



HAL
open science

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

Tung Duy Luu, Jalal M. Fadili, Christophe Chesneau

► **To cite this version:**

Tung Duy Luu, Jalal M. Fadili, Christophe Chesneau. PAC-Bayesian risk bounds for group-analysis sparse regression by exponential weighting. 2017. hal-01367742v3

HAL Id: hal-01367742

<https://hal.science/hal-01367742v3>

Preprint submitted on 10 Aug 2017 (v3), last revised 22 May 2018 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Tung Duy Luu*

Jalal Fadili*

Christophe Chesneau[†]

Abstract

In this paper, we consider a high-dimensional non-parametric regression model with fixed design and i.i.d. random errors. We propose a powerful 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 guarantees. In turn, this allows us to implement our estimator and validate it on some numerical experiments.

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

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) , $i = 1, \dots, n$, is generated according to the non-parametric regression model

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

where x_1, \dots, x_n are deterministic in an arbitrary set \mathcal{X} , $f : \mathcal{X} \rightarrow \mathbb{R}$ is the unknown regression function and (ξ_1, \dots, ξ_n) are random errors. (1.1) is equivalently written in vector form $\mathbf{Y} = \mathbf{f} + \boldsymbol{\xi}$. Assume that there exists a dictionary $\mathcal{H} = \{f_j : \mathcal{X} \rightarrow \mathbb{R}, j \in \{1, \dots, p\}\}$ such that f is well approximated by a linear combination of elements in \mathcal{H} , i.e., there exists $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \dots, \tilde{\theta}_p)^T \in \mathbb{R}^p$ such that $f_{\tilde{\boldsymbol{\theta}}} = \sum_{j=1}^p \tilde{\theta}_j f_j$ 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 $\tilde{\boldsymbol{\theta}}$ is obviously not applicable.

The idea of aggregating elements in a dictionary has been introduced in machine learning to combine different techniques (see [36, 52]) with some procedures such as bagging [5], boosting [28, 50] and random forests [1–4, 6, 29]. In the recent years, there has been a flurry of research on the use of the concept of

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

[†]Normandie Univ, UNICAEN, CNRS, LMNO, France, Email: christophe.chesneau@unicaen.fr.

sparsity in various areas 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 $\tilde{\theta}$ is assumed to be much smaller than n . This makes it possible to build an estimate $f_{\tilde{\theta}}$ 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 [22–24] on orthogonal wavelet thresholding estimators. Oracle inequalities (according to the terminology introduced in e.g. [23]), 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 $g : \mathcal{X} \rightarrow \mathbb{R}$, and denote

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

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

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

We aim to find an estimator \hat{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:

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

where $C \geq 1$ and the remainder term $\Delta_{n,p}(\theta)$ depends on the performance of the estimator, the complexity of θ , the dimension p and the sample size n . 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 θ 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”), $\Delta_{n,p}(\theta)$ is small even if $n \ll p$ and decreases rapidly to 0 as $n \rightarrow +\infty$. 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 $Y = \mathbf{y}$, The penalization approach considers the minimization problem

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

where $\text{pen} : \mathbb{R}^p \rightarrow \mathbb{R}^+$ is a sparsity promoting penalty function, see e.g. [7]. Our work focuses on EWA approach that we briefly describe now.

1.2.2 Exponential Weighted Aggregation (EWA)

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

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

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

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

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

$$\hat{f}_n(x) = \int_\Lambda f_\lambda(x)\mu_n(d\lambda). \quad (1.4)$$

This idea was initially proposed in [32, 36, 52] with a uniform prior on a finite set Λ .

In the literature, the works in [15, 16, 18, 20] consider deterministic dictionaries. These papers proposed several PAC-Bayesian type of oracle inequalities under different assumptions. Especially, the assumptions in [20] 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 $O(\|\boldsymbol{\theta}\|_0 \log(p)/n)$, which scales linearly with the sparsity level and increases in p only logarithmically.

The random dictionary case are tackled in [41]. The initial idea is to obtain two independent samples from the initial sample by randomization or sample splitting (see [34, 45, 54]). 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 [35]. More recent work such as [17] generalizes the idea where the pre-estimators are affine and the priors are continuous.

A shortcoming of EWA is the suboptimality in deviation. Namely, in some context, the quality of estimators is measured by the fluctuations in $\|\hat{f} - f\|_n^2$ under a large probability instead of summarized by the expectation $\mathbb{E}\left[\|\hat{f} - f\|_n^2\right]$. The works in [14, Section 2] show that the EWA estimators lead a suboptimal remainder term for such type of inequalities.

1.2.3 Generalization of sparsity assumption

Analysis sparsity Let $q \geq p$ and $\mathbf{D}^T \in \mathbb{R}^{q \times p}$ be a linear analysis operator. The analysis sparsity assumption means that $\|\mathbf{D}^T \tilde{\boldsymbol{\theta}}\|_0 \ll n$. A typical example is total variation [49] where the operator \mathbf{D}^T corresponds to “finite differences” (i.e., $(\mathbf{D}^T \boldsymbol{\theta})_1 = \theta_1$, $(\mathbf{D}^T \boldsymbol{\theta})_j = \theta_j - \theta_{j-1}$, $\forall j \geq 2$). Another example is the fused Lasso [51] where \mathbf{D}^T is a positive combination of the identity and finite differences.

Group sparsity Group sparsity corresponds to saying that the aggregator θ is block sparse, see Section 4.1 for formal details. Group sparsity is at the heart of the group Lasso and related methods [31, 33, 39, 40, 44, 55]. In the EWA context, the group sparsity prior is considered in [46] 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 D is a frame, and thus is not necessarily invertible unlike previous work. In a finite space, that equivalents to the fact that DD^T is invertible. In addition, our prior class (see (4.2)) is much more general than previously proposed ones [20] 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 $g : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ 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 $D^T\theta$ 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 $O\left(\|D^T\theta\|_{0,g} \log(L)/n\right)$, where $\|D^T\theta\|_{0,g}$ is the number of active groups in $D^T\theta$, and L is the total number of groups (see Corollary 5.2). This rate coincides with the classical one $O(\|\theta\|_0 \log(p)/n)$ under the sparsity scenario, i.e. $D = I_p$ and $L = p$.
- We also propose a forward-backward proximal Langevin Monte-Carlo (LMC) algorithm to sample from the target distribution (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 [20] 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 $\|\cdot\|_2$. I_d is the identity matrix on \mathbb{R}^d . For any $k > 0$ and $\mathbf{x} \in \mathbb{R}^d$, we also define the ℓ_k function $\|\mathbf{x}\|_k = \left(\sum_{j=1}^d |x_j|^k\right)^{1/k}$ with the usual adaptation $\|\mathbf{x}\|_\infty = \max_{j \in \{1, \dots, d\}} |x_j|$. It is a norm for $k \geq 1$, and a quasi-norm for $k \in]0, 1[$. $\|\mathbf{x}\|_0$ is the ℓ_0 pseudo-norm which counts the number of non-zero elements in \mathbf{x} . Let $M \in \mathbb{R}^{d \times d}$ symmetric positive definite, we denote $\langle \cdot, \cdot \rangle_M = \langle \cdot, M \cdot \rangle$ and $\|\cdot\|_M$ its associated norm. Of course, $\|\cdot\|_M$ and $\|\cdot\|_2$ are equivalent.

For a matrix $M \in \mathbb{R}^{d \times r}$, we set $\sigma(M) = (\sigma_1(M), \dots, \sigma_r(M))^T \in \mathbb{R}^r$ be the vector of singular values of M in non-increasing order. Note that, when M is symmetric semi-definite positive, $\sigma(M)$ is also the ordered vector of positive eigenvalues of M . We denote $\|M\|$ the spectral norm of M .

For $\mathcal{I} \subset \{1, \dots, d\}$, $\mathbf{x} \in \mathbb{R}^d$ and $M \in \mathbb{R}^{d \times d}$, $\mathbf{x}_{\mathcal{I}}$ is the subvector whose entries are those of \mathbf{x} restricted to the indices in \mathcal{I} , and $M_{\mathcal{I}}$ the submatrix whose rows and columns are those of M indexed by \mathcal{I} .

Let us denote $\mathbb{R}^{d \times d} \rightarrow \mathbb{R}^{d^2}$ the vectorization operator of a matrix. For any matrix M , M^T denotes its transpose. For any square matrix M , $\det(M)$ denotes its determinant.

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

By the Courant-Fischer theorem, Definition 2.1 is equivalent to the fact that μ (resp. ν) is a lower (resp. upper) bound of the eigenvalues of MM^T . Moreover, since $\mu > 0$, we have that MM^T is bijective and M^T is injective. The frame is said tight when $\mu = \nu$. Typical examples of (tight) frames that have been used in statistics are translation invariant wavelets [13], ridgelets [9] and curvelets [8] (example of groups and what they represent for wavelets/ridgelets/curvelets in applications are discussed in [12]). Let $\tilde{M} \in \mathbb{R}^{d \times r}$ be the canonical dual frame associated to M , i.e. $\tilde{M} = (MM^T)^{-1}M$. We know that

$$\tilde{M}M^T = I_d \quad (2.1)$$

and

$$\frac{1}{\mu} \geq \sigma_1(\tilde{M}^T \tilde{M}) \geq \dots \geq \sigma_d(\tilde{M}^T \tilde{M}) \geq \frac{1}{\nu}. \quad (2.2)$$

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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we define $\mathbf{f} = (f(x_1), \dots, f(x_d))^T$ its d -sample vector form, $\|f\|_d = \left(\frac{1}{d} \sum_{j=1}^d h^2(x_j)\right)^{1/2}$ and $\|f\|_\infty = \sup_{\mathbf{x} \in \mathbb{R}^d} |f(\mathbf{x})|$.

For a function $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$, its effective domain is $\text{dom}(f) = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) < +\infty\}$ and f is proper if $f(\mathbf{x}) > -\infty$ for all \mathbf{x} and $\text{dom}(f) \neq \emptyset$ as is the case when it is finite-valued. A function $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{-\infty, +\infty\}$ is lower semi continuous (lsc) at \mathbf{x}_0 if $\liminf_{\mathbf{x} \rightarrow \mathbf{x}_0} f(\mathbf{x}) \geq f(\mathbf{x}_0)$. 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 $\text{sgn} : \mathbb{R} \mapsto \{-1, 1\}$ be the sign function, i.e. $\text{sgn}(x) = 1$ when $x \geq 0$, and $\text{sgn}(x) = -1$ otherwise.

Let $x \in \mathbb{R}$, define $\lfloor x \rfloor$ be the stands of integer part of x .

Recall the Gamma function $\Gamma :]0, +\infty[\rightarrow]0, +\infty[$, $\Gamma : t \mapsto \int_0^\infty x^{t-1} \exp(-x) dx$.

Definition 2.2 (Proximal mapping and Moreau envelope). *Let $\mathbf{M} \in \mathbb{R}^{d \times d}$ be symmetric positive definite. For a proper lsc function f and $\gamma > 0$, the proximal mapping and Moreau envelope in the metric \mathbf{M} are defined respectively by*

$$\begin{aligned} \text{prox}_{\gamma f}^{\mathbf{M}}(\mathbf{x}) &\stackrel{\text{def}}{=} \underset{\mathbf{w} \in \mathbb{R}^d}{\text{Argmin}} \frac{1}{2\gamma} \|\mathbf{w} - \mathbf{x}\|_{\mathbf{M}}^2 + f(\mathbf{w}) \\ \mathbf{M}, \gamma f(\mathbf{x}) &\stackrel{\text{def}}{=} \inf_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2\gamma} \|\mathbf{w} - \mathbf{x}\|_{\mathbf{M}}^2 + f(\mathbf{w}) \end{aligned}$$

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

Some useful tools The following lemmas contain useful formula used throughout the paper.

Lemma 2.1 ([30, 3.251.11]). *Let $p, \gamma, \nu, \eta > 0$. If $\gamma/\nu < \eta + 1$ we have*

$$\int_0^\infty \frac{x^{\gamma-1}}{(p+x^\nu)^{\eta+1}} dx = \frac{1}{\nu p^{\eta+1-\gamma/\nu}} \frac{\Gamma(\gamma/\nu)\Gamma(1+\eta-\gamma/\nu)}{\Gamma(1+\eta)}, \quad (2.3)$$

otherwise this integral is not definite.

Lemma 2.2 (Cartesian to spherical coordinates [27]). *Let $d \geq 1$ and a mapping $h : \mathbb{R}^+ \rightarrow \mathbb{R}$ such that $\mathbf{u} \rightarrow h(\|\mathbf{u}\|_2)$ is measurable in \mathbb{R}^d . We obtain*

$$\int_{\mathbb{R}^d} h(\|\mathbf{u}\|_2) d\mathbf{u} = C_d \int_0^\infty x^{d-1} h(x) dx, \quad (2.4)$$

where $C_d = 2\pi^{d/2}/\Gamma(d/2)$ is the surface area of a d -dimensional 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 $\mathbf{M} \in \mathbb{R}^{r \times d}$ corresponding to the analysis operator of a finite frame of \mathbb{R}^d , let $u : \mathbb{R}^r \rightarrow \mathbb{R}$ such that the mapping $\mathbf{x} \mapsto u(\mathbf{M}\mathbf{x})$ is measurable on Θ . We have*

$$\int_{\Theta} u(\mathbf{M}\mathbf{x}) d\mathbf{x} = \frac{1}{\sqrt{\det(\mathbf{M}\mathbf{M}^T)}} \int_{\mathbf{M}\Theta} u(\mathbf{v}) d\mathbf{v} \quad (2.5)$$

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.

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 [20] 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 $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^T$ has zero mean.

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

- (a) $\boldsymbol{\xi}'$ has the same distribution as $\boldsymbol{\xi}$.
- (b) $\boldsymbol{\xi}' + \zeta$ has the same distribution as $(1+\gamma)\boldsymbol{\xi}'$ and the conditional expectation satisfies $\mathbb{E}[\zeta|\boldsymbol{\xi}'] = 0$.
- (c) There exist $t_0 \in]0, \infty]$ and a bounded Borel function $v : \mathbb{R}^p \rightarrow \mathbb{R}^+$ such that

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

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

Assumption (P.2) is based on [20, 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 noise having variance $\boldsymbol{\Sigma}$, with $t_0 = \infty$ and $v(\mathbf{a}) \equiv \boldsymbol{\Sigma}$,
- $\boldsymbol{\xi}$ is a Laplace noise having variance $\boldsymbol{\Sigma}$, with $t_0 < \sqrt{2/\|\boldsymbol{\Sigma}\|}$ and $v(\mathbf{a}) \equiv \frac{\|\boldsymbol{\Sigma}\|}{1-t_0^2\|\boldsymbol{\Sigma}\|^2/2}$,
- $\boldsymbol{\xi}$ is a bounded symmetric noise, i.e. $\mathbb{P}[\|\xi_i\| \leq B_i]$ for some $\mathbf{B} \in \mathbb{R}^n$, with $t_0 = \infty$ and $v(\mathbf{a}) = \|\mathbf{a}\|_2 \leq \|\mathbf{B}\|_2$.

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

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

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*

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

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

The proof of Theorem 3.1 is a mild adaptation of the original one in [20, Section 2], where we used directly Assumption (P.2)-3 in the vector ζ instead of splitting it into $\zeta_{i, i \in \{1, \dots, n\}}$ (that are no longer i.i.d.).

Related work The work of [16] proposed three types of oracle inequalities which are similar to (3.2) under different assumptions. The first type (see [16, Theorem 1]) holds under a restrictive condition on the noise. The second (see [16, Theorem 2]) involves conditions depending on the noise and also on the dictionary. The last (see [16, 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 [17]:

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

where \hat{f}_{λ} , $\lambda \in \Lambda$ are affine estimators satisfying some conditions imposed in [17, Theorem 1] which yield the definition of \hat{r}_{λ} , $\lambda \in \Lambda$. 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 [19] 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 $\{\mathcal{G}_l\}_{1 \leq l \leq L}$ such that

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

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

$$\|\mathbf{v}\|_{k, \mathcal{G}} = \left(\sum_{l=1}^L \|\mathbf{v}_{\mathcal{G}_l}\|_2^k \right)^{1/k}, \quad k > 0,$$

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

(H.1) Let $\mathbf{D} \in \mathbb{R}^{p \times q}$, there exists $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^p$ such that $f_{\tilde{\boldsymbol{\theta}}}$ is close to f and $\|\mathbf{D}^T \tilde{\boldsymbol{\theta}}\|_{0, \mathcal{G}} \ll n$.

In plain words, Assumption (H.1) says that the number of active groups of $\mathbf{D}^T \tilde{\boldsymbol{\theta}}$ 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 $(\mathbf{D}^T \tilde{\boldsymbol{\theta}})_{\mathcal{G}_l}$ are nearly zero. We also impose the following assumption on \mathbf{D} .

(H.2) \mathbf{D} is a frame.

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

Example 4.1 (2-D piecewise constant image). Let $\theta_0 \in \mathbb{R}^{\sqrt{p} \times \sqrt{p}}$ be a 2-D piecewise constant image. In this framework, a popular analysis operator is the isotropic total variation called D_{TV} (see [49]). Namely, let $D_c : \mathbb{R}^{\sqrt{p} \times \sqrt{p}} \rightarrow \mathbb{R}^{\sqrt{p} \times \sqrt{p}}$ and $D_r : \mathbb{R}^{\sqrt{p} \times \sqrt{p}} \rightarrow \mathbb{R}^{\sqrt{p} \times \sqrt{p}}$ the finite difference operators along, respectively, the columns and rows of an image, with Neumann boundary conditions. We define D_{TV} as

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

By vectorizing $\tilde{\theta}$ in Assumption (H.1), D^T can be considered as the linearized version of D_{TV} , called \mathbf{D}_{TV} . Here, $\mathbf{D}_{\text{TV}} = [\mathbf{D}_r^T \mathbf{D}_c^T]^T$ where $\mathbf{D}_r \in \mathbb{R}^{p \times p}$ (resp. $\mathbf{D}_c \in \mathbb{R}^{p \times p}$) is the linearized counterpart of D_r (resp. D_c). With Neumann boundary conditions, \mathbf{D}_r and \mathbf{D}_c are bijective implying injectivity of \mathbf{D}_{TV} . Thus, \mathbf{D} is a frame in view of Courant-Fisher theorem.

The isotropic total variation prior on θ promotes sparsity of $\text{vec} \left(\left\| [D_{\text{TV}}(\theta)]_{i,j} \right\|_2 \right)_{1 \leq i,j \leq \sqrt{p}}$. 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 $D_{\text{TV}}(\theta_0)$ with groups of size 2 along the third dimension.

Example 4.2 (Signal with overlapping groups). Consider $\theta_0 \in \mathbb{R}^p$ generated from L groups which overlap. The analysis operator acts as a group extractor (see [11, 43]). In this framework, $\mathbf{D}\mathbf{D}^T = \text{diag}(\mathbf{B}_1^T \mathbf{B}_1, \dots, \mathbf{B}_L^T \mathbf{B}_L)$ where $\mathbf{B}_l \in \mathbb{R}^{q_l \times p}$, for $l \in \{1, \dots, L\}$, is a countable collection of localization operators, and then $q = \sum_{l=1}^L q_l$. Since the localization operators are injective, $\mathbf{D}\mathbf{D}^T$ is bijective. Thus \mathbf{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 $\mathbf{X} \in \mathbb{R}^{n \times p}$ where $\mathbf{X}_{i,j} = f_j(x_i)$, $f_j \in \mathcal{H}$. We impose the following standard normalization assumption on \mathbf{X} .

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

Now, let us introduce our dictionary of aggregation:

$$\mathcal{F}_\Theta = \left\{ f_\theta = \ell \left(\sum_{j=1}^p \theta_j f_j \right) : \theta \in \Theta = \left\{ \theta \in \mathbb{R}^p : \left\| \mathbf{D}^T \theta \right\|_{a, \mathcal{G}}^a \leq R \right\} \right\}, \quad (4.1)$$

where $a \in]0, 1]$, $R \in]0, +\infty]$ and $\ell : \mathbb{R} \rightarrow \mathbb{R}$ is twice continuously differentiable and known depending on the regression problem (for example: $\ell(x) = e^x$ for the exponential regression, $\ell(x) = e^x / (e^x + 1)$ for the logistic regression and $\ell(x) = x$ for the linear regression). This dictionary of aggregation is similar to the one proposed in [18–20]. However, the set of index is modified to adapt 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}_Θ in (4.1) is established in the following result.

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

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

where $\mathcal{B} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 \leq \frac{\|\mathbf{X}\|_R^{1/a}}{\sqrt{\mu}}\}$ and $\mathbf{L} : \mathbf{x} \in \mathbb{R}^n \rightarrow (\ell(x_1), \dots, \ell(x_n))$.

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

Remark 4.1. By choosing $H = 2 \max_{\mathbf{x} \in \mathcal{B}} \|\mathbf{L}(\mathbf{x})\|_2$ with $\mathcal{B} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 \leq \frac{\|\mathbf{X}\|_R^{1/a}}{\sqrt{\mu}}\}$, H depends on \mathbf{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, ξ must satisfy Assumption (P.2) with $t_0 = \infty$. In view of Proposition 3.1, we can consider ξ as a Gaussian or a bounded symmetric noise.

4.3 Choice of prior

We choose a general prior of the form

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

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

- (G.1) $g : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is a bounded function such that $g \not\equiv 0$, $\theta \mapsto g(\|[D^T \theta]_{\mathcal{G}_l}\|_2)$ is measurable on \mathbb{R}^p , $\forall l \in \{1, \dots, L\}$.
- (G.2) The integrability condition: $\int_{\mathbb{R}^p} \prod_{l=1}^L g(\|[D^T \mathbf{u}]_{\mathcal{G}_l}\|_2) d\mathbf{u} < \infty$.
- (G.3) The moment condition: $\int_{\mathbb{R}^p} \|[D^T \mathbf{u}]_{\mathcal{G}_l}\|_2^2 \prod_{k=1}^L g(\|[D^T \mathbf{u}]_{\mathcal{G}_k}\|_2) d\mathbf{u} < \infty$, $\forall l \in \{1, \dots, L\}$.
- (G.4) There exist $\lambda \geq 0$ and $h : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ such that $\frac{g(\|\mathbf{t} - \mathbf{t}^*\|_2)}{g(\|\mathbf{t}\|_2)} \leq h(\|\mathbf{t}^*\|_2)^\lambda$, $\forall (\mathbf{t}, \mathbf{t}^*) \in \mathbb{R}^G \times \mathbb{R}^G$.

Assumptions (G.2) and (G.3) lead the following remark that is a core part for the construction of the general group-analysis SOI in Theorem 5.1.

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

$$\frac{\int_{\mathbb{R}^p} \|[D^T \mathbf{u}]_{\mathcal{G}_l}\|_2^{2a} \prod_{k=1}^L g(\|[D^T \mathbf{u}]_{\mathcal{G}_k}\|_2) d\mathbf{u}}{\int_{\mathbb{R}^p} \prod_{k=1}^L g(\|[D^T \mathbf{v}]_{\mathcal{G}_k}\|_2) d\mathbf{v}} \leq K_{a, g}^{\mathbf{D}}. \quad (4.3)$$

Proof. Remark 4.2 is briefly proved as follows. 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), we have

$$\mathbf{u} \rightarrow \frac{\prod_{l=1}^L g(\|[D^T \mathbf{u}]_{\mathcal{G}_l}\|_2) d\mathbf{u}}{\int_{\mathbb{R}^p} \prod_{k=1}^L g(\|[D^T \mathbf{v}]_{\mathcal{G}_k}\|_2) d\mathbf{v}}$$

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

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

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

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

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

In the two following remarks, we consider the case where \mathbf{D} is invertible. Remark 4.3 provides a simple and explicit form for $K_{a,g}^{\mathbf{D}}$ and Remark 4.4 shows that condition (4.4) is necessary for g to obey Assumption **(G.3)**.

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

$$K_{a,g}^{\mathbf{D}} = \frac{\int_0^\infty x^{G-1+2a} g(x) dx}{\int_0^\infty z^{G-1} g(z) dz}. \quad (4.5)$$

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

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

□

Remark 4.4. *When \mathbf{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

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

□

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

Example 4.3. *Consider $g : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ 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 [20] for the individual sparsity scenario, i.e. with $\mathbf{D} = \mathbf{I}_p$, $G = 1$ and $a = 1$.

Example 4.4. Consider $g : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ 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 $G \geq 1$, \mathbf{X} satisfying Assumption **(H.3)**, and \mathbf{D} satisfying Assumption **(H.2)** with $\mu > 0$. 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 $K_{a,g}^{\mathbf{D}}$, as defined in (4.3), and assume that $R > 3\sqrt{K_{a,g}^{\mathbf{D}}}L$. Then the following group-analysis SOI holds,*

$$\mathbb{E} \left[\|\widehat{f}_n - f\|_n^2 \right] \leq \inf_{\Theta_{\mu_n, L, R}^{\mathbf{D}}} \left(\|f_{\boldsymbol{\theta}} - f\|_n^2 + \Phi_{\mu_n, n, L}^{\mathbf{D}}(\boldsymbol{\theta}) + \Omega_{\mu_n, n, L, \lambda}^{\mathbf{D}}(\boldsymbol{\theta}) \right) + \Psi_{\mu_n, L, p}^{\mathbf{D}}, \quad (5.1)$$

$$\begin{aligned} \text{with } \Theta_{\mu_n, L, R}^{\mathbf{D}} &= \left\{ \boldsymbol{\theta} \in \mathbb{R}^p : \|\mathbf{D}^T \boldsymbol{\theta}\|_{a, g}^a \leq R - 3\sqrt{K_{a, g}^{\mathbf{D}}}L \right\}, & \Phi_{\mu_n, n, L}^{\mathbf{D}}(\boldsymbol{\theta}) &= \frac{\beta}{n} \left(1 + 3\sqrt{K_{a, g}^{\mathbf{D}}}L\alpha^a + \alpha^a \|\mathbf{D}^T \boldsymbol{\theta}\|_{a, g}^a \right), \\ \Omega_{\mu_n, n, L, \lambda}^{\mathbf{D}}(\boldsymbol{\theta}) &= \frac{\lambda\beta}{n} \sum_{l=1}^L \log h \left(\|\mathbf{D}^T \boldsymbol{\theta}\|_{g_l} \right), & \Psi_{\mu_n, L, p}^{\mathbf{D}} &= \frac{2K_{1, g}^{\mathbf{D}} \exp(3\sqrt{K_{a, g}^{\mathbf{D}}}L\alpha^a) p C_{f, \ell}}{\mu}, \\ \text{and } C_{f, \ell} &= \|\ell'\|_\infty^2 + \|\ell''\|_\infty (\|\ell\|_\infty + \|f\|_\infty). \end{aligned}$$

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

1. *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 $\left\{ \boldsymbol{\theta} \in \mathbb{R}^p : \|\mathbf{D}^T \boldsymbol{\theta}\|_{a, g}^a \leq R - 3\sqrt{K_{a, g}^{\mathbf{D}}}L \right\}$. Then R must be sufficiently large to cover the “good” model $f_{\widehat{\boldsymbol{\theta}}}$ in Assumption **(H.1)**. Moreover, since $R > 3\sqrt{K_{a, g}^{\mathbf{D}}}L$, $R \sim \sqrt{K_{a, g}^{\mathbf{D}}}L$ is the smallest rate we can choose to reduce the rate of $\Phi_{\mu_n, n, L}^{\mathbf{D}}(\boldsymbol{\theta})$ as $\|\mathbf{D}^T \boldsymbol{\theta}\|_{a, g}^a \leq R$.*
2. *The parameter α is used to cancel the effect of $L\sqrt{K_{a, g}^{\mathbf{D}}}$ in the remainder terms. By choosing $\alpha \leq (3K\sqrt{K_{a, g}^{\mathbf{D}}})^{-1/\alpha}$, $\Phi_{\mu_n, n, L}^{\mathbf{D}}(\boldsymbol{\theta}) \leq \frac{\beta}{n} (1 + 3\sqrt{K_{a, g}^{\mathbf{D}}}L\alpha^a + \alpha^a R) \sim \frac{1}{n}$ and $\Psi_{\mu_n, L, p}^{\mathbf{D}} \leq \frac{2eC_{f, \ell}}{\mu} K_{1, g}^{\mathbf{D}} p$.*
3. *The parameter $K_{a, g}^{\mathbf{D}}$ and the function h depend on the choice of g . They also control the rate of $\Omega_{\mu_n, n, L, \lambda}^{\mathbf{D}}(\boldsymbol{\theta})$ and $\Psi_{\mu_n, L, p}^{\mathbf{D}}$.*

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 about the rate of $\Omega_{\mu_n, n, L, \lambda}^D(\boldsymbol{\theta})$ and $\Psi_{\mu_n, L, p}^D$.

We first consider the prior (4.2) in Example 4.3, under the individual sparsity scenario ($\mathbf{D} = \mathbf{I}_p$, $G = 1$) and the choice $a = 1$ (i.e., $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta}\|_1 \leq R\}$). This is the setting considered in [20]. We obtain the following SOI as a corollary of our main result.

Corollary 5.1. *Let \mathbf{X} satisfying Assumption (H.3), $\mathbf{D} = \mathbf{I}_p$ 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 $\Theta_{\mu_n, L, R}^D = \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\boldsymbol{\theta}\|_1 \leq R - 3p\tau\}$,*

$$\Omega_{\mu_n, n, L, \lambda}^D(\boldsymbol{\theta}) \sim \frac{\|\boldsymbol{\theta}\|_0 \log(p)}{n}, \quad \text{and} \quad \Psi_{\mu_n, L, p}^D \sim \frac{1}{n}.$$

The order of $\Omega_{\mu_n, n, L, \lambda}^D(\boldsymbol{\theta})$ is the classical rate under the sparsity scenario. This scaling is similar as in [20] 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 $\int_0^\infty \frac{x^{G+1}}{(\tau^2 + x^2)^2} dx$ 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 \mathbf{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 $\mathbf{D} \in \mathbb{R}^{p \times q}$ satisfying Assumption (H.2) and any $a \in]0, 1]$.

Corollary 5.2. *Let \mathbf{X} satisfying Assumption (H.3), $G \geq 1$ and \mathbf{D} satisfying Assumption (H.2) with $\mu > 0$. 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 $K_{a, g}^D$ as defined in (4.3), and assume that $R > 3\sqrt{K_{a, g}^D}L$. 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 $K_{a, g}^D$. This can be done for instance when \mathbf{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 \mathbf{D} invertible. For $a \in]0, 1]$, let $\tilde{K}_{a, g}^D = \frac{\Gamma((2a+G)/b)\Gamma(c-(2a+G)/b)}{\Gamma(G/b)\Gamma(c-G/b)}$, and set $\alpha \leq 1/\left(3\tau^a \sqrt{\tilde{K}_{a, g}^D}L\right)^{1/a}$. Choosing $\tau^2 \sim 1/(pn)$ and $R \sim L\tau^a$, the group-analysis SOI (5.1) holds with $\Theta_{\mu_n, L, R}^D = \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\mathbf{D}^T \boldsymbol{\theta}\|_{a, G}^a \leq R - 3\tau^a \sqrt{\tilde{K}_{a, g}^D}L\}$,*

$$\Omega_{\mu_n, n, L, \lambda}^D(\boldsymbol{\theta}) \sim \frac{\|\mathbf{D}^T \boldsymbol{\theta}\|_{0, G} \log(L)}{n}, \quad \text{and} \quad \Psi_{\mu_n, L, p}^D \sim \frac{1}{n}.$$

By Assumption (H.1), $\|\mathbf{D}^T \boldsymbol{\theta}\|_{0, G}$ is small when $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$ (with R must be sufficiently large to cover $\tilde{\boldsymbol{\theta}}$). Thus, $\|\mathbf{D}^T \tilde{\boldsymbol{\theta}}\|_{0, G} \log(L)$ is small compared to n . Under the sparsity scenario, the order of $\Omega_{\mu_n, n, L, \lambda}^D(\boldsymbol{\theta})$ becomes $O(\|\boldsymbol{\theta}\|_0 \log(p)/n)$ 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 $\ell(x) = x$, and thus

$$\widehat{f}_n = \mathbf{X}\widehat{\boldsymbol{\theta}}_n,$$

where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the design matrix described in Section 4.2. Recall the EWA estimator

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

with the measure

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

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 [20], we here use the Langevin diffusion process; see [47].

6.1 The Langevin diffusion

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

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

where $\boldsymbol{\rho} = \nabla \log \nu$, ν is everywhere non-zero and a suitably smooth target density function on \mathbb{R}^p , \mathbf{W} is a p -dimensional Brownian process and $\mathbf{l}_0 \in \mathbb{R}^p$ is the initial value. Under mild assumptions, the SDE (6.2) has a unique strong solution and, moreover, $\mathbf{L}(t)$ has a stationary distribution with density precisely ν [47, Theorem 2.1]. $\mathbf{L}(t)$ is therefore interesting for sampling from ν . In particular, this opens the door to approximating integrals $\int_{\mathbb{R}^p} F(\boldsymbol{\theta})\nu(\boldsymbol{\theta})d\boldsymbol{\theta}$ by the average value of a Langevin diffusion, i.e., $\frac{1}{T} \int_0^T F(\mathbf{L}(t))dt$ for a large enough T . Under additional assumptions on ν and F in a proper functional class, the expected squared error of the approximation can be controlled [53].

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

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

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

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

It is then natural 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 ν (hence ρ) to be sufficiently smooth, which limits their applicability in our context. To cope with this difficulty, several works have proposed to replace $\log \nu$ with a smoothed version (typically involving the Moreau-Yosida regularization/envelope, see Definition 2.2) [20, 25, 26, 42]. In [26, 42] for instance, the authors proposed proximal-type algorithms to sample from possibly non-smooth log-concave densities ν using the forward Euler discretization and the Moreau-Yosida regularization. In [42]¹, $-\log \nu$ is replaced with its Moreau envelope, while in [26], it is assumed that $-\log \nu = F + G$, 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 [10], though without any guarantees.

6.2 Forward-Backward proximal LMC 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 involved arguments from variational analysis [48]. For ease of notation, in the following, we denote with the same symbol the measure and its density with respect to the Lebesgue measure. Assume that μ_n is structured as

$$\mu_n(\boldsymbol{\theta}) \propto \exp\left(-\left(F_\beta(\boldsymbol{\theta}) + W_\lambda \circ \mathbf{D}^T(\boldsymbol{\theta})\right)\right) \quad (6.3)$$

where $F_\beta(\boldsymbol{\theta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2/\beta$ and $W_\lambda(\mathbf{u}) = \sum_{l=1}^L w_\lambda(\|\mathbf{u}\|_{2,G})$ with

$$w_\lambda(x) = \alpha^a x^a + c \log(\tau^b + x^b), \quad (6.4)$$

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

Lemma 6.1. *Let w_λ given by (6.4). On $]0, +\infty[$, the function w_λ is bounded from below, non-decreasing and continuously differentiable. Fix $a = 1$ and $b = 1$, the problem $\min_{t \in [0, +\infty[} \{t + \gamma w_\lambda'(t)\}$ has a unique solution at 0 for a given $0 < \gamma \leq \tau^2/c$.*

From the properties of w_λ stated in Lemma 6.1, the works in [37] proposed 2 algorithms of type LMC based on forward-backward proximal splitting for which the guarantees of consistence have been proven.

Semi-Forward-Backward LMC (Semi-FBLMC) Define the following SDE with the Moreau-Yosida regularized version of W_λ

$$d\mathbf{L}(t) = \boldsymbol{\psi}(\mathbf{L}(t))dt + d\mathbf{W}(t), \quad t > 0, \quad (6.5)$$

where $\boldsymbol{\psi} : \boldsymbol{\theta} \in \mathbb{R}^p \mapsto -\frac{1}{2}\nabla(F_\beta + (\gamma W_\lambda) \circ \mathbf{D}^T)(\boldsymbol{\theta})$.

$\boldsymbol{\psi}$ is the drift coefficient. [37, Lemma 3.2] allows to compute $\nabla^\gamma W_\lambda$ by exploiting the relation between $\nabla^\gamma W_\lambda$ and $\text{prox}_{\gamma W_\lambda}$, indeed $\nabla^\gamma W_\lambda = \gamma^{-1}(\mathbf{I}_q - \text{prox}_{\gamma W_\lambda})$. Inserting this expression into (6.5), the Euler discretization of (6.5) is given by

$$\mathbf{L}_{k+1} = \mathbf{L}_k - \frac{\delta}{2}\nabla F_\beta(\mathbf{L}_k) - \frac{\delta}{2\gamma}\mathbf{D}(\mathbf{D}^T \mathbf{L}_k - \text{prox}_{\gamma W_\lambda}(\mathbf{D}^T \mathbf{L}_k)) + \sqrt{\delta}\mathbf{Z}_k, \quad t > 0, \quad \mathbf{L}_0 = \mathbf{l}_0. \quad (6.6)$$

¹The author however applied it to problems where $-\log \nu = F + G$. But the gradient of the Moreau envelope of a sum, which amounts to computing the proximity operator of $-\log \nu$ does not have an easily implementable expression even if those of F and G do.

Forward-Backward LMC (FBLMC) Define $V \stackrel{\text{def}}{=} F_\beta + W_\lambda \circ D^T$ and $M_\gamma \stackrel{\text{def}}{=} I_p - (2\gamma/\beta)X^T X$ which is symmetric positive definite. One can consider an alternative version of the SDE (6.5) with the Moreau-Yosida regularized version of V in the metric M_γ , i.e.,

$$d\mathbf{L}(t) = -\frac{1}{2}\nabla\left({}^{M_\gamma, \gamma}V \circ M_\gamma^{-1/2}\right)(\mathbf{L}(t))dt + M_\gamma^{1/2}d\mathbf{W}(t), \quad t > 0. \quad (6.7)$$

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

$$d\mathbf{U}(t) = -\frac{1}{2}M_\gamma^{-1}\nabla{}^{M_\gamma, \gamma}V(\mathbf{U}(t))dt + d\mathbf{W}(t), \quad t > 0. \quad (6.8)$$

The relation between $\nabla{}^{M_\gamma, \gamma}V$ and $\text{prox}_{\gamma W_\lambda}$ is shown by combining the results in [37, Lemma 3.2] and [37, Lemma 6.1], indeed $\nabla{}^{M_\gamma, \gamma}V = \gamma^{-1}M_\gamma\left(I_p - \text{prox}_{\gamma W_\lambda \circ D^T}(I_p - \gamma\nabla F_\beta)\right)$. With this expression at hand, the Euler discretization of (6.8) is given by

$$\mathbf{U}_{k+1} = \left(1 - \frac{\delta}{2\gamma}\right)\mathbf{U}_k + \frac{\delta}{2\gamma}\text{prox}_{\gamma W_\lambda \circ D^T}(\mathbf{U}_k - \gamma\nabla F_\beta(\mathbf{U}_k)) + \sqrt{\delta}\mathbf{Z}_k, \quad t > 0, \quad \mathbf{U}_0 = \mathbf{l}_0. \quad (6.9)$$

The guarantees of convergence of these algorithms have been proven in [37, Section 4]. We report the details in the proof of the following proposition.

Proposition 6.1. *Let $N_{\delta, T} = \lfloor T/\delta \rfloor$, $\{\mathbf{L}_k\}_{k \in \{0, \dots, N_{\delta, T}-1\}}$ and $\{\mathbf{U}_k\}_{k \in \{0, \dots, N_{\delta, T}-1\}}$ generated respectively by the schemes (6.6) and (6.9). Let $\bar{\mathbf{L}}_{\delta, T, \gamma} = \frac{1}{N_{\delta, T}} \sum_{k=0}^{N_{\delta, T}-1} \mathbf{L}_k$ and $\bar{\mathbf{U}}_{\delta, T, \gamma} = \frac{1}{N_{\delta, T}} \sum_{k=0}^{N_{\delta, T}-1} \mathbf{U}_k$, we obtain that $\bar{\mathbf{L}}_{\delta, T, \gamma} \xrightarrow{\delta \rightarrow 0, \gamma \rightarrow 0, T \rightarrow \infty} \hat{\boldsymbol{\theta}}_n$, and $\bar{\mathbf{U}}_{\delta, T, \gamma} \xrightarrow{\delta \rightarrow 0, \gamma \rightarrow 0, T \rightarrow \infty} \hat{\boldsymbol{\theta}}_n$.*

It remains now to compute the proximal mapping in the schemes of discretization above. Since $W_\lambda \circ D^T$ is analysis group-separable, i.e. $W_\lambda(\mathbf{u}) = \sum_{l=1}^L w_\lambda(\|\mathbf{u}_{\mathcal{G}_l}\|_2)$. Moreover, $w_\lambda : \mathbb{R}^+ \rightarrow \mathbb{R}$ has the properties stated in Lemma 6.1, we get then $\text{prox}_{\gamma W_\lambda}$ is also separable in view of [37, Lemma 7.1], i.e.

$$\text{prox}_{\gamma W_\lambda}(\mathbf{u}) = \left(\text{prox}_{\gamma w_\lambda}(\|\mathbf{u}_{\mathcal{G}_1}\|_2) \frac{\mathbf{u}_{\mathcal{G}_1}^T}{\|\mathbf{u}_{\mathcal{G}_1}\|_2}, \dots, \text{prox}_{\gamma w_\lambda}(\|\mathbf{u}_{\mathcal{G}_L}\|_2) \frac{\mathbf{u}_{\mathcal{G}_L}^T}{\|\mathbf{u}_{\mathcal{G}_L}\|_2}\right)^T, \quad (6.10)$$

and the formula of $\text{prox}_{\gamma w_\lambda}$ is given by [37, Lemma 7.2], i.e.

$$\text{prox}_{\gamma w_\lambda/\beta}(t) = \begin{cases} 0 & \text{if } t \leq \frac{\gamma}{\beta}w_\lambda'(0^+), \\ t - \frac{\gamma}{\beta}w_\lambda'\left(\text{prox}_{\gamma w_\lambda/\beta}(t)\right) & \text{if } t > \frac{\gamma}{\beta}w_\lambda'(0^+), \end{cases} \quad (6.11)$$

for any $t \geq 0$.

Consider now the mapping $\text{prox}_{\gamma W_\lambda \circ D^T}$ in (6.9) which is no longer separable. Owing to Assumption (H.2), one can show quite immediately that

$$\text{prox}_{\gamma W_\lambda \circ D^T} = \tilde{\mathbf{D}} \circ \text{prox}_{\gamma W_\lambda + \iota_{\text{Im}(D^T)}} \circ D^T,$$

where $\iota_{\text{Im}(D^T)}(\mathbf{u}) = 0$ if $\mathbf{u} \in \text{Im}(D^T)$ and $+\infty$ otherwise, and $D^T \tilde{\mathbf{D}}$ is indeed definite positive on $\text{Im}(D^T)$. This means that, in general, $W_\lambda \circ D^T$ does not have an easy-to-compute expression, but rather necessitates to solve an optimization subproblem. To overcome this difficulty, we consider the case where D is invertible. In this case, one can operate a simple change of variable $\mathbf{u} = D^T \boldsymbol{\theta}$ and replace F_β with $F_\beta \circ D^{T-1}$, and $W_\lambda \circ D^T$ with W_λ .

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

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

where $\boldsymbol{\xi}$ is the noise, \mathbf{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 [20, 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: $a = 1, b = 1, c > (2 + G)/b, \tau \sim 1/(pn)$ and $\alpha \leq 1/\left(3\sqrt{K_{a,g}^D L}\right)^{1/a}$, where $K_{a,g}^D$ 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.

7.1 1-D signal recovery under group sparsity

In this example, we set $\mathbf{D} = \mathbf{I}_p$, 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 $\{1, \dots, p\}$.

The experiments are performed by fixing $p = 128$, and taking $S = 4, 8, \dots, p, n = 8, 16, \dots, p$, step-size $\delta = 4\sigma^2/(np)$ and integration time $T = 3500$. The parameters in the prior are chosen to minimize the remainder term in the oracle inequality (5.1). For each (S, n) , and each value of $G_{\boldsymbol{\theta}_0}$, $N_{\text{rep}} = 20$ instances of the problem suite $(\mathbf{X}, \boldsymbol{\theta}_0, \mathbf{Y})$ 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

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

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 [21]. 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 than with $G = 4$. The

situation reverses for Figures 1(c)-(d) where $G_{\theta_0} = 4$, 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 $G = G_{\theta_0}$. This is also confirmed by visual inspection of Figures 1(c')-(d'), where we plotted instances of recovered vectors $\widehat{\theta}_n^{(j,S,n)}$ when $(S, G_{\theta_0}) \in \{4, 8\} \times \{1, 4\}$ 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 $S/p = \|\theta_0\|_{0,G} G_{\theta_0}/p$. As far as the expected phase transition curve is concerned, one has from Corollary 5.3 that it is expected to occur for

$$n/p = C_\epsilon \|\theta_0\|_{0,G} G_{\theta_0}/p (\log(p/G_{\theta_0})/G_{\theta_0}) = C_\epsilon S/p (\log(p/G_{\theta_0})/G_{\theta_0})$$

for some constant $C_\epsilon > 0$ depending on ϵ . That is, the phase transition curve is linear (p and G_{θ_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.

7.2 2-D image recovery under analysis group-sparsity

In the second numerical experiment, θ_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 $\text{vec}(\theta_0)$ is vector in \mathbb{R}^p with $p = 160^2$, and our goal is to recover θ_0 with values in $[0, 1]$ from

$$Y = X \text{vec}(\theta_0) + \xi,$$

where $\xi \sim \mathcal{N}(0, \sigma^2 I_n)$ and $X \in \mathbb{R}^{n \times p}$ 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 [49] 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.

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 $\|D_{\text{TV}}(\theta_0)\|_{0,G} = 1376 \ll n$. 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 consistency 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 to the target distribution and/or to relax some of the assumptions to cover an even larger class of distributions.

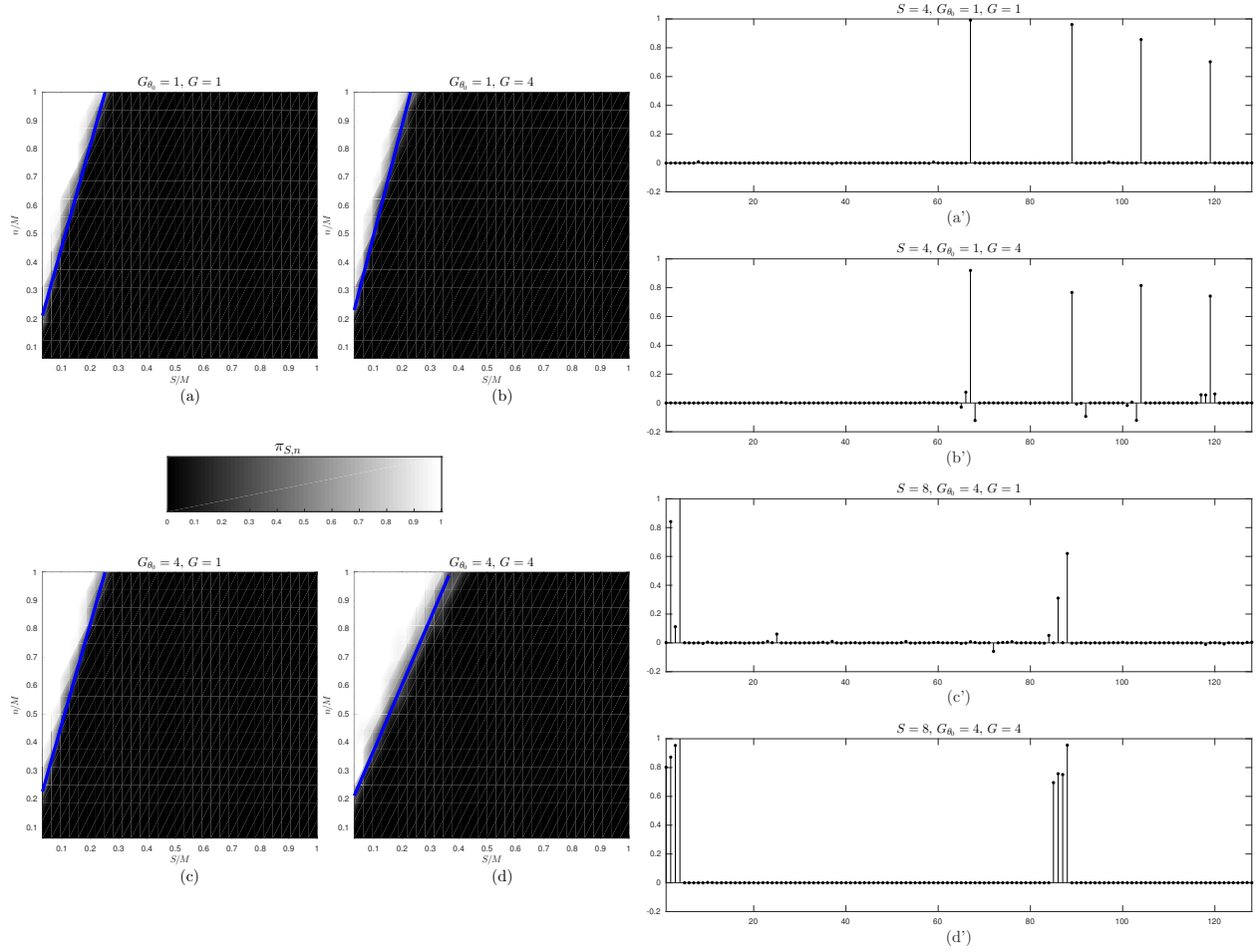


Figure 1: (a)-(d): Phase diagrams of EWA for $D = I_p$, the color bar ranges from dark ($\pi_{S,n} = 0$) to bright ($\pi_{S,n} = 1$). The blue line is the fitted phase transition curve. (a')-(d'): Examples of vectors $\hat{\theta}_n^{(j,S,n)}$ recovered by EWA with $n/p = 1/2$, two sparsity levels $S = 4$ and $S = 8$ and two group sizes $G_{\theta_0} = 1$ and $G_{\theta_0} = 4$.

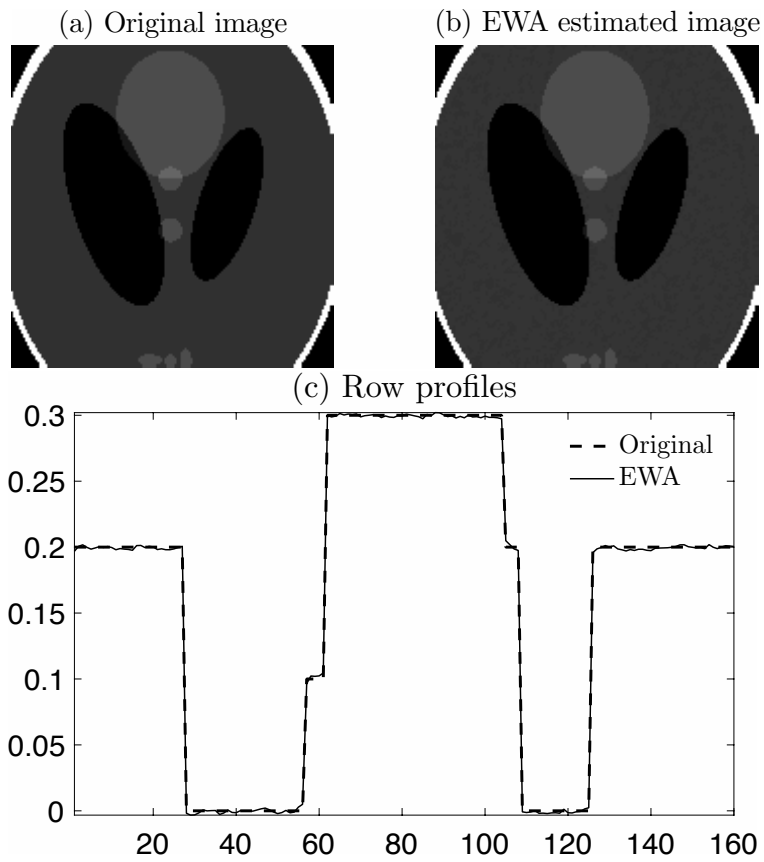


Figure 2: (a): Original close-up of Shepp-Logan phantom image. (b): Image recovered by EWA with $\delta = 2 \cdot 10^{-8}$ and $T = 10^4$. (c) Profiles of a row extracted from each image.

9 Proofs

9.1 Proofs of auxiliary results

Proof of Lemma 2.3 Consider the linear mapping $M : \mathbf{x} \in \mathbb{R}^d \mapsto M\mathbf{x} \in \mathbb{R}^r$, $r \geq d$. The Jacobian matrix of this mapping is obviously M for any $\mathbf{x} \in \mathbb{R}^d$. Since M is a frame, it is injective, hence so-called d -regular (see [38, Section 1.5]). In particular, $\det(MM^T) > 0$. Thus combining [38, Theorems 1.12 and 3.4] and the Cauchy-Binet formula [38, Theorem 3.3]), we have the change of variables formula

$$\int_{\Theta} u(M\mathbf{x})d\mathbf{x} = \frac{\int_{\mathbb{R}^r} \sum_{\mathbf{x} \in \Theta \cap \{\boldsymbol{\omega} : M\boldsymbol{\omega} = \mathbf{v}\}} u(M\mathbf{x})d\mathbf{v}}{\sqrt{\det(MM^T)}} = \frac{\int_{\text{Im}(M)} \sum_{\mathbf{x} \in \Theta \cap \{\boldsymbol{\omega} : M\boldsymbol{\omega} = \mathbf{v}\}} u(M\mathbf{x})d\mathbf{v}}{\sqrt{\det(MM^T)}}. \quad (9.1)$$

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

Proof of Proposition 3.1

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

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

Thus, let $\mathbf{a} \in \mathbb{R}^n$ and $v(\mathbf{a}) \equiv \|\boldsymbol{\Sigma}\|$, we get $\frac{\log \mathbb{E} [e^{\mathbf{t}^T \boldsymbol{\zeta}} | \boldsymbol{\xi} = \mathbf{a}]}{\|\mathbf{t}\|_2^2 \gamma v(\mathbf{a})} \leq 1 + \frac{\gamma}{2} \xrightarrow{\gamma \rightarrow 0} 1 \leq 1$.

- **Laplace noise:** Let $\boldsymbol{\xi} \sim \mathcal{L}(0, \boldsymbol{\Sigma})$, i.e. its associated characteristic function is $\varphi_{\boldsymbol{\xi}}(\mathbf{t}) = \frac{1}{1 + \mathbf{t}^T \boldsymbol{\Sigma} \mathbf{t}/2}$, we choose $\boldsymbol{\zeta}$ according to the distribution associated to the characteristic function $\varphi_{\boldsymbol{\zeta}}(\mathbf{t}) = \frac{1}{(1+\gamma)^2} \left(1 + \frac{2\gamma + \gamma^2}{1 + (1+\gamma)^2 \mathbf{t}^T \boldsymbol{\Sigma} \mathbf{t}/2}\right)$. For any $\mathbf{t} \in \mathbb{R}^n$, we get that

$$\varphi_{\boldsymbol{\xi} + \boldsymbol{\zeta}}(\mathbf{t}) = \varphi_{\boldsymbol{\xi}}(\mathbf{t}) \varphi_{\boldsymbol{\zeta}}(\mathbf{t}) = \frac{1}{1 + (1 + \gamma)^2 \mathbf{t}^T \boldsymbol{\Sigma} \mathbf{t}/2} = \varphi_{(1+\gamma)\boldsymbol{\xi}}(\mathbf{t}). \quad (9.2)$$

Thus, $\boldsymbol{\zeta} + \boldsymbol{\xi}$ has the same distribution as $(1 + \gamma)\boldsymbol{\xi}$. We also obtain $\mathbb{E} [\boldsymbol{\zeta} | \boldsymbol{\xi}] = \mathbb{E} [\boldsymbol{\zeta}] = (-i) \nabla \varphi_{\boldsymbol{\zeta}}(0) = 0$. It suffices to check the condition (3) of Assumption (P.2). We know that $\mathbb{E} [e^{\mathbf{t}^T \boldsymbol{\zeta}} | \boldsymbol{\xi}] = \mathbb{E} [e^{\mathbf{t}^T \boldsymbol{\zeta}}] = \varphi_{\boldsymbol{\zeta}}(-i\mathbf{t}) = \frac{1}{(1+\gamma)^2} \left(1 + \frac{2\gamma + \gamma^2}{1 - (1+\gamma)^2 \mathbf{t}^T \boldsymbol{\Sigma} \mathbf{t}/2}\right)$. Using Taylor's formula, we have

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

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

$$\frac{\log \mathbb{E} [e^{\mathbf{t}^T \boldsymbol{\zeta}} | \boldsymbol{\xi} = \mathbf{a}]}{\|\mathbf{t}\|_2^T \gamma v(\mathbf{a})} \xrightarrow{\gamma \rightarrow 0} \frac{\mathbf{t}^T \boldsymbol{\Sigma} \mathbf{t} / (1 - \mathbf{t}^T \boldsymbol{\Sigma} \mathbf{t} / 2)}{\|\mathbf{t}\|_2^2 \|\boldsymbol{\Sigma}\| / (1 - t_0^2 \|\boldsymbol{\Sigma}\| / 2)} \leq \frac{1 - t_0^2 \|\boldsymbol{\Sigma}\| / 2}{1 - \mathbf{t}^T \boldsymbol{\Sigma} \mathbf{t} / 2} \leq \frac{1 - t_0^2 \|\boldsymbol{\Sigma}\| / 2}{1 - \|\mathbf{t}\|_2^2 \|\boldsymbol{\Sigma}\| / 2} \leq 1.$$

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

- **Bounded symmetric noise:** Let ξ are symmetric and $\mathbb{P} [|\xi_i| \leq B_i] = 1$ for some $\mathbf{B} \in \mathbb{R}^n$, we set $\zeta = (\zeta_1, \dots, \zeta_n)^T$ such that $\zeta_i = (1 + \gamma)|\xi_i| \operatorname{sgn}(\operatorname{sgn}(\xi_i) - (1 + \gamma)U_i) - \xi_i$, $U_i \sim \mathcal{U}([-1, 1])$ for any $i \in \{1, \dots, n\}$. Using [20, Equation (22)], for any $\mathbf{t} \in \mathbb{R}^n$ and $\mathbf{a} \in \{\mathbf{x} \in \mathbb{R}^n : x_i \in [-B_i, B_i], \forall i \in \{1, \dots, n\}\}$, we get that

$$\mathbb{E} \left[e^{\mathbf{t}^T \zeta} | \boldsymbol{\xi} = \mathbf{a} \right] = \prod_{j=1}^n \mathbb{E} \left[e^{t_j \zeta_j} | \xi_j = a_j \right] = e^{-\mathbf{t}^T \mathbf{a}} \left(e^{(1+\gamma)\mathbf{t}^T \mathbf{a}} \frac{2+\gamma}{2+2\gamma} + e^{-(1+\gamma)\mathbf{t}^T \mathbf{a}} \frac{\gamma}{2+2\gamma} \right). \quad (9.3)$$

From (9.3) and the symmetry of ξ , we obtain

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

Thus, $\zeta + \xi$ has the same distribution as $(1 + \gamma)\xi$. Since $\mathbb{E} [\zeta | \boldsymbol{\xi} = \mathbf{a}]$ equals to the gradient of $\mathbb{E} \left[e^{\mathbf{t}^T \zeta} | \boldsymbol{\xi} = \mathbf{a} \right]$ at $\mathbf{t} = 0$, from (9.3) we have then $\mathbb{E} [\zeta | \boldsymbol{\xi} = \mathbf{a}] = 0$, $\forall \mathbf{a} \in [-\mathbf{B}, \mathbf{B}]$. It suffices to check the condition (c) of Assumption (P.2). Owing to [16, Lemma 3] and [20, Equation (22)], we get that $\log \left(\mathbb{E} \left[e^{t_i \zeta_i} | \xi_i = a_i \right] \right) \leq (t_i a_i)^2 \gamma (1 + \gamma)$. Thus, let $\mathbf{t} \in \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 \in [-\infty, \infty]\}$ and $v(\mathbf{a}) = \|\mathbf{a}\|_2^2$, we get that

$$\frac{\log \mathbb{E} \left[e^{\mathbf{t}^T \zeta} | \boldsymbol{\xi} = \mathbf{a} \right]}{\|\mathbf{t}\|_2^2 \gamma v(\mathbf{a})} = \frac{\sum_{i=1}^n \log \mathbb{E} \left[e^{t_i \zeta_i} | \xi_i = a_i \right]}{\|\mathbf{t}\|_2^2 \gamma \|\mathbf{a}\|_2^2} \leq \frac{\sum_{i=1}^n t_i^2 a_i^2 \gamma (1 + \gamma)}{\|\mathbf{t}\|_2^2 \gamma \|\mathbf{a}\|_2^2} \leq 1 + \gamma \xrightarrow{\gamma \rightarrow 0} 1 \leq 1.$$

Proof of Proposition 4.1 Let $\boldsymbol{\theta} \in \Theta$ and $i \in \{1, \dots, n\}$. Setting $u_i^\boldsymbol{\theta} = \sum_{j=1}^p \theta_j f_j(x_i)$ and $\mathbf{u}^\boldsymbol{\theta} = (u_1^\boldsymbol{\theta}, \dots, u_n^\boldsymbol{\theta})^T$, and by virtue of (2.1), (2.2), (4.1) and the fact that $a \in]0, 1]$, we have $\|\mathbf{u}^\boldsymbol{\theta}\|_2 = \|\mathbf{X}\boldsymbol{\theta}\|_2 \leq \|\mathbf{X}\| \|\boldsymbol{\theta}\|_2 \leq \frac{\|\mathbf{X}\| \|\mathbf{D}^T \boldsymbol{\theta}\|_2}{\sqrt{\mu}} \leq \frac{\|\mathbf{X}\| R^{1/\alpha}}{\sqrt{\mu}}$. Which in turn implies $\mathbf{u}^\boldsymbol{\theta} \in \mathcal{B}$. Therefore, for any $(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \Theta^2$, $\|\mathbf{f}_\boldsymbol{\theta} - \mathbf{f}_{\boldsymbol{\theta}'}\|_2 = \|\mathbf{L}(\mathbf{u}^\boldsymbol{\theta}) - \mathbf{L}(\mathbf{u}^{\boldsymbol{\theta}'})\|_2 \leq 2 \max_{\mathbf{x} \in \mathcal{B}} \|\mathbf{L}(\mathbf{x})\|_2$. \square

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

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

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

$$\begin{aligned} \int_{\mathbb{R}^p} \prod_{l=1}^L g(\|[\mathbf{D}^T \mathbf{u}]_{\mathcal{G}_l}\|_2) d\mathbf{u} &\leq \frac{C_G^L \left(\int_0^1 z^{G-1} g(z) dz + \int_1^\infty w^{G-1} g(w) dw \right)^L}{\sqrt{\det(\mathbf{D}\mathbf{D}^T)}} \\ &\leq \frac{C_G^L \left(\sup_{z \in [0,1]} g(z) + \int_1^\infty w^{G+1} g(w) dw \right)^L}{\sqrt{\det(\mathbf{D}\mathbf{D}^T)}} < \infty. \end{aligned} \quad (9.4)$$

Therefore, g satisfies Assumption (G.2). Now, we check the moment condition (G.3). Using similar arguments to the bound (9.4), we have

$$\begin{aligned} \int_{\mathbb{R}^p} \|[\mathbf{D}^T \mathbf{u}]_{\mathcal{G}_i}\|_2^2 \prod_{k=1}^L g(\|[\mathbf{D}^T \mathbf{u}]_{\mathcal{G}_k}\|_2) d\mathbf{u} &\leq \frac{\int_{\mathbb{R}^q} \|\mathbf{v}_{\mathcal{G}_i}\|_2^2 \prod_{k=1}^L g(\|\mathbf{v}_{\mathcal{G}_k}\|_2) d\mathbf{v}}{\sqrt{\det(\mathbf{D}\mathbf{D}^T)}} \\ &= \frac{\left(\int_{\mathbb{R}^G} \|\mathbf{u}\|_2^2 g(\|\mathbf{u}\|_2) d\mathbf{u} \right) \left(\int_{\mathbb{R}^G} g(\|\mathbf{v}\|_2) d\mathbf{v} \right)^{L-1}}{\sqrt{\det(\mathbf{D}\mathbf{D}^T)}} \\ &= \frac{C_G^L \int_0^\infty z^{G+1} g(z) dz \left(\int_0^\infty w^{G-1} g(w) dw \right)^{L-1}}{\sqrt{\det(\mathbf{D}\mathbf{D}^T)}} < \infty, \end{aligned} \quad (9.5)$$

whence we conclude that g satisfies Assumption (G.3). \square

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

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

Therefore, g satisfies Assumptions (G.1)-(G.4) for $G = 1$. Owing to Remark 4.3 and Lemma 2.1, we obtain $K_{1,g}^{\mathbf{D}} = \int_0^\infty \frac{x^2}{(\tau^2 + x^2)^2} dx / \int_0^\infty \frac{1}{(\tau^2 + y^2)^2} dy = \tau^2$. We are now in position to apply Theorem 5.1 with $\mathbf{D} = \mathbf{I}_p$ (then $q = p$), $G = 1$ (then $L = q$), $a = 1$ and $\alpha \leq 1/(3p\tau)$ to conclude. Namely, since $\tau^2 \sim \frac{1}{pn}$ and $R \sim p\tau$, we get that $\Psi_{\mu_n, L, p}^{\mathbf{D}} \leq 2eC_{f, \ell} \tau^2 p \sim \frac{1}{n}$, and $\Omega_{\mu_n, n, L, \lambda}^{\mathbf{D}}(\boldsymbol{\theta}) = \frac{4\beta}{n} \sum_{j=1}^p \log(1 + \frac{|\theta_j|}{\tau}) \leq \frac{4\beta}{n} \|\boldsymbol{\theta}\|_0 \log(1 + \frac{R}{\tau}) \sim \frac{\|\boldsymbol{\theta}\|_0 \log(p)}{n}$. This completes the proof. \square

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 $(\mathbf{t}, \mathbf{t}^*) \in \mathbb{R}^G \times \mathbb{R}^G$, we have

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

Therefore, g satisfies Assumptions (G.1)-(G.4) with any $G \geq 1$. Applying Theorem 5.1, we conclude the proof. \square

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

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

Since $\tau^2 \sim \frac{1}{pn}$ and $R \sim L\tau^a$, we get that $\Psi_{\mu_n, L, p}^{\mathbf{D}} \leq \frac{2C_{f,\ell} \tilde{K}_{1,g}^{\mathbf{D}}}{\mu} p\tau^2 \sim \frac{1}{n}$, and $\Omega_{\mu_n, n, L, \lambda}^{\mathbf{D}}(\boldsymbol{\theta}) = \frac{c\beta}{n} \sum_{l=1}^L \log \left(1 + \left[\frac{\|\mathbf{D}^T \boldsymbol{\theta}\|_{0,l}}{\tau} \right]^b \right) \leq \frac{c\beta}{n} \|\mathbf{D}^T \boldsymbol{\theta}\|_{0,g} \log \left(1 + \left[\frac{R^{1/a}}{\tau} \right]^b \right) \sim \frac{\|\mathbf{D}^T \boldsymbol{\theta}\|_{0,g} \log(L)}{n}$. This ends the proof. \square

Proof of Lemma 6.1 w_λ is clearly nondecreasing, bounded from below by $w_\lambda(0)$, and continuously differentiable (in fact even C^∞) on $]0, +\infty[$. Let us set $a = 1$ and $b = 1$, and $u(x) = x + \gamma w_\lambda'(x) = x + \frac{\gamma c}{\tau+x} + \gamma \alpha$. One can see that u admits a local maximum at $x_0 = -\sqrt{\gamma c} - \tau \notin [0, +\infty[$ and a local minimum at $x_1 = \sqrt{\gamma c} - \tau$. Thus, the problem $\min_{t \in [0, +\infty[} u(t)$ has a unique solution at 0 when $x_1 \leq 0$ equivalent to $\gamma \leq \tau^2/c$. \square

Proof of Proposition 6.1 Let us begin by the convergence of $\bar{\mathbf{L}}_{\delta, T, \gamma}$. In view of Lemma 6.1, [37, Lemma 7.3] and [37, Sub Section 6.2], we get that [37, (H.1)-(H.4)] hold with $H \stackrel{\text{def}}{=} W_\lambda$ and $\mathbf{M} \stackrel{\text{def}}{=} \mathbf{I}_q$. Thus, owing to [37, Proposition 4.1], $\psi \stackrel{\text{def}}{=} -\frac{1}{2} \nabla (F_\beta + \gamma W_\lambda \circ \mathbf{D}^T)$ is locally Lipschitz continuous, and $\langle \psi(\boldsymbol{\theta}), \boldsymbol{\theta} \rangle \leq K(1 + \|\boldsymbol{\theta}\|_2^2)$ for some $K > 0$.

From [37, Proposition 4.2] and [37, Proposition 3.1], the SDE (6.5) admits a unique solution which is strongly Markovian and the diffusion is non explosive, and $\mathbf{L}(t)$ has a distribution $\mu_\gamma \propto \exp(-(F_\beta + \gamma W_\lambda \circ \mathbf{D}^T))$ which converges to μ_n in total variation (TV) as $\gamma \rightarrow 0$.

Consider an Euler approximate solution of the scheme (6.6) defined as

$$\mathbf{L}^\delta(t) \stackrel{\text{def}}{=} \mathbf{L}_0 + \int_0^t \psi(\bar{\mathbf{L}}(s)) ds + \int_0^t d\mathbf{W}(s),$$

where $\bar{\mathbf{L}}(t) = \mathbf{L}_k$ for $t \in [k\delta, (k+1)\delta[$. Owing to [37, Theorem 4.1], for any $t \in [0, T]$, $\mathbb{E}[\mathbf{L}^\delta(t)]$ converges in mean square to $\mathbb{E}[\mathbf{L}(t)]$ as $\delta \rightarrow 0$. According to the convergence in TV of μ_γ to μ_n , we also get that

$$\mathbb{E}[\mathbf{L}^\delta(t)] \xrightarrow{\delta \rightarrow 0} \mathbb{E}[\mathbf{L}(t)] \xrightarrow{\gamma \rightarrow 0} \hat{\boldsymbol{\theta}}_n, \quad \forall t \in [0, T]. \quad (9.6)$$

By Young's inequality, the fact that $\langle \psi(\boldsymbol{\theta}), \boldsymbol{\theta} \rangle \leq K(1 + \|\boldsymbol{\theta}\|_2^2)$ for some $K > 0$ implies that $\|\psi\|_2 \leq K(1 + \|\boldsymbol{\theta}\|_2)$ for some $K > 0$. Using the reasoning in [20, Step 1, Proposition 2] and (9.6), we get that

$$\bar{\mathbf{L}}_{\delta, T, \gamma} \xrightarrow{\delta \rightarrow 0} \frac{1}{T} \int_0^T L^\delta(t) dt \xrightarrow{T \rightarrow \infty} \mathbb{E} \left[\mathbf{L}^\delta(t) \right] \xrightarrow{\delta \rightarrow 0} \mathbb{E} [\mathbf{L}(t)] \xrightarrow{\gamma \rightarrow 0} \hat{\boldsymbol{\theta}}_n. \quad (9.7)$$

That concludes the proof of the convergence of $\bar{\mathbf{L}}_{\delta, T, \gamma}$ to $\hat{\boldsymbol{\theta}}_n$.

We prove now the convergence of $\bar{\mathbf{U}}_{\delta, T, \gamma}$. Indeed, the SDE (6.7) can be identified by [37, SDE (4.7)] with $H \stackrel{\text{def}}{=} F_\beta + W_\lambda \circ \mathbf{D}^T$, $\mathbf{D} \stackrel{\text{def}}{=} \mathbf{I}_p$ and $L \stackrel{\text{def}}{=} 0$. In view of Lemma 6.1, [37, Lemma 7.3] and [37, Sub Section 6.1], we get that [37, (H.1)-(H.4)] hold. Then, by [37, Approach 2, Sub Section 4.2], we also get the wellposedness of the SDE (6.7) with the unique solution $\mathbf{U}(t)$ has a distribution $\tilde{\mu}_\gamma$ converging in TV to μ_n as $\gamma \rightarrow 0$, and the mean square convergence of $\mathbb{E} [\mathbf{U}^\delta]$ vers $\mathbb{E} [\mathbf{U}]$ where \mathbf{U}^δ is the continuous extension of (6.9). Thus, using the same reasoning, we get the same result as (9.7) for $\bar{\mathbf{U}}_{\delta, T, \gamma}$.

9.2 Proofs of the SOI results

Proof of Theorem 5.1 Remind the prior $\pi(d\boldsymbol{\theta})$ from (4.2), where $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\mathbf{D}^T \boldsymbol{\theta}\|_{a, \mathcal{G}}^a \leq R\}$.

Let $r_L = 3\sqrt{K_{a, \mathcal{G}}^D} L$, $\Theta_{p_0^D} = \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\mathbf{D}^T \boldsymbol{\theta} - \mathbf{D}^T \boldsymbol{\theta}^*\|_{a, \mathcal{G}}^a \leq r_L\}$ and

$$\boldsymbol{\theta}^* \in \{\boldsymbol{\theta} \in \mathbb{R}^p : \|\mathbf{D}^T \boldsymbol{\theta}\|_{a, \mathcal{G}}^a \leq R - 3\sqrt{K_{a, \mathcal{G}}^D} L = R - r_L\}. \quad (9.8)$$

We define the probability measure

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

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

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

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

$$(f_{\boldsymbol{\theta}}(x_i) - f(x_i))^2 \leq (f_{\boldsymbol{\theta}^*}(x_i) - f(x_i))^2 + C_{f, \ell} [\mathbf{X}_i^T (\boldsymbol{\theta} - \boldsymbol{\theta}^*)]^2 + 2(f_{\boldsymbol{\theta}^*}(x_i) - f(x_i)) \ell'(\mathbf{X}_i^T \boldsymbol{\theta}^*) \mathbf{X}_i^T (\boldsymbol{\theta} - \boldsymbol{\theta}^*), \quad (9.9)$$

where $C_{f, \ell} = \|\ell'\|_\infty^2 + \|\ell''\|_\infty (\|\ell'\|_\infty + \|f\|_\infty)$. By summing over i from 1 to n , normalizing by $1/n$, taking the integral in Θ w.r.t. p_0^D , inequality (9.9) becomes

$$\begin{aligned} \int_{\Theta} \|f_{\boldsymbol{\theta}} - f\|_n^2 p_0^D(d\boldsymbol{\theta}) &\leq \|f_{\boldsymbol{\theta}^*} - f\|_n^2 + C_{f, \ell} \int_{\mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n [\mathbf{X}_i^T (\boldsymbol{\theta} - \boldsymbol{\theta}^*)]^2 p_0^D(d\boldsymbol{\theta}) \\ &\quad + \frac{2}{n} \sum_{i=1}^n (f_{\boldsymbol{\theta}^*}(x_i) - f(x_i)) \ell'(\mathbf{X}_i^T \boldsymbol{\theta}^*) \mathbf{X}_i^T \int_{\Theta} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) p_0^D(d\boldsymbol{\theta}). \quad (9.10) \end{aligned}$$

Note that, the right term of inequality (9.10) 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 $\boldsymbol{\theta} \in \Theta_{p_0^D}$, from (9.8) and the fact that $a \in]0, 1]$, we have

$$\begin{aligned} \|\mathbf{D}^T \boldsymbol{\theta}\|_{a, \mathcal{G}}^a &= \sum_{l=1}^L \|\mathbf{D}^T \boldsymbol{\theta}\|_{\mathcal{G}_l}^a \leq \sum_{l=1}^L (\|\mathbf{D}^T \boldsymbol{\theta} - \mathbf{D}^T \boldsymbol{\theta}^*\|_{\mathcal{G}_l} + \|\mathbf{D}^T \boldsymbol{\theta}^*\|_{\mathcal{G}_l})^a \\ &\leq \|\mathbf{D}^T \boldsymbol{\theta} - \mathbf{D}^T \boldsymbol{\theta}^*\|_{a, \mathcal{G}}^a + \|\mathbf{D}^T \boldsymbol{\theta}^*\|_{a, \mathcal{G}}^a \\ &\leq r_L + \|\mathbf{D}^T \boldsymbol{\theta}^*\|_{a, \mathcal{G}}^a \leq R. \end{aligned}$$

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

$$\Theta_{p_0^D} \subseteq \Theta. \quad (9.11)$$

In what follows, we denote $\mathbb{B}_{a, \mathcal{G}}^a(x) = \{\mathbf{z} \in \mathbb{R}^q : \|\mathbf{z}\|_{a, \mathcal{G}}^a \leq x\}$, $\forall x > 0$ for brevity. By (9.11), property (2.1), Lemma 2.3 and symmetry of $\mathbb{B}_{a, \mathcal{G}}^a(r_L) \cap \text{Im}(\mathbf{D}^T)$, we obtain

$$\begin{aligned} &\int_{\Theta} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) p_0^D(d\boldsymbol{\theta}) \\ &\propto \int_{\Theta \cap \Theta_{p_0^D}} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \prod_{l=1}^L \exp(-\alpha^a \|\mathbf{D}^T \boldsymbol{\theta} - \mathbf{D}^T \boldsymbol{\theta}^*\|_{\mathcal{G}_l}^a) g(\|\mathbf{D}^T \boldsymbol{\theta} - \mathbf{D}^T \boldsymbol{\theta}^*\|_{\mathcal{G}_l}) d\boldsymbol{\theta} \\ &= \int_{\Theta_{p_0^D}} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \prod_{l=1}^L \exp(-\alpha^a \|\mathbf{D}^T \boldsymbol{\theta} - \mathbf{D}^T \boldsymbol{\theta}^*\|_{\mathcal{G}_l}^a) g(\|\mathbf{D}^T \boldsymbol{\theta} - \mathbf{D}^T \boldsymbol{\theta}^*\|_{\mathcal{G}_l}) d\boldsymbol{\theta} \\ &= \frac{\tilde{\mathbf{D}}}{\sqrt{\det(\mathbf{D}\mathbf{D}^T)}} \int_{\mathbb{B}_{a, \mathcal{G}}^a(r_L) \cap \text{Im}(\mathbf{D}^T)} \mathbf{z} \prod_{l=1}^L \exp(-\alpha^a \|\mathbf{z}_{\mathcal{G}_l}\|_2^a) g(\|\mathbf{z}_{\mathcal{G}_l}\|_2) d\mathbf{z} = 0, \end{aligned} \quad (9.12)$$

which is the desired claim.

We now bound the second term in the right hand side of (9.10). Define

$$p_0(d\mathbf{u}) = \frac{1}{C_L \sqrt{\det(\mathbf{D}\mathbf{D}^T)}} \prod_{l=1}^L \exp(-\alpha^a \|\mathbf{u}_{\mathcal{G}_l}\|_2^a) g(\|\mathbf{u}_{\mathcal{G}_l}\|_2) I_{\text{Im}(\mathbf{D}^T) \cap \mathbb{B}_{a, \mathcal{G}}^a(r_L)}(\mathbf{u}) d\mathbf{u}. \quad (9.13)$$

One can see that p_0 coincides with the probability measure p_0^D on \mathbb{R}^p via a change of variables of type (2.5). So, p_0 is a probability measure on \mathbb{R}^q . For any $i, j \in \{1, \dots, L\}$, $i \neq j$, by a change of variables, we get $\int_{\mathbb{R}^q} \mathbf{u}_{\mathcal{G}_i} \mathbf{u}_{\mathcal{G}_j}^T p_0(d\mathbf{u}) = -\int_{\mathbb{R}^q} \mathbf{u}_{\mathcal{G}_i} \mathbf{u}_{\mathcal{G}_j}^T p_0(d\mathbf{u})$, so

$$\int_{\mathbb{R}^q} \mathbf{u}_{\mathcal{G}_i} \mathbf{u}_{\mathcal{G}_j}^T p_0(d\mathbf{u}) = 0. \quad (9.14)$$

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

$$\int_{\mathbb{R}^q} \mathbf{u}_{\mathcal{G}_j} \mathbf{u}_{\mathcal{G}_j}^T p_0(d\mathbf{u}) = \int_{\mathbb{R}^q} \mathbf{u}_{\mathcal{G}_1} \mathbf{u}_{\mathcal{G}_1}^T p_0(d\mathbf{u}). \quad (9.15)$$

We obtain

$$\begin{aligned}
& \int_{\mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n [\mathbf{X}_i^T (\boldsymbol{\theta} - \boldsymbol{\theta}^*)]^2 p_0^D(d\boldsymbol{\theta}) \\
\stackrel{\text{(2.1) and Lemma 2.3}}{=} & \frac{1}{n} \int_{\mathbb{R}^q} [\mathbf{X} \tilde{\mathbf{D}} \mathbf{u}]^T \mathbf{X} \tilde{\mathbf{D}} \mathbf{u} p_0(d\mathbf{u}) \\
& = \frac{1}{n} \int_{\mathbb{R}^q} \text{tr} \left(\mathbf{u}^T \tilde{\mathbf{D}}^T \mathbf{X}^T \mathbf{X} \tilde{\mathbf{D}} \mathbf{u} \right) p_0(d\mathbf{u}) \\
& = \frac{1}{n} \text{tr} \left(\left(\mathbf{X} \tilde{\mathbf{D}} \right)^T \mathbf{X} \tilde{\mathbf{D}} \int_{\mathbb{R}^q} \mathbf{u} \mathbf{u}^T p_0(d\mathbf{u}) \right) \\
\stackrel{\text{(9.14) and (9.15)}}{=} & \frac{1}{n} \sum_{l=1}^L \text{tr} \left(\left[\left(\mathbf{X} \tilde{\mathbf{D}} \right)^T \mathbf{X} \tilde{\mathbf{D}} \right]_{\mathcal{G}_l} \int_{\mathbb{R}^q} \mathbf{u}_{\mathcal{G}_l} \mathbf{u}_{\mathcal{G}_l}^T p_0(d\mathbf{u}) \right) \\
\stackrel{\text{(Von Neumann's trace inequality)}}{\leq} & \frac{1}{n} \sum_{l=1}^L \sum_{j=1}^G \sigma_j \left(\left[\left(\mathbf{X} \tilde{\mathbf{D}} \right)^T \mathbf{X} \tilde{\mathbf{D}} \right]_{\mathcal{G}_l} \right) \sigma_j \left(\int_{\mathbb{R}^q} \mathbf{u}_{\mathcal{G}_l} \mathbf{u}_{\mathcal{G}_l}^T p_0(d\mathbf{u}) \right) \\
& \leq \frac{1}{n} \int_{\mathbb{R}^q} \|\mathbf{u}_{\mathcal{G}_1}\|_2^2 p_0(d\mathbf{u}) \sum_{l=1}^L \text{tr} \left(\left[\left(\mathbf{X} \tilde{\mathbf{D}} \right)^T \mathbf{X} \tilde{\mathbf{D}} \right]_{\mathcal{G}_l} \right) \\
& = \frac{1}{n} \text{tr} \left(\left(\mathbf{X} \tilde{\mathbf{D}} \right)^T \mathbf{X} \tilde{\mathbf{D}} \right) \int_{\mathbb{R}^q} \|\mathbf{u}_{\mathcal{G}_1}\|_2^2 p_0(d\mathbf{u}). \tag{9.16}
\end{aligned}$$

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

$$\frac{\text{tr} \left(\left(\mathbf{X} \tilde{\mathbf{D}} \right)^T \mathbf{X} \tilde{\mathbf{D}} \right)}{n} \leq \sum_{j=1}^p \sigma_j \left(\frac{\mathbf{X}^T \mathbf{X}}{n} \right) \sigma_j \left(\tilde{\mathbf{D}} \tilde{\mathbf{D}}^T \right) \leq \sigma_1 \left(\tilde{\mathbf{D}} \tilde{\mathbf{D}}^T \right) \sum_{j=1}^p \sigma_j \left(\frac{\mathbf{X}^T \mathbf{X}}{n} \right) \leq \frac{p}{\mu}. \tag{9.17}$$

Putting together (9.16) and (9.17), we get the bound

$$C_{f,\ell} \int_{\mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n [\mathbf{X}_i^T (\boldsymbol{\theta} - \boldsymbol{\theta}^*)]^2 p_0^D(d\boldsymbol{\theta}) \leq C_{f,\ell} \frac{p}{\mu} \int_{\mathbb{R}^q} \|\mathbf{u}_{\mathcal{G}_1}\|_2^2 p_0(d\mathbf{u}). \tag{9.18}$$

Thanks to (9.12) and (9.18), inequality (9.10) becomes

$$\int_{\Theta} \|f_{\boldsymbol{\theta}} - f\|_n^2 p_0^D(d\boldsymbol{\theta}) \leq \|f_{\boldsymbol{\theta}^*} - f\|_n^2 + C_{f,\ell} \frac{p}{\mu} \int_{\mathbb{R}^q} \|\mathbf{u}_{\mathcal{G}_1}\|_2^2 p_0(d\mathbf{u}). \tag{9.19}$$

Now, inserting (9.19) into Theorem 3.1 (with $p = p_0^D$), we arrive at

$$\mathbb{E} \left[\|\hat{f}_n - f\|_n^2 \right] \leq \|f_{\boldsymbol{\theta}^*} - f\|_n^2 + C_{f,\ell} \frac{p}{\mu} \int_{\mathbb{R}^q} \|\mathbf{u}_{\mathcal{G}_1}\|_2^2 p_0(d\mathbf{u}) + \frac{\beta \text{KL}(p_0^D, \pi)}{n}. \tag{9.20}$$

To complete the proof, it remains to bound the last two terms in the right hand side of (9.20). This is the goal of the following lemma.

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

$$\int_{\mathbb{R}^q} \|\mathbf{u}_{\mathcal{G}_1}\|_2^2 p_0(d\mathbf{u}) \leq 2K_{1,g}^D e^{r_L \alpha^a}, \tag{9.21}$$

and

$$\text{KL}(p_0^D, \pi) \leq 1 + r_L \alpha^a + \lambda \sum_{l=1}^L \log \left\{ h \left(\left\| \left[\mathbf{D}^T \boldsymbol{\theta}^* \right]_{\mathcal{G}_l} \right\|_2 \right) \right\} + \alpha^a \|\mathbf{D}^T \boldsymbol{\theta}^*\|_{a,\mathcal{G}}^a. \tag{9.22}$$

With $r_L = 3\sqrt{K_{a,g}^D}L$, it follows from (9.20) and Lemma 9.1 that

$$\begin{aligned} \mathbb{E} \left[\|\widehat{f}_n - f\|_n^2 \right] &\leq \|f\theta^* - f\|_n^2 + \frac{\beta}{n} \left(1 + 3\sqrt{K_{a,g}^D}L\alpha^a + \alpha^a \|\mathbf{D}^T \theta^*\|_{a,g}^a \right) \\ &\quad + \frac{\lambda\beta}{n} \sum_{l=1}^L \log \left\{ h \left(\|\mathbf{D}^T \theta^*\|_{\mathcal{G}_l} \right) \right\} + \frac{2K_{1,g}^D e^{3\sqrt{K_{a,g}^D}L\alpha^a} pC_{f,\ell}}{\mu}. \end{aligned}$$

According to (9.8), this completes the proof of Theorem 5.1. \square

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

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

$$\frac{1}{T} \int_{\{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{D}^T \mathbf{u}\|_{a,g}^a > s\}} \prod_{l=1}^L g(\|\mathbf{D}^T \mathbf{u}\|_{\mathcal{G}_l}) d\mathbf{u} \leq \frac{L^2 K_{a,g}^D}{\left(s - L\sqrt{K_{a,g}^D}\right)^2},$$

where $T = \int_{\mathbb{R}^p} \prod_{l=1}^L g(\|\mathbf{D}^T \mathbf{u}\|_{\mathcal{G}_l}) d\mathbf{u}$.

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

$$\begin{aligned} &\frac{1}{T} \int_{\{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{D}^T \mathbf{u}\|_{a,g}^a > s\}} \prod_{l=1}^L g(\|\mathbf{D}^T \mathbf{u}\|_{\mathcal{G}_l}) d\mathbf{u} \\ &= \mathbb{P} \left[\sum_{l=1}^L \|\mathbf{D}^T \mathbf{U}\|_{\mathcal{G}_l}^a > s \right] \\ &= \mathbb{P} \left[\sum_{l=1}^L \|\mathbf{D}^T \mathbf{U}\|_{\mathcal{G}_l}^a - \mathbb{E} [\|\mathbf{D}^T \mathbf{U}\|_{\mathcal{G}_l}^a] > s - \sum_{l=1}^L \mathbb{E} [\|\mathbf{D}^T \mathbf{U}\|_{\mathcal{G}_l}^a] \right] \\ &\leq \mathbb{E} \left[\left(\sum_{l=1}^L \|\mathbf{D}^T \mathbf{U}\|_{\mathcal{G}_l}^a - \mathbb{E} [\|\mathbf{D}^T \mathbf{U}\|_{\mathcal{G}_l}^a] \right)^2 \right] / \left(s - \sum_{l=1}^L \mathbb{E} [\|\mathbf{D}^T \mathbf{U}\|_{\mathcal{G}_l}^a] \right)^2 \\ &= \text{var} \left(\sum_{l=1}^L \|\mathbf{D}^T \mathbf{U}\|_{\mathcal{G}_l}^a \right) / \left(s - \sum_{l=1}^L \mathbb{E} [\|\mathbf{D}^T \mathbf{U}\|_{\mathcal{G}_l}^a] \right)^2 \\ &\leq \mathbb{E} \left[\left(\sum_{l=1}^L \|\mathbf{D}^T \mathbf{U}\|_{\mathcal{G}_l}^a \right)^2 \right] / \left(s - \sum_{l=1}^L \mathbb{E} [\|\mathbf{D}^T \mathbf{U}\|_{\mathcal{G}_l}^a] \right)^2. \end{aligned} \tag{9.23}$$

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

$$\mathbb{E} \left[\left(\sum_{l=1}^L \|\mathbf{D}^T \mathbf{U}\|_{\mathcal{G}_l}^a \right)^2 \right] \leq \mathbb{E} \left[L \sum_{l=1}^L \|\mathbf{D}^T \mathbf{U}\|_{\mathcal{G}_l}^{2a} \right] \leq L^2 K_{a,g}^D \tag{9.24}$$

and by Jensen inequality

$$s - \sum_{l=1}^L \mathbb{E} [\|\mathbf{D}^T \mathbf{U}\|_{\mathcal{G}_l}^a] \geq s - \sum_{l=1}^L \sqrt{\mathbb{E} [\|\mathbf{D}^T \mathbf{U}\|_{\mathcal{G}_l}^{2a}]} \geq s - L\sqrt{K_{a,g}^D} > 0. \tag{9.25}$$

Thus, combining (9.23), (9.24) and (9.25), we get

$$\frac{1}{T} \int_{\{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{D}^T \mathbf{u}\|_{a,g}^a > s\}} \prod_{l=1}^L g(\|\mathbf{D}^T \mathbf{u}\|_{\mathcal{G}_l}) d\mathbf{u} \leq \frac{L^2 K_{a,g}^D}{\left(s - L\sqrt{K_{a,g}^D}\right)^2}.$$

\square

We now turn to the proof of Lemma 9.1

Proof of Lemma 9.1. Let us begin by the proof of inequality (9.21). We have

$$\begin{aligned} \int_{\mathbb{R}^q} \|\mathbf{u}_{\mathcal{G}_1}\|_2^2 p_0(d\mathbf{u}) &= \frac{1}{C_L \sqrt{\det(\mathbf{D}\mathbf{D}^T)}} \int_{\mathbb{B}_{a,g}^a(r_L) \cap \text{Im}(\mathbf{D}^T)} \|\mathbf{u}_{\mathcal{G}_1}\|_2^2 \prod_{l=1}^L e^{-\alpha^a \|\mathbf{u}_{\mathcal{G}_l}\|_2^a} g(\|\mathbf{u}_{\mathcal{G}_l}\|_2) d\mathbf{u} \\ &\leq \frac{1}{C_L \sqrt{\det(\mathbf{D}\mathbf{D}^T)}} \int_{\mathbb{B}_{a,g}^a(r_L) \cap \text{Im}(\mathbf{D}^T)} \|\mathbf{u}_{\mathcal{G}_1}\|_2^2 \prod_{l=1}^L g(\|\mathbf{u}_{\mathcal{G}_l}\|_2) d\mathbf{u}. \end{aligned} \quad (9.26)$$

In the following, we show inequality (9.21) by bounding the right term of inequality (9.26). By Lemma 2.3 and Remark 4.2, we get

$$\begin{aligned} \frac{\int_{\mathbb{B}_{a,g}^a(r_L) \cap \text{Im}(\mathbf{D}^T)} \|\mathbf{u}_{\mathcal{G}_1}\|_2^2 \prod_{l=1}^L g(\|\mathbf{u}_{\mathcal{G}_l}\|_2) d\mathbf{u}}{\sqrt{\det(\mathbf{D}\mathbf{D}^T)} \int_{\mathbb{R}^p} \prod_{l=1}^L g(\|[\mathbf{D}^T \mathbf{u}]_{\mathcal{G}_l}\|_2) d\mathbf{u}} &\leq \frac{\int_{\text{Im}(\mathbf{D}^T)} \|\mathbf{u}_{\mathcal{G}_1}\|_2^2 \prod_{l=1}^L g(\|\mathbf{u}_{\mathcal{G}_l}\|_2) d\mathbf{u}}{\sqrt{\det(\mathbf{D}\mathbf{D}^T)} \int_{\mathbb{R}^p} \prod_{l=1}^L g(\|[\mathbf{D}^T \mathbf{u}]_{\mathcal{G}_l}\|_2) d\mathbf{u}} \\ &= \frac{\int_{\mathbb{R}^p} \|[\mathbf{D}^T \mathbf{u}]_{\mathcal{G}_1}\|_2^2 \prod_{l=1}^L g(\|[\mathbf{D}^T \mathbf{u}]_{\mathcal{G}_l}\|_2) d\mathbf{u}}{\int_{\mathbb{R}^p} \prod_{l=1}^L g(\|[\mathbf{D}^T \mathbf{u}]_{\mathcal{G}_l}\|_2) d\mathbf{u}} \leq K_{1,g}^D. \end{aligned}$$

Then

$$\frac{1}{\sqrt{\det(\mathbf{D}\mathbf{D}^T)}} \int_{\mathbb{B}_{a,g}^a(r_L) \cap \text{Im}(\mathbf{D}^T)} \|\mathbf{u}_{\mathcal{G}_1}\|_2^2 \prod_{l=1}^L g(\|\mathbf{u}_{\mathcal{G}_l}\|_2) d\mathbf{u} \leq K_{1,g}^D T. \quad (9.27)$$

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

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

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

$$\begin{aligned} C_L^{-1} &\leq e^{r_L \alpha^a} \left[T \left(1 - \frac{1}{T} \int_{\{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{D}^T \mathbf{u}\|_{a,g}^a > r_L\}} \prod_{l=1}^L g(\|[\mathbf{D}^T \mathbf{u}]_{\mathcal{G}_l}\|_2) d\mathbf{u} \right) \right]^{-1} \\ &\leq e^{r_L \alpha^a} T^{-1} \left(1 - \frac{L^2 K_{a,g}^D}{(r_L - L\sqrt{K_{a,g}^D})^2} \right)^{-1} \\ &= e^{r_L \alpha^a} T^{-1} \left(1 - \frac{1}{4} \right)^{-1} \leq 2e^{r_L \alpha^a} T^{-1}. \end{aligned} \quad (9.28)$$

Combining (9.27) and (9.28), (9.26) becomes $\int_{\mathbb{R}^q} \|\mathbf{u}_{\mathcal{G}_1}\|_2^2 p_0(d\mathbf{u}) \leq 2K_{1,g}^D e^{r_L \alpha^a}$. That concludes the proof of inequality (9.21) in Lemma 9.1.

Next, we prove inequality (9.22). Remind that $\text{supp}(\pi) = \Theta$, $\text{supp}(p_0^D) = \Theta_{p_0^D}$. By (9.11), we get $\Theta_{p_0^D} \subseteq \Theta$ implying that p_0^D is absolutely continuous w.r.t. π . So $\text{KL}(p_0^D, \pi) < \infty$ which can be bounded. The bound in (9.22) can be proved as follows. By Lemma 2.3, we have

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

where p_0 is a probability measure in \mathbb{R}^q defined in (9.13). We know that $\mathbf{t}^* = D^T \boldsymbol{\theta}^*$, according to the fact that $\|\mathbf{t}_{\mathcal{G}_l}\|_2^a - \|\mathbf{t}_{\mathcal{G}_l} - \mathbf{t}_{\mathcal{G}_l}^*\|_2^a \leq \|\mathbf{t}_{\mathcal{G}_l}^*\|_2^a$ and Assumption (G.4), we get

$$\text{KL}(p_0^D, \pi) \leq \log \left(\frac{C_{\alpha,g,R}}{C_L} \right) + \alpha^a \|D^T \boldsymbol{\theta}^*\|_{a,\mathcal{G}}^a + \lambda \sum_{l=1}^L \log \left\{ h(\|[D^T \boldsymbol{\theta}^*]_{\mathcal{G}_l}\|_2) \right\}. \quad (9.29)$$

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

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

Combining this with the bound of C_L^{-1} in (9.28), we obtain

$$\log \left(\frac{C_{\alpha,g,R}}{C_L} \right) \leq r_L \alpha^a + \log(2) \leq 1 + r_L \alpha^a. \quad (9.30)$$

Inserting (9.30) into (9.29), we get inequality (9.22). This completes the proof. \square

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

References

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

- [2] G. Biau. Analysis of a random forests model. *J. Mach. Learn. Res.*, 13(1):1063–1095, Apr. 2012.
- [3] 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.
- [4] G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.*, 9:2015–2033, June 2008.
- [5] L. Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, Aug. 1996.
- [6] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, Oct. 2001.
- [7] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.
- [8] 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.
- [9] E. J. Candès. Ridgelets: Estimating with Ridge Functions. *Annals of Statistics*, 31, 1999. 1561–1599.
- [10] 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.
- [11] 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.
- [12] C. Chesneau, M. Fadili, and J.-L. Starck. Stein block thresholding for image denoising. *Applied and Computational Harmonic Analysis*, 28(1):67–88, 2010.
- [13] R. Coifman and D. Donoho. Translation invariant de-noising. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*. Springer-Verlag, 1995. 125–150.
- [14] D. Dai, P. Rigollet, and T. Zhang. Deviation optimal learning using greedy q -aggregation. *Ann. Statist.*, 40(3):1878–1905, 06 2012.
- [15] 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.
- [16] 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.
- [17] A. S. Dalalyan and J. Salmon. Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.*, 40(4):2327–2355, 08 2012.
- [18] 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.
- [19] A. S. Dalalyan and A. B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 18(3):914–944, 08 2012.

- [20] 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.
- [21] 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.
- [22] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.
- [23] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, pages 1200–1224, 1995.
- [24] 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.
- [25] A. Durmus and E. Moulines. Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm. Preprint hal-01176132, July 2015.
- [26] A. Durmus, E. Moulines, and M. Pereyra. Sampling from convex non continuously differentiable functions, when Moreau meets Langevin. Preprint hal-01267115, Feb. 2016.
- [27] K. Fang, S. Kotz, and K. Ng. *Symmetric multivariate and related distributions*. Monographs on statistics and applied probability. Chapman and Hall, 1990.
- [28] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256–285, 1995.
- [29] R. Genuer. *Random Forests: elements of theory, variable selection and applications*. Theses, Université Paris Sud - Paris XI, Nov. 2010.
- [30] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, London, 4th edition, 1965.
- [31] J. Huang, J. L. Horowitz, and F. Wei. Variable selection in nonparametric additive models. *Ann. Statist.*, 38(4):2282–2313, 08 2010.
- [32] J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Inf. Comput.*, 132(1):1–63, Jan. 1997.
- [33] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines, 2008.
- [34] G. Lecué. Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.*, 35(4):1698–1721, 08 2007.
- [35] G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006.
- [36] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Inf. Comput.*, 108(2):212–261, Feb. 1994.

- [37] T. D. Luu, J. Fadili, and C. Chesneau. Sampling from non-smooth distribution through Langevin diffusion. Technical report, hal-01492056, May 2017.
- [38] J. Maly. Lectures on change of variables in integral. Preprint 305, Department of Mathematics, University of Helsinki, 2001.
- [39] L. Meier, S. V. D. Geer, P. Bühlmann, and E. T. H. Zehrfich. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 2008.
- [40] L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *Ann. Statist.*, 37(6B):3779–3821, 12 2009.
- [41] A. Nemirovski. Topics in non-parametric statistics, 2000.
- [42] M. Pereyra. Proximal markov chain monte carlo algorithms. *Statistics and Computing*, 26(4):745–760, 2016.
- [43] G. Peyré, J. Fadili, and C. Chesneau. Group sparsity with overlapping partition functions. In *EUSIPCO*, Barcelona, Spain, Aug. 2011.
- [44] 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.
- [45] P. Rigollet and A. Tsybakov. Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, 16(3):260–280, 2007.
- [46] P. Rigollet and A. B. Tsybakov. Sparse estimation by exponential weighting. *Statist. Sci.*, 27(4):558–575, 11 2012.
- [47] G. O. Roberts and R. L. Tweedie. Exponential Convergence of Langevin Distributions and Their Discrete Approximations. *Bernoulli*, 2(4):341–363, 1996.
- [48] R. T. Rockafellar and R. Wets. *Variational analysis*, volume 317. Springer Verlag, 1998.
- [49] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, Nov. 1992.
- [50] R. E. Schapire. The strength of weak learnability. *Mach. Learn.*, 5(2):197–227, July 1990.
- [51] 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.
- [52] 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.
- [53] M. Xuerong. *Stochastic differential equations and applications*. Woodhead Publishing, 2007.
- [54] Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10(1):25–47, 2004.
- [55] 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.