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

Tung Duy Luu, Jalal Fadili, Christophe Chesneau

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

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

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

The idea of aggregating elements in a dictionary has been introduced in machine learning to combine different techniques (see [40, 55]) with some procedures such as bagging [8], boosting [28, 52] and random forests [1, 5–7, 9, 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 $\widetilde{\boldsymbol{\theta}}$ is assumed much smaller than n. This makes it possible to build an estimate $f_{\widetilde{\boldsymbol{\theta}}}$ with good provable performance guarantees under appropriate conditions on the dictionary and noise.

1.2 Overview of previous work

1.2.1 Oracle inequalities

This type of guarantees dates back, for instance, to the work [22–24] on orthogonal wavelet thresholding estimators. Oracle inequalities, which are at the heart of our work, quantify the quality of an estimator compared to the best possible one that could only be given with an oracle. These inequalities are well adapted under the sparsity scenario. Formally, let $g : \mathcal{X} \to \mathbb{R}$, and denote

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

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

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

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

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

where $C \geq 1$ and the remainder term $\Delta_{n,M}(\boldsymbol{\theta})$ depends on the performance of the estimator, the complexity of $\boldsymbol{\theta}$, the dimension M and the sample size n. Such type of inequality is called balanced oracle inequality. Under the sparsity scenario, the complexity of $\boldsymbol{\theta}$ is characterized by the number of its non-zero components, in which case inequalities of type (1.2) 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"), $\Delta_{n,M}(\boldsymbol{\theta})$ is small even if $n \ll M$ and decreases rapidly to 0 as $n \to +\infty$. Besides, the choice of Θ is crucial: on the one hand, a non suitable choice can lead a large bias term in (1.2). On the other hand, if Θ is too complex, the remainder term becomes large. Then, a suitable choice for Θ must achieve a good bias-complexity trade-off.

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

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

where $\mathrm{pen} : \mathbb{R}^M \to \mathbb{R}^+$ is a sparsity promoting penalty function, see e.g. [10]. Our work focuses on EWA approach that we briefly describe now.

1.2.2 Exponential Weighted Aggregation (EWA)

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

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

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

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

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

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

This idea was initially proposed in [34, 40, 55] with a uniform prior on a finite set Λ.

In the literature, the fact that the elements in the dictionary may be either deterministic or random splits the EWA approach into two corresponding cases. The deterministic case is considered in [15, 16, 18, 20]. 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 $O\big(\|\boldsymbol{\theta}\|_0 \log(M)/n\big)$, which scales linearly with the sparsity level and increases in M only logarithmically.

The random case is tackled in [44]. The initial idea is to obtain two independent samples from the initial sample by randomization or sample splitting (see [38, 47, 57]). 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 [39]. More recent work such as [17] generalizes the idea where the pre-estimators are affine and the priors are continuous.

1.2.3 Generalization of sparsity assumption

Analysis sparsity Let $P \geq M$ and $\boldsymbol{D} : \mathbb{R}^M \to \mathbb{R}^P$ be a linear analysis operator. The analysis sparsity assumption means that $\|\boldsymbol{D}\widetilde{\boldsymbol{\theta}}\|_0 \ll n$. A typical example is total variation [51] where the operator $\boldsymbol{D}$ corresponds to "finite differences" (i.e., $(\boldsymbol{D\theta})_1 = \theta_1$, $(\boldsymbol{D\theta})_j = \theta_j - \theta_{j-1}$, $\forall j \geq 2$). Another example is the fused Lasso [54] where $\boldsymbol{D}$ is a positive combination of the identity and finite differences.

Group sparsity Group sparsity corresponds to saying that the aggregator $\boldsymbol{\theta}$ is block sparse, see Section 4.1 for formal details. Group sparsity is at the heart of the group Lasso and related methods [32, 36, 42, 43, 46, 58]. In the EWA context, the group sparsity prior is considered in [48] as an application of the aggregation of orthogonal projectors.

1.3 Contributions

Our main contributions are summarized as follows:

- We propose an EWA estimator, with a deterministic dictionary, under a group-analysis sparsity prior (see Section 1.2.3). More precisely, we assume that $\boldsymbol{D}$ is the analysis operator corresponding to a frame of $\mathbb{R}^M$, and thus is not necessarily invertible unlike previous work. In addition, our prior class (see (5.2)) 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 $g : \mathbb{R}^+ \to \mathbb{R}^+$ that satisfies mild conditions (see Assumption 5.2).
- We establish a group-analysis SOI where the remainder term depends on the number of active groups in $\boldsymbol{D\theta}$ and on the function g (see Theorem 6.1)
- 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 $O\Big(\big\|\boldsymbol{D\theta}\big\|_{0,\mathcal{G}} \log(L)/n\Big)$, where $\big\|\boldsymbol{D\theta}\big\|_{0,\mathcal{G}}$ is the number of active groups in $\boldsymbol{D\theta}$, and L is the total number of groups (see Corollary 6.2 and Remark 6.2). This rate coincides with the classical one $O\big(\big\|\boldsymbol{\theta}\big\|_0 \log(M)/n\big)$ under the sparsity scenario, i.e. $\boldsymbol{D} = \boldsymbol{I}_M$ and $L = M$.
- 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. Section 4 consists in specifying our framework, i.e., group-analysis sparsity scenario where the analysis operator $\boldsymbol{D}$ constitute a frame. In Section 5, we describe our EWA procedure after specifying the aggregation dictionary and our prior family. In Section 6, we establish our main results, namely group-analysis SOI. Section 7 is devoted to the forward-backward proximal LMC algorithm that implements EWA, and the numerical experiments on several numerical settings are described in Section 8. The proofs of all results are collected in Section 10.

2 Preliminaries

Throughout the paper, we will use the following notations.

For each function $h : \mathcal{X} \to \mathbb{R}$, we define $\boldsymbol{h} = (h(x_1), \dots, h(x_n))^T$ its n-sample vector form and

$$\big\|h\big\|_n = \left(\frac{1}{n}\sum_{j=1}^{n} h^2(x_j)\right)^{1/2}$$

its empirical norm. Moreover, for each function $h : \mathbb{R}^M \to \mathbb{R}$, we denote

$$\big\|h\big\|_\infty = \sup_{\boldsymbol{x}\in\mathbb{R}^M} \big|h(\boldsymbol{x})\big|.$$

For all $\boldsymbol{v} \in \mathbb{R}^d$, $d \geq 1$, we also define the ℓ_p function

$$\left\|\boldsymbol{v}\right\|_p = \left(\sum_{j=1}^{d} \left|v_j\right|^p\right)^{1/p}, \quad p > 0,$$

with the usual adaptation $\left\|\boldsymbol{v}\right\|_\infty = \max_{j\in\{1,\dots,d\}} \left|v_j\right|$. It is a norm for $p \geq 1$, and a quasi-norm for $p \in]0,1[$. $\left\|\boldsymbol{v}\right\|_0$ is the ℓ_0 pseudo-norm which counts the number of non-zero elements in $\boldsymbol{v}$.

Let $\boldsymbol{A} \in \mathbb{R}^{p\times q}$, we set $\boldsymbol{\sigma}(\boldsymbol{A}) = (\sigma_1(\boldsymbol{A}), \dots, \sigma_q(\boldsymbol{A}))^T \in \mathbb{R}^q$ be the vector of singular values of $\boldsymbol{A}$ in non-increasing order. Note that, when $\boldsymbol{A}$ is symmetric semi-definite positive, $\boldsymbol{\sigma}(\boldsymbol{A})$ is also the ordered vector or positive eigenvalues of $\boldsymbol{A}$. The spectral norm of $\boldsymbol{A}$ is

$$|||\boldsymbol{A}||| = \sqrt{\sigma_1(\boldsymbol{A}^T\boldsymbol{A})}.$$

For a set Ω, I_Ω is its characteristic function, i.e. 1 if the argument is in Ω and 0 otherwise.

Recall the Gamma function $\Gamma :]0,+\infty] \to]0,+\infty]$

$$\Gamma(t) = \int_0^\infty x^{t-1} \exp(-x) dx.$$

The two 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*

$$\int_0^\infty \frac{x^{\gamma-1}}{(p+x^\nu)^{\eta+1}} dx = \frac{1}{\nu p^{\eta+1-\gamma/\nu}} \frac{\Gamma(\gamma/\nu)\Gamma(1+\eta-\gamma/\nu)}{\Gamma(1+\eta)}, \tag{2.1}$$

otherwise this integral is not definite.

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

$$\int_{\mathbb{R}^d} h(\left\|\boldsymbol{u}\right\|_2) d\boldsymbol{u} = C_d \int_0^\infty x^{d-1} h(x) dx, \tag{2.2}$$

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

3 PAC-Bayesian type oracle inequalities

This section recalls a PAC-Bayesian type oracle inequality which holds for the EWA procedure of type (1.3)-(1.4) with any deterministic aggregation dictionary, any prior and a large class of noises. Such type of oracle inequalities was introduced in [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.

Assumption 3.1 (Standard assumption). *The noise vector $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^T$ has zero mean.*

Assumption 3.2 (Main assumption). *For any $\gamma > 0$ small enough, there exist a probability space and two random variables $\boldsymbol{\xi}'$ and $\boldsymbol{\zeta}$ defined on this probability space satisfying the three following points:*

1. *$\boldsymbol{\xi}'$ has the same distribution as $\boldsymbol{\xi}$.*
2. *$\boldsymbol{\xi}' + \boldsymbol{\zeta}$ has the same distribution as $(1+\gamma)\boldsymbol{\xi}'$ and the conditional expectation satisfies $\mathbb{E}\left[\boldsymbol{\zeta}|\boldsymbol{\xi}'\right] = 0$.*
3. *There exist $t_0 \in (0, \infty]$ and a bounded Borel function $v : \mathbb{R}^M \to \mathbb{R}^+$ such that*

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

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

Assumption 3.2 is based on [20, Assumption N], and can be shown to be fulfilled for a large class of noise such as Gaussian, Laplace and any bounded symmetric noise.

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

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

Note that (3.1) is always satisfied since H is allowed to be infinite. However, for the sake of sharpness in our theoretical results, we wish to choose H as small as possible. We are now ready to state the PAC-Bayesian type oracle inequalities.

Theorem 3.1. *Let Assumptions 3.1 and 3.2 be satisfied with some function v and let (3.1) holds. Then for any prior π, any probability measure p on Λ and any $\beta \geq \max(4\|v\|_\infty, 2H/t_0)$ or $\beta \geq 4\|v\|_\infty$ when $H = \infty$, $t_0 = \infty$, we have*

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

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

The proof of Theorem 3.1 is a mild adaptation of the original one in [20, Section 2], where we used directly Assumption 3.2-3 in the vector $\boldsymbol{\zeta}$ instead of splitting it into $\zeta_{i, i \in \{1, \ldots, n\}}$ (that are no longer i.i.d.).

Related work The work of [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 3.2) 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 $\xi_i \sim \mathcal{N}(0, \sigma_i^2)$, $i = 1, \ldots, n$, the following aggregate was proposed in [17]:

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

where $\widehat{f}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda} \in \Lambda$ are affine estimators satisfying some conditions imposed in [17, Theorem 1] which yield the definition of $\widehat{r}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda} \in \Lambda$. This aggregate satisfies oracle inequalities defined therein which are the counterparts of (3.2) for the aggregation of estimators. In addition, in the case of random design (i.e., $x_1, \ldots, x_n$ 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 Group-analysis sparsity

4.1 Group 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 $P \geq M$. We partition the index set $\{1, \dots, P\}$ into L non-overlapping groups/blocks of indices $\{\mathcal{G}_l\}_{1 \leq l \leq L}$ such that

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

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

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

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

Assumption 4.1 (Group-analysis sparsity assumption)**.** *Let $\boldsymbol{D} : \mathbb{R}^M \to \mathbb{R}^P$ be a linear analysis operator. There exists $\widetilde{\boldsymbol{\theta}} \in \mathbb{R}^M$ such that $f_{\widetilde{\boldsymbol{\theta}}}$ is close to f and $\|\boldsymbol{D}\widetilde{\boldsymbol{\theta}}\|_{0,\mathcal{G}} \ll n$.*

In plain words, Assumption 4.1 says that the number of active groups of $\boldsymbol{D}\widetilde{\boldsymbol{\theta}}$ is much smaller than the sample size. Note that this is a strict notion of group-analysis sparsity, and a weaker one could be also considered where most $\left(\boldsymbol{D}\widetilde{\boldsymbol{\theta}}\right)_{\mathcal{G}_l}$ are nearly zero.

4.2 Frame assumption

Let $\boldsymbol{D} : \mathbb{R}^M \to \mathbb{R}^P$ be a linear operator seen as a matrix in $\mathbb{R}^{P \times M}$. We assume the following throughout this paper.

Assumption 4.2 (Frame assumption)**.** *$\boldsymbol{D}$ corresponds to the analysis operator of a finite frame of $\mathbb{R}^M$, i.e. there exist two constants ν and μ with $\nu \geq \mu > 0$, called frame bounds, such that the generalized Parseval relation is satisfied*

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

By the Courant-Fischer theorem, Assumption 4.2 is equivalent to the fact that μ (resp. ν) is a lower (resp. upper) bound of the eigenvalues of $\boldsymbol{D}^T\boldsymbol{D}$. Moreover, since $\mu > 0$, we have that

$$\boldsymbol{D}^T\boldsymbol{D} \text{ is bijective and } \boldsymbol{D} \text{ is injective.}$$

The frame is said tight when $\mu = \nu$. Typical examples of (tight) frames that have been used in statistics are translation invariant wavelets [14], ridgelets [12] and curvelets [11].

Let $\widetilde{\boldsymbol{D}^T} \in \mathbb{R}^{M\times P}$ be the matrix whose columns form the canonical dual frame associated to $\boldsymbol{D}^T$. We know that

$$\widetilde{\boldsymbol{D}^T}\boldsymbol{D} = \boldsymbol{I}_M \tag{4.1}$$

and

$$\frac{1}{\mu} \geq \sigma_1\left(\widetilde{\boldsymbol{D}^T}\widetilde{\boldsymbol{D}^T}^T\right) \geq \cdots \geq \sigma_M\left(\widetilde{\boldsymbol{D}^T}\widetilde{\boldsymbol{D}^T}^T\right) \geq \frac{1}{\nu}. \tag{4.2}$$

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.

Under the frame Assumption 4.2, 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 6.1).

Lemma 4.1. *Let $\Theta \subseteq \mathbb{R}^M$ be a measurable set. For any $\boldsymbol{D} \in \mathbb{R}^{P\times M}$ satisfying Assumption 4.2, let $u : \mathbb{R}^P \to \mathbb{R}$ such that the mapping $\boldsymbol{\theta} \mapsto u(\boldsymbol{D\theta})$ is measurable on Θ. We have*

$$\int_\Theta u(\boldsymbol{D\theta})d\boldsymbol{\theta} = \frac{1}{\sqrt{\det(\boldsymbol{D}^T\boldsymbol{D})}}\int_{\boldsymbol{D}\Theta} u(\boldsymbol{v})d\boldsymbol{v} \tag{4.3}$$

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

Though quite natural, proving Lemma 4.1 rigorously requires nontrivial arguments from geometric measure theory (see Section 10.1.1 for details).

5 EWA estimator

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.

5.1 Choice of dictionary

Let $\boldsymbol{X} \in \mathbb{R}^{n\times M}$ where $\boldsymbol{X}_{i,j} = f_j(x_i)$, $f_j \in \mathcal{H}$. We impose the following standard normalization assumption on $\boldsymbol{X}$.

Assumption 5.1. *$\boldsymbol{X}$ is normalized such that all the diagonal entries of $\boldsymbol{X}^T\boldsymbol{X}/n$ are 1.*

Now, let us introduce our dictionary of aggregation:

$$\mathcal{F}_\Theta = \left\{ f_{\boldsymbol{\theta}} = \ell\left(\sum_{j=1}^M \theta_j f_j\right) : \boldsymbol{\theta} \in \Theta \right\}, \tag{5.1}$$

where $\Theta = \left\{\boldsymbol{\theta} \in \mathbb{R}^M : \left\|\boldsymbol{D\theta}\right\|_{a,\mathcal{G}}^a \leq R\right\}$, $a \in]0,1]$, $R \in]0,+\infty]$ and $\ell : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable and known depending on the regression problem (for example: $\ell(x) = e^x$ for the exponential regression, $\ell(x) = e^x/(e^x+1)$ for the logistic regression and $\ell(x) = x$ for the linear regression). This dictionary of aggregation is similar to the one proposed in [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 $\mathcal{F}_\Theta$ in (5.1) is established in the following result.

Proposition 5.1. *Let $\mathcal{F}_\Theta = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ defined in (5.1) with some $R > 0$, $a \in]0, 1]$ and $\ell : \mathbb{R} \to \mathbb{R}$ twice continuously differentiable. Let Assumption 4.2 holds for some $\mu > 0$. We get that*

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

where $\mathcal{B} = \{\boldsymbol{x} \in \mathbb{R}^n : \left\|\boldsymbol{x}\right\|_2 \leq |||\boldsymbol{X}||| R^{1/a}/\sqrt{\mu}\}$ and
$\boldsymbol{L} : \boldsymbol{x} \in \mathbb{R}^n \to (\ell(x_1), \ldots, \ell(x_n))$.

From Proposition 5.1, one can choose $H = 2 \max_{\boldsymbol{x}\in\mathcal{B}} \left\|\boldsymbol{L}(\boldsymbol{x})\right\|_2$.

5.2 Choice of prior

We choose a general prior of the form

$$\pi(d\boldsymbol{\theta}) = \frac{1}{C_{\alpha,g,R}} \prod_{l=1}^{L} \exp\left(-\alpha^a \left\|[\boldsymbol{D\theta}]_{\mathcal{G}_l}\right\|_2^a\right) g\left(\left\|[\boldsymbol{D\theta}]_{\mathcal{G}_l}\right\|_2\right) I_\Theta(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{5.2}$$

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

Assumption 5.2.

1. *$g : \mathbb{R}^+ \to \mathbb{R}^+$ is a bounded function such that $g \not\equiv 0$,*
   $\boldsymbol{\theta} \mapsto g(\left\|[\boldsymbol{D\theta}]_{\mathcal{G}_l}\right\|_2)$ is measurable on $\mathbb{R}^M$, $\forall l \in \{1, \ldots, L\}$.
2. *The integrability condition:*

$$\int_{\mathbb{R}^M} \prod_{l=1}^{L} g(\left\|[\boldsymbol{Du}]_{\mathcal{G}_l}\right\|_2) d\boldsymbol{u} < \infty. \tag{5.3}$$

3. *The moment condition:*

$$\int_{\mathbb{R}^M} \left\|[\boldsymbol{Du}]_{\mathcal{G}_l}\right\|_2^2 \prod_{k=1}^{L} g(\left\|[\boldsymbol{Du}]_{\mathcal{G}_k}\right\|_2) d\boldsymbol{u} < \infty, \quad \forall l \in \{1, \ldots, L\}. \tag{5.4}$$

4. *There exist $\lambda \geq 0$ and a function $h : \mathbb{R}^+ \to \mathbb{R}^+$ such that $\forall (\boldsymbol{t}, \boldsymbol{t}^*) \in \mathbb{R}^G \times \mathbb{R}^G$,*

$$\frac{g(\left\|\boldsymbol{t} - \boldsymbol{t}^*\right\|_2)}{g(\left\|\boldsymbol{t}\right\|_2)} \leq h(\left\|\boldsymbol{t}^*\right\|_2)^\lambda.$$

Assumptions 5.2-2 and 5.2-3 lead the following remark that is a core part for the construction of the general group-analysis SOI in Theorem 6.1.

Remark 5.1. *Let $G \geq 1$ and $\boldsymbol{D} \in \mathbb{R}^{P\times M}$ satisfying Assumption 4.2. For any function g satisfying Assumptions 5.2-1, 5.2-2 and 5.2-3, and any $a \in]0, 1]$, there exists $K_{a,g} \in]0, \infty[$ such that, $\forall l \in \{1, \ldots, L\}$,*

$$\frac{\int_{\mathbb{R}^M} \left\|[\boldsymbol{Du}]_{\mathcal{G}_l}\right\|_2^{2a} \prod_{k=1}^{L} g(\left\|[\boldsymbol{Du}]_{\mathcal{G}_k}\right\|_2) d\boldsymbol{u}}{\int_{\mathbb{R}^M} \prod_{k=1}^{L} g(\left\|[\boldsymbol{Dv}]_{\mathcal{G}_k}\right\|_2) d\boldsymbol{v}} \leq K_{a,g}. \tag{5.5}$$

Proof. Remark 5.1 is briefly proved as follows. From Assumption 5.2-3 and the fact that $g \not\equiv 0$, one can show that (5.5) holds for $a = 1$. Moreover, since $g \not\equiv 0$ and g satisfies Assumption 5.2-2, we have

$$\boldsymbol{u} \to \frac{\prod_{l=1}^{L} g(\left\|[\boldsymbol{D}\boldsymbol{u}]_{\mathcal{G}_l}\right\|_2) d\boldsymbol{u}}{\int_{\mathbb{R}^M} \prod_{k=1}^{L} g(\left\|[\boldsymbol{D}\boldsymbol{v}]_{\mathcal{G}_k}\right\|_2) d\boldsymbol{v}}$$

is a probability measure. Therefore, (5.5) holds for any a in $]0, 1]$ by Hölder's inequality. □

At first glance, Assumptions 5.2-2 and 5.2-3 may seem cumbersome. However the following lemma gives a simple condition on g that implies them.

Lemma 5.1. *Let $G \geq 1$ and $\boldsymbol{D} \in \mathbb{R}^{P \times M}$ satisfying Assumption 4.2. Suppose that g satisfies Assumption 5.2-1 and*

$$\int_0^\infty z^{G+1} g(z) dz < \infty. \tag{5.6}$$

Then Assumptions 5.2-2 and 5.2-3 are in force.

In the two following remarks, we consider the case where $\boldsymbol{D}$ is invertible. Remark 5.2 provides a simple and explicit form for $K_{a,g}$ and Remark 5.3 shows that condition (5.6) is necessary for g to obey Assumption 5.2.

Remark 5.2. *Let $G \geq 1$ and $\boldsymbol{D} \in \mathbb{R}^{M \times M}$ be invertible. For any function g satisfying Assumptions 5.2-1, 5.2-2 and 5.2-3, and for any $a \in]0, 1]$, one can choose $K_{a,g}$ in (5.5) as*

$$K_{a,g} = \frac{\int_0^\infty x^{G-1+2a} g(x) dx}{\int_0^\infty z^{G-1} g(z) dz}. \tag{5.7}$$

Proof. The proof follows by combining Lemmas 4.1 and 2.2, i.e.,

$$\begin{aligned} \frac{\int_{\mathbb{R}^M} \left\|[\boldsymbol{D}\boldsymbol{u}]_{\mathcal{G}_l}\right\|_2^{2a} \prod_{k=1}^{L} g(\left\|[\boldsymbol{D}\boldsymbol{u}]_{\mathcal{G}_k}\right\|_2) d\boldsymbol{u}}{\int_{\mathbb{R}^M} \prod_{k=1}^{L} g(\left\|[\boldsymbol{D}\boldsymbol{v}]_{\mathcal{G}_k}\right\|_2) d\boldsymbol{v}} &= \frac{\int_{\mathbb{R}^G} \left\|\boldsymbol{u}\right\|_2^{2a} g(\left\|\boldsymbol{u}\right\|_2) d\boldsymbol{u}}{\int_{\mathbb{R}^G} g(\left\|\boldsymbol{v}\right\|_2) d\boldsymbol{v}} \\ &= \frac{\int_0^\infty x^{G-1+2a} g(x) dx}{\int_0^\infty y^{G-1} g(y) dy}. \end{aligned}$$

□

Remark 5.3. *When $\boldsymbol{D}$ is invertible, if g does not satisfy (5.6) then g cannot fulfill Assumption 5.2-3. Consequently, Assumption 5.2-3 and condition (5.6) are equivalent in the invertible case.*

Proof. By Lemmas 4.1 and 2.2, we get

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

□

Let us now discuss some choices of g. The goal is to find a prior leading an oracle inequality with a small remainder term while promoting group-analysis sparsity. Namely, g should be peaked around 0 with heavy tails.

Example 5.1. *Consider $g : \mathbb{R}^+ \to \mathbb{R}^+$ defined by*

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

This choice of g yields a prior that specializes to the one in [20] for the individual sparsity scenario, i.e. with $\boldsymbol{D} = \boldsymbol{I}_M$, $G = 1$ and $a = 1$.

Example 5.2. *Consider $g : \mathbb{R}^+ \to \mathbb{R}^+$ defined by*

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

where $\tau > 0$, $b \in]0, 1]$ and $c > (2+G)/b$. The choice of c guarantees the validity of Assumption 5.2. Thanks to the parameters b and c, this choice of g offers more flexibility than the one of the previous example. This allows for example to optimize the performance of EWA by tuning these parameters for the particular dataset at hand.

6 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 (5.1) and the prior (5.2) with a function g obeying Assumptions 5.2-1 to 4.

Theorem 6.1 (General group-analysis sparse oracle inequality)**.** *Let $G \geq 1$ and $\boldsymbol{D}$ satisfying Assumption 4.2 with $\mu > 0$. Let Assumptions 3.1 and 3.2 be satisfied with some function v, (3.1) holds and $\beta \geq \max\left(4\|v\|_\infty, 2H/t_0\right)$. For some $a \in]0, 1]$, take the dictionary (5.1) and the prior (5.2) with g satisfying Assumptions 5.2-1 to 4. Let $K_{a,g}$, as defined in (5.5), and assume that $R > 3\sqrt{K_{a,g}}L$. Then the following group-analysis SOI holds,*

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{f}_n - f\|_n^2\right] \leq & \inf_{\boldsymbol{\theta}\in\mathbb{R}^M : \|\boldsymbol{D}\boldsymbol{\theta}\|_{a,\mathcal{G}}^a \leq R - 3\sqrt{K_{a,g}}L} \Bigg[\|f_{\boldsymbol{\theta}} - f\|_n^2 \\
& + \frac{\beta}{n}\left(1 + 3\sqrt{K_{a,g}}L\alpha^a + \alpha^a \|\boldsymbol{D}\boldsymbol{\theta}\|_{a,\mathcal{G}}^a\right) \\
& + \frac{\lambda\beta}{n}\sum_{l=1}^{L} \log\left\{h\left(\|[\boldsymbol{D}\boldsymbol{\theta}]_{\mathcal{G}_l}\|_2\right)\right\}\Bigg] \\
& + \frac{2K_{1,g}e^{3\sqrt{K_{a,g}}L\alpha^a} M C_{f,\ell}}{\mu},
\end{aligned} \tag{6.1}$$

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

See Section 10.2.1 for the proof.

In this group-analysis SOI, the EWA estimator $\widehat{f}_n$ mimics the best aggregate $f_{\boldsymbol{\theta}}$ for all possible weights belonging to $\left\{\boldsymbol{\theta} \in \mathbb{R}^M : \|\boldsymbol{D}\boldsymbol{\theta}\|_{a,\mathcal{G}}^a \leq R - 3\sqrt{K_{a,g}}L\right\}$. The remainder term of the inequality (6.1) depends

on the choice of g via the function h and $K_{a,g}$. Theorem 6.1 has several consequences that we report now, for instance with the choices of g in Example 5.1 and 5.2.

We first consider the prior (5.2) in Example 5.1, under the individual sparsity scenario ($\boldsymbol{D} = \boldsymbol{I}_M$, $G = 1$) and the choice $a = 1$ (i.e., $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^M : \|\boldsymbol{\theta}\|_1 \leq R\}$). This is the setting considered in [20]. We obtain the following SOI as a corollary of our main result.

Corollary 6.1. *Let $\boldsymbol{D} = \boldsymbol{I}_M$ and fix $G = 1$. Suppose that Assumptions 3.1 and 3.2 hold with some function v, (3.1) holds and $\beta \geq \max\left(4\|v\|_\infty, 2H/t_0\right)$. Fix $a = 1$, take the dictionary (5.1) and the prior (5.2) with g defined in Example 5.1 and $\alpha \leq 1/(3M\tau)$. Assume that $R > 3M\tau$. Then,*

$$\mathbb{E}\left[\|\widehat{f}_n - f\|_n^2\right] \leq \inf_{\boldsymbol{\theta}\in\mathbb{R}^M : \|\boldsymbol{\theta}\|_1 \leq R - 3M\tau} \left[\|f_{\boldsymbol{\theta}} - f\|_n^2 + \frac{2\beta}{n}\left(1 + \frac{1}{6M\tau}\|\boldsymbol{\theta}\|_1\right) + \frac{4\beta}{n}\sum_{j=1}^{M} \log\left\{1 + \frac{|\theta_j|}{\tau}\right\}\right] + 2\tau^2 C_{f,\ell} M e, \tag{6.2}$$

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

SOI (6.2) is similar to the one in [20, Theorem 2]. By choosing $\tau^2 \sim 1/(Mn)$ and $R \sim M\tau$, its remainder term is of order $O\big(\|\boldsymbol{\theta}\|_0 \log(M)/n\big)$ which is the classical rate under the sparsity scenario. 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 6.1. *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*

$$\int_0^\infty \frac{x^{G+1}}{(\tau^2 + x^2)^2} dx \text{ is not definite.}$$

Consequently, condition (5.6) is not fulfilled with g defined in Example 5.1 when $G \geq 2$.

According to Remark 5.3, Remark 6.1 implies that Assumption 5.2 is not fulfilled for g in Example 5.1 when the group size $G \geq 2$ and $\boldsymbol{D}$ invertible. Thus one cannot construct a group-analysis SOI from Theorem 6.1 to guarantee the quality of the corresponding estimator. Overcoming this limitation was yet another motivation behind the choice of g in Example 5.2, which turns out to work well under the group-analysis sparsity scenario. In a nutshell, an aggregate with g in Example 5.2 exhibits the group-analysis SOI defined in the following corollary with any $G \geq 1$, any $\boldsymbol{D} \in \mathbb{R}^{P\times M}$ satisfying Assumption 4.2 and any $a \in]0, 1]$.

Corollary 6.2. *Let $G \geq 1$ and $\boldsymbol{D}$ satisfying Assumption 4.2 with $\mu > 0$. Let Assumptions 3.1 and 3.2 be satisfied with some function v, (3.1) holds and $\beta \geq \max\left(4\|v\|_\infty, 2H/t_0\right)$. Take the dictionary (5.1) and the prior (5.2) with $a \in]0, 1]$, $\alpha \geq 0$ and g defined in Example 5.2. We get that g satisfies Assumption 5.2. Then,*

let $K_{a,g}$ as defined in (5.5), and assume that $R > 3\sqrt{K_{a,g}}L$. Then the following group-analysis SOI holds,

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{f}_n - f\right\|_n^2\right] \leq & \inf_{\boldsymbol{\theta}\in\mathbb{R}^M:\left\|\boldsymbol{D}\boldsymbol{\theta}\right\|_{a,\mathcal{G}}^a \leq R-3\sqrt{K_{a,g}}L} \Bigg[\left\|f_{\boldsymbol{\theta}} - f\right\|_n^2 \\ & + \frac{\beta}{n}\left(1 + 3\sqrt{K_{a,g}}L\alpha^a + \alpha^a\left\|\boldsymbol{D}\boldsymbol{\theta}\right\|_{a,\mathcal{G}}^a\right) \\ & + \frac{c\beta}{n}\sum_{l=1}^{L}\log\left(1 + \left[\frac{\left\|[\boldsymbol{D}\boldsymbol{\theta}]_{\mathcal{G}_l}\right\|_2}{\tau}\right]^b\right)\Bigg] \\ & + \frac{2K_{1,g}e^{3\sqrt{K_{a,g}}L\alpha^a}MC_{f,\ell}}{\mu}, \end{aligned} \tag{6.3}$$

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

To get an explicit control of the remainder term, it is instructive to have a closed-form of $K_{a,g}$. This can be done for instance when $\boldsymbol{D}$ is invertible, see (5.7). The obtained group-analysis SOI is stated as follows.

Corollary 6.3. *Consider the same framework as Corollary 6.2 with $\boldsymbol{D}$ invertible. For $a \in]0,1]$, let $\widetilde{K}_{a,g} = \frac{\Gamma((2a+G)/b)\Gamma(c-(2a+G)/b)}{\Gamma(G/b)\Gamma(c-G/b)}$, and set $\alpha \leq 1/\left(3\tau^a\sqrt{\widetilde{K}_{a,g}}L\right)^{1/a}$. Then we obtain the following group-analysis SOI*

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{f}_n - f\right\|_n^2\right] \leq & \inf_{\boldsymbol{\theta}\in\mathbb{R}^M:\left\|\boldsymbol{D}\boldsymbol{\theta}\right\|_{a,\mathcal{G}}^a \leq R-3\tau^a\sqrt{\widetilde{K}_{a,g}}L} \Bigg[\left\|f_{\boldsymbol{\theta}} - f\right\|_n^2 \\ & + \frac{c\beta}{n}\sum_{l=1}^{L}\log\left(1 + \left[\frac{\left\|[\boldsymbol{D}\boldsymbol{\theta}]_{\mathcal{G}_l}\right\|_2}{\tau}\right]^b\right) \\ & + \frac{\beta}{n}\left(2 + \frac{\left\|\boldsymbol{D}\boldsymbol{\theta}\right\|_{a,\mathcal{G}}^a}{3L\tau^a\sqrt{\widetilde{K}_{a,g}}}\right)\Bigg] + \frac{2C_{f,\ell}M\tau^2\widetilde{K}_{1,g}e}{\mu}, \end{aligned} \tag{6.4}$$

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

For an efficient aggregate, the parameters should be chosen such that the remainder term is small and decreases rapidly to 0 as $n \to \infty$. A suitable choice is described in the following remark.

Remark 6.2. *Set $\tau^2 \sim 1/(Mn)$ and $R \sim L\tau^a$, we have*

$$\begin{aligned}\frac{\beta}{n}\left(2 + \frac{\left\|\boldsymbol{D}\boldsymbol{\theta}\right\|_{a,\mathcal{G}}^a}{3L\tau^a\sqrt{\widetilde{K}_{a,g}}}\right) + \frac{2C_{f,\ell}M\tau^2\widetilde{K}_{1,g}e}{\mu} &\leq \frac{\beta}{n}\left(2 + \frac{R}{3L\tau^a\sqrt{\widetilde{K}_{a,g}}}\right) + \frac{2C_{f,\ell}M\tau^2\widetilde{K}_{1,g}e}{\mu} \\ &= O\left(\frac{1}{n}\right),\end{aligned}$$

and

$$\frac{c\beta}{n}\sum_{l=1}^{L}\log\left(1+\left[\frac{\left\|[\boldsymbol{D\theta}]_{\mathcal{G}_l}\right\|_2}{\tau}\right]^b\right) \leq \frac{c\beta}{n}\sum_{l=1}^{L}\log\left(1+\left[\frac{R^{1/a}}{\tau}\right]^b\right) = O\left(\frac{\left\|\boldsymbol{D\theta}\right\|_{0,\mathcal{G}}\log(L)}{n}\right). \tag{6.5}$$

In (6.5), $\left\|\boldsymbol{D\theta}\right\|_{0,\mathcal{G}}$ is small for instance when $\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}$, by the group-analysis sparsity assumption (R must be sufficiently large to cover $\widetilde{\boldsymbol{\theta}}$).

From Remark 6.2, the aggregate with g in Example 5.2 performs well under the group-analysis sparsity scenario. Under the sparsity scenario, this rate becomes $O(\|\boldsymbol{\theta}\|_0\log(M)/n)$ which is the same rate as the aggregate with g in Example 5.1.

7 Forward-Backward proximal LMC algorithm

The goal of this section is to implement our EWA estimator (with g defined in Example 5.2) 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

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

where $\boldsymbol{X}\in\mathbb{R}^{n\times M}$ is the design matrix described in Section 5.1. We also consider the EWA estimator

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

with the prior (5.2), i.e.

$$\mu_n(d\boldsymbol{\theta}) \propto \exp\left(\frac{-\|\boldsymbol{y}-\boldsymbol{X\theta}\|_2^2}{\beta}\right)\pi(d\boldsymbol{\theta}). \tag{7.1}$$

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 by constructing an appropriate Markov chain whose stationary distribution is μ, 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 [49].

7.1 The Langevin diffusion

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

$$d\boldsymbol{L}(t) = \frac{1}{2}\boldsymbol{\rho}(\boldsymbol{L}(t))dt + d\boldsymbol{W}(t),\ t>0,\ \boldsymbol{L}(0)=\boldsymbol{l}_0, \tag{7.2}$$

where $\boldsymbol{\rho}=\nabla\log\nu$, ν is everywhere non-zero and suitably smooth target density function on $\mathbb{R}^M$, $\boldsymbol{W}$ is a M-dimensional Brownian process and $\boldsymbol{l}_0\in\mathbb{R}^M$ is the initial value. Under mild assumptions, the SDE

(7.2) has a unique strong solution and, moreover, $\boldsymbol{L}(t)$ has a stationary distribution with density precisely ν [49, Theorem 2.1]. $\boldsymbol{L}(t)$ is therefore interesting for sampling from ν. In particular, this opens the door to approximating integrals $\int_{\mathbb{R}^M} F(\boldsymbol{\theta})\nu(\boldsymbol{\theta})d\boldsymbol{\theta}$ by the average value of a Langevin diffusion, i.e., $\frac{1}{T}\int_0^T F(\boldsymbol{L}(t))dt$ for a large enough T. Under additional assumptions on ν and G in a proper functional class, the expected squared error of the approximation can be controlled [56].

Forward Euler discretization In practice, in simulating the diffusion sample path, we cannot follow exactly the dynamic defined by the SDE (7.2). Instead, we must discretize it. A popular discretization is given by the forward (Euler) scheme, which reads

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

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

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

It is then natural to approximate $\widehat{\boldsymbol{\theta}}_n$ by applying this discretization strategy to the Langevin diffusion with μ in (7.1) as the target density. However, quantitative consistency guarantees of this discretization require ν (hence $\boldsymbol{\rho}$) 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 7.2) [20, 25, 26, 45]. In [26, 45] 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 [45][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 [13], though without any guarantees.

7.2 Forward-Backward proximal LMC algorithm

Recall that in our context, μ_n in (1.3) with the prior (5.2) is nor differentiable nor log-concave. To overcome these difficulties, we will exploit the structure of μ_n and involved arguments from variational analysis [50]. For ease of notation, in the following, we denote with the same symbol the measure and its density with respect to the Lebesgue measure. Let

$$V(\boldsymbol{\theta}) \overset{\text{def}}{=} -\log \mu_n(\boldsymbol{\theta}) = F(\boldsymbol{\theta}) + G(\boldsymbol{\theta}), \tag{7.3}$$

where

$$F(\boldsymbol{\theta}) = \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2/\beta \quad \text{and} \quad G(\boldsymbol{\theta}) = -\log \pi(\boldsymbol{\theta}).$$

[1]The 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.

It is important to emphasize that the results we derive in the rest of the section apply to a general class of potentials G that encompasses as a special case the one corresponding to the prior π (5.2). In Section 7.3, we show that these results indeed apply to (5.2).

7.2.1 A glimpse of variational analysis

Before proceeding, we need a bit of notations and definitions. A more comprehensive account on variational analysis in finite-dimensional Eudlidean spaces can be found in [50].

Definition 7.1 (Subdifferential)**.** *Given a point $\boldsymbol{\theta} \in \mathbb{R}^M$ where a function $V : \mathbb{R}^M \to \mathbb{R} \cup \{+\infty\}$ is finite, the subdifferential of V at $\boldsymbol{\theta}$ is defined as,*

$$\partial V(\boldsymbol{\theta}) = \{\boldsymbol{v} \in \mathbb{R}^M : \exists \boldsymbol{\theta}_k \to \boldsymbol{\theta}, V(\boldsymbol{\theta}_k) \to V(\boldsymbol{\theta}), \boldsymbol{v} \leftarrow \boldsymbol{v}_k \in \partial^F V(\boldsymbol{\theta}_k)\},$$

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

$$V(\boldsymbol{w}) \geq V(\boldsymbol{\theta}) + \big(\boldsymbol{v}\big)^T \big(\boldsymbol{w} - \boldsymbol{\theta}\big) + o\big(\big\|\boldsymbol{w} - \boldsymbol{\theta}\big\|_2\big).$$

V is subdifferentially regular at $\boldsymbol{\theta}$ if and only if V is locally lower-semicontinuous (lsc) there with $\partial V(\boldsymbol{\theta}) = \partial^F V(\boldsymbol{\theta})$.

$\partial V(\boldsymbol{\theta})$ and $\partial^F V(\boldsymbol{\theta})$ and closed, with $\partial^F V(\boldsymbol{\theta})$ convex and $\partial^F V(\boldsymbol{\theta}) \subset \partial V(\boldsymbol{\theta})$ [50, Theorem 8.6]. A proper lsc convex function is subdifferentially regular.

A function V is proper if it is not identically $+\infty$ and $V(\boldsymbol{\theta}) > -\infty$ for all $\boldsymbol{\theta}$.

Definition 7.2 (Proximal mapping and Moreau envelope)**.** *Let $\boldsymbol{M} \in \mathbb{R}^{M \times M}$ be symmetric positive definite. For a proper lsc function V and $\gamma > 0$, the proximal mapping and Moreau enveloppe in the metric $\boldsymbol{M}$ are defined respectively by*

$$\mathrm{prox}_{\gamma V}^{\boldsymbol{M}}(\boldsymbol{\theta}) \overset{def}{=} \operatorname*{Argmin}_{\boldsymbol{w} \in \mathbb{R}^M} \frac{1}{2\gamma}\big\|\boldsymbol{w} - \boldsymbol{\theta}\big\|_{\boldsymbol{M}}^2 + V(\boldsymbol{w})$$

$$^{\boldsymbol{M},\gamma}V(\boldsymbol{\theta}) \overset{def}{=} \inf_{\boldsymbol{w} \in \mathbb{R}^M} \frac{1}{2\gamma}\big\|\boldsymbol{w} - \boldsymbol{\theta}\big\|_{\boldsymbol{M}}^2 + V(\boldsymbol{w})$$

$\mathrm{prox}_{\gamma V}^{\boldsymbol{M}}$ here is a set-valued operator since the minimizer, if it exsists, is not necessarily unique. When $\boldsymbol{M} = \boldsymbol{I}_M$, we simply write $\mathrm{prox}_{\gamma V}$ and ${}^{\gamma}V$.
V is prox-bounded if there exists $\gamma > 0$ such that ${}^{\boldsymbol{M},\gamma}V(\boldsymbol{\theta}) > -\infty$ for some $\boldsymbol{\theta}$. The supremum of the set of all such γ is the threshold of prox-boundedness for V.

Definition 7.3 (Hypomonotone and monotone operators)**.** *A set-valued operator $S : \mathbb{R}^M \rightrightarrows \mathbb{R}^M$ is hypomonotone of modulus $r > 0$ if*

$$\big(\boldsymbol{\theta}' - \boldsymbol{\theta}\big)^T \big(\boldsymbol{\eta}' - \boldsymbol{\eta}\big) \geq -r\big\|\boldsymbol{\theta}' - \boldsymbol{\theta}\big\|_2^2, \quad \forall \boldsymbol{\theta}', \boldsymbol{\theta} \in \mathbb{R}^M \text{ and } \boldsymbol{\eta}' \in S(\boldsymbol{\theta}'), \boldsymbol{\eta} \in S(\boldsymbol{\theta}).$$

It is monotone if the inequality holds with $r = 0$.

7.2.2 Key properties

We start by collecting some important properties on the potential V in (7.3) that will be crucial on our way to design our algorithm. To avoid trivialities, from now on, we assume that $\mathrm{Argmin}(V) \neq \emptyset$. We also state the following assumptions.

(H.1) $G : \mathbb{R}^M \to \mathbb{R} \cup \{+\infty\}$ is proper, lsc and bounded from below.

(H.2) $\gamma \in]0, \frac{\beta}{2|||\boldsymbol{X}|||^2}[$.

Lemma 7.1. *Assume that (H.1) and (H.2) hold. Define* $\boldsymbol{M}_\gamma \overset{def}{=} \boldsymbol{I}_M - (2\gamma/\beta)\boldsymbol{X}^T\boldsymbol{X}$*, which is symmetric positive definite. Then* $\mathrm{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}(\boldsymbol{\theta})$ *and* $\mathrm{prox}_{\gamma G}(\boldsymbol{\theta})$ *are non-empty compact sets for any* $\boldsymbol{\theta}$*. Moreover,*

$$\mathrm{prox}_{\gamma V}^{\boldsymbol{M}_\gamma} = \mathrm{prox}_{\gamma G} \circ (\boldsymbol{I}_M - \gamma \nabla F),$$

and

$$\boldsymbol{\theta} \in \mathrm{Argmin}(V) \Rightarrow \boldsymbol{\theta} \in \mathrm{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}(\boldsymbol{\theta}).$$

The careful reader may have recognized from this lemma the classical forward-backward operator associated to $F + G$. This lemma states that this operator coincides with the proximal mapping of γV in the metric $\boldsymbol{M}_\gamma$. Moreover, fixed points of this proximal mapping include minimizers of V. They are not equal however in general, unless for instance V is convex.

Lemma 7.2. *Assume that (H.1) and (H.2) hold. Then* ${}^{\boldsymbol{M}_\gamma,\gamma}V(\theta)$ *is finite and depends continuously on* $(\boldsymbol{\theta}, \gamma) \in \mathbb{R}^M \times]0, \frac{\beta}{2|||\boldsymbol{X}|||^2}[$*, and* $\left({}^{\boldsymbol{M}_\gamma,\gamma}V(\boldsymbol{\theta})\right)_{\gamma \in]0, \frac{\beta}{2|||\boldsymbol{X}|||^2}[}$ *is a decreasing net. More precisely,*

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

In the following lemma, we need the additional assumption:

(H.3) $\mathrm{prox}_{\gamma G}$ is single-valued on $\mathbb{R}^M$, for any γ satisfying (H.2).

Lemma 7.3. *Assume that (H.1), (H.2) and (H.3) hold. Then* $\mathrm{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}$ *is single-valued, continuous in* $(\boldsymbol{\theta}, \gamma) \in \mathbb{R}^M \times]0, \frac{\beta}{2|||\boldsymbol{X}|||^2}[$*, and* ${}^{\boldsymbol{M}_\gamma,\gamma}V \in C^1(\mathbb{R}^M)$ *with gradient*

$$\nabla {}^{\boldsymbol{M}_\gamma,\gamma}V = \gamma^{-1}\boldsymbol{M}_\gamma\Big(\boldsymbol{I}_M - \mathrm{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}\Big).$$

In plain words, Lemma 7.1 and 7.3 tell us that the action of the operator $\mathrm{prox}_{\gamma G} \circ (\boldsymbol{I}_M - \gamma \nabla F)$ is equivalent to a gradient descent on the Moreau envelope of V in the metric $\boldsymbol{M}_\gamma$ with step-size γ.

To afford even better properties of ${}^{\boldsymbol{M}_\gamma,\gamma}V$, we need to strengthen our assumptions on G and γ.

(H.4) ∂G is hypomonotone of modulus r.

(H.5) $\gamma \in]0, \frac{1}{r + 2|||\boldsymbol{X}|||^2/\beta}[$.

It is obvious that (H.5) implies (H.2). The hypomonotonicity assumption is equivalent to saying that $G + \frac{r}{2}\|\cdot\|_2^2$ is convex, or that G has a supporting quadratic minorant at each point of its domain. Such functions are closely related to lower-C^2 and $1/r$-proximal functions in variational analysis; see [50, Theorem 10.33 and Example 11.26(d)].

Lemma 7.4. *Assume that* (H.1), (H.4) *and* (H.5) *hold. Then*

(i) $\mathrm{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}$ *is single-valued and Lipschitz continuous with constant* $\big(1-\gamma\big(r+2|||\boldsymbol{X}|||^2/\beta\big)\big)^{-1}$. *Moreover,* $\mathrm{prox}_{\gamma V}^{\boldsymbol{M}_\gamma} \circ \boldsymbol{M}_\gamma^{-1}$ *is monotone.*

(ii) ${}^{\boldsymbol{M}_\gamma,\gamma}V$ *is* $C^1(\mathbb{R}^M)$ *and its gradient is Lipschitz-continuous and hypomonotone.*

Hypomonotonicity of the drift coefficient has been exploited for backward (implicit) Euler discretization of SDEs in [31] (see also references therein).

7.2.3 Algorithm and guarantees

We are now ready to state our scheme to sample from μ_n. Define the following SDE

$$d\boldsymbol{L}(t) = -\frac{1}{2}\nabla\Big({}^{\boldsymbol{M}_\gamma,\gamma}V \circ \boldsymbol{M}_\gamma^{-1/2}\Big)(\boldsymbol{L}(t))dt + \boldsymbol{M}_\gamma^{1/2}d\boldsymbol{W}(t),\ t>0. \tag{7.4}$$

where G fulfills (H.1) and (H.4), and γ obeys (H.5). Owing to Lemma 7.4, it follows from [33, Theorem 2.4-6.1, Chapter IV] that for every initial point $\boldsymbol{L}(0)$ such that $\mathbb{E}\left[\|\boldsymbol{L}(0)\|_2^2\right] < \infty$, there exists a unique solution to the SDE (7.4) which is strongly Markovian, and the diffusion is non-explosive, i.e. $\mathbb{E}\left[\|\boldsymbol{L}(t)\|_2^2\right] < \infty$ for all $t>0$. Moreover, by [49, Theorem 2.1], $\boldsymbol{L}$ admits an (unique) invariant measure μ_γ having a density $\boldsymbol{\theta} \mapsto \exp\Big(-{}^{\boldsymbol{M}_\gamma,\gamma}V(\boldsymbol{M}_\gamma^{-1/2}\boldsymbol{\theta})\Big)/Z_\gamma$, where

$$Z_\gamma = \sqrt{\det(\boldsymbol{M}_\gamma)}\int_{\mathbb{R}^M} \exp\big(-{}^{\boldsymbol{M}_\gamma,\gamma}V(\boldsymbol{u})\big)d\boldsymbol{u}.$$

The following proposition answers the natural question on the behaviour of $\mu_\gamma - \mu_n$ as a function of γ.

Proposition 7.1. *Assume that* (H.1), (H.4) *and* (H.5) *hold. Then,* μ_γ *converges to* μ_n *in total variation as* $\gamma \to 0$.

In fact, the proof of this proposition only needs (H.1) and (H.2). Assumptions (H.4) and (H.5) are required only for (7.4) to be well-posed.

Inserting the identities of Lemma 7.1 and 7.3 into (7.4), and applying the change of variable $\boldsymbol{U}(t) = \boldsymbol{M}_\gamma^{-1/2}$, we get the SDE

$$d\boldsymbol{U}(t) = -\frac{1}{2}\left(\frac{\boldsymbol{I}_M - \mathrm{prox}_{\gamma G}\circ(\boldsymbol{I}_M - \gamma\nabla F)}{\gamma}\right)(\boldsymbol{U}(t))dt + d\boldsymbol{W}(t),\ \boldsymbol{U}(0)=\boldsymbol{u}_0,\ t>0. \tag{7.5}$$

With the same arguments as above, this SDE is well-posed (has a unique strong solution and the diffusion is non-explosive).

Consider now the forward Euler discretization of (7.5) with step-size $\delta>0$, which can be rearranged as

$$\boldsymbol{U}_{k+1} = \Big(1-\tfrac{\delta}{2\gamma}\Big)\boldsymbol{U}_k + \tfrac{\delta}{2\gamma}\mathrm{prox}_{\gamma G}(\boldsymbol{U}_k - \gamma\nabla F(\boldsymbol{U}_k)) + \sqrt{\delta}\boldsymbol{Z}_k,\ t>0,\ \boldsymbol{U}_0 = \boldsymbol{u}_0. \tag{7.6}$$

Clearly, without the stochastic term $\sqrt{\delta}\boldsymbol{Z}_k$, this would be a relaxed version of the classical Forward-Backward proximal splitting algorithm with relaxation parameter $\delta/(2\gamma)$ and descent step-size γ [4]. Hence, we coin the scheme (7.6) FB-LMC. Taking $\delta = 2\gamma$, we obtain its "unrelaxed" version. Observe again that by Lemma 7.1 and 7.3, this is also equivalent to a gradient descent on the Moreau envelope of V in the metric $\boldsymbol{M}_\gamma$ with step-size δ.

Remark 7.1. *One can propose yet another algorithm to approximately sample from μ_n according to the following strategy. Consider smoothing G (resp. J if $G = J \circ \boldsymbol{D}$ as for analysis-type prior as in (5.2)) by replacing it with its Moreau envelope ${}^{\boldsymbol{I}_M,\gamma}G$ (resp. with ${}^{\boldsymbol{I}_M,\gamma}J \circ \boldsymbol{D}$), where $\gamma \in]0, 1/r[$. In this case Lemma 7.4 applies to G (resp. J), whence we deduce that ${}^{\boldsymbol{I}_M,\gamma}G$ (resp. ${}^{\boldsymbol{I}_M,\gamma}J \circ \boldsymbol{D}$) is $C^1(\mathbb{R}^M)$ with Lipschitz-continuous and hypomonotone gradient. Proposition 7.1 can also be modified correspondingly to show that $\pi_\gamma^G(d\boldsymbol{\theta}) \propto \exp\left(-{}^{\boldsymbol{I}_M,\gamma}G\right)(\theta)$ converges in total variation to π as $\gamma \to 0$. Thus one can replace $\boldsymbol{M}_\gamma$ by $\boldsymbol{I}_M$ in (7.4), and $\nabla\left({}^{\boldsymbol{M}_\gamma,\gamma}V \circ \boldsymbol{M}_\gamma^{-1/2}\right)$ with $\nabla F + \nabla {}^{\boldsymbol{I}_M,\gamma}G = \nabla F + \gamma^{-1}\left(\boldsymbol{I}_M - \mathrm{prox}_{\gamma G}\right)$ (resp. $\nabla F + \gamma^{-1}\boldsymbol{D}^T \circ \left(\boldsymbol{I}_M - \mathrm{prox}_{\gamma J}\right) \circ \boldsymbol{D}$). This is a generalization beyond the convex case to the scheme described in [26]. We do not pursue this further here for space limitation and leave the details to a forthcoming work.*

From (7.6), an Euler approximate solution is defined as

$$\boldsymbol{U}^\delta(t) \stackrel{\text{def}}{=} \boldsymbol{U}_0 - \frac{1}{2\gamma}\int_0^t \left(\overline{\boldsymbol{U}}(s) - \mathrm{prox}_{\gamma G} \circ \left(\overline{\boldsymbol{U}}(s) - \gamma \nabla F(\overline{\boldsymbol{U}}(s))\right)\right) ds + \int_0^t d\boldsymbol{W}(s) ds,$$

where $\overline{\boldsymbol{U}}(t) = \boldsymbol{U}_k$ for $t \in [k\delta, (k+1)\delta[$. Observe that $\boldsymbol{U}^\delta(k\delta) = \overline{\boldsymbol{U}}(k\delta) = \boldsymbol{U}_k$, hence $\boldsymbol{U}^\delta(t)$ and $\overline{\boldsymbol{U}}(t)$ are continuous-time extensions to the discrete-time chain $\{\boldsymbol{U}_k\}_k$.

Mean square convergence of the pathwise approximation (7.6) and of its first-order moment can be established as follows.

Theorem 7.1. *Assume that* **(H.1)**, **(H.4)** *and* **(H.5)** *hold. Then*

$$\left\|\mathbb{E}\left[\boldsymbol{U}^\delta(T)\right] - \mathbb{E}\left[\boldsymbol{U}(T)\right]\right\|_2 \leq \mathbb{E}\left[\sup_{0\leq t\leq T}\left\|\boldsymbol{U}^\delta(t) - \boldsymbol{U}(t)\right\|_2\right] = O(\delta^{1/2}) \underset{\delta\to 0}{\longrightarrow} 0.$$

7.3 Application to group analysis sparsity prior

Let's now consider the prior (5.2) with g chosen according to Example 5.2. For the sake of simplicity, we take $R = +\infty$ (implying $H = +\infty$ by Proposition 5.1)[2]. Thus

$$G(\boldsymbol{\theta}) = \sum_{l=1}^{L}\left[\alpha^a \left\|[\boldsymbol{D}\boldsymbol{\theta}]_{\mathcal{G}_l}\right\|_2^a + c\log\left(\tau^b + \left\|[\boldsymbol{D}\boldsymbol{\theta}]_{\mathcal{G}_l}\right\|_2^b\right)\right],$$

and G is proper, continuous and bounded below, i.e. it fullfils **(H.1)**. This function also has a supporting quadratic minorant at each point, and thus verifies **(H.4)**.

It remains now to compute the proximal mapping of G. Owing to Assumption 4.2, one can show quite immediately that

$$\mathrm{prox}_{\gamma G} = \widetilde{\boldsymbol{D}^T} \circ \mathrm{prox}^{\boldsymbol{D}\widetilde{\boldsymbol{D}^T}}_{\gamma J + \iota_{\mathrm{Im}(\boldsymbol{D})}} \circ \boldsymbol{D},$$

where $G = J \circ \boldsymbol{D}$, $\iota_{\mathrm{Im}(\boldsymbol{D})}(\boldsymbol{u}) = 0$ if $\boldsymbol{u} \in \mathrm{Im}(\boldsymbol{D})$ and $+\infty$ otherwise, and $\boldsymbol{D}\widetilde{\boldsymbol{D}^T}$ is indeed definite positive on $\mathrm{Im}(\boldsymbol{D})$. This means that, in general, G 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 $\boldsymbol{D}$ is invertible. Observe that this assumption can be removed using the algorithmic alternative outlined in Remark 7.1. In

[2]The exposition can be generalized to the case of bounded R, in which case G will involve the indicator function of Θ, and still verifies **(H.1)**.

this case, one can operate a simple change of variable $\boldsymbol{u} = \boldsymbol{D}\boldsymbol{\theta}$ and replace F with $F \circ \boldsymbol{D}^{-1}$, and G with $J = G \circ \boldsymbol{D}^{-1}$. The latter has a closed form expression as shown in the following two lemmas.

First, observe that J is separable into L independent functions (one for each group $\mathcal{G}_l$), and each of these functions depends only on the ℓ_2 norm of the block, whence the proximal mapping has the useful form.

Lemma 7.5. *For any $\boldsymbol{u} \in \mathbb{R}^P$ and $\gamma > 0$, we have*

$$\operatorname{prox}_{\gamma J}(\boldsymbol{u}) = \begin{pmatrix} \operatorname{prox}_{\gamma w}(\|\boldsymbol{u}_{\mathcal{G}_1}\|_2) \frac{\boldsymbol{u}_{\mathcal{G}_1}}{\|\boldsymbol{u}_{\mathcal{G}_1}\|_2} \\ \vdots \\ \operatorname{prox}_{\gamma w}\big(\|\boldsymbol{u}_{\mathcal{G}_P}\|_2\big) \frac{\boldsymbol{u}_{\mathcal{G}_P}}{\|\boldsymbol{u}_{\mathcal{G}_P}\|_2} \end{pmatrix}$$

where $w : t \in \mathbb{R}^+ \mapsto w(t) = \alpha^a t^a + c \log(\tau^b + t^b)$.

Remark 7.2. *Lemma 7.5 remains valid for any prior that depends only on $\boldsymbol{u}$ through the norms of its groups, e.g. (5.2), in which case $w(t) = \alpha^a t^a - \log(g(t))$, with the proviso that $-\log \circ g$ is lsc.*

We now turn to the computation of $\operatorname{prox}_{\gamma w}$. This problem is treated in the following lemma.

Lemma 7.6. *Let w as defined in Lemma 7.5 and $\gamma > 0$. Then*

$$\operatorname{prox}_{\gamma w}(t) = \begin{cases} 0 & \text{if } |t| \leq p_0, \\ t - \operatorname{sgn}(t)\gamma w'\big(\big|\operatorname{prox}_{\gamma w}(t)\big|\big) & \text{if } |t| > p_0, \end{cases} \tag{7.7}$$

where $p_0 = \min_{t \geq 0} t + \gamma w'(t)$.

Clearly, the proximal mapping of $J = G \circ \boldsymbol{D}^{-1}$ is single-valued hence conforming with (**H.3**). In addition, computing it boils down to solving the simple one-dimensional non-linear equation (7.7), which can be done very efficiently.

8 Numerical experiments

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

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

where $\boldsymbol{\xi}$ is the noise, $\boldsymbol{X}$ is the design matrix. $\boldsymbol{\theta}_0 \in \mathbb{R}^M$ is the unknown regression vector of interest assumed to obey the group-analysis sparsity assumption (Assumption 4.1), with a given invertible dictionary $\boldsymbol{D}$. We take the prior (5.2) where g is given in Example 5.2 with $H = +\infty$. 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 main steps of FB-LM applied to compute EWA with our prior are summarized in Algorithm 8.1. It takes as inputs the vector of observations $\boldsymbol{y}$, the design matrix $\boldsymbol{X}$, the noise level σ, the analysis operator $\boldsymbol{D}$ and the group size G. 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: $a \in]0, 1]$, $b \in]0, 1]$, $c > (2+G)/b$, $\tau \sim 1/(Mn)$ and $\alpha \leq 1/\big(3\sqrt{K_{a,g}}L\big)^{1/a}$, where $K_{a,g}$ is given in Corollary 6.3. 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.

Algorithm 8.1 Pseudo-code of the FB-LMC algorithm.

Input: The given data $(\boldsymbol{y}, \boldsymbol{X}, \sigma, \boldsymbol{D}, G)$ and the parameters $(a, b, c, \alpha, \beta, \tau, \delta, T)$
Output: The vector $\widehat{\boldsymbol{\theta}}_n \in \mathbb{R}^M$.
Pre-computations: `P = size(D,1);N=floor(T/`δ`);invD = inv(D); A = X*invD; AA = A'*A; Ay=A'*y;`
Initialization: `Li = zeros(P,1); u_bar = zeros(P,1);`
for all `i in 1:1:N` **do**
  `nablaFD = (2/`β`) * (AA*L-Ay);`
  `epsilon = sqrt(`δ`) * randn(P,1);`
  `Li = (1-`$\delta/(2\gamma)$`)+`$\delta/(2\gamma) * \mathrm{prox}_{\gamma G \circ D^{-1}}$`(Li - `$\gamma$`*nablaFD + epsilon);`
  `u_bar` $\leftarrow$ `u_bar + Li;`
end for
return `invD*u_bar/N;`

8.1 1-D signal recovery under group sparsity

In this example, we set $\boldsymbol{D} = \boldsymbol{I}_M$, which corresponds to the classical group sparsity. The design matrix is drawn uniformly at random from the Rademacher ensemble, i.e. its entries are i.i.d. variates valued in $\{-1, 1\}$ with equal probabilities. The non-zero entries of $\boldsymbol{\theta}_0$ are equal to 1 and we denote $S = \|\boldsymbol{\theta}_0\|_0$ the sparsity level of $\boldsymbol{\theta}_0$. Two types of sparsity behavior are considered: individual sparsity where $G_{\boldsymbol{\theta}_0} = 1$; group structured sparsity with $G_{\boldsymbol{\theta}_0} = 4$. Besides, the positions of the non-zero/active entries (for $G_{\boldsymbol{\theta}_0} = 1$) or groups (for $G_{\boldsymbol{\theta}_0} = 4$) are chosen randomly uniformly on $\{1, \ldots, M\}$.

The experiments are performed by fixing $M = 128$, and taking $S = 4, 8, \ldots, M$, $n = 8, 16, \ldots, M$, step-size $\delta = 4\sigma^2/(nM)$ and integration time $T = 3500$. The parameters in the prior are chosen to minimize the remainder term in the oracle inequality (6.4). For each (S, n), and each value of $G_{\boldsymbol{\theta}_0}$, $N_{\mathrm{rep}} = 20$ instances of the problem suite $(\boldsymbol{X}, \boldsymbol{\theta}_0, \boldsymbol{Y})$ are generated, and EWA is applied with a chosen G and the other parameters as detailed above. The estimation quality/success is then assessed by

$$\pi_{S,n} = \frac{1}{N_{\mathrm{rep}}} \sum_{j=1}^{N_{\mathrm{rep}}} I\Big(\big\| \widehat{\boldsymbol{\theta}}_n^{(j,S,n)} - \boldsymbol{\theta}_0^{(j,S,n)} \big\|_n \leq \epsilon \Big), \tag{8.1}$$

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/M and n/M are respectively normalized measures of sparsity and problem indeterminacy. We get a two-dimensional phase space $(S/M, n/M) \in [0, 1]^2$ 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 (8.1) as a function $(S/M, n/M)$ 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 that with $G = 4$. The situation reverses for Figures 1(c)-(d) where $G_{\boldsymbol{\theta}_0} = 4$, and one observes that the success area is significantly better using $G = 4$ than $G = 1$. This is expected as it reveals better performance of EWA when used with the choice $G = G_{\boldsymbol{\theta}_0}$. This is also confirmed by visual inspection of Figures 1(c')-(d'), where we plotted instances of recovered vectors $\widehat{\boldsymbol{\theta}}_n^{(j,S,n)}$ when $(S, G_{\boldsymbol{\theta}_0}) \in \{4, 8\} \times \{1, 4\}$ and $n/M = 1/2$. 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 $S/M = \|\boldsymbol{\theta}_0\|_{0,\mathcal{G}} G_{\boldsymbol{\theta}_0}/M$. As far as the expected phase transition curve is concerned, one has from the oracle inequality (6.4) and Remark 6.2 that it is expected to occur for

$$n/M = C_\epsilon \|\boldsymbol{\theta}_0\|_{0,\mathcal{G}} G_{\boldsymbol{\theta}_0}/M(\log(M/G_{\boldsymbol{\theta}_0})/G_{\boldsymbol{\theta}_0}) = C_\epsilon S/M(\log(M/G_{\boldsymbol{\theta}_0})/G_{\boldsymbol{\theta}_0})$$

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

8.2 2-D image recovery under analysis group-sparsity

In the second numerical experiment, $\boldsymbol{\theta}_0$ is a 2-D image which is a matrix in $\mathbb{R}^{160\times 160}$ (a close-up of the known Shepp-Logan phantom, see Figure 2(a)). Let us denote vec the vectorization operator. Thus $\mathrm{vec}(\boldsymbol{\theta}_0)$ is vector in $\mathbb{R}^M$ with $M = 160^2$, and our goal is to recover $\boldsymbol{\theta}_0$ from

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

where $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$ and $\boldsymbol{X} \in \mathbb{R}^{n\times M}$ 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 [51]. It turns out that this can be cast in our analysis group-sparsity framework as a special case.

Let $\boldsymbol{D}_1 : \mathbb{R}^{\sqrt{M}\times\sqrt{M}} \to \mathbb{R}^{\sqrt{M}\times\sqrt{M}}$ and $\boldsymbol{D}_2 : \mathbb{R}^{\sqrt{M}\times\sqrt{M}} \to \mathbb{R}^{\sqrt{M}\times\sqrt{M}}$ be the finite difference operators along, respectively, the columns and rows, with appropriate boundary conditions that we will specify shortly. We define the linear analysis operator $\boldsymbol{D}$ as

$$\boldsymbol{D}\colon \boldsymbol{\theta} \in \mathbb{R}^{\sqrt{M}\times\sqrt{M}} \mapsto (\boldsymbol{D}_1(\boldsymbol{\theta}), \boldsymbol{D}_2(\boldsymbol{\theta})) \in \mathbb{R}^{\sqrt{M}\times\sqrt{M}} \times \mathbb{R}^{\sqrt{M}\times\sqrt{M}},$$

The isotropic total variation prior on $\boldsymbol{\theta}$ promotes sparsity of
$\mathrm{vec}\left(\left\|[\boldsymbol{D}(\boldsymbol{\theta}_0)]_{i,j}\right\|_2\right)_{1\le i,j\le\sqrt{M}}$. By defining the set of groups by

$$\mathcal{G} = \bigcup_{(i,j)\in\{1,\dots,s_{\boldsymbol{p}_0}\}^2} \{(i,j,1),(i,j,2)\},$$

one immediately realizes that measuring sparsity of the above vectorized form is equivalent to group sparsity of $\boldsymbol{D}(\boldsymbol{\theta}_0)$ with groups of size 2 along the third dimension. One can then use FB-LMC Algorithm 8.1 to get the EWA estimator. We have however to ensure that $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are invertible, which is achieved by a proper choice of the boundary conditions (not periodic of course).

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

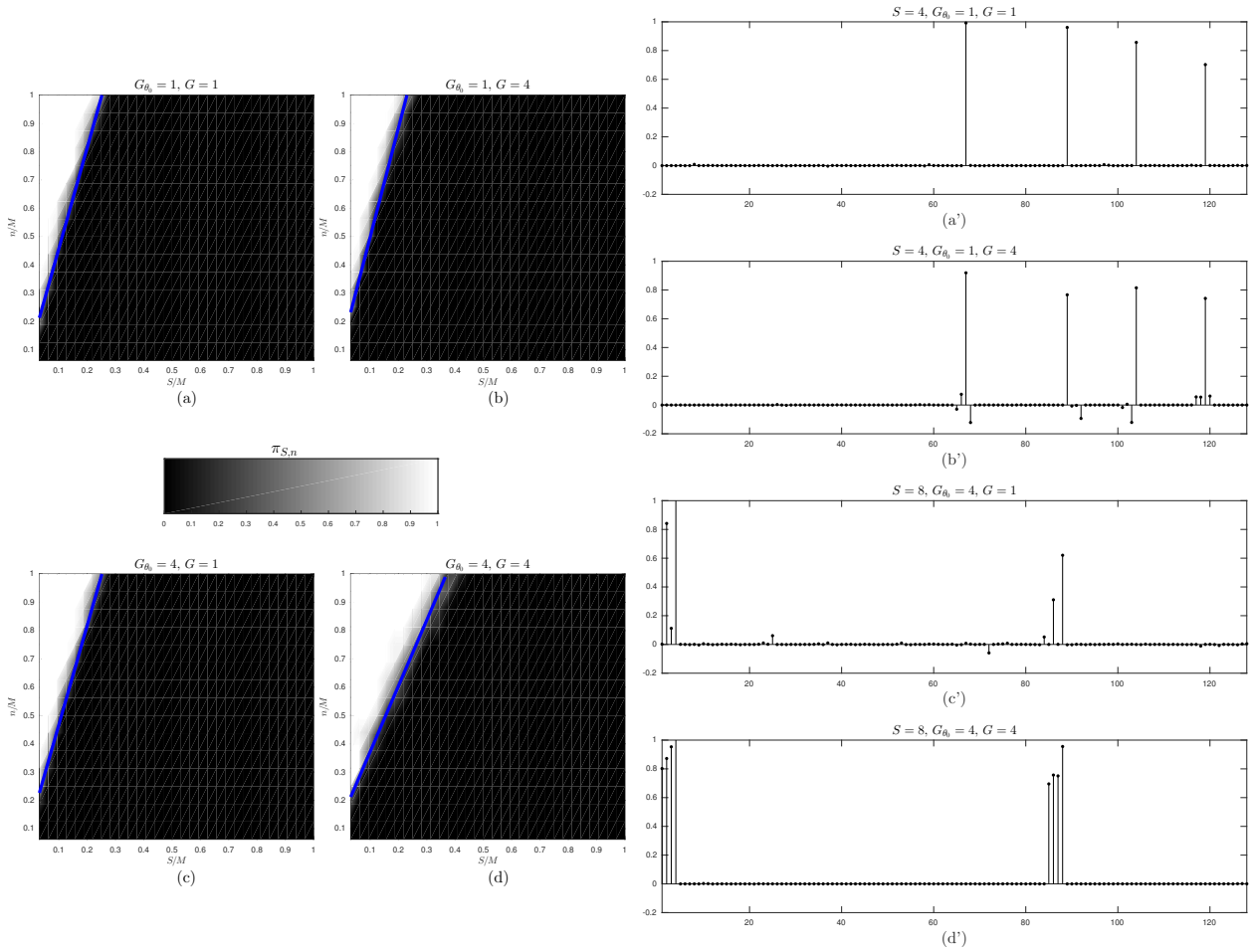
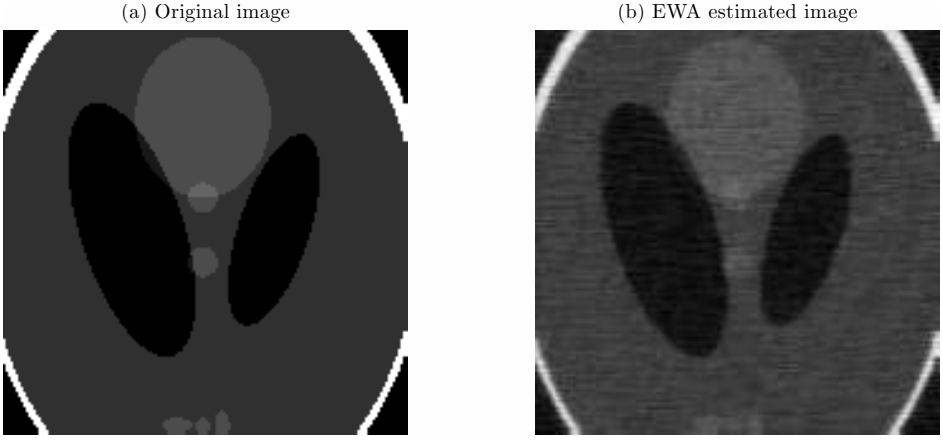


Figure 1: (a)-(d): Phase diagrams of EWA for $\boldsymbol{D} = \boldsymbol{I}_M$, the color bar ranges from dark ($\pi_{S,n} = 0$) to bright ($\pi_{S,n} = 1$). The blue line is the fitted phase transition curve. (a')-(d'): Examples of vectors $\widehat{\boldsymbol{\theta}}_n^{(j,S,n)}$ recovered by EWA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



(a) Original image (b) EWA estimated image

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

9 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-bakcward 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.

10 Proofs

10.1 Proofs of auxiliary results

10.1.1 Proof of Lemma 4.1

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

$$\begin{aligned}\int_\Theta u(\boldsymbol{D\theta})d\boldsymbol{\theta} &= \frac{1}{\sqrt{\det(\boldsymbol{D}^T\boldsymbol{D})}} \int_{\mathbb{R}^P} \Big(\sum_{\boldsymbol{\theta}\in\Theta\cap\{\boldsymbol{\omega}:\boldsymbol{D\omega}=\boldsymbol{v}\}} u(\boldsymbol{D\theta}) \Big) d\boldsymbol{v} \\ &= \frac{1}{\sqrt{\det(\boldsymbol{D}^T\boldsymbol{D})}} \int_{\mathrm{Im}(\boldsymbol{D})} \Big(\sum_{\boldsymbol{\theta}\in\Theta\cap\{\boldsymbol{\omega}:\boldsymbol{D\omega}=\boldsymbol{v}\}} u(\boldsymbol{D\theta}) \Big) d\boldsymbol{v}. \end{aligned} \tag{10.1}$$

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

10.1.2 Proof of Proposition 5.1

Let $\boldsymbol{\theta} \in \Theta$ and $i \in \{1, \dots, n\}$. Setting $u_i^{\boldsymbol{\theta}} = \sum_{j=1}^{M} \theta_j f_j(x_i)$ and $\boldsymbol{u}^{\boldsymbol{\theta}} = (u_1^{\boldsymbol{\theta}}, \dots, u_n^{\boldsymbol{\theta}})^T$, and by virtue of (4.1), we have

$$\left\|\boldsymbol{u}^{\boldsymbol{\theta}}\right\|_2 = \left\|\boldsymbol{X}\boldsymbol{\theta}\right\|_2 \leq |||\boldsymbol{X}||| \left|\left|\left|\widetilde{\boldsymbol{D}^T}\right|\right|\right| \left\|\boldsymbol{D}\boldsymbol{\theta}\right\|_2. \tag{10.2}$$

From (4.2), we get also

$$\left|\left|\left|\widetilde{\boldsymbol{D}^T}\right|\right|\right| = \sqrt{\sigma_1\left(\widetilde{\boldsymbol{D}^T}\widetilde{\boldsymbol{D}^T}^T\right)} \leq \frac{1}{\sqrt{\mu}}. \tag{10.3}$$

Moreover, since $a \in]0, 1]$, by the triangle inequality, we get

$$\left\|\boldsymbol{D}\boldsymbol{\theta}\right\|_2 \leq \sum_{l=1}^{L} \left\|[\boldsymbol{D}\boldsymbol{\theta}]_{\mathcal{G}_l}\right\|_2 \leq \left[\sum_{l=1}^{L} \left\|[\boldsymbol{D}\boldsymbol{\theta}]_{\mathcal{G}_l}\right\|_2^a\right]^{1/a} = \left\|\boldsymbol{D}\boldsymbol{\theta}\right\|_{a,\mathcal{G}} \leq R^{1/a}. \tag{10.4}$$

Combining (10.2), (10.3) and (10.4), we obtain

$$\left\|\boldsymbol{u}^{\boldsymbol{\theta}}\right\|_2 \leq \frac{|||\boldsymbol{X}|||R^{1/a}}{\sqrt{\mu}},$$

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

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

□

10.1.3 Proof of Lemma 5.1

Let us first check the integrability condition (5.3). By Lemmas 4.1 and 2.2, we obtain

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

Since $G \geq 1$ and $g : \mathbb{R}^+ \to \mathbb{R}^+$, by (5.6), we get

$$\begin{aligned}
&\int_{\mathbb{R}^M} \prod_{l=1}^{L} g(\left\|[\boldsymbol{D}\boldsymbol{u}]_{\mathcal{G}_l}\right\|_2) d\boldsymbol{u} \\
\leq\ & \frac{C_G^L}{\sqrt{\det(\boldsymbol{D}^T\boldsymbol{D})}} \left(\int_0^1 z^{G-1} g(z) dz + \int_1^\infty w^{G-1} g(w) dw \right)^L \\
\leq\ & \frac{C_G^L}{\sqrt{\det(\boldsymbol{D}^T\boldsymbol{D})}} \left(\sup_{z\in[0,1]} g(z) + \int_1^\infty w^{G+1} g(w) dw \right)^L < \infty. \qquad (10.5)
\end{aligned}$$

Therefore, g satisfies Assumption 5.2-2.

Now, we check the moment condition (5.4). Using similar arguments to the bound (10.5), we have

$$\begin{aligned}
&\int_{\mathbb{R}^M} \left\|[\boldsymbol{D}\boldsymbol{u}]_{\mathcal{G}_l}\right\|_2^2 \prod_{k=1}^{L} g(\left\|[\boldsymbol{D}\boldsymbol{u}]_{\mathcal{G}_k}\right\|_2) d\boldsymbol{u} \\
\leq\ & \frac{1}{\sqrt{\det(\boldsymbol{D}^T\boldsymbol{D})}} \int_{\mathbb{R}^P} \left\|\boldsymbol{v}_{\mathcal{G}_l}\right\|_2^2 \prod_{k=1}^{L} g(\left\|\boldsymbol{v}_{\mathcal{G}_k}\right\|_2) d\boldsymbol{v} \\
=\ & \frac{1}{\sqrt{\det(\boldsymbol{D}^T\boldsymbol{D})}} \left(\int_{\mathbb{R}^G} \left\|\boldsymbol{u}\right\|_2^2 g(\left\|\boldsymbol{u}\right\|_2) d\boldsymbol{u} \right) \left(\int_{\mathbb{R}^G} g(\left\|\boldsymbol{v}\right\|_2) d\boldsymbol{v} \right)^{L-1} \\
=\ & \frac{C_G^L}{\sqrt{\det(\boldsymbol{D}^T\boldsymbol{D})}} \int_0^\infty z^{G+1} g(z) dz \left(\int_0^\infty w^{G-1} g(w) dw \right)^{L-1} < \infty, \qquad (10.6)
\end{aligned}$$

whence we conclude that g satisfies Assumption 5.2-3. □

10.1.4 Proof of Corollary 6.1

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 5.1, g satisfies Assumptions 5.2-2 and 5.2-3. Moreover, taking $h(t) = 1 + t/\tau$ and $\lambda = 4$, for all $(t, t^*) \in \mathbb{R}^2$, we have by Young's inequality

$$\begin{aligned}
\frac{g(|t-t^*|)}{g(|t|)} &= \left[\frac{\tau^2 + t^2}{\tau^2 + (t-t^*)^2} \right]^2 = \left[1 + \frac{2\tau(t-t^*)t^*/\tau + t^{*2}}{\tau^2 + (t-t^*)^2} \right]^2 \\
&\leq \left[1 + \frac{|t^*|}{\tau} + \frac{t^{*2}}{\tau^2} \right]^2 \leq \left[1 + \frac{|t^*|}{\tau} \right]^4 = h(|t^*|)^\lambda.
\end{aligned}$$

Therefore, g satisfies Assumption 5.2 for $G = 1$.

Owing to Remark 5.2 and Lemma 2.1, we obtain

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

We are now in position to apply Theorem 6.1 with $\boldsymbol{D} = \boldsymbol{I}_M$ (then $P = M$), $G = 1$ (then $L = P$), $a = 1$ and $\alpha \leq 1/(3M\tau)$ to conclude. This completes the proof. □

10.1.5 Proof of Corollary 6.2

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 5.1, g satisfies Assumptions 5.2-2 and 5.2-3. Recall that $b \in]0, 1]$. Taking $h(x) = 1 + (x/\tau)^b$ and $\lambda = c$, for all $(\boldsymbol{t}, \boldsymbol{t}^*) \in \mathbb{R}^G \times \mathbb{R}^G$, we have

$$\begin{aligned}\frac{g(\|\boldsymbol{t} - \boldsymbol{t}^*\|_2)}{g(\|\boldsymbol{t}\|_2)} &= \left[\frac{\tau^b + \|\boldsymbol{t}\|_2^b}{\tau^b + \|\boldsymbol{t} - \boldsymbol{t}^*\|_2^b}\right]^c \leq \left[\frac{\tau^b + \|\boldsymbol{t} - \boldsymbol{t}^*\|_2^b + \|\boldsymbol{t}^*\|_2^b}{\tau^b + \|\boldsymbol{t} - \boldsymbol{t}^*\|_2^b}\right]^c \\ &\leq \left[1 + \frac{\|\boldsymbol{t}^*\|_2^b}{\tau^b + \|\boldsymbol{t} - \boldsymbol{t}^*\|_2^b}\right]^c \leq \left[1 + \left(\|\boldsymbol{t}^*\|_2/\tau\right)^b\right]^c.\end{aligned}$$

Therefore, g satisfies Assumption 5.2 with any $G \geq 1$. Applying Theorem 6.1, we obtain the claimed SOI. The proof is complete. □

10.1.6 Proof of Corollary 6.3

Since g satisfies Assumptions 5.2-2, 5.2-3 and $\boldsymbol{D}$ is invertible, by Remark 5.2 and Lemma 2.1, we get

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

According to (6.3) and the fact that $\alpha \leq 1/\left(3\tau^a\sqrt{\widetilde{K}_{a,g}}L\right)^{1/a}$ $= 1/\left(3\sqrt{K_{a,g}}L\right)^{1/a}$, we obtain (6.4). This ends the proof. □

10.2 Proofs of the SOI results

10.2.1 Proof of Theorem 6.1

Remind the prior $\pi(d\boldsymbol{\theta})$ from (5.2), where $\Theta = \left\{\boldsymbol{\theta} \in \mathbb{R}^M : \|\boldsymbol{D}\boldsymbol{\theta}\|_{a,\mathcal{G}}^a \leq R\right\}$.

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

$$\boldsymbol{\theta}^* \in \left\{\boldsymbol{\theta} \in \mathbb{R}^M : \|\boldsymbol{D}\boldsymbol{\theta}\|_{a,\mathcal{G}}^a \leq R - 3\sqrt{K_{a,g}}L = R - r_L\right\}. \tag{10.7}$$

We define the probability measure

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

where $C_L > 0$ is the normalization factor for $p_0{}^{\boldsymbol{D}}$.

Since $r_L < R$, $\boldsymbol{\theta} \in \Theta_{p_0{}^{\boldsymbol{D}}}$ implies that $\boldsymbol{\theta} - \boldsymbol{\theta}^* \in \Theta$. Therefore,

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

For any $i \in \{1, \dots, n\}$, with $\boldsymbol{X}_i = (f_1(x_i), \dots, f_M(x_i))^T$, one can write

$$f_{\boldsymbol{\theta}}(x_i) = \ell\left(\sum_{j=1}^{M} \theta_j f_j(x_i)\right) = \ell\big(\boldsymbol{X}_i^T \boldsymbol{\theta}\big).$$

Taylor-Lagrange formula then gives us

$$\begin{aligned} (f_{\boldsymbol{\theta}}(x_i) - f(x_i))^2 \leq \quad & (f_{\boldsymbol{\theta}^*}(x_i) - f(x_i))^2 + C_{f,\ell}\big[\boldsymbol{X}_i^T(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\big]^2 \\ + \quad & 2(f_{\boldsymbol{\theta}^*}(x_i) - f(x_i))\ell'\big(\boldsymbol{X}_i^T\boldsymbol{\theta}^*\big)\boldsymbol{X}_i^T(\boldsymbol{\theta}-\boldsymbol{\theta}^*), \end{aligned} \tag{10.8}$$

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

By summing, normalizing by $1/n$, taking the integral in Θ w.r.t. $p_0{}^{\boldsymbol{D}}$, inequality (10.8) becomes

$$\begin{aligned} \int_\Theta \left\|f_{\boldsymbol{\theta}} - f\right\|_n^2 p_0{}^{\boldsymbol{D}}(d\boldsymbol{\theta}) \leq \quad & \left\|f_{\boldsymbol{\theta}^*} - f\right\|_n^2 \\ + \quad & \frac{2}{n}\sum_{i=1}^{n}(f_{\boldsymbol{\theta}^*}(x_i) - f(x_i))\ell'\big(\boldsymbol{X}_i^T\boldsymbol{\theta}^*\big) \\ & \boldsymbol{X}_i^T \int_\Theta (\boldsymbol{\theta}-\boldsymbol{\theta}^*) p_0{}^{\boldsymbol{D}}(d\boldsymbol{\theta}) \\ + \quad & C_{f,\ell}\int_{\mathbb{R}^M} \frac{1}{n}\sum_{i=1}^{n}\big[\boldsymbol{X}_i^T(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\big]^2 p_0{}^{\boldsymbol{D}}(d\boldsymbol{\theta}). \end{aligned} \tag{10.9}$$

Note that, the right term of inequality (10.9) 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 second component vanishes. Indeed, let $\boldsymbol{\theta} \in \Theta_{p_0{}^{\boldsymbol{D}}}$, from (10.7) and the fact that $a \in]0, 1]$, we have

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

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

$$\Theta_{p_0{}^{\boldsymbol{D}}} \subseteq \Theta. \tag{10.10}$$

In what follows, we denote $\mathbb{B}_{a,\mathcal{G}}^a(x) = \left\{\boldsymbol{z} \in \mathbb{R}^P : \left\|\boldsymbol{z}\right\|_{a,\mathcal{G}}^a \leq x\right\}, \forall x > 0$ for brevity. By (10.10), property (4.1), Lemma 4.1 and symmetry of $\mathbb{B}_{a,\mathcal{G}}^a(r_L) \cap \mathrm{Im}(\boldsymbol{D})$, we obtain

$$\begin{aligned}
&\int_{\Theta} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) p_0{}^{\boldsymbol{D}}(d\boldsymbol{\theta}) \\
\propto\ &\int_{\Theta \cap \Theta_{p_0{}^{\boldsymbol{D}}}} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \prod_{l=1}^{L} \exp\left(-\alpha^a \left\|[\boldsymbol{D}\boldsymbol{\theta} - \boldsymbol{D}\boldsymbol{\theta}^*]_{\mathcal{G}_l}\right\|_2^a\right) g\left(\left\|[\boldsymbol{D}\boldsymbol{\theta} - \boldsymbol{D}\boldsymbol{\theta}^*]_{\mathcal{G}_l}\right\|_2\right) d\boldsymbol{\theta} \\
=\ &\int_{\Theta_{p_0{}^{\boldsymbol{D}}}} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \prod_{l=1}^{L} \exp\left(-\alpha^a \left\|[\boldsymbol{D}\boldsymbol{\theta} - \boldsymbol{D}\boldsymbol{\theta}^*]_{\mathcal{G}_l}\right\|_2^a\right) g\left(\left\|[\boldsymbol{D}\boldsymbol{\theta} - \boldsymbol{D}\boldsymbol{\theta}^*]_{\mathcal{G}_l}\right\|_2\right) d\boldsymbol{\theta} \\
=\ &\frac{\widetilde{\boldsymbol{D}^T}}{\sqrt{\det(\boldsymbol{D}^T\boldsymbol{D})}} \int_{\mathbb{B}_{a,\mathcal{G}}^a(r_L) \cap \mathrm{Im}(\boldsymbol{D})} \boldsymbol{z} \prod_{l=1}^{L} \exp\left(-\alpha^a \left\|\boldsymbol{z}_{\mathcal{G}_l}\right\|_2^a\right) g\left(\left\|\boldsymbol{z}_{\mathcal{G}_l}\right\|_2\right) d\boldsymbol{z} = 0,
\end{aligned} \tag{10.11}$$

which is the desired claim.

We now bound the last term in the right hand side of (10.9). Define

$$p_0(d\boldsymbol{u}) = \frac{1}{C_L \sqrt{\det(\boldsymbol{D}^T\boldsymbol{D})}} \prod_{l=1}^{L} \exp\left(-\alpha^a \left\|\boldsymbol{u}_{\mathcal{G}_l}\right\|_2^a\right) g\left(\left\|\boldsymbol{u}_{\mathcal{G}_l}\right\|_2\right) I_{\mathrm{Im}(\boldsymbol{D}) \cap \mathbb{B}_{a,\mathcal{G}}^a(r_L)}(\boldsymbol{u}) d\boldsymbol{u}. \tag{10.12}$$

One can see that p_0 coincides with the probability measure $p_0{}^{\boldsymbol{D}}$ on $\mathbb{R}^M$ via a change of variables of type (4.3). So, p_0 is a probability measure on $\mathbb{R}^P$.

For any $i, j \in \{1, \ldots, L\}$, $i \neq j$, by a change of variables, we get

$$\int_{\mathbb{R}^P} \boldsymbol{u}_{\mathcal{G}_i} \boldsymbol{u}_{\mathcal{G}_j}^T p_0(d\boldsymbol{u}) = -\int_{\mathbb{R}^P} \boldsymbol{u}_{\mathcal{G}_i} \boldsymbol{u}_{\mathcal{G}_j}^T p_0(d\boldsymbol{u}),$$

so

$$\int_{\mathbb{R}^P} \boldsymbol{u}_{\mathcal{G}_i} \boldsymbol{u}_{\mathcal{G}_j}^T p_0(d\boldsymbol{u}) = 0. \tag{10.13}$$

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

$$\int_{\mathbb{R}^P} \boldsymbol{u}_{\mathcal{G}_j} \boldsymbol{u}_{\mathcal{G}_j}^T p_0(d\boldsymbol{u}) = \int_{\mathbb{R}^P} \boldsymbol{u}_{\mathcal{G}_1} \boldsymbol{u}_{\mathcal{G}_1}^T p_0(d\boldsymbol{u}). \tag{10.14}$$

For a matrix $\boldsymbol{A} \in \mathbb{R}^{M \times M}$ and $l \in \{1, \ldots, L\}$, $[\boldsymbol{A}]_{\mathcal{G}_l, \mathcal{G}_l}$ stands for the restriction of $\boldsymbol{A}$ to rows and columns

indexed by $\mathcal{G}_l$. Combining (4.1), Lemma 4.1, (10.13), (10.14) and Von Neumann's trace inequality, we obtain

$$\begin{aligned}
&\int_{\mathbb{R}^M} \frac{1}{n}\sum_{i=1}^n \left[\boldsymbol{X}_i^T(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\right]^2 p_0{}^{\boldsymbol{D}}(d\boldsymbol{\theta}) = \frac{1}{n}\int_{\mathbb{R}^P} \left[\boldsymbol{X}\widetilde{\boldsymbol{D}^T}\boldsymbol{u}\right]^T \boldsymbol{X}\widetilde{\boldsymbol{D}^T}\boldsymbol{u} p_0(d\boldsymbol{u}) \\
= &\frac{1}{n}\int_{\mathbb{R}^P} \operatorname{tr}\left(\boldsymbol{u}^T\widetilde{\boldsymbol{D}^T}^T\boldsymbol{X}^T\boldsymbol{X}\widetilde{\boldsymbol{D}^T}\boldsymbol{u}\right) p_0(d\boldsymbol{u}) \\
= &\frac{1}{n}\operatorname{tr}\left(\left(\boldsymbol{X}\widetilde{\boldsymbol{D}^T}\right)^T\boldsymbol{X}\widetilde{\boldsymbol{D}^T}\int_{\mathbb{R}^P}\boldsymbol{u}\boldsymbol{u}^T p_0(d\boldsymbol{u})\right) \\
= &\frac{1}{n}\sum_{l=1}^L \operatorname{tr}\left(\left[\left(\boldsymbol{X}\widetilde{\boldsymbol{D}^T}\right)^T\boldsymbol{X}\widetilde{\boldsymbol{D}^T}\right]_{\mathcal{G}_l,\mathcal{G}_l}\int_{\mathbb{R}^P}\boldsymbol{u}_{\mathcal{G}_1}\boldsymbol{u}_{\mathcal{G}_1}^T p_0(d\boldsymbol{u})\right) \\
\leq &\frac{1}{n}\sum_{l=1}^L\sum_{j=1}^G \sigma_j\left(\left[\left(\boldsymbol{X}\widetilde{\boldsymbol{D}^T}\right)^T\boldsymbol{X}\widetilde{\boldsymbol{D}^T}\right]_{\mathcal{G}_l,\mathcal{G}_l}\right)\sigma_j\left(\int_{\mathbb{R}^P}\boldsymbol{u}_{\mathcal{G}_1}\boldsymbol{u}_{\mathcal{G}_1}^T p_0(d\boldsymbol{u})\right) \\
\leq &\frac{1}{n}\int_{\mathbb{R}^P}\left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2 p_0(d\boldsymbol{u})\sum_{l=1}^L \operatorname{tr}\left(\left[\left(\boldsymbol{X}\widetilde{\boldsymbol{D}^T}\right)^T\boldsymbol{X}\widetilde{\boldsymbol{D}^T}\right]_{\mathcal{G}_l,\mathcal{G}_l}\right) \\
= &\frac{1}{n}\operatorname{tr}\left(\left(\boldsymbol{X}\widetilde{\boldsymbol{D}^T}\right)^T\boldsymbol{X}\widetilde{\boldsymbol{D}^T}\right)\int_{\mathbb{R}^P}\left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2 p_0(d\boldsymbol{u}). \qquad (10.15)
\end{aligned}$$

Moreover, by inequality (4.2), Assumption 5.1 and Von Neumann's trace inequality, we obtain

$$\begin{aligned}
&\frac{1}{n}\operatorname{tr}\left(\left(\boldsymbol{X}\widetilde{\boldsymbol{D}^T}\right)^T\boldsymbol{X}\widetilde{\boldsymbol{D}^T}\right) = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{X}^T\boldsymbol{X}\widetilde{\boldsymbol{D}^T}\widetilde{\boldsymbol{D}^T}^T\right) \\
\leq &\sum_{j=1}^M \sigma_j\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)\sigma_j\left(\widetilde{\boldsymbol{D}^T}\widetilde{\boldsymbol{D}^T}^T\right) \leq \sigma_1\left(\widetilde{\boldsymbol{D}^T}\widetilde{\boldsymbol{D}^T}^T\right)\sum_{j=1}^M\sigma_j\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right) \leq \frac{M}{\mu}. \qquad (10.16)
\end{aligned}$$

Putting together (10.15) and (10.16), we get the bound

$$C_{f,\ell}\int_{\mathbb{R}^M}\frac{1}{n}\sum_{i=1}^n\left[\boldsymbol{X}_i^T(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\right]^2 p_0{}^{\boldsymbol{D}}(d\boldsymbol{\theta}) \leq C_{f,\ell}\frac{M}{\mu}\int_{\mathbb{R}^P}\left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2 p_0(d\boldsymbol{u}). \qquad (10.17)$$

Thanks to (10.11) and (10.17), inequality (10.9) becomes

$$\int_\Theta \left\|f_{\boldsymbol{\theta}}-f\right\|_n^2 p_0{}^{\boldsymbol{D}}(d\boldsymbol{\theta}) \leq \left\|f_{\boldsymbol{\theta}^*}-f\right\|_n^2 + C_{f,\ell}\frac{M}{\mu}\int_{\mathbb{R}^P}\left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2 p_0(d\boldsymbol{u}). \qquad (10.18)$$

Now, inserting (10.18) into Theorem 3.1 (with $p = p_0{}^{\boldsymbol{D}}$), we arrive at

$$\mathbb{E}\left[\left\|\widehat{f}_n - f\right\|_n^2\right] \leq \left\|f_{\boldsymbol{\theta}^*}-f\right\|_n^2 + C_{f,\ell}\frac{M}{\mu}\int_{\mathbb{R}^P}\left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2 p_0(d\boldsymbol{u}) + \frac{\beta\,\mathrm{KL}(p_0{}^{\boldsymbol{D}},\pi)}{n}. \qquad (10.19)$$

To complete the proof, it remains to bound the last two terms in the right hand side of (10.19). This is the goal of the following lemma whose proof will be detailed in Section 10.2.2.

Lemma 10.1. *Consider the same framework as the one in Theorem 6.1, we have*

$$\int_{\mathbb{R}^P}\left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2 p_0(d\boldsymbol{u}) \leq 2K_{1,g}e^{r_L\alpha^a}, \qquad (10.20)$$

and

$$\mathrm{KL}(p_0^{\boldsymbol{D}}, \pi) \leq 1 + r_L \alpha^a + \lambda \sum_{l=1}^{L} \log\left\{h\Big(\big\|[\boldsymbol{D}\boldsymbol{\theta}^*]_{\mathcal{G}_l}\big\|_2\Big)\right\} + \alpha^a \big\|\boldsymbol{D}\boldsymbol{\theta}^*\big\|_{a,\mathcal{G}}^a. \tag{10.21}$$

With $r_L = 3\sqrt{K_{a,g}}L$, it follows from (10.19) and Lemma 10.1 that

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

According to (10.7), this completes the proof of Theorem 6.1. □

10.2.2 Proof of Lemma 10.1

To prove Lemma 10.1, we need an intermediate result.

Lemma 10.2. *Let $s > L\sqrt{K_{a,g}}$. The following inequality holds*

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

where $T = \int_{\mathbb{R}^M}\prod_{l=1}^{L} g(\big\|[\boldsymbol{D}\boldsymbol{u}]_{\mathcal{G}_l}\big\|_2)d\boldsymbol{u}$.

Proof of Lemma 10.2. Let $\boldsymbol{U}$ be a random vector in $\mathbb{R}^M$ with density

$$\boldsymbol{u} \mapsto \frac{1}{T}\prod_{l=1}^{L} g(\big\|[\boldsymbol{D}\boldsymbol{u}]_{\mathcal{G}_l}\big\|_2),$$

where $T < \infty$ by Assumption 5.2-2. By Chebyshev inequality, we have

$$\begin{aligned}&\frac{1}{T}\int_{\left\{\boldsymbol{u}\in\mathbb{R}^M:\big\|\boldsymbol{D}\boldsymbol{u}\big\|_{a,\mathcal{G}}^a > s\right\}}\prod_{l=1}^{L} g(\big\|[\boldsymbol{D}\boldsymbol{u}]_{\mathcal{G}_l}\big\|_2)d\boldsymbol{u} = \mathbb{P}\left[\sum_{l=1}^{L}\big\|[\boldsymbol{D}\boldsymbol{U}]_{\mathcal{G}_l}\big\|_2^a > s\right] \\ =\ &\mathbb{P}\left[\sum_{l=1}^{L}(\big\|[\boldsymbol{D}\boldsymbol{U}]_{\mathcal{G}_l}\big\|_2^a - \mathbb{E}\left[\big\|[\boldsymbol{D}\boldsymbol{U}]_{\mathcal{G}_l}\big\|_2^a\right]) > s - \sum_{l=1}^{L}\mathbb{E}\left[\big\|[\boldsymbol{D}\boldsymbol{U}]_{\mathcal{G}_l}\big\|_2^a\right]\right] \\ \leq\ &\frac{\mathbb{E}\left[\left[\sum_{l=1}^{L}(\big\|[\boldsymbol{D}\boldsymbol{U}]_{\mathcal{G}_l}\big\|_2^a - \mathbb{E}\left[\big\|[\boldsymbol{D}\boldsymbol{U}]_{\mathcal{G}_l}\big\|_2^a\right])\right]^2\right]}{(s - \sum_{l=1}^{L}\mathbb{E}\left[\big\|[\boldsymbol{D}\boldsymbol{U}]_{\mathcal{G}_l}\big\|_2^a\right])^2} \\ =\ &\frac{\mathrm{var}\left(\sum_{l=1}^{L}\big\|[\boldsymbol{D}\boldsymbol{U}]_{\mathcal{G}_l}\big\|_2^a\right)}{(s - \sum_{l=1}^{L}\mathbb{E}\left[\big\|[\boldsymbol{D}\boldsymbol{U}]_{\mathcal{G}_l}\big\|_2^a\right])^2} \\ \leq\ &\frac{\mathbb{E}\left[\left(\sum_{l=1}^{L}\big\|[\boldsymbol{D}\boldsymbol{U}]_{\mathcal{G}_l}\big\|_2^a\right)^2\right]}{(s - \sum_{l=1}^{L}\mathbb{E}\left[\big\|[\boldsymbol{D}\boldsymbol{U}]_{\mathcal{G}_l}\big\|_2^a\right])^2}.\end{aligned} \tag{10.22}$$

Next, by Cauchy-Schwartz inequality and Remark 5.1, we obtain

$$\mathbb{E}\left[\left(\sum_{l=1}^{L}\left\|[\boldsymbol{DU}]_{\mathcal{G}_l}\right\|_2^a\right)^2\right] \leq \mathbb{E}\left[L\sum_{l=1}^{L}\left\|[\boldsymbol{DU}]_{\mathcal{G}_l}\right\|_2^{2a}\right] \leq L^2 K_{a,g} \tag{10.23}$$

and by Jensen inequality

$$s - \sum_{l=1}^{L}\mathbb{E}\left[\left\|[\boldsymbol{DU}]_{\mathcal{G}_l}\right\|_2^a\right] \geq s - \sum_{l=1}^{L}\sqrt{\mathbb{E}\left[\left\|[\boldsymbol{DU}]_{\mathcal{G}_l}\right\|_2^{2a}\right]} \geq s - L\sqrt{K_{a,g}} > 0. \tag{10.24}$$

Thus, combining (10.22), (10.23) and (10.24), we get

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

□

We now turn to the proof of Lemma 10.1

Proof of Lemma 10.1. Let us begin by the proof of inequality (10.20). We have

$$\begin{aligned}
&\int_{\mathbb{R}^P}\left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2 p_0(d\boldsymbol{u}) \\
=\ & \frac{1}{C_L\sqrt{\det(\boldsymbol{D}^T\boldsymbol{D})}}\int_{\mathbb{B}_{a,\mathcal{G}}^a(r_L)\cap\mathrm{Im}(\boldsymbol{D})}\left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2\prod_{l=1}^{L} e^{-\alpha^a\left\|\boldsymbol{u}_{\mathcal{G}_l}\right\|_2^a} g(\left\|\boldsymbol{u}_{\mathcal{G}_l}\right\|_2)d\boldsymbol{u} \\
\leq\ & \frac{1}{C_L\sqrt{\det(\boldsymbol{D}^T\boldsymbol{D})}}\int_{\mathbb{B}_{a,\mathcal{G}}^a(r_L)\cap\mathrm{Im}(\boldsymbol{D})}\left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2\prod_{l=1}^{L} g\left(\left\|\boldsymbol{u}_{\mathcal{G}_l}\right\|_2\right)d\boldsymbol{u}.
\end{aligned} \tag{10.25}$$

In the following, we show inequality (10.20) by bounding the right term of inequality (10.25). By Lemma 4.1 and Remark 5.1, we get

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

Then

$$\frac{1}{\sqrt{\det(\boldsymbol{D}^T\boldsymbol{D})}}\int_{\mathbb{B}_{a,\mathcal{G}}^a(r_L)\cap\mathrm{Im}(\boldsymbol{D})}\left\|\boldsymbol{u}_{\mathcal{G}_1}\right\|_2^2\prod_{l=1}^{L} g(\left\|\boldsymbol{u}_{\mathcal{G}_l}\right\|_2)d\boldsymbol{u} \leq K_{1,g}T. \tag{10.26}$$

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

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

Since $r_L = 3\sqrt{K_{a,g}}L > \sqrt{K_{a,g}}L$, Lemma 10.2 gives us

$$\begin{aligned}
{C_L}^{-1} &\le e^{r_L\alpha^a}\left[T\left(1-\frac{1}{T}\int_{\{\boldsymbol{u}\in\mathbb{R}^M:\left\|\boldsymbol{Du}\right\|_{a,\mathcal{G}}^a > r_L\}} \prod_{l=1}^{L} g(\left\|[\boldsymbol{Du}]_{\mathcal{G}_l}\right\|_2) d\boldsymbol{u}\right)\right]^{-1} \\
&\le e^{r_L\alpha^a} T^{-1}\left(1-\frac{L^2 K_{a,g}}{(r_L - L\sqrt{K_{a,g}})^2}\right)^{-1} \\
&= e^{r_L\alpha^a} T^{-1}\left(1-\frac{1}{4}\right)^{-1} \le 2e^{r_L\alpha^a} T^{-1}. \qquad (10.27)
\end{aligned}$$

Combining (10.26) and (10.27), (10.25) becomes

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

That concludes the proof of inequality (10.20) in Lemma 10.1.

Next, we prove inequality (10.21). Remind that $\text{supp}(\pi) = \Theta$, $\text{supp}({p_0}^{\boldsymbol{D}}) = \Theta_{{p_0}^{\boldsymbol{D}}}$. By (10.10), we get $\Theta_{{p_0}^{\boldsymbol{D}}} \subseteq \Theta$ implying that ${p_0}^{\boldsymbol{D}}$ is absolutely continuous w.r.t. π. So $\text{KL}({p_0}^{\boldsymbol{D}}, \pi) < \infty$ which can be bounded. The bound in (10.21) can be proved as follows. By Lemma 4.1, we have

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

where p_0 is a probability measure in $\mathbb{R}^P$ defined in (10.12). We know that $\boldsymbol{t}^* = \boldsymbol{D}\boldsymbol{\theta}^*$, according to the fact that $\left\|\boldsymbol{t}_{\mathcal{G}_l}\right\|_2^a - \left\|\boldsymbol{t}_{\mathcal{G}_l} - \boldsymbol{t}^*_{\mathcal{G}_l}\right\|_2^a \leq \left\|\boldsymbol{t}^*_{\mathcal{G}_l}\right\|_2^a$ and Assumption 5.2-4, we get

$$\mathrm{KL}(p_0{}^{\boldsymbol{D}}, \pi) \leq \log\left(\frac{C_{\alpha,g,R}}{C_L}\right) + \alpha^a \left\|\boldsymbol{D}\boldsymbol{\theta}^*\right\|_{a,\mathcal{G}}^a + \lambda \sum_{l=1}^{L} \log\left\{h\Big(\left\|[\boldsymbol{D}\boldsymbol{\theta}^*]_{\mathcal{G}_l}\right\|_2\Big)\right\}. \tag{10.28}$$

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 we get

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

Combining this with the bound of ${C_L}^{-1}$ in (10.27), we obtain

$$\log\left(\frac{C_{\alpha,g,R}}{C_L}\right) \leq r_L \alpha^a + \log(2) \leq 1 + r_L \alpha^a. \tag{10.29}$$

Inserting (10.29) into (10.28), we get inequality (10.21). This completes the proof. □

10.3 Proofs of Section 7

10.3.1 Proof of Lemma 7.1

In view of **(H.1)**, V is also proper lsc and bounded from below. This together with **(H.2)** shows that V is prox-bounded for all $\gamma \in]0, \frac{\beta}{2\|\boldsymbol{X}\|^2}[$, and for any $\boldsymbol{\theta}$, $\frac{1}{2\gamma}\left\|\boldsymbol{\theta} - \cdot\right\|_{\boldsymbol{M}_\gamma}^2 + V$ is proper lsc and level-bounded uniformly in $(\boldsymbol{\theta}, \gamma)$. This entails that the set of minimizers of this function, i.e. $\mathrm{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}(\boldsymbol{\theta})$, is a non-empty compact set. The same reasoning leads to compactness of $\mathrm{prox}_{\gamma J}(\boldsymbol{\theta})$.

Now, we have

$$\begin{aligned}
\mathrm{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}(\boldsymbol{\theta}) &= \underset{\boldsymbol{w}\in\mathbb{R}^M}{\mathrm{Argmin}} \frac{1}{2\gamma}\left\|\boldsymbol{w} - \boldsymbol{\theta}\right\|_{\boldsymbol{M}_\gamma}^2 + V(\boldsymbol{w}) \\
&= \underset{\boldsymbol{w}\in\mathbb{R}^M}{\mathrm{Argmin}} \frac{1}{2}\left\|\boldsymbol{w} - \boldsymbol{\theta}\right\|_2^2 - \frac{\gamma}{\beta}\left\|\boldsymbol{X}(\boldsymbol{w} - \boldsymbol{\theta})\right\|_2^2 + \frac{\gamma}{\beta}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\right\|_2^2 + \gamma G(\boldsymbol{w}).
\end{aligned}$$

By the Pythagoras relation, we then get

$$\begin{aligned}
\mathrm{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}(\boldsymbol{\theta}) &= \underset{\boldsymbol{w}\in\mathbb{R}^M}{\mathrm{Argmin}} \frac{1}{2}\left\|\boldsymbol{w} - \boldsymbol{\theta}\right\|_2^2 + \frac{\gamma}{\beta}\left(\frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2 - \big(\boldsymbol{X}(\boldsymbol{\theta} - \boldsymbol{w})\big)^T\big(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\big)\right) + \gamma G(\boldsymbol{w}) \\
&= \underset{\boldsymbol{w}\in\mathbb{R}^M}{\mathrm{Argmin}} \frac{1}{2}\left\|\boldsymbol{w} - \boldsymbol{\theta}\right\|_2^2 - \frac{\gamma}{\beta}\big(\boldsymbol{w} - \boldsymbol{\theta}\big)^T\big(\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})\big) + \gamma G(\boldsymbol{w}) \\
&= \underset{\boldsymbol{w}\in\mathbb{R}^M}{\mathrm{Argmin}} \frac{1}{2}\left\|\boldsymbol{w} - \big(\boldsymbol{\theta} - 2\gamma/\beta\boldsymbol{X}^T\big(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\big)\big)\right\|_2^2 + \gamma G(\boldsymbol{w}) \\
&= \mathrm{prox}_{\gamma G}(\boldsymbol{\theta} - \gamma\nabla F(\boldsymbol{\theta})).
\end{aligned}$$

For the last claim, suppose that $\boldsymbol{\theta} \in \mathrm{Argmin}(V) \neq \emptyset$ and bounded but $\boldsymbol{\theta} \notin \mathrm{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}(\boldsymbol{\theta})$. Thus, for any $\boldsymbol{p} \in \mathrm{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}(\boldsymbol{\theta})$, we have $\boldsymbol{p} \neq \boldsymbol{\theta}$ and

$$V(\boldsymbol{p}) < \frac{1}{2\gamma}\left\|\boldsymbol{p} - \boldsymbol{\theta}\right\|_{\boldsymbol{M}_\gamma}^2 + V(\boldsymbol{p}) \leq V(\boldsymbol{\theta}),$$

leading to a constradiction with $\boldsymbol{\theta}$ is a minimizer of V. □

10.3.2 Proof of Lemma 7.2

The continuity and finiteness properties follow from [50, Theorem 1.17(c)] (see also[50, Theorem 1.25]), where we use (H.1) and (H.2). For the second claim, we have $\forall \boldsymbol{\theta} \in \mathbb{R}^M$

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

Moreover, let $\boldsymbol{p} \in \operatorname{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}(\boldsymbol{\theta})$. Then, $\forall \delta]\gamma, \frac{\beta}{2|||\boldsymbol{X}|||^2}[$,

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

This together with continuity concludes the proof. □

10.3.3 Proof of Lemma 7.3

By virtue of Lemma 7.1 and (H.3), $\operatorname{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}$ is clearly non-empty and single valued. The continuity property follows from [50, Theorem 1.17(b)] (see also[50, Theorem 1.25]) and single-valuedness. By Lemma 7.2, V is prox-bounded with threshold $\beta/(2|||X|||^2)$ and ${}^{\boldsymbol{M}_\gamma, \gamma} V(\boldsymbol{\theta})$ is finite. Invoking [50, Example 10.32], we get that $-{}^{\boldsymbol{M}_\gamma, \gamma} V$ is strictly continuous, subdifferentially regular and

$$\partial\big(-{}^{\boldsymbol{M}_\gamma, \gamma} V\big)(\boldsymbol{\theta}) = \Big\{\gamma^{-1} \boldsymbol{M}_\gamma \Big(\operatorname{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}(\boldsymbol{\theta}) - \boldsymbol{\theta}\Big)\Big\}, \quad \forall \boldsymbol{\theta} \in \mathbb{R}^M.$$

Combining this with [50, Theorem 9.18] applied to $-{}^{\boldsymbol{M}_\gamma, \gamma} V$, we obtain that ${}^{\boldsymbol{M}_\gamma, \gamma} V$ is differentiable and its gradient is precisely as given. □

10.3.4 Proof of Lemma 7.4

(i) Assumption (H.4), [50, Exercise 8.8(c)] and convexity of F imply that ∂V is hypomonotone of modulus r. In turn $S = \partial V + r' \boldsymbol{M}_\gamma = \partial\Big(V + \frac{r'}{2} \|\cdot\|_{\boldsymbol{M}_\gamma}^2\Big)$ is monotone with $r' = \frac{r}{1 - 2\gamma |||\boldsymbol{X}|||^2/\beta}$, or equivalently that $V + \frac{r'}{2} \|\cdot\|_{\boldsymbol{M}_\gamma}^2$ is convex. Let $\delta = \gamma^{-1} - r'$ and $W(\boldsymbol{w}, \boldsymbol{\theta}) = V(\boldsymbol{w}) + \frac{r'}{2} \|\boldsymbol{w} - \boldsymbol{\theta}\|_{\boldsymbol{M}_\gamma}^2$. Thus

$$V(\boldsymbol{w}) + \frac{1}{2\gamma} \|\boldsymbol{w} - \boldsymbol{\theta}\|_{\boldsymbol{M}_\gamma}^2 = W(\boldsymbol{w}, \boldsymbol{\theta}) + \frac{\delta}{2} \|\boldsymbol{w} - \boldsymbol{\theta}\|_{\boldsymbol{M}_\gamma}^2.$$

$W(\cdot, \boldsymbol{\theta})$ is a convex function on $\mathbb{R}^M$ and assumption (H.5) is equivalent to $\delta > 0$. Altogether, this entails that $W(\cdot, \boldsymbol{\theta}) + \frac{\delta}{2} \|\cdot - \boldsymbol{\theta}\|_{\boldsymbol{M}_\gamma}^2$ is strongly convex uniformly in $\boldsymbol{\theta}$ and γ complying with (H.5). It then follows that $\operatorname{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}$ is single-valued. We have

$$\boldsymbol{M}_\gamma + \gamma \partial V = \gamma(\delta \boldsymbol{M}_\gamma + S) = \gamma\delta\big(\boldsymbol{M}_\gamma + \delta^{-1} S\big).$$

By Fermat's rule, we then get

$$\operatorname{prox}_{\gamma V}^{\boldsymbol{M}_\gamma} = (\boldsymbol{M}_\gamma + \gamma \partial V)^{-1} \circ \boldsymbol{M}_\gamma = \big(\boldsymbol{M}_\gamma + \delta^{-1} S\big)^{-1} \circ (\gamma\delta)^{-1} \boldsymbol{M}_\gamma,$$

and the latter operator is well-defined as a single-valued operator since S is maximal monotone; see [3, Proposition 3.22 (ii)(d)]. Let $\boldsymbol{p} = \operatorname{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}(\boldsymbol{\theta})$ and $\boldsymbol{p}' = \operatorname{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}(\boldsymbol{\theta}')$. It then follows that

$$\boldsymbol{M}_\gamma\boldsymbol{\theta} - \gamma\delta\boldsymbol{M}_\gamma\boldsymbol{p} \in \gamma S(\boldsymbol{p}) \text{ and } \boldsymbol{M}_\gamma\boldsymbol{\theta}' - \gamma\delta\boldsymbol{M}_\gamma\boldsymbol{p}' \in \gamma S(\boldsymbol{p}'),$$

and monotonicity of S yields

$$\big(\boldsymbol{p}' - \boldsymbol{p}\big)^T\big(\boldsymbol{M}_\gamma(\boldsymbol{\theta}' - \boldsymbol{\theta})\big) \geq \gamma\delta\big\|\boldsymbol{p}' - \boldsymbol{p}\big\|_{\boldsymbol{M}_\gamma}^2 \geq \gamma\delta\big(1 - 2\gamma|||\boldsymbol{X}|||^2/\beta\big)\big\|\boldsymbol{p}' - \boldsymbol{p}\big\|_2^2.$$

Using Cauchy-Schwartz inequality and $|||\boldsymbol{M}_\gamma||| \leq 1$, we then obtain

$$\big\|\boldsymbol{p}' - \boldsymbol{p}\big\|_2 \leq \kappa\big\|\boldsymbol{\theta}' - \boldsymbol{\theta}\big\|_2,$$

where $\kappa^{-1} = \gamma\delta\Big(1 - 2\gamma|||\boldsymbol{X}|||^2/\beta\Big) = 1 - \gamma\Big(r + 2|||\boldsymbol{X}|||^2/\beta\Big)$.

Let now $\boldsymbol{p} = \operatorname{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}(\boldsymbol{M}_\gamma^{-1}\boldsymbol{\theta})$ and $\boldsymbol{p}' = \operatorname{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}(\boldsymbol{M}_\gamma^{-1}\boldsymbol{\theta}')$. With the same arguments as before, we get

$$\big(\boldsymbol{p}' - \boldsymbol{p}\big)^T\big(\boldsymbol{\theta}' - \boldsymbol{\theta}\big) \geq \gamma\delta\big\|\boldsymbol{p}' - \boldsymbol{p}\big\|_{\boldsymbol{M}_\gamma}^2 \geq 0,$$

which shows monotonicity of $\operatorname{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}$ (in fact even co-coercivity).

(ii) Since $\operatorname{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}$ is single-valued and Lipschitz continuous, C^1 smoothness and Lipschiz continuity of the gradient follow from Lemma 7.3. Let's now turn to hypomonotonicity. We have

$$\begin{aligned}
&\boldsymbol{p} = \operatorname{prox}_{\gamma V}^{\boldsymbol{M}_\gamma}(\boldsymbol{\theta}) = \big(\boldsymbol{M}_\gamma + \delta^{-1}S\big)^{-1}\Big((\gamma\delta)^{-1}\boldsymbol{M}_\gamma\boldsymbol{\theta}\Big)\\
\iff &\gamma^{-1}\boldsymbol{M}_\gamma\boldsymbol{\theta} - \delta\boldsymbol{M}_\gamma\boldsymbol{p} \in S(\boldsymbol{p})\\
\iff &\boldsymbol{p} \in S^{-1}\big(\gamma^{-1}\boldsymbol{M}_\gamma\boldsymbol{\theta} - \delta\boldsymbol{M}_\gamma\boldsymbol{p}\big)\\
\iff &(\delta\gamma)^{-1}\boldsymbol{M}_\gamma\boldsymbol{\theta} \in \boldsymbol{M}_\gamma\Big((\delta\boldsymbol{M}_\gamma)^{-1} + S^{-1}\Big)\big(\gamma^{-1}\boldsymbol{M}_\gamma\boldsymbol{\theta} - \delta\boldsymbol{M}_\gamma\boldsymbol{p}\big).
\end{aligned}$$

Since S is monotone, so is S^{-1}. In turn, with similar arguments as for S above, $\Big((\delta\boldsymbol{M}_\gamma)^{-1} + S^{-1}\Big)^{-1}$ is well-defined as a single-valued operator, and is monotone. Therefore,

$$\gamma^{-1}\boldsymbol{M}_\gamma\boldsymbol{\theta} - \delta\boldsymbol{M}_\gamma\boldsymbol{p} = \Big((\delta\boldsymbol{M}_\gamma)^{-1} + S^{-1}\Big)^{-1}\Big((\delta\gamma)^{-1}\boldsymbol{\theta}\Big).$$

Using Lemma 7.3, we obtain

$$\delta\gamma\nabla^{\boldsymbol{M}_\gamma}V(\boldsymbol{\theta}) = \big(\delta - \gamma^{-1}\big)\boldsymbol{M}_\gamma\boldsymbol{\theta} + \Big((\delta\boldsymbol{M}_\gamma)^{-1} + S^{-1}\Big)^{-1}\Big((\delta\gamma)^{-1}\boldsymbol{\theta}\Big).$$

As we argued above, one the one hand, $\Big((\delta\boldsymbol{M}_\gamma)^{-1} + S^{-1}\Big)^{-1}$ is monotone. On the other hand, $\delta - \gamma^{-1} = -r'$. Clearly $\big(\delta - \gamma^{-1}\big)\boldsymbol{M}_\gamma$ is symmetric definite negative, whence we deduce the hypomonotonicity claim. □

10.3.5 Proof of Proposition 7.1

Recall that we denote with the same symbol the measure and its density with respect to the Lebesgue measure. Thus

$$\left\|\mu_\gamma - \mu_n\right\|_{\mathrm{TV}} = \int_{\mathbb{R}^M} \left|\mu_\gamma(\boldsymbol{\theta}) - \mu_n\boldsymbol{\theta}\right| d\boldsymbol{\theta},$$

where

$$\mu_\gamma(\boldsymbol{\theta}) = \exp\left(-^{\boldsymbol{M}_\gamma,\gamma}V(\boldsymbol{M}_\gamma^{-1/2}\boldsymbol{\theta})\right)/Z_\gamma \text{ and } \mu_n(\boldsymbol{\theta}) = \exp\left(-V(\boldsymbol{\theta})\right)/Z,$$

and $Z = \int_{\mathbb{R}^M} \exp\left(-V(\boldsymbol{u})\right) d\boldsymbol{u}$. We have $\det(\boldsymbol{M}_\gamma) \to 1$ as $\gamma \to 0$. In view of Lemma 7.2, applying the monotone convergence theorem, we conclude that $Z_\gamma \to Z$. This together with Lemma 7.2 yield that μ_γ converges to μ_n pointwise. We conclude using Scheffé(-Riesz) theorem [37, 53]. □

10.3.6 Proof of Theorem 7.1

By Lemma 7.4(i) and Lemma 7.1, we have Lipschitz continuity of the operator $S = \frac{\boldsymbol{I}_M - \mathrm{prox}_{\gamma G} \circ (\boldsymbol{I}_M - \gamma \nabla F)}{\gamma}$. Let κ be this Lipschitz constant (which depends on γ). By these two lemmas again, we have for any bounded $\theta \in \mathrm{Argmin}(V)$

$$\left\|S(\boldsymbol{w})\right\|_2^2 = \left\|S(\boldsymbol{w}) - S(\boldsymbol{\theta})\right\|_2^2 \leq \kappa^2 \left\|\boldsymbol{w} - \boldsymbol{\theta}\right\|_2^2 \leq 2\kappa^2 \left\|\boldsymbol{w}\right\|_2^2 + 2\kappa^2 \left\|\boldsymbol{\theta}\right\|_2^2 \leq c(1 + \left\|\boldsymbol{w}\right\|_2^2),$$

where $c = 2\kappa^2 \max(1, \left\|\boldsymbol{\theta}\right\|_2^2)$. Thus the global Lipschitz and growth conditions required on the drift coefficient are verified, and the bound follows from [56, Theorem 7.3, Chapter II] or [35, Theorem 10.2.2 and Remark 10.2.3]. Using Jensen's inequality we get the bound on the moment. □

10.3.7 Proof of Lemma 7.5

This is a probably known result, for which we provide a simple proof. Since $J = G \circ \boldsymbol{D}^{-1}$ is separable and w is continuous and lower-bounded, we have

$$\min_{\boldsymbol{w} \in \mathbb{R}^P} \frac{1}{2} \left\|\boldsymbol{w} - \boldsymbol{u}\right\|_2^2 + \gamma J(\boldsymbol{w}) = \sum_{l=1}^{L} \min_{\boldsymbol{v} \in \mathbb{R}^G} \frac{1}{2} \left\|\boldsymbol{v} - \boldsymbol{u}_{\mathcal{G}_l}\right\|_2^2 + \gamma w(\left\|\boldsymbol{v}\right\|_2),$$

and thus, $\forall l \in \{1, \ldots, L\}$,

$$\left[\mathrm{prox}_{\gamma J}(\boldsymbol{u})\right]_{\mathcal{G}_l} = \mathrm{Argmin}_{\boldsymbol{v} \in \mathbb{R}^G} \frac{1}{2} \left\|\boldsymbol{v} - \boldsymbol{u}_{\mathcal{G}_l}\right\|_2^2 + \gamma w(\left\|\boldsymbol{v}\right\|_2). \tag{10.30}$$

If $\boldsymbol{u}_{\mathcal{G}_l} = 0$, then as w is an increasing function, $\left[\mathrm{prox}_{\gamma J}(\boldsymbol{u})\right]_{\mathcal{G}_l} = 0$. For $\boldsymbol{u}_{\mathcal{G}_l} \neq 0$, by isotropy of problem (10.30), we can write

$$\min_{\boldsymbol{v} \in \mathbb{R}^G} \frac{1}{2} \left\|\boldsymbol{v} - \boldsymbol{u}_{\mathcal{G}_l}\right\|_2^2 + \gamma w(\left\|\boldsymbol{v}\right\|_2) = \min_{t \geq 0} \gamma w(t) + \left(\min_{\left\|\boldsymbol{v}\right\|_2 = t} \frac{1}{2} \left\|\boldsymbol{v} - \boldsymbol{u}_{\mathcal{G}_l}\right\|_2^2 \right). \tag{10.31}$$

The inner minimization problem amounts to solving for the orthogonal projector on the ℓ_2 sphere in $\mathbb{R}^G$ of radius t, which is obviously $\boldsymbol{v} = t \frac{\boldsymbol{u}_{\mathcal{G}_l}}{\left\|\boldsymbol{u}_{\mathcal{G}_l}\right\|_2}$ since $\boldsymbol{u}_{\mathcal{G}_l} \neq 0$. Inserting this into (10.31) and rearranging the terms, (10.30) becomes

$$\left[\mathrm{prox}_{\gamma J}(\boldsymbol{u})\right]_{\mathcal{G}_l} = \frac{\boldsymbol{u}_{\mathcal{G}_l}}{\left\|\boldsymbol{u}_{\mathcal{G}_l}\right\|_2} \mathrm{Argmin}_{t \geq 0} \frac{1}{2} \left(t - \left\|\boldsymbol{u}_{\mathcal{G}_l}\right\|_2\right)^2 + \gamma w(t) = \frac{\boldsymbol{u}_{\mathcal{G}_l}}{\left\|\boldsymbol{u}_{\mathcal{G}_l}\right\|_2} \mathrm{prox}_{\gamma w}(\left\|\boldsymbol{u}_{\mathcal{G}_l}\right\|_2),$$

where we used even-symmetry of w. □

10.3.8 Proof of Lemma 7.6

We can see directly that

$$\widetilde{w}(t) = \alpha^a t^a + c\log(\tau^b + t^b) - c\log(\tau^b)$$

is non negative, differentiable and non decreasing on $]0, +\infty[$. Let $u : t \in]0, +\infty[\mapsto t + \widetilde{w}'(t)$, it suffices to prove that this function is strictly unimodal to apply [2, Theorem 1]. Indeed, one can show by calculating the first and second derivatives of u that

$$\lim_{t\to 0^+} u'(t) = -\infty, \ \lim_{t\to+\infty} u'(t) = 1,$$

and

$$u''(t) > 0, \ \forall t \in]0, +\infty[.$$

These two arguments show that there exists a unique minimum of u on $]0, +\infty[$. So, we can apply [2, Theorem 1] to get the proximity of $\widetilde{w}$ which is the same as that of w since the minimization problem is unchanged by adding a constant. □

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] A. Antoniadis and J. Fan. Regularization of Wavelet Approximations. *Journal of the American Statistical Association*, 96:939–967, 2001.

[3] H. H. Bauschke, J. M. Borwein, and P. L. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on Control and Optimization*, 42(2):596–636, 2003.

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

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

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

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

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

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

[10] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.

[11] 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.

[12] E. J. Candès. Ridgelets: Estimating with Ridge Functions. *Annals of Statistics*, 31, 1999. 1561–1599.

[13] 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.

[14] R. Coifman and D. Donoho. Translation invariant de-noising. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*. Springer-Verlag, 1995. 125–150.

[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. Ryshik. *Table of Integrals, Series, and Products*. Academic Press, London, 4th edition, 1965.

[31] D. Higham, X. Mao, and A. Stuart. Strong convergence of euler-type methods for nonlinear stochastic differential equations. *SIAM J. Numer. Anal.*, 40(3):1041–1063, 2003.

[32] J. Huang, J. L. Horowitz, and F. Wei. Variable selection in nonparametric additive models. *Ann. Statist.*, 38(4):2282–2313, 08 2010.

[33] N. Ikeda and S. Watanabe. *Stochastic differential equations and diffusion processes*. NH, 1989.

[34] J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Inf. Comput.*, 132(1):1–63, Jan. 1997.

[35] P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*. Stochastic Modelling and Applied Probability. Springer, 1995.

[36] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines, 2008.

[37] N. Kusolitsch. Why the theorem of scheffé should be rather called a theorem of riesz. *Periodica Mathematica Hungarica*, 61(1):225–229, 2010.

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

[39] G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006.

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

[41] J. Maly. Lectures on change of variables in integral. Preprint 305, Department of Mathematics, University of Helsinki, 2001.

[42] L. Meier, S. V. D. Geer, P. BÃijhlmann, and E. T. H. ZÃijrich. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 2008.

[43] L. Meier, S. van de Geer, and P. BÃijhlmann. High-dimensional additive modeling. *Ann. Statist.*, 37(6B):3779–3821, 12 2009.

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

[45] M. Pereyra. Proximal markov chain monte carlo algorithms. *Statistics and Computing*, 26(4):745–760, 2016.

[46] 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.

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

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

[49] G. O. Roberts and R. L. Tweedie. Exponential Convergence of Langevin Distributions and Their Discrete Approximations. *Bernoulli*, 2(4):341–363, 1996.

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

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

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

[53] H. Scheffe. A useful convergence theorem for probability distributions. *Ann. Math. Statist.*, 18(3):434–438, 09 1947.

[54] 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.

[55] 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.

[56] M. Xuerong. *Stochastic differential equations and applications*. Woodhead Publishing, 2007.

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

[58] 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.