$$

In view of Lemma 6.1(vi), the Euler discretization of (6.8) is given by

$$\boldsymbol{U}_{k+1} = (1 - \tfrac{\delta}{2\gamma})\boldsymbol{U}_k + \frac{\delta}{2\gamma}\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}\circ\boldsymbol{D}^\top}(\boldsymbol{U}_k - \gamma\nabla F_{\boldsymbol{\beta}}(\boldsymbol{U}_k)) + \sqrt{\delta}\boldsymbol{Z}_k,\ t > 0,\ \boldsymbol{U}_0 = \boldsymbol{l}_0. \tag{6.9}$$

Remark 6.1. *Observe that* $\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}\circ\boldsymbol{D}^\top}$ *in* (6.9) *is no separable in general. Owing to Assumption* (**H.2**)*, one can show quite immediately that*

$$\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}\circ\boldsymbol{D}^\top} = \widetilde{\boldsymbol{D}}\circ\mathrm{prox}^{\boldsymbol{D}^\top\widetilde{\boldsymbol{D}}}_{\gamma W_{\boldsymbol{\lambda}}+\iota_{\mathrm{Im}(\boldsymbol{D}^\top)}}\circ\boldsymbol{D}^\top,$$

where $\iota_{\mathrm{Im}(\boldsymbol{D}^\top)}(\boldsymbol{u}) = 0$ *if* $\boldsymbol{u}\in\mathrm{Im}(\boldsymbol{D}^\top)$ *and* $+\infty$ *otherwise, and* $\boldsymbol{D}^\top\widetilde{\boldsymbol{D}}$ *is indeed definite positive on* $\mathrm{Im}(\boldsymbol{D}^\top)$*. This means that, unless* $\boldsymbol{D}$ *is orthogonal,* $W_{\boldsymbol{\lambda}}\circ\boldsymbol{D}^\top$ *does not have an easy-to-compute expression, but rather necessitates to solve an optimization subproblem. Thus, from a computational perspective, for a general frame* $\boldsymbol{D}$*, Semi-FBLMC is more efficient. For the case where* $\boldsymbol{D}$ *is invertible, one can operate a simple change of variable* $\boldsymbol{u} = \boldsymbol{D}^\top\boldsymbol{\theta}$ *and replace* $F_{\boldsymbol{\beta}}$ *with* $F_{\boldsymbol{\beta}}\circ\boldsymbol{D}^{\top-1}$*, and* $W_{\boldsymbol{\lambda}}\circ\boldsymbol{D}^\top$ *with* $W_{\boldsymbol{\lambda}}$*. In this case, FBLMC can be applied efficiently.*

6.3. Convergence guarantees

By virtue of Lemma 6.1(v)-(vi), standard results (see, e.g., [51, Theorem 2.1]) show that for any initial point $\boldsymbol{L}(0)$ such that $\mathbb{E}\left\{\|\boldsymbol{L}(0)\|_2^2\right\} < \infty$, (6.5) has a unique solution which is strongly Markovian, $\mathbb{E}\left\{\|\boldsymbol{L}(t)\|_2^2\right\} < \infty$ for all $t > 0$, and $\boldsymbol{L}$ admits an (unique) invariant measure having the density μ_γ

$$\mu_\gamma(\boldsymbol{\theta}) \propto \exp\left(-V_\gamma(\boldsymbol{\theta})\right).$$

The same also holds for (6.7) with the corresponding invariant measure. The following claim is a consequence of [40, Proposition 3.1] and Lemma 6.1. In the sequel, $\|\nu\|_{\mathrm{TV}}$ stands for the total variation norm of a signed measure ν.

Proposition 6.1. *Let* μ_γ *be the invariant measure of either* (6.5) *or* (6.7)*. Then,* $\left\|\mu_\gamma - \mu_n\right\|_{\mathrm{TV}} \to 0$ *as* $\gamma \to 0$*.*

We consider the Semi-FBLMC discretization (6.6) of the Langevin diffusion (6.5). Let $\tilde{\boldsymbol{L}}_\delta$ be the continuous-time extension of the scheme (6.6)

$$\tilde{\boldsymbol{L}}_\delta(t) \overset{\mathrm{def}}{=} \boldsymbol{L}_0 - \frac{1}{2}\int_0^t \nabla V_\gamma(\bar{\boldsymbol{L}}(t))dt + \int_0^t d\boldsymbol{W}(s),$$

where $\bar{\boldsymbol{L}}(t) = \boldsymbol{L}_k$ for $t\in[k\delta,(k+1)\delta[$. We write $\mathbf{P}_{\boldsymbol{L}}^{t,\gamma}(\boldsymbol{l}_0,\Omega) = \Pr\{\boldsymbol{L}(t)\in\Omega|\boldsymbol{L}(0) = \boldsymbol{l}_0\}$ for all Borel sets Ω and initial condition $\boldsymbol{l}_0$. Similarly, we denote $\mathbf{P}_{\tilde{\boldsymbol{L}}_\delta}^{t,\delta,\gamma}(\boldsymbol{l}_0,\Omega) = \Pr\left\{\tilde{\boldsymbol{L}}_\delta(t)\in\Omega|\tilde{\boldsymbol{L}}_\delta(0) = \boldsymbol{l}_0\right\}$. The superscripts stress the dependence on the parameters.

We also define the sample average

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

and the EWA estimate

$$\widehat{\boldsymbol{\theta}}_\gamma = \int_{\mathbb{R}^p}\boldsymbol{\theta}\mu_\gamma(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

Theorem 6.1. *The following holds:*

(i) $\lim_{\gamma\to0}\lim_{T\to+\infty}\lim_{\delta\to0}\left\|\mathbf{P}_{\tilde{\boldsymbol{L}}_\delta}^{T,\delta,\gamma}(\boldsymbol{l}_0,\cdot) - \mu_n\right\|_{\mathrm{TV}} = 0.$

(ii) Suppose that $\boldsymbol{l}_0 = 0$*, then* $\lim_{T\to+\infty}\lim_{\delta\to0}\mathbb{E}\left\{\left\|\bar{\boldsymbol{L}}_{\delta,T,\gamma} - \widehat{\boldsymbol{\theta}}_\gamma\right\|_2\right\} = 0.$

To prove these claims, we first show that the Langevin diffusion in (6.5) is uniformly geometrically ergodic. Then we follow standard arguments, invoking Lemma 6.1, Proposition 6.1, the Girsanov formula and Pinsker inequality[4]. Similarly to Theorem 6.1, convergence guarantees of the FBLMC discretization scheme (6.9) can be established. We omit the details here for the sake of brevity.

[4]We have made no effort to sharpen the constants and the rates, and for instance their dependence on the dimension. This is beyond the scope of this paper.

7. Numerical experiments

In this section, some numerical experiments are conducted to illustrate and validate the numerical performance of the proposed EWA estimator. We consider a linear regression problem

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\theta}_0 + \boldsymbol{\xi},$$

where $\boldsymbol{\xi}$ is the noise, $\boldsymbol{X}$ is the design matrix. $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ is the unknown regression vector of interest assumed to obey Assumption **(H.1)**. Since $H = +\infty$, we then have to choose a distribution on the noise $\boldsymbol{\xi}$ such that β is independent of H. Such type of distributions is specified in [23, Section 2]. For our implementation, we assume $\boldsymbol{\xi}$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$. The noise level σ is chosen according to the simulated $\boldsymbol{\theta}_0$.

The parameters of EWA were chosen as prescribed in our theoretical analysis. For instance, the temperature parameter is set to $\beta = 4\sigma^2$, the parameters of the prior: $a = b = 1$, $c > (2+G)/b$, $\tau \sim 1/(pn)$ and $\alpha \leq 1/\left(3\sqrt{K^{\boldsymbol{D}}_{a,g} L}\right)^{1/a}$, where $K^{\boldsymbol{D}}_{a,g}$ is given in Corollary 5.3. The number of iterations N and the step-size δ are chosen respectively large and small enough to guarantee convergence and discretization consistency of the algorithm.

Following the philosophy of reproducible research, all the code implementing our EWA algorithm and reproducing the experiments of this paper are made publicly available for download at `https://github.com/luuduytung/GroupAnalyseEWAToolbox`.

7.1. 1-D signal recovery under group sparsity

In this example, 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 $G_{\boldsymbol{\theta}_0} = 1$; group structured sparsity with $G_{\boldsymbol{\theta}_0} = 4$. Besides, the positions of the non-zero/active entries (for $G_{\boldsymbol{\theta}_0} = 1$) or groups (for $G_{\boldsymbol{\theta}_0} = 4$) are chosen randomly uniformly on $\{1, \ldots, p\}$.

The experiments are performed by fixing $p = 128$, and taking $S \in 2^{\{2,\ldots,7\}}$, $n \in 2^{\{3,\ldots,7\}}$, 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 (5.1). For each (S, n), and each value of $G_{\boldsymbol{\theta}_0}$, $N_{\text{rep}} = 20$ instances of the problem suite $(\boldsymbol{X}, \boldsymbol{\theta}_0, \boldsymbol{Y})$ are generated, and EWA is applied with a chosen G and the other parameters as detailed above. The estimation quality/success is then assessed by

$$\pi_{S,n} = \frac{1}{N_{\text{rep}}} \sum_{j=1}^{N_{\text{rep}}} I\left(\left\| \widehat{\boldsymbol{\theta}}_n^{(j,S,n)} - \boldsymbol{\theta}_0^{(j,S,n)} \right\|_n \leq \epsilon \right), \tag{7.1}$$

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., problems are easier as one moves up (more measurements) and to the left (sparser $\boldsymbol{\theta}_0$). Phase diagrams plotting $\pi_{S,n}$ in (7.1) as a function $(S/p, n/p)$ were widely advocated by Donoho and co-authors for ℓ_1 minimization [24]. Such diagrams often have an interesting two-phase structure (as displayed in Figures 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 1(a) (resp. (b)) shows the phase diagrams when $G_{\boldsymbol{\theta}_0} = 1$ and $G = 1$ (resp. $G = 4$) in EWA. In this case, the phase transition curve for $G = 1$, the correct group size, is slightly better that with $G = 4$. The situation reverses for Figures 1(c)-(d) where $G_{\boldsymbol{\theta}_0} = 4$, and one observes that the success area is significantly better using $G = 4$ than $G = 1$. This is expected as it reveals better performance of EWA when used with the choice $G = G_{\boldsymbol{\theta}_0}$. This is also confirmed by visual inspection of Figures 1(c')-(d'), where we plotted instances of recovered vectors $\widehat{\boldsymbol{\theta}}_n^{(j,S,n)}$ when $(S, G_{\boldsymbol{\theta}_0}) \in \{4, 8\} \times \{1, 4\}$ and $n/p = 1/2$. EWA was again applied with $G = 1$ and $G = 4$ 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 $G = 1$.

It is worth observing that $S/p = \left\|\boldsymbol{\theta}_0\right\|_{0,\mathcal{G}} G_{\boldsymbol{\theta}_0}/p$. As far as the expected phase transition curve is concerned, one has from Corollary 5.3 that it is expected to occur for

$$n/p = C_\epsilon \left\|\boldsymbol{\theta}_0\right\|_{0,\mathcal{G}} G_{\boldsymbol{\theta}_0}/p(\ln(p/G_{\boldsymbol{\theta}_0})/G_{\boldsymbol{\theta}_0}) = C_\epsilon S/p(\ln(p/G_{\boldsymbol{\theta}_0})/G_{\boldsymbol{\theta}_0})$$

for some constant $C_\epsilon > 0$ depending on ϵ. That is, the phase transition curve is linear (p and $G_{\boldsymbol{\theta}_0}$ are fixed for each diagram), which is confirmed by visual inspection of Figures 1(a)-(d), where the overlaid blue line is the fitted linear phase transition curve.

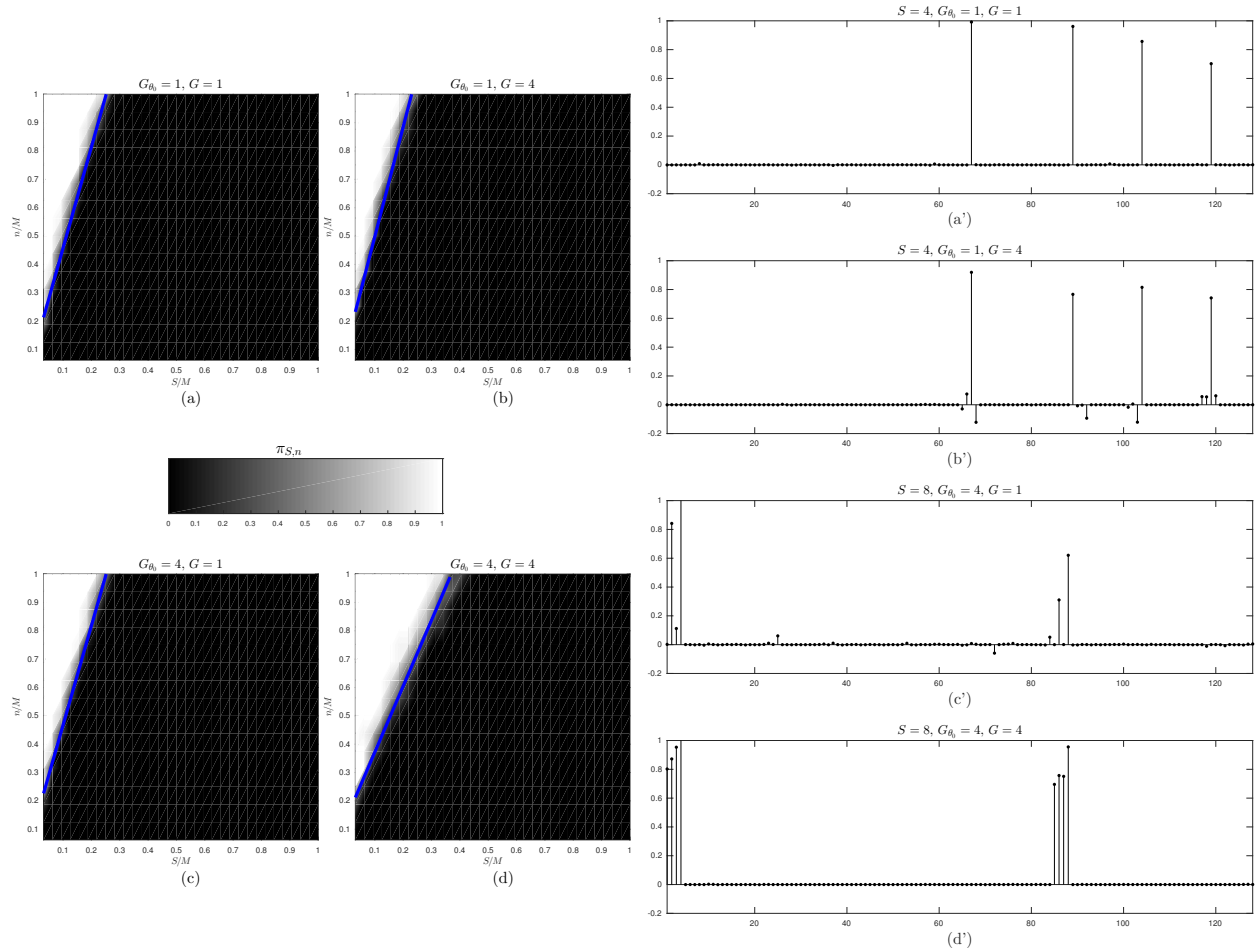

Figure 1: (a)-(d): Phase diagrams of EWA for $\boldsymbol{D} = \mathbf{I}_p$, the color bar ranges from dark ($\pi_{S,n} = 0$) to bright ($\pi_{S,n} = 1$). The blue line is the fitted phase transition curve. (a')-(d'): Examples of vectors $\widetilde{\boldsymbol{\theta}}_n^{(j,S,n)}$ recovered by EWA with $n/p = 1/2$, two sparsity levels $S = 4$ and $S = 8$ and two group sizes $G_{\boldsymbol{\theta}_0} = 1$ and $G_{\boldsymbol{\theta}_0} = 4$.

7.2. 2-D image recovery under analysis group-sparsity

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 2(a)). Thus $\mathrm{vec}(\boldsymbol{\theta}_0)$ is vector in $\mathbb{R}^p$ with $p = 160^2$, and our goal is to recover $\boldsymbol{\theta}_0$ with values in $[0, 1]$ from

$$\boldsymbol{Y} = \boldsymbol{X}\,\mathrm{vec}(\boldsymbol{\theta}_0) + \boldsymbol{\xi},$$

where $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$ and $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is again random whose entries are i.i.d. from the Rademacher distribution. Since the targeted image is piecewise-constant, a popular prior is the so-called isotropic total variation [53] which is

described in Example 4.1. It turns out that this can be cast in our analysis group-sparsity framework as a special case. In this experiment, we use Semi FBLMC Algorithm to compute the EWA estimator.

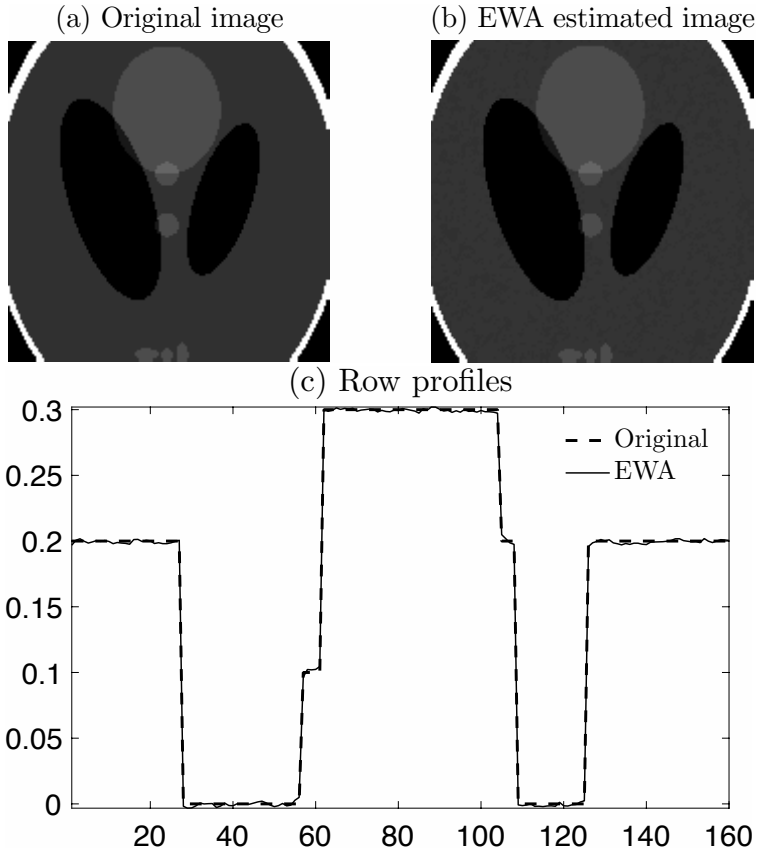

Figure 2: (a): Original close-up of Shepp-Logan phantom image. (b): Image recovered by EWA with $\delta = 2 \cdot 10^{-8}$ and $T = 10^4$. (c) Profiles of a row extracted from each image.

The results are depicted in Figure 2. $\sigma = 0.525$ in this experiment, the number of observations is $n = 9p/16 = 14400$, and we have $\left\|D_{\mathrm{TV}}(\theta_0)\right\|_{0,\mathcal{G}} = 1376 \ll n$. A notable property of the EWA estimate is that it does not suffer from the stair-casing effect, unlike total variation minimization.

8. Conclusion

In this paper, we proposed a class of EWA estimators constructed from a novel and versatile family of priors which promotes analysis group-sparsity, where the analysis operator corresponds to a frame. Its quality is guaranteed by establishing a sharp SOI with a small remainder term in high-dimension. We also described a forward-backward proximal LMC algorithm, which is an implementation of EWA and can be viewed as a Forward Euler discretization of a Langevin diffusion involving the Moreau-Envelope of the potential in a proper metric. We derived convergence guarantees of this discretization. The performance of the estimator was illustrated on some numerical experiments which support our theoretical findings. There are still open problems that we leave to a future work. More precisely, one direction is to investigate how to remove the frame assumption. Another one would be to derive further/better quantitative convergence bounds of the proposed discretization.

9. Proofs

9.1. Proofs of Section 2

Proof of Lemma 2.3 Consider the linear mapping $\boldsymbol{M} : \boldsymbol{x} \in \mathbb{R}^d \mapsto \boldsymbol{M}\boldsymbol{x} \in \mathbb{R}^r$, $r \geq d$. The Jacobian matrix of this mapping is obviously $\boldsymbol{M}$ for any $\boldsymbol{x} \in \mathbb{R}^d$. Since $\boldsymbol{M}$ is a frame, it is injective, hence so-called d-regular (see [41, Section 1.5]). In particular, $\det(\boldsymbol{M}\boldsymbol{M}^\top) > 0$. Thus combining [41, Theorems 1.12 and 3.4] and the Cauchy-Binet formula [41, Theorem 3.3]), we have the change of variables formula

$$\int_\Theta u(\boldsymbol{M}\boldsymbol{x})d\boldsymbol{x} = \frac{\int_{\mathbb{R}^r} \sum_{\boldsymbol{x}\in\Theta\cap\{\boldsymbol{\omega} \,:\, \boldsymbol{M}\boldsymbol{\omega}=\boldsymbol{v}\}} u(\boldsymbol{M}\boldsymbol{x})d\boldsymbol{v}}{\sqrt{\det(\boldsymbol{M}\boldsymbol{M}^\top)}} = \frac{\int_{\mathrm{Im}(\boldsymbol{M})} \sum_{\boldsymbol{x}\in\Theta\cap\{\boldsymbol{\omega} \,:\, \boldsymbol{M}\boldsymbol{\omega}=\boldsymbol{v}\}} u(\boldsymbol{M}\boldsymbol{x})d\boldsymbol{v}}{\sqrt{\det(\boldsymbol{M}\boldsymbol{M}^\top)}}.$$

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

9.2. Proofs of Section 3

Proof of Proposition 3.1

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

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

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

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

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

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

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

$$\varphi_{\xi+\zeta}(\boldsymbol{t}) = \varphi_\xi(\boldsymbol{t})\varphi_\zeta(\boldsymbol{t}) = \frac{1}{1+(1+\gamma)^2\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}/2} = \varphi_{(1+\gamma)\xi}(\boldsymbol{t}). \tag{9.1}$$

Thus, $\boldsymbol{\zeta} + \boldsymbol{\xi}$ has the same distribution as $(1+\gamma)\boldsymbol{\xi}$. We also obtain $\mathbb{E}\{\boldsymbol{\zeta}|\boldsymbol{\xi}\} = \mathbb{E}\{\boldsymbol{\zeta}\} = (-i)\nabla\varphi_\zeta(0) = 0$. It suffices to check condition (c) of Assumption (P.2). We know that

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

Using Taylor's formula, we have

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

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

$$\frac{\ln \mathbb{E}\left\{e^{\boldsymbol{t}^\top\boldsymbol{\zeta}}|\boldsymbol{\xi} = \boldsymbol{a}\right\}}{\|\boldsymbol{t}\|_2^\top\,\gamma v(\boldsymbol{a})} \underset{\gamma\to 0}{\to} \frac{\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}/(1-\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}/2)}{\|\boldsymbol{t}\|_2^2\left|\!\left|\!\left|\boldsymbol{\Sigma}\right|\!\right|\!\right|/(1-t_0^2\left|\!\left|\!\left|\boldsymbol{\Sigma}\right|\!\right|\!\right|/2)} \leq \frac{1-t_0^2\left|\!\left|\!\left|\boldsymbol{\Sigma}\right|\!\right|\!\right|/2}{1-\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}/2} \leq \frac{1-t_0^2\left|\!\left|\!\left|\boldsymbol{\Sigma}\right|\!\right|\!\right|/2}{1-\|\boldsymbol{t}\|_2^2\left|\!\left|\!\left|\boldsymbol{\Sigma}\right|\!\right|\!\right|/2} \leq 1.$$

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

- **Bounded symmetric noise:** Let $\boldsymbol{\xi}$ are symmetric and $\Pr\left\{\left|\xi_i\right| \leq B_i\right\} = 1$ for some $\boldsymbol{B} \in \mathbb{R}^n$, we set $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_n)^\top$ such that $\zeta_i = (1+\gamma)\left|\xi_i\right| \operatorname{sgn}(\operatorname{sgn}(\xi_i) - (1+\gamma)U_i) - \xi_i$, $U_i \sim \mathcal{U}([-1,1])$ for any $i \in \{1, \ldots, n\}$. Using [23, Equation (22)], for any $\boldsymbol{t} \in \mathbb{R}^n$ and $\boldsymbol{a} \in \{\boldsymbol{x} \in \mathbb{R}^n \ : \ x_i \in [-B_i, B_i],\ \forall i \in \{1, \ldots, n\}\}$, we get that

$$\mathbb{E}\left\{e^{\boldsymbol{t}^\top \boldsymbol{\zeta}} | \boldsymbol{\xi} = \boldsymbol{a}\right\} = \prod_{j=1}^n \mathbb{E}\left\{e^{t_i \zeta_i} | \xi_i = a_i\right\} = e^{-\boldsymbol{t}^\top \boldsymbol{a}}\left(e^{(1+\gamma)\boldsymbol{t}^\top \boldsymbol{a}} \frac{2+\gamma}{2+2\gamma} + e^{-(1+\gamma)\boldsymbol{t}^\top \boldsymbol{a}} \frac{\gamma}{2+2\gamma}\right). \tag{9.2}$$

From (9.2) and the symmetry of $\boldsymbol{\xi}$, we obtain

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

Thus, $\boldsymbol{\zeta} + \boldsymbol{\xi}$ has the same distribution as $(1+\gamma)\boldsymbol{\xi}$. Since $\mathbb{E}\{\boldsymbol{\zeta} | \boldsymbol{\xi} = \boldsymbol{a}\}$ equals to the gradient of $\mathbb{E}\left\{e^{\boldsymbol{t}^\top \boldsymbol{\zeta}} | \boldsymbol{\xi} = \boldsymbol{a}\right\}$ at $\boldsymbol{t} = 0$, from (9.2) we have then $\mathbb{E}\{\boldsymbol{\zeta} | \boldsymbol{\xi} = \boldsymbol{a}\} = 0$, $\forall \boldsymbol{a} \in [-\boldsymbol{B}, \boldsymbol{B}]$. It suffices to check the condition (c) of Assumption (**P.2**). Owing to [19, Lemma 3] and [23, Equation (22)], we get that $\ln\left(\mathbb{E}\left\{e^{t_i \zeta_i | \xi_i = a_i}\right\}\right) \leq (t_i a_i)^2 \gamma(1+\gamma)$. Thus, let $\boldsymbol{t} \in \{\boldsymbol{x} \in \mathbb{R}^n \ : \ \|\boldsymbol{x}\|_2 \in [-\infty, \infty]\}$ and $v(\boldsymbol{a}) = \|\boldsymbol{a}\|_2^2$, we get that

$$\frac{\ln \mathbb{E}\left\{e^{\boldsymbol{t}^\top \boldsymbol{\zeta}} | \boldsymbol{\xi} = \boldsymbol{a}\right\}}{\|\boldsymbol{t}\|_2^2 \gamma v(\boldsymbol{a})} = \frac{\sum_{i=1}^n \ln \mathbb{E}\left\{e^{t_i \zeta_i} | \xi_i = a_i\right\}}{\|\boldsymbol{t}\|_2^2 \gamma \|\boldsymbol{a}\|_2^2} \leq \frac{\sum_{i=1}^n t_i^2 a_i^2 \gamma(1+\gamma)}{\|\boldsymbol{t}\|_2^2 \gamma \|\boldsymbol{a}\|_2^2} \leq 1 + \gamma \underset{\gamma \to 0}{\to} 1 \leq 1.$$

9.3. Proofs of Section 4

Proof of Proposition 4.1 Let $\boldsymbol{\theta} \in \Theta$ and $i \in \{1, \ldots, n\}$. Setting $u_i^{\boldsymbol{\theta}} = \sum_{j=1}^p \theta_j f_j(x_i)$ and $\boldsymbol{u}^{\boldsymbol{\theta}} = (u_1^{\boldsymbol{\theta}}, \ldots, u_n^{\boldsymbol{\theta}})^\top$, and by virtue of (2.1), (2.2), (4.1) and the fact that $a \in]0, 1]$, we have

$$\left\|\boldsymbol{u}^{\boldsymbol{\theta}}\right\|_2 = \left\|\boldsymbol{X}\boldsymbol{\theta}\right\|_2 \leq \left|\!\left|\!\left|\boldsymbol{X}\right|\!\right|\!\right| \left|\!\left|\!\left|\widetilde{\boldsymbol{D}}\right|\!\right|\!\right| \left\|\boldsymbol{D}^\top \boldsymbol{\theta}\right\|_2 \leq \frac{\left|\!\left|\!\left|\boldsymbol{X}\right|\!\right|\!\right| \left\|\boldsymbol{D}^\top \boldsymbol{\theta}\right\|_{a,\mathcal{G}}}{\sqrt{\kappa}} \leq \frac{\left|\!\left|\!\left|\boldsymbol{X}\right|\!\right|\!\right|_{R^{1/a}}}{\sqrt{\kappa}},$$

which in turn implies $\boldsymbol{u}^{\boldsymbol{\theta}} \in \mathcal{B}$. Therefore, for any $(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \Theta^2$, $\left\|\boldsymbol{f}_{\boldsymbol{\theta}} - \boldsymbol{f}_{\boldsymbol{\theta}'}\right\|_2 = \left\|\mathscr{L}(\boldsymbol{u}^{\boldsymbol{\theta}}) - \mathscr{L}(\boldsymbol{u}^{\boldsymbol{\theta}'})\right\|_2 \leq 2 \max_{\boldsymbol{x} \in \mathcal{B}} \left\|\mathscr{L}(\boldsymbol{x})\right\|_2$. □

Proof of Lemma 4.1 Let us first check the integrability condition (**G.2**). By Lemmas 2.3 and 2.2, we obtain

$$\begin{aligned}
\int_{\mathbb{R}^p} \prod_{\ell=1}^L g(\left\|[\boldsymbol{D}^\top \boldsymbol{u}]_{\mathcal{G}_\ell}\right\|_2) d\boldsymbol{u} = \frac{\int_{\operatorname{Im}(\boldsymbol{D}^\top)} \prod_{\ell=1}^L g(\left\|\boldsymbol{v}_{\mathcal{G}_\ell}\right\|_2) d\boldsymbol{v}}{\sqrt{\det(\boldsymbol{D}\boldsymbol{D}^\top)}} &\leq \frac{\int_{\mathbb{R}^q} \prod_{\ell=1}^L g(\left\|\boldsymbol{v}_{\mathcal{G}_\ell}\right\|_2) d\boldsymbol{v}}{\sqrt{\det(\boldsymbol{D}\boldsymbol{D}^\top)}} \\
&= \frac{\left(\int_{\mathbb{R}^G} g(\left\|\boldsymbol{u}\right\|_2) d\boldsymbol{u}\right)^L}{\sqrt{\det(\boldsymbol{D}\boldsymbol{D}^\top)}} \\
&= \frac{C_G^L \left(\int_0^\infty z^{G-1} g(z) dz\right)^L}{\sqrt{\det(\boldsymbol{D}\boldsymbol{D}^\top)}}.
\end{aligned}$$

Since $G \geq 1$ and $g : \mathbb{R}^+ \to \mathbb{R}^+$, by (4.4), we get

$$\begin{aligned}
\int_{\mathbb{R}^p} \prod_{\ell=1}^L g(\left\|[\boldsymbol{D}^\top \boldsymbol{u}]_{\mathcal{G}_\ell}\right\|_2) d\boldsymbol{u} &\leq \frac{C_G^L \left(\int_0^1 z^{G-1} g(z) dz + \int_1^\infty w^{G-1} g(w) dw\right)^L}{\sqrt{\det(\boldsymbol{D}\boldsymbol{D}^\top)}} \\
&\leq \frac{C_G^L \left(\sup_{z \in [0,1]} g(z) + \int_1^\infty w^{G+1} g(w) dw\right)^L}{\sqrt{\det(\boldsymbol{D}\boldsymbol{D}^\top)}} < \infty.
\end{aligned} \tag{9.3}$$

Therefore, g satisfies Assumption (**G.2**). Now, we check the moment condition (**G.3**). Using similar arguments to the bound (9.3), we have

$$\begin{aligned}\int_{\mathbb{R}^p} \left\|[\boldsymbol{D}^\top \boldsymbol{u}]_{\mathcal{G}_\ell}\right\|_2^2 \prod_{k=1}^L g(\left\|[\boldsymbol{D}^\top \boldsymbol{u}]_{\mathcal{G}_k}\right\|_2) d\boldsymbol{u} &\leq \frac{\int_{\mathbb{R}^q} \left\|\boldsymbol{v}_{\mathcal{G}_\ell}\right\|_2^2 \prod_{k=1}^L g(\left\|\boldsymbol{v}_{\mathcal{G}_k}\right\|_2) d\boldsymbol{v}}{\sqrt{\det(\boldsymbol{D}\boldsymbol{D}^\top)}} \\ &= \frac{\left(\int_{\mathbb{R}^G} \left\|\boldsymbol{u}\right\|_2^2 g(\left\|\boldsymbol{u}\right\|_2) d\boldsymbol{u}\right)\left(\int_{\mathbb{R}^G} g(\left\|\boldsymbol{v}\right\|_2) d\boldsymbol{v}\right)^{L-1}}{\sqrt{\det(\boldsymbol{D}\boldsymbol{D}^\top)}} \\ &= \frac{C_G^L \int_0^\infty z^{G+1} g(z) dz \left(\int_0^\infty w^{G-1} g(w) dw\right)^{L-1}}{\sqrt{\det(\boldsymbol{D}\boldsymbol{D}^\top)}} < \infty, \end{aligned} \tag{9.4}$$

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

9.4. *Proofs of Section 5*

9.4.1. *Proof of Theorem 5.1*

Remind the prior $\pi(d\boldsymbol{\theta})$ from (4.2), where $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \left\|\boldsymbol{D}^\top \boldsymbol{\theta}\right\|_{a,\mathcal{G}}^a \leq R\}$. Let $r_L = 3\sqrt{K_{a,g}^{\boldsymbol{D}}} L$, $\Theta_{p_0{}^{\boldsymbol{D}}} = \{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \left\|\boldsymbol{D}^\top \boldsymbol{\theta} - \boldsymbol{D}^\top \boldsymbol{\theta}^*\right\|_{a,\mathcal{G}}^a \leq r_L\}$ and

$$\boldsymbol{\theta}^* \in \{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \left\|\boldsymbol{D}^\top \boldsymbol{\theta}\right\|_{a,\mathcal{G}}^a \leq R - 3\sqrt{K_{a,g}^{\boldsymbol{D}}} L = R - r_L\}. \tag{9.5}$$

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,

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

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

$$(f_{\boldsymbol{\theta}}(x_i) - f(x_i))^2 \leq (f_{\boldsymbol{\theta}^*}(x_i) - f(x_i))^2 + C_{f,\mathscr{L}}\left[\boldsymbol{X}_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right]^2 + 2(f_{\boldsymbol{\theta}^*}(x_i) - f(x_i))\mathscr{L}'\left(\boldsymbol{X}_i^\top \boldsymbol{\theta}^*\right) \boldsymbol{X}_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*), \tag{9.6}$$

where $C_{f,\mathscr{L}} = \left\|\mathscr{L}'\right\|_\infty^2 + \left\|\mathscr{L}''\right\|_\infty \left(\left\|\mathscr{L}\right\|_\infty + \left\|f\right\|_\infty\right)$. 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 (9.6) becomes

$$\begin{aligned} \int_\Theta \left\|f_{\boldsymbol{\theta}} - f\right\|_n^2 p_0{}^{\boldsymbol{D}}(d\boldsymbol{\theta}) \leq \left\|f_{\boldsymbol{\theta}^*} - f\right\|_n^2 + C_{f,\mathscr{L}} \int_{\mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \left[\boldsymbol{X}_i^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right]^2 p_0{}^{\boldsymbol{D}}(d\boldsymbol{\theta}) \\ + \frac{2}{n} \sum_{i=1}^n (f_{\boldsymbol{\theta}^*}(x_i) - f(x_i)) \mathscr{L}'\left(\boldsymbol{X}_i^\top \boldsymbol{\theta}^*\right) \boldsymbol{X}_i^\top \int_\Theta (\boldsymbol{\theta} - \boldsymbol{\theta}^*) p_0{}^{\boldsymbol{D}}(d\boldsymbol{\theta}). \end{aligned} \tag{9.7}$$

Note that, the right term of inequality (9.7) 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 (9.5) and the fact that $a \in]0, 1]$, we have

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

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

$$\Theta_{{p_0}^{\boldsymbol{D}}} \subseteq \Theta. \tag{9.8}$$

In what follows, we denote $\mathbb{B}_{a,\mathcal{G}}^a(x) = \left\{z \in \mathbb{R}^q : \left\|z\right\|_{a,\mathcal{G}}^a \leq x\right\}, \forall x > 0$ for brevity. By (9.8), property (2.1), Lemma 2.3 and symmetry of $\mathbb{B}_{a,\mathcal{G}}^a(r_L) \cap \mathrm{Im}(\boldsymbol{D}^\top)$, we obtain

$$\begin{aligned}&\int_\Theta (\boldsymbol{\theta} - \boldsymbol{\theta}^*) {p_0}^{\boldsymbol{D}}(d\boldsymbol{\theta}) \\ &\propto \int_{\Theta \cap \Theta_{{p_0}^{\boldsymbol{D}}}} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \prod_{\ell=1}^L \exp\left(-\alpha^a \left\|[\boldsymbol{D}^\top\boldsymbol{\theta} - \boldsymbol{D}^\top\boldsymbol{\theta}^*]_{\mathcal{G}_\ell}\right\|_2^a\right) g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{\theta} - \boldsymbol{D}^\top\boldsymbol{\theta}^*]_{\mathcal{G}_\ell}\right\|_2\right) d\boldsymbol{\theta} \\ &= \int_{\Theta_{{p_0}^{\boldsymbol{D}}}} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \prod_{\ell=1}^L \exp\left(-\alpha^a \left\|[\boldsymbol{D}^\top\boldsymbol{\theta} - \boldsymbol{D}^\top\boldsymbol{\theta}^*]_{\mathcal{G}_\ell}\right\|_2^a\right) g\left(\left\|[\boldsymbol{D}^\top\boldsymbol{\theta} - \boldsymbol{D}^\top\boldsymbol{\theta}^*]_{\mathcal{G}_\ell}\right\|_2\right) d\boldsymbol{\theta} \\ &= \frac{\widetilde{\boldsymbol{D}}}{\sqrt{\det(\boldsymbol{D}\boldsymbol{D}^\top)}} \int_{\mathbb{B}_{a,\mathcal{G}}^a(r_L) \cap \mathrm{Im}(\boldsymbol{D}^\top)} z \prod_{\ell=1}^L \exp\left(-\alpha^a \left\|z_{\mathcal{G}_\ell}\right\|_2^a\right) g\left(\left\|z_{\mathcal{G}_\ell}\right\|_2\right) dz = 0,\end{aligned} \tag{9.9}$$

which is the desired claim.

We now bound the second term in the right hand side of (9.7). Define

$$p_0(d\boldsymbol{u}) = \frac{1}{C_L \sqrt{\det(\boldsymbol{D}\boldsymbol{D}^\top)}} \prod_{\ell=1}^L \exp\left(-\alpha^a \left\|\boldsymbol{u}_{\mathcal{G}_\ell}\right\|_2^a\right) g\left(\left\|\boldsymbol{u}_{\mathcal{G}_\ell}\right\|_2\right) I_{\mathrm{Im}(\boldsymbol{D}^\top) \cap \mathbb{B}_{a,\mathcal{G}}^a(r_L)}(\boldsymbol{u}) d\boldsymbol{u}. \tag{9.10}$$

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 (2.3). So, p_0 is a probability measure on $\mathbb{R}^q$. For any $i, j \in \{1, \ldots, L\}$, $i \neq j$, by a change of variables, we get $\int_{\mathbb{R}^q} \boldsymbol{u}_{\mathcal{G}_i} \boldsymbol{u}_{\mathcal{G}_j}^\top p_0(d\boldsymbol{u}) = -\int_{\mathbb{R}^q} \boldsymbol{u}_{\mathcal{G}_i} \boldsymbol{u}_{\mathcal{G}_j}^\top p_0(d\boldsymbol{u})$, so

$$\int_{\mathbb{R}^q} \boldsymbol{u}_{\mathcal{G}_i} \boldsymbol{u}_{\mathcal{G}_j}^\top p_0(d\boldsymbol{u}) = 0. \tag{9.11}$$

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

$$\int_{\mathbb{R}^q} \boldsymbol{u}_{\mathcal{G}_j} \boldsymbol{u}_{\mathcal{G}_j}^\top p_0(d\boldsymbol{u}) = \int_{\mathbb{R}^q} \boldsymbol{u}_{\mathcal{G}_1} \boldsymbol{u}_{\mathcal{G}_1}^\top p_0(d\boldsymbol{u}). \tag{9.12}$$

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 2.3)} &= \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} \operatorname{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}\operatorname{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{((9.11) and (9.12))} &= \frac{1}{n}\sum_{\ell=1}^L \operatorname{tr}\left(\left[\left(\boldsymbol{X}\widetilde{\boldsymbol{D}}\right)^\top \boldsymbol{X}\widetilde{\boldsymbol{D}}\right]_{\mathcal{G}_\ell} \int_{\mathbb{R}^q} \boldsymbol{u}_{\mathcal{G}_1}\boldsymbol{u}_{\mathcal{G}_1}^\top p_0(d\boldsymbol{u})\right) \\
\text{(Von Neumann's trace inequality)} &\le \frac{1}{n}\sum_{\ell=1}^L \sum_{j=1}^G \sigma_j\left(\left[\left(\boldsymbol{X}\widetilde{\boldsymbol{D}}\right)^\top \boldsymbol{X}\widetilde{\boldsymbol{D}}\right]_{\mathcal{G}_\ell}\right)\sigma_j\left(\int_{\mathbb{R}^q} \boldsymbol{u}_{\mathcal{G}_1}\boldsymbol{u}_{\mathcal{G}_1}^\top p_0(d\boldsymbol{u})\right) \\
&\le \frac{1}{n}\int_{\mathbb{R}^q} \left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2 p_0(d\boldsymbol{u}) \sum_{\ell=1}^L \operatorname{tr}\left(\left[\left(\boldsymbol{X}\widetilde{\boldsymbol{D}}\right)^\top \boldsymbol{X}\widetilde{\boldsymbol{D}}\right]_{\mathcal{G}_\ell}\right) \\
&= \frac{1}{n}\operatorname{tr}\left(\left(\boldsymbol{X}\widetilde{\boldsymbol{D}}\right)^\top \boldsymbol{X}\widetilde{\boldsymbol{D}}\right) \int_{\mathbb{R}^q} \left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2 p_0(d\boldsymbol{u}).
\end{aligned} \tag{9.13}
$$

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

$$
\frac{\operatorname{tr}\left(\left(\boldsymbol{X}\widetilde{\boldsymbol{D}}\right)^\top \boldsymbol{X}\widetilde{\boldsymbol{D}}\right)}{n} \le \sum_{j=1}^p \sigma_j\left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\right)\sigma_j\left(\widetilde{\boldsymbol{D}}\widetilde{\boldsymbol{D}}^\top\right) \le \sigma_1\left(\widetilde{\boldsymbol{D}}\widetilde{\boldsymbol{D}}^\top\right)\sum_{j=1}^p \sigma_j\left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\right) \le \frac{p}{\kappa}. \tag{9.14}
$$

Putting together (9.13) and (9.14), we get the bound

$$
C_{f,\mathscr{L}} \int_{\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \left[\boldsymbol{X}_i^\top(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\right]^2 {p_0}^{\boldsymbol{D}}(d\boldsymbol{\theta}) \le C_{f,\mathscr{L}} \frac{p}{\kappa} \int_{\mathbb{R}^q} \left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2 p_0(d\boldsymbol{u}). \tag{9.15}
$$

Thanks to (9.9) and (9.15), inequality (9.7) becomes

$$
\int_\Theta \left\|f_{\boldsymbol{\theta}} - f\right\|_n^2 {p_0}^{\boldsymbol{D}}(d\boldsymbol{\theta}) \le \left\|f_{\boldsymbol{\theta}^*} - f\right\|_n^2 + C_{f,\mathscr{L}} \frac{p}{\kappa} \int_{\mathbb{R}^q} \left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2 p_0(d\boldsymbol{u}). \tag{9.16}
$$

Now, inserting (9.16) into Theorem 3.1 (with $p = {p_0}^{\boldsymbol{D}}$), we arrive at

$$
\mathbb{E}\left\{\left\|\widehat{f_n} - f\right\|_n^2\right\} \le \left\|f_{\boldsymbol{\theta}^*} - f\right\|_n^2 + C_{f,\mathscr{L}} \frac{p}{\kappa} \int_{\mathbb{R}^q} \left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2 p_0(d\boldsymbol{u}) + \frac{\beta\,\mathrm{KL}({p_0}^{\boldsymbol{D}}, \pi)}{n}. \tag{9.17}
$$

To complete the proof, it remains to bound the last two terms in the right hand side of (9.17). This is the goal of the following lemma.

Lemma 9.1. *Consider the same framework as the one in Theorem 5.1, we have*

$$
\int_{\mathbb{R}^q} \left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2 p_0(d\boldsymbol{u}) \le 2K_{1,g}^{\boldsymbol{D}} e^{r_L \alpha^a}, \tag{9.18}
$$

and

$$
\mathrm{KL}({p_0}^{\boldsymbol{D}}, \pi) \le 1 + r_L\alpha^a + \lambda\sum_{\ell=1}^L \ln\left\{h\left(\left\|\left[\boldsymbol{D}^\top\boldsymbol{\theta}^*\right]_{\mathcal{G}_\ell}\right\|_2\right)\right\} + \alpha^a\left\|\boldsymbol{D}^\top\boldsymbol{\theta}^*\right\|_{a,\mathcal{G}}^a. \tag{9.19}
$$

With $r_L = 3\sqrt{K_{a,g}^{\boldsymbol{D}} L}$, it follows from (9.17) and Lemma 9.1 that

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

According to (9.5), this completes the proof of Theorem 5.1. □

Proof of Lemma 9.1 To prove Lemma 9.1, we need an intermediate result.

Lemma 9.2. *Let $s > L\sqrt{K^{\boldsymbol{D}}_{a,g}}$. The following inequality holds*

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

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

Proof of Lemma 9.2. Let $\boldsymbol{U}$ be a random vector in $\mathbb{R}^p$ with density $\boldsymbol{u}\mapsto\frac{1}{T}\prod_{\ell=1}^L g(\left\|[\boldsymbol{D}^\top\boldsymbol{u}]_{\mathcal{G}_\ell}\right\|_2)$, where $T<\infty$ by Assumption **(G.2)**. By Chebyshev inequality, we have

$$\begin{aligned}
&\frac{1}{T}\int_{\{\boldsymbol{u}\in\mathbb{R}^p\,:\,\left\|\boldsymbol{D}^\top\boldsymbol{u}\right\|^a_{a,\mathcal{G}}>s\}}\prod_{\ell=1}^L g(\left\|[\boldsymbol{D}^\top\boldsymbol{u}]_{\mathcal{G}_\ell}\right\|_2)d\boldsymbol{u}\\
&=\Pr\left\{\textstyle\sum_{\ell=1}^L\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{\mathcal{G}_\ell}\right\|_2^a>s\right\}\\
&=\Pr\left\{\textstyle\sum_{\ell=1}^L\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{\mathcal{G}_\ell}\right\|_2^a-\mathbb{E}\left\{\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{\mathcal{G}_\ell}\right\|_2^a\right\}>s-\textstyle\sum_{\ell=1}^L\mathbb{E}\left\{\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{\mathcal{G}_\ell}\right\|_2^a\right\}\right\}\\
&\leq\mathbb{E}\left\{\left[\textstyle\sum_{\ell=1}^L\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{\mathcal{G}_\ell}\right\|_2^a-\mathbb{E}\left\{\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{\mathcal{G}_\ell}\right\|_2^a\right\}\right]^2\right\}/(s-\textstyle\sum_{\ell=1}^L\mathbb{E}\left\{\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{\mathcal{G}_\ell}\right\|_2^a\right\})^2\\
&=\operatorname{var}\left(\textstyle\sum_{\ell=1}^L\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{\mathcal{G}_\ell}\right\|_2^a\right)/(s-\textstyle\sum_{\ell=1}^L\mathbb{E}\left\{\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{\mathcal{G}_\ell}\right\|_2^a\right\})^2\\
&\leq\mathbb{E}\left\{\left(\textstyle\sum_{\ell=1}^L\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{\mathcal{G}_\ell}\right\|_2^a\right)^2\right\}/(s-\textstyle\sum_{\ell=1}^L\mathbb{E}\left\{\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{\mathcal{G}_\ell}\right\|_2^a\right\})^2.
\end{aligned}\tag{9.20}$$

Next, by Cauchy-Schwartz inequality and Remark 4.2, we obtain

$$\mathbb{E}\left\{\left(\textstyle\sum_{\ell=1}^L\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{\mathcal{G}_\ell}\right\|_2^a\right)^2\right\}\leq\mathbb{E}\left\{L\textstyle\sum_{\ell=1}^L\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{\mathcal{G}_\ell}\right\|_2^{2a}\right\}\leq L^2K^{\boldsymbol{D}}_{a,g}\tag{9.21}$$

and by Jensen inequality

$$s-\textstyle\sum_{\ell=1}^L\mathbb{E}\left\{\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{\mathcal{G}_\ell}\right\|_2^a\right\}\geq s-\textstyle\sum_{\ell=1}^L\sqrt{\mathbb{E}\left\{\left\|[\boldsymbol{D}^\top\boldsymbol{U}]_{\mathcal{G}_\ell}\right\|_2^{2a}\right\}}\geq s-L\sqrt{K^{\boldsymbol{D}}_{a,g}}>0.\tag{9.22}$$

Thus, combining (9.20), (9.21) and (9.22), we get

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

□

We now turn to the proof of Lemma 9.1

Proof of Lemma 9.1. Let us begin by the proof of inequality (9.18). We have

$$\begin{aligned}
\int_{\mathbb{R}^q}\left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2p_0(d\boldsymbol{u})&=\frac{1}{C_L\sqrt{\det(\boldsymbol{D}\boldsymbol{D}^\top)}}\int_{\mathbb{B}^a_{a,\mathcal{G}}(r_L)\cap\operatorname{Im}(\boldsymbol{D}^\top)}\left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2\prod_{\ell=1}^L e^{-\alpha^a\left\|\boldsymbol{u}_{\mathcal{G}_\ell}\right\|_2^a}g(\left\|\boldsymbol{u}_{\mathcal{G}_\ell}\right\|_2)d\boldsymbol{u}\\
&\leq\frac{1}{C_L\sqrt{\det(\boldsymbol{D}\boldsymbol{D}^\top)}}\int_{\mathbb{B}^a_{a,\mathcal{G}}(r_L)\cap\operatorname{Im}(\boldsymbol{D}^\top)}\left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2\prod_{\ell=1}^L g\left(\left\|\boldsymbol{u}_{\mathcal{G}_\ell}\right\|_2\right)d\boldsymbol{u}.
\end{aligned}\tag{9.23}$$

In the following, we show inequality (9.18) by bounding the right term of inequality (9.23). By Lemma 2.3 and Remark 4.2, we get

$$\frac{\int_{\mathbb{B}^a_{a,\mathcal{G}}(r_L)\cap \mathrm{Im}(\boldsymbol{D}^\top)} \left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2 \prod_{\ell=1}^L g(\left\|\boldsymbol{u}_{\mathcal{G}_\ell}\right\|_2) d\boldsymbol{u}}{\sqrt{\det(\boldsymbol{D}\boldsymbol{D}^\top)} \int_{\mathbb{R}^p} \prod_{\ell=1}^L g(\left\|[\boldsymbol{D}^\top \boldsymbol{u}]_{\mathcal{G}_\ell}\right\|_2) d\boldsymbol{u}} \leq \frac{\int_{\mathrm{Im}(\boldsymbol{D}^\top)} \left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2 \prod_{\ell=1}^L g(\left\|\boldsymbol{u}_{\mathcal{G}_\ell}\right\|_2) d\boldsymbol{u}}{\sqrt{\det(\boldsymbol{D}\boldsymbol{D}^\top)} \int_{\mathbb{R}^p} \prod_{\ell=1}^L g(\left\|[\boldsymbol{D}^\top \boldsymbol{u}]_{\mathcal{G}_\ell}\right\|_2) d\boldsymbol{u}}$$
$$= \frac{\int_{\mathbb{R}^p} \left\|[\boldsymbol{D}^\top \boldsymbol{u}]_{\mathcal{G}_1}\right\|_2^2 \prod_{\ell=1}^L g(\left\|[\boldsymbol{D}^\top \boldsymbol{u}]_{\mathcal{G}_\ell}\right\|_2) d\boldsymbol{u}}{\int_{\mathbb{R}^p} \prod_{\ell=1}^L g(\left\|[\boldsymbol{D}^\top \boldsymbol{u}]_{\mathcal{G}_\ell}\right\|_2) d\boldsymbol{u}} \leq K^{\boldsymbol{D}}_{1,g}.$$

Then

$$\frac{1}{\sqrt{\det(\boldsymbol{D}\boldsymbol{D}^\top)}} \int_{\mathbb{B}^a_{a,\mathcal{G}}(r_L)\cap \mathrm{Im}(\boldsymbol{D}^\top)} \left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2 \prod_{\ell=1}^L g(\left\|\boldsymbol{u}_{\mathcal{G}_\ell}\right\|_2) d\boldsymbol{u} \leq K^{\boldsymbol{D}}_{1,g} T. \tag{9.24}$$

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

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

Since $r_L = 3\sqrt{K^{\boldsymbol{D}}_{a,g}} L > \sqrt{K^{\boldsymbol{D}}_{a,g}} L$, Lemma 9.2 gives us

$$\begin{aligned}
{C_L}^{-1} &\leq e^{r_L \alpha^a} \left[T \left(1 - \frac{1}{T} \int_{\left\{\boldsymbol{u}\in\mathbb{R}^p \,:\, \left\|\boldsymbol{D}^\top \boldsymbol{u}\right\|_{a,\mathcal{G}}^a > r_L\right\}} \prod_{\ell=1}^L g(\left\|[\boldsymbol{D}^\top \boldsymbol{u}]_{\mathcal{G}_\ell}\right\|_2) d\boldsymbol{u} \right) \right]^{-1} \\
&\leq e^{r_L \alpha^a} T^{-1} \left(1 - \frac{L^2 K^{\boldsymbol{D}}_{a,g}}{(r_L - L\sqrt{K^{\boldsymbol{D}}_{a,g}})^2} \right)^{-1} \\
&= e^{r_L \alpha^a} T^{-1} \left(1 - \frac{1}{4} \right)^{-1} \leq 2 e^{r_L \alpha^a} T^{-1}.
\end{aligned} \tag{9.25}$$

Combining (9.24) and (9.25), (9.23) becomes $\int_{\mathbb{R}^q} \left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2 p_0(d\boldsymbol{u}) \leq 2K^{\boldsymbol{D}}_{1,g} e^{r_L \alpha^a}$. That concludes the proof of inequality (9.18) in Lemma 9.1.

Next, we prove inequality (9.19). Remind that $\mathrm{supp}(\pi) = \Theta$, $\mathrm{supp}({p_0}^{\boldsymbol{D}}) = \Theta_{{p_0}^{\boldsymbol{D}}}$. By (9.8), 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 (9.19)

can be proved as follows. By Lemma 2.3, we have

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

where p_0 is a probability measure in $\mathbb{R}^q$ defined in (9.10). We know that $\boldsymbol{t}^* = \boldsymbol{D}^\top\boldsymbol{\theta}^*$, according to the fact that $\left\|\boldsymbol{t}_{\mathcal{G}_\ell}\right\|_2^a - \left\|\boldsymbol{t}_{\mathcal{G}_\ell}-\boldsymbol{t}^*_{\mathcal{G}_\ell}\right\|_2^a \le \left\|\boldsymbol{t}^*_{\mathcal{G}_\ell}\right\|_2^a$ and Assumption **(G.4)**, we get

$$\mathrm{KL}({p_0}^{\boldsymbol{D}}, \pi) \le \ln\left(\frac{C_{\alpha,g,R}}{C_L}\right) + \alpha^a \left\|\boldsymbol{D}^\top\boldsymbol{\theta}^*\right\|_{a,\mathcal{G}}^a + \lambda \textstyle\sum_{\ell=1}^L \ln\left\{h\left(\left\|\left[\boldsymbol{D}^\top\boldsymbol{\theta}^*\right]_{\mathcal{G}_\ell}\right\|_2\right)\right\}. \tag{9.26}$$

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

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

Combining this with the bound of ${C_L}^{-1}$ in (9.25), we obtain

$$\ln\left(\frac{C_{\alpha,g,R}}{C_L}\right) \le r_L \alpha^a + \ln(2) \le 1 + r_L \alpha^a. \tag{9.27}$$

Inserting (9.27) into (9.26), we get inequality (9.19). This completes the proof. □

9.4.2. Proofs of corollaries

Proof of Corollary 5.1 Let $\gamma = 3$, $\nu = 2$ and $\eta = 1$. We have $\gamma/\nu < \eta + 1$ so that Lemma 2.1 applies. We thus obtain

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

From Lemma 4.1, 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

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

Therefore, g satisfies Assumptions **(G.1)**-**(G.4)** for $G = 1$. Owing to Remark 4.3 and Lemma 2.1, we obtain

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

We are now in position to apply Theorem 5.1 with $\boldsymbol{D} = \mathbf{I}_p$ (then $q = p$), $G = 1$ (then $L = q$), $a = 1$ and $\alpha \le 1/(3p\tau)$ to conclude. Namely, since $\tau^2 \sim (pn)^{-1}$ and $R \sim p\tau$, we get that $\Psi_{\mu_n,L,p}^{\boldsymbol{D}} \le 2eC_{f,\mathscr{L}}\tau^2 p \sim n^{-1}$, and

$$\Omega_{\mu_n,n,L,\lambda}^{\boldsymbol{D}}(\boldsymbol{\theta}) = \frac{4\beta}{n} \textstyle\sum_{j=1}^p \ln(1+|\theta_j|/\tau) \le \frac{4\beta}{n} \|\boldsymbol{\theta}\|_0 \ln(1+R/\tau) \sim \frac{\|\boldsymbol{\theta}\|_0 \ln(p)}{n}.$$

This completes the proof. □

Proof of Corollary 5.2 Let $\gamma = 2+G$, $\nu = b$ and $\eta = c-1$. We have $\gamma/\nu < \eta+1$ and thus Lemma 2.1 applies, whence we obtain

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

From Lemma 4.1, 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 $(\boldsymbol{t}, \boldsymbol{t}^*) \in \mathbb{R}^G \times \mathbb{R}^G$, 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.4)** with any $G \geq 1$. Applying Theorem 5.1, we conclude the proof. □

Proof of Corollary 5.3 Since g satisfies Assumptions **(G.2)**, **(G.3)** and $\boldsymbol{D}$ is invertible, by Remark 4.3 and Lemma 2.1, we get

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

Since $\tau^2 \sim (pn)^{-1}$ and $R \sim L\tau^a$, we get that $\Psi_{\mu_n,L,p}^{\boldsymbol{D}} \leq 2C_{f,\mathscr{L}} \widetilde{K}_{1,g}^{\boldsymbol{D}} e\kappa^{-1} p\tau^2 \sim n^{-1}$, and

$$\Omega_{\mu_n,n,L,\lambda}^{\boldsymbol{D}}(\boldsymbol{\theta}) = \frac{c\beta}{n} \sum_{\ell=1}^L \ln\left(1 + \left[\frac{\left\|[\boldsymbol{D}^\top \boldsymbol{\theta}]_{\mathcal{G}_\ell}\right\|_2}{\tau}\right]^b\right) \leq \frac{c\beta}{n} \left\|\boldsymbol{D}^\top \boldsymbol{\theta}\right\|_{0,\mathcal{G}} \ln\left(1 + \left[\frac{R^{1/a}}{\tau}\right]^b\right) \sim \frac{\left\|\boldsymbol{D}^\top \boldsymbol{\theta}\right\|_{0,\mathcal{G}} \ln(L)}{n}.$$

This ends the proof. □

9.5. Proofs of Section 6

9.5.1. Proof of Lemma 6.1

w_λ is clearly increasing, bounded from below by $w_\lambda(0)$, and continuously differentiable (in fact even C^∞) on $]0,+\infty[$.

(i) The expression of the proximal mapping follows from [40, Lemma 7.2] provided that it is single-valued. To show the latter statement, it is sufficient to prove that the function $u : x \in [0,+\infty[\mapsto x + \gamma w_\lambda'(x) = x + \gamma c/(\tau + x) + \gamma\alpha$ has a unique minimizer occurring at 0. One can see that u admits a local maximum at $\underline{x} = -\sqrt{\gamma c} - \tau \notin [0,+\infty[$ and a local minimum at $\overline{x} = \sqrt{\gamma c} - \tau$. Thus, the problem $\min_{x\in[0,+\infty[} u(x)$ has a unique solution at 0 when $\overline{x} \leq 0$ equivalent to $\gamma \leq \tau^2/c$.

(ii) For $\gamma \leq \tau^2/c$, it is immediate to see that the function u defined in (i) is nondecreasing on $[0,+\infty[$, and thus so is $\mathrm{prox}_{\gamma w_\lambda}$. It then follows that $\mathrm{prox}_{\gamma w_\lambda}(x) \geq \mathrm{prox}_{\gamma w_\lambda}(0) = 0$ for any $x \in [0,+\infty[$. Since w_λ is increasing, the second inequality follows.

(iii) See [40, Lemma 7.1].

(iv) Denote the function $x \in]0,+\infty[\mapsto v(x) = \frac{x}{r} + w_\lambda'(x)$ for $r > 0$. For $x, z \geq 0$, we have

$$(v(z) - v(x))(z-x) = (z-x)^2/(r) - c\frac{(z-x)^2}{(\tau+z)(\tau+x)} \geq (z-x)^2(r^{-1} - c\tau^{-2}) \geq 0, \quad \forall r \in]0, \tau^2/c].$$

This shows that v is nondecreasing, or equivalently, that $w_\lambda(x) + \frac{1}{2r}x^2$ is convex (i.e., w_λ is semi-convex). Thus

$$\frac{\|\boldsymbol{u}\|_2^2}{2r} + W_\lambda(\boldsymbol{u}) = \sum_{\ell=1}^L \left(\frac{\left\|\boldsymbol{u}_{\mathcal{G}_\ell}\right\|_2^2}{2r} + w_\lambda\left(\left\|\boldsymbol{u}_{\mathcal{G}_\ell}\right\|_2\right)\right) = \sum_{\ell=1}^L \left\{\left(\frac{(\cdot)^2}{2r} + w_\lambda\right) \circ \|\cdot\|_2 \left(\boldsymbol{u}_{\mathcal{G}_\ell}\right)\right\}.$$

Since $x \mapsto \frac{x^2}{2r} + w_\lambda(x)$ is convex and increasing, and the norm is also convex, we deduce that $\frac{\|\cdot\|_2^2}{2r} + W_\lambda$ is convex, and so is $\left(\frac{\|\cdot\|_2^2}{2r'} + W_\lambda\right) \circ \boldsymbol{D}^\top$ for $r' = r/\nu$ (recall ν from **(H.2)**). It then follows from [40, Lemma 5.3] that $\mathrm{prox}_{\gamma W_\lambda \circ \boldsymbol{D}^\top}$ is Lipschitz continuous for any $\gamma \in]0, r]$.

(v) [40, Lemma 3.2], which applies thanks to (i), yields the expression of $\mathrm{prox}_{\gamma W_\lambda}$. $\nabla(^\gamma W_\lambda \circ \boldsymbol{D}^\top)$ is Lipschitz continuous thanks to (iii), which, together with Lipschitz continuity of ∇F_β yields that of ∇V_γ. Let us turn to uniform boundedness of $\nabla(^\gamma W_\lambda \circ \boldsymbol{D}^\top)$. For any $\boldsymbol{\theta} \in \mathbb{R}^p$, denote $\boldsymbol{u} = \boldsymbol{D}^\top\boldsymbol{\theta}$. Thus, using (i) and (iii) and that $w_\lambda{}'$ is a decreasing function on $]0, +\infty[$, we have $\forall \boldsymbol{\theta} \in \mathbb{R}^p$,

$$\begin{aligned}\left\|\nabla(^\gamma W_\lambda \circ \boldsymbol{D}^\top)(\boldsymbol{\theta})\right\|_2^2 &\leq \gamma^{-2}\nu\textstyle\sum_{\ell=1}^L \left\{\left\|\boldsymbol{u}_{\mathcal{G}_\ell}\right\|_2 - \mathrm{prox}_{\gamma w_\lambda}\left(\left\|\boldsymbol{u}_{\mathcal{G}_\ell}\right\|_2\right)\right\}^2 \\ &\leq \gamma^{-2}\nu\textstyle\sum_{\ell=1}^L \begin{cases}\gamma w_\lambda{}'\left(\mathrm{prox}_{\gamma w_\lambda}\left(\left\|\boldsymbol{u}_{\mathcal{G}_\ell}\right\|_2\right)\right) & \text{if } \left\|\boldsymbol{u}_{\mathcal{G}_\ell}\right\|_2 > \gamma w_\lambda{}'(0^+) \\ \gamma w_\lambda{}'(0^+) & \text{otherwise}\end{cases} \\ &\leq \gamma^{-1}\nu L w_\lambda{}'(0^+) = \gamma^{-1}\nu L(\alpha + c/\tau).\end{aligned}$$

(vi) Combine (i), [40, Lemma 3.2 and Lemma 6.1], (iii) and that $\boldsymbol{M}_\gamma$ is positive definite.

□

9.5.2. Proof of Theorem 6.1

Uniform geometric ergodicity We start by showing that $\boldsymbol{L}(t)$ in (6.5) is H-uniformly geometrically ergodic, where $H : \mathbb{R}^p \to [1, +\infty[$ is a measurable function. Let $|f|_H \overset{\text{def}}{=} \sup_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{|f(\boldsymbol{\theta})|}{H(\boldsymbol{\theta})}$. Recall that H-uniform geometric ergodicity requires that for all $\boldsymbol{l}_0$

$$\left|\mathbb{E}\left\{f(\boldsymbol{L}(t))|\boldsymbol{L}_0 = \boldsymbol{l}_0\right\} - \int_{\mathbb{R}^p} f(\boldsymbol{\theta})\mu_\gamma(\boldsymbol{\theta})d\boldsymbol{\theta}\right| \leq KH(\boldsymbol{l}_0)\rho^t \tag{9.28}$$

for some $K < +\infty$ and $\rho \in [0, 1[$ and any function f satisfying $|f|_H \leq 1$.

Denote $\mathscr{A} \overset{\text{def}}{=} -\left\langle\nabla V_\gamma, \nabla\right\rangle + \Delta$ is the μ_γ symmetric natural operator. By [51, Theorem 2.2], $\boldsymbol{L}(t)$ is H-uniformly geometrically ergodic if H is a Lyapunov function for the generator $\mathscr{A}$, i.e. H is a C^2 function and there exist $\delta > 0$, $\vartheta \geq 0$ and some $R > 0$ such that for all $\boldsymbol{\theta} \in \mathbb{R}^p$

$$\mathscr{A}H(\boldsymbol{\theta}) \leq -\delta H(\boldsymbol{\theta}) + \vartheta I_{\mathbb{B}_R}(\boldsymbol{\theta}) \tag{9.29}$$

where $\mathbb{B}_R \overset{\text{def}}{=} \{\boldsymbol{\theta} \in \mathbb{R}^p \ : \ \|\boldsymbol{\theta}\|_2 \leq R\}$. We denote $\mathbb{B}_R^c \overset{\text{def}}{=} \mathbb{R}^p \setminus \mathbb{B}_R$. To prove geometric ergodicity, we use the following fact. Since F_β is differentiable, convex and e^{-F_β} is integrable, it follows from [2, Lemma 2.2] that there exist $\eta > 0$ and $r > 0$ such that for all $\boldsymbol{\theta} \in \mathbb{B}_r^c$,

$$\left\langle\nabla F_\beta(\boldsymbol{\theta}), \boldsymbol{\theta}\right\rangle \geq \eta\left\|\boldsymbol{\theta}\right\|_2. \tag{9.30}$$

Lemma 9.3. *The drift condition (9.29) is satisfied with the Lyapunov function* $H(\boldsymbol{\theta}) = \exp\left(\eta/4\sqrt{1 + \|\boldsymbol{\theta}\|_2^2}\right)$, $\delta = \eta^2/16$, $\vartheta = \eta/4H(R)\left(\eta/4 + d + 2\beta^{-1}\left|\!\left|\!\left|\boldsymbol{X}\right|\!\right|\!\right|^2R^2 + 2\beta^{-1}\left\|\boldsymbol{X}^\top\boldsymbol{y}\right\|_2 R\right) + \eta^2/16$ *and* $R = \max(r, 1 + 2p/\eta)$. *In turn,* $\boldsymbol{L}(t)$ *in (6.5) satisfies (9.28).*

Proof of Lemma 9.3. Elementary derivations give

$$\mathscr{A}H(\boldsymbol{\theta}) = -\eta/4\frac{H(\boldsymbol{\theta})}{\sqrt{1 + \|\boldsymbol{\theta}\|_2^2}}\left(\left\langle\nabla F_\beta(\boldsymbol{\theta}), \boldsymbol{\theta}\right\rangle + \left\langle\nabla(^\gamma W_\lambda)(\boldsymbol{D}^\top\boldsymbol{\theta}), \boldsymbol{D}^\top\boldsymbol{\theta}\right\rangle - \eta/4\frac{\|\boldsymbol{\theta}\|_2^2}{\sqrt{1 + \|\boldsymbol{\theta}\|_2^2}} - p + \frac{\|\boldsymbol{\theta}\|_2^2}{1 + \|\boldsymbol{\theta}\|_2^2}\right).$$

Denote $\boldsymbol{u} = \boldsymbol{D}^\top\boldsymbol{\theta}$. In view of Lemma 6.1(ii), (iii) and (v), we have

$$\langle\nabla(^\gamma W_\lambda)(\boldsymbol{u}), \boldsymbol{u}\rangle = \gamma^{-1}\left(\|\boldsymbol{u}\|_2^2 - \left\langle\mathrm{prox}_{\gamma W_\lambda}(\boldsymbol{u}), \boldsymbol{u}\right\rangle\right) = \gamma^{-1}\sum_{\ell=1}^L \left\|\boldsymbol{u}_{\mathcal{G}_\ell}\right\|_2\left\{\left\|\boldsymbol{u}_{\mathcal{G}_\ell}\right\|_2 - \mathrm{prox}_{\gamma w_\lambda}\left(\left\|\boldsymbol{u}_{\mathcal{G}_\ell}\right\|_2\right)\right\} \geq 0. \tag{9.31}$$

This, together with (9.30) yields the bound

$$\mathscr{A}H(\boldsymbol{\theta}) \leq -\eta/4H(\boldsymbol{\theta})\left(\frac{\eta R - p}{1 + R} - \eta/4\right)I_{\mathbb{B}_R^c}(\boldsymbol{\theta}) + \eta/4H(R)\left(\eta/4 + d + 2\beta^{-1}\left|\!\left|\!\left|\boldsymbol{X}\right|\!\right|\!\right|^2R^2 + 2\beta^{-1}\left\|\boldsymbol{X}^\top\boldsymbol{y}\right\|_2 R\right)I_{\mathbb{B}_R}(\boldsymbol{\theta}),$$

whence we deduce the desired claim. □

Let us now turn to the proof of the theorem.

(i) By the triangle inequality, we have

$$\left\|\mathbf{P}^{T,\delta,\gamma}_{\tilde{\boldsymbol{L}}_\delta}(\boldsymbol{l}_0,\cdot)-\mu_n\right\|_{\mathrm{TV}} \le \left\|\mathbf{P}^{T,\delta,\gamma}_{\tilde{\boldsymbol{L}}_\delta}(\boldsymbol{l}_0,\cdot)-\mathbf{P}^{T,\gamma}_{\boldsymbol{L}}(\boldsymbol{l}_0,\cdot)\right\|_{\mathrm{TV}} + \left\|\mathbf{P}^{T,\gamma}_{\boldsymbol{L}}(\boldsymbol{l}_0,\cdot)-\mu_\gamma\right\|_{\mathrm{TV}} + \left\|\mu_\gamma-\mu_n\right\|_{\mathrm{TV}}.$$

In view of Lemma 6.1(v) and [17, Lemma 2] and the Pinkser inequality, the first term in the right hand side goes to 0 as $\delta \to 0$. The second term converges to 0 as $T \to +\infty$ thanks to uniform geometric ergodicity (see Lemma 9.3), where we apply (9.28) with $\|f\|_\infty \le 1$. The last term vanishes as $\gamma \to 0$ by virtue of Proposition 6.1.

(ii) In the following, C is any positive constant that does not depend on δ and T. Let $\delta \in \,]0, 2\beta/\left\|\left|\boldsymbol{X}\right|\right\|^2[$. Since $\mathbb{E}\{\boldsymbol{Z}_k\} = 0$ and $\nabla({}^\gamma W_\lambda)\circ \boldsymbol{D}^\top$ is uniformly bounded, we have

$$\begin{aligned}\left\|\mathbb{E}\{\boldsymbol{L}_{k+1}\}\right\|_2 &\le \left\|\mathbb{E}\left\{\boldsymbol{L}_k - \delta/2\nabla F_\beta(\boldsymbol{L}_k)\right\}\right\|_2 + C\delta\\ &\le \mathbb{E}\left\{\left\|(\mathbf{I}_p-\delta/2\nabla F_\beta)(\boldsymbol{L}_k)-(\mathbf{I}_p-\delta/2\nabla F_\beta)(0)\right\|_2\right\} + (C+\left\|\nabla F_\beta(0)\right\|_2/2)\delta\\ &\le \mathbb{E}\{\boldsymbol{L}_k\} + (C+\left\|\boldsymbol{X}^\top\boldsymbol{y}\right\|_2/2)\delta \le (C+\left\|\boldsymbol{X}^\top\boldsymbol{y}\right\|_2/2)T, \quad \forall k[0,\lfloor T/\delta\rfloor],\end{aligned}$$

where we used the fact that F_β is a differentiable convex function whose gradient is $2\left\|\left|\boldsymbol{X}\right|\right\|^2/\beta$-Lipschitz, and thus, $\mathbf{I}_p - \delta\nabla F_\beta$ is non-expansive for the prescribed choice of δ [3]. In addition, by independence of $\boldsymbol{Z}_k$ from $\boldsymbol{L}_k$ and $\boldsymbol{Y}$, and in view of (9.31) and Lipschitz continuity of ∇F_β, we have

$$\begin{aligned}\mathbb{E}\left\{\left\|\boldsymbol{L}_{k+1}\right\|_2^2\right\} &\le \mathbb{E}\left\{\left\|\boldsymbol{L}_k-\delta/2\nabla V_\gamma(\boldsymbol{L}_k)\right\|_2^2\right\}+\delta p\\ &\le \mathbb{E}\left\{\left\|\boldsymbol{L}_k-\delta/2\nabla F_\beta(\boldsymbol{L}_k)+\delta/2\nabla F_\beta(0)\right\|_2^2 - \delta\left\langle \boldsymbol{L}_k-\delta/2(\nabla F_\beta(\boldsymbol{L}_k)-\nabla F_\beta(0)), \nabla({}^\gamma W_\lambda\circ\boldsymbol{D}^\top)(\boldsymbol{L}_k)\right\rangle\right\}\\ &\quad + C\delta^2+\delta p\\ &\le \mathbb{E}\left\{\left\|\boldsymbol{L}_k\right\|_2^2 - \delta\left\langle \boldsymbol{L}_k-\delta/2(\nabla F_\beta(\boldsymbol{L}_k)-\nabla F_\beta(0)), \nabla({}^\gamma W_\lambda\circ\boldsymbol{D}^\top)(\boldsymbol{L}_k)\right\rangle\right\}+C\delta^2+\delta p\\ &\le \mathbb{E}\left\{\left\|\boldsymbol{L}_k\right\|_2^2\right\} + C\delta^2\left\|\mathbb{E}\{\boldsymbol{L}_k\}\right\|_2 + C\delta^2+\delta p\\ &\le \mathbb{E}\left\{\left\|\boldsymbol{L}_k\right\|_2^2\right\} + C\delta^2\left\|\mathbb{E}\{\boldsymbol{L}_k\}\right\|_2 + C\delta^2+\delta p\\ &\le \mathbb{E}\left\{\left\|\boldsymbol{L}_k\right\|_2^2\right\} + C\delta^2 T + C\delta^2+\delta p \le C\delta T^2 + C\delta T + pT = O(T^2), \quad \forall k[0,\lfloor T/\delta\rfloor]. \end{aligned} \tag{9.32}$$

For $T > 0$ and $R > 0$, denote $\tilde{\boldsymbol{L}}^T_\delta \overset{\text{def}}{=} 1/T\int_0^T \tilde{\boldsymbol{L}}_\delta(t)dt$, $\tilde{\boldsymbol{L}}^{T,R}_\delta \overset{\text{def}}{=} 1/T\int_0^T \tilde{\boldsymbol{L}}_\delta(t)I_{\mathbb{B}_R}(\tilde{\boldsymbol{L}}_\delta(t))dt$, $\boldsymbol{L}^{T,R} \overset{\text{def}}{=} 1/T\int_0^T \boldsymbol{L}(t)I_{\mathbb{B}_R}(\boldsymbol{L}(t))dt$ and $\widehat{\boldsymbol{\theta}}^R_\gamma \overset{\text{def}}{=} \int_{\mathbb{B}_R}\boldsymbol{\theta}\mu_\gamma(\boldsymbol{\theta})d\boldsymbol{\theta}$. The triangle and Jensen inequalities yield

$$\begin{aligned}\mathbb{E}\left\{\left\|\bar{\boldsymbol{L}}_{\delta,T,\gamma}-\widehat{\boldsymbol{\theta}}_\gamma\right\|_2\right\} &\le \mathbb{E}\left\{\left\|\bar{\boldsymbol{L}}_{\delta,T,\gamma}-\tilde{\boldsymbol{L}}^T_\delta\right\|_2\right\} + \mathbb{E}\left\{\left\|\tilde{\boldsymbol{L}}^{T,R}_\delta-\boldsymbol{L}^{T,R}\right\|_2\right\} + \mathbb{E}\left\{\left\|\boldsymbol{L}^{T,R}-\widehat{\boldsymbol{\theta}}^R_\gamma\right\|_2\right\}\\ &\quad + \frac{1}{T}\int_0^T \mathbb{E}\left\{\left\|\tilde{\boldsymbol{L}}_\delta(t)\right\|_2 I_{\mathbb{B}^c_R(\tilde{\boldsymbol{L}}_\delta(t))}\right\}dt + \int_{\mathbb{B}^c_R}\left\|\boldsymbol{\theta}\right\|_2\mu_\gamma(\boldsymbol{\theta})d\boldsymbol{\theta}.\end{aligned}$$

For the last two terms, we have the bounds

$$\frac{1}{T}\int_0^T \mathbb{E}\left\{\left\|\tilde{\boldsymbol{L}}_\delta(t)\right\|_2 I_{\mathbb{B}^c_R(\tilde{\boldsymbol{L}}_\delta(t))}\right\}dt \le R^{-1}\frac{1}{T}\int_0^T \mathbb{E}\left\{\left\|\tilde{\boldsymbol{L}}_\delta(t)\right\|_2^2\right\}dt \underset{(9.32)}{\le} CT^2/R \quad \text{and}$$

$$\int_{\mathbb{B}^c_R}\left\|\boldsymbol{\theta}\right\|_2\mu_\gamma(\boldsymbol{\theta})d\boldsymbol{\theta} \le R^{-1}\int_{\mathbb{R}^p}\left\|\boldsymbol{\theta}\right\|_2^2\mu_\gamma(\boldsymbol{\theta})d\boldsymbol{\theta} \underset{(\mathbf{G.3})}{\le} C/R.$$

Choosing, e.g., $R = 1/T^3$, these terms converge to 0 as $T \to \infty$. Using (9.32) and arguing as in [23, Step 1, Proposition 2], we have

$$\mathbb{E}\left\{\left\|\bar{\boldsymbol{L}}_{\delta,T,\gamma}-\tilde{\boldsymbol{L}}^T_\delta\right\|_2\right\} \le C\delta(1+\delta T^2) \to 0 \text{ as } \delta \to 0.$$

Using Girsanov formula and Pinsker inequality, as in [23, Step 2, Proposition 2], the distribution of $\left\{\tilde{\boldsymbol{L}}_\delta(t)\right\}_{t\in[0,T]}$ converges to that of $\{\boldsymbol{L}(t)\}_{t\in[0,T]}$ in total variation as $\delta \to 0$. In turn,

$$\mathbb{E}\left\{\left\|\tilde{\boldsymbol{L}}_\delta^{T,R} - \boldsymbol{L}^{T,R}\right\|_2\right\} \to 0 \text{ as } \delta \to 0.$$

Using Lemma 9.3 with $f(\boldsymbol{\theta}) = \boldsymbol{\theta}_i I_{\mathbb{B}_R}(\boldsymbol{\theta})$, and arguing as in [23, Step 4, Proposition 2], we have

$$\mathbb{E}\left\{\left\|\boldsymbol{L}^{T,R} - \widehat{\boldsymbol{\theta}}_\gamma^R\right\|_2\right\} \leq CT^{-1/2} \to 0 \text{ as } T \to \infty.$$

This completes the proof.

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

References

[1] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Comput.*, 9(7):1545–1588, Oct. 1997.
[2] D. Bakry, F. Barthe, P. Cattiaux, and A. Guillin. A simple proof of the Poincaré inequality for a large class of probability measures. *Electronic Communications in Probability*, 13:60–66, 2008.
[3] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.
[4] G. Biau. Analysis of a random forests model. *J. Mach. Learn. Res.*, 13(1):1063–1095, Apr. 2012.
[5] 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.
[6] G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.*, 9:2015–2033, June 2008.
[7] L. Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, Aug. 1996.
[8] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, Oct. 2001.
[9] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.
[10] 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.
[11] E. J. Candès. Ridgelets: Estimating with Ridge Functions. *Annals of Statistics*, 31, 1999. 1561–1599.
[12] 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.
[13] 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.
[14] C. Chesneau, M. Fadili, and J.-L. Starck. Stein block thresholding for image denoising. *Applied and Computational Harmonic Analysis*, 28(1):67–88, 2010.
[15] R. Coifman and D. Donoho. Translation invariant de-noising. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*. Springer-Verlag, 1995. 125–150.
[16] D. Dai, P. Rigollet, and T. Zhang. Deviation optimal learning using greedy q-aggregation. *Ann. Statist.*, 40(3):1878–1905, 06 2012.
[17] A. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2016.
[18] A. Dalalyan and A. Tsybakov. Pac-bayesian bounds for the expected error of aggregation by exponential weights. Technical report, Université Paris 6, CREST and CERTIS, Ecole des Ponts ParisTech, 2009. personal communication.
[19] A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Mach. Learn.*, 72(1-2):39–61, Aug. 2008.
[20] A. S. Dalalyan and J. Salmon. Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.*, 40(4):2327–2355, 08 2012.
[21] 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.
[22] A. S. Dalalyan and A. B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 18(3):914–944, 08 2012.
[23] A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and langevin monte-carlo. *J. Comput. Syst. Sci.*, 78(5):1423–1443, Sept. 2012.
[24] 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.
[25] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.
[26] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, pages 1200–1224, 1995.
[27] 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.
[28] A. Durmus and E. Moulines. Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm. Preprint hal-01176132, July 2015.

[29] A. Durmus, E. Moulines, and M. Pereyra. Sampling from convex non continuously differentiable functions, when Moreau meets Langevin. Preprint hal-01267115, Feb. 2016.
[30] K. Fang, S. Kotz, and K. Ng. *Symmetric multivariate and related distributions*. Monographs on statistics and applied probability. Chapman and Hall, 1990.
[31] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256 – 285, 1995.
[32] R. Genuer. *Random Forests: elements of theory, variable selection and applications*. Theses, Université Paris Sud - Paris XI, Nov. 2010.
[33] I. S. Gradshteyn and I. M. Ryshik. *Table of Integrals, Series, and Products*. Academic Press, London, 4th edition, 1965.
[34] J. Huang, J. L. Horowitz, and F. Wei. Variable selection in nonparametric additive models. *Ann. Statist.*, 38(4):2282–2313, 08 2010.
[35] J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Inf. Comput.*, 132(1):1–63, Jan. 1997.
[36] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines, 2008.
[37] G. Lecué. Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.*, 35(4):1698–1721, 08 2007.
[38] G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006.
[39] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Inf. Comput.*, 108(2):212–261, Feb. 1994.
[40] T. D. Luu, J. Fadili, and C. Chesneau. Sampling from non-smooth distribution through Langevin diffusion. Technical report, hal-01492056, May 2017.
[41] J. Maly. Lectures on change of variables in integral. Preprint 305, Department of Mathematics, University of Helsinki, 2001.
[42] L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 2008.
[43] L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *Ann. Statist.*, 37(6B):3779–3821, 12 2009.
[44] P. Müller and B. Vidakovic, editors. *Bayesian Inference in Wavelet-Based Models*, volume 141 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1999.
[45] A. Nemirovski. Topics in non-parametric statistics, 2000.
[46] M. Pereyra. Proximal markov chain monte carlo algorithms. *Statistics and Computing*, 26(4):745–760, 2016.
[47] G. Peyré, J. Fadili, and C. Chesneau. Group sparsity with overlapping partition functions. In *EUSIPCO*, Barcelona, Spain, Aug. 2011.
[48] 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.
[49] P. Rigollet and A. Tsybakov. Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, 16(3):260–280, 2007.
[50] P. Rigollet and A. B. Tsybakov. Sparse estimation by exponential weighting. *Statist. Sci.*, 27(4):558–575, 11 2012.
[51] G. O. Roberts and R. L. Tweedie. Exponential Convergence of Langevin Distributions and Their Discrete Approximations. *Bernoulli*, 2(4):341–363, 1996.
[52] R. T. Rockafellar and R. Wets. *Variational analysis*, volume 317. Springer Verlag, 1998.
[53] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, Nov. 1992.
[54] R. E. Schapire. The strength of weak learnability. *Mach. Learn.*, 5(2):197–227, July 1990.
[55] 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*, pages 91–108, 2005.
[56] 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.
[57] M. Xuerong. *Stochastic differential equations and applications*. Woodhead Publishing, 2007.
[58] Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10(1):25–47, 2004.
[59] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.