Sampling from non-smooth distribution through Langevin diffusion

Tung Duy Luu, Jalal Fadili, Christophe Chesneau

Sampling from non-smooth distribution through Langevin diffusion

Tung Duy Luu* Jalal Fadili* Christophe Chesneau†

Abstract

In this paper, we propose algorithms for sampling from the distributions whose density is non-smoothed nor log-concave. Our algorithms are based on the Langevin diffusion on the regularized counterpart of density by the Moreau-Yosida regularization. These results are then applied to compute the exponentially weighted aggregates for high dimensional framework with a general class of priors encouraging objects which conform to some notion of simplicity/complexity. Some popular priors are detailed and implemented on some numerical experiments.

1 Introduction

1.1 Problem statement

We consider the following linear regression

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

where $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector, $\boldsymbol{X} \in \mathbb{R}^{n\times M}$ is a design matrix, $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$ is a random vector of errors, and $\boldsymbol{\theta}_0 \in \mathbb{R}^M$ is the unknown regression vector of interest. When the dismension of $\boldsymbol{\theta}_0$ is larger than the number of observations (i.e., $M \geq n$), (1.1) becomes ill-posed and can not be estimated by the classic Least-Square estimators.

The idea of aggregating elements in a dictionary has been introduced in machine learning to combine different techniques (see [39, 64]) with some procedures such as bagging [10], boosting [32, 54] and random forests [1, 6–8, 11, 33]. In the recent years, there has been a flurry of research on the use of low-complexity regularization (among which sparsity and low-rank are the most popular) in various areas including statistics and machine learning in high dimension. The idea is that even if the ambient dimension M of $\boldsymbol{\theta}_0$ is very large, its intrinsic dimension is much smaller than the sample size n. This makes it possible to build an estimate $\boldsymbol{X}\widehat{\boldsymbol{\theta}}$ with good provable performance guarantees under appropriate conditions. In literature, the information of sparsity/low-complexity has been taken into account through two families of estimators: Penalized Estimators and Exponentially Weighted Aggregates (EWA).

1.2 Variational/Penalized Estimators

The penalized approach consists in imposing on the set of candidate solutions some prior structure on the object to be estimated. The class of estimators are obtained by solving the convex optimization problem

$$\hat{\boldsymbol{\theta}}^{\mathrm{PEN}} \in \operatorname*{Argmin}_{\boldsymbol{\theta}\in\mathbb{R}^M} \Big\{ V(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) + G_\lambda(\boldsymbol{\theta}) \Big\}, \tag{1.2}$$

*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.

where $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ is a general loss function, $G_\lambda : \mathbb{R}^M \to \mathbb{R}$ is the regularizing penalty promoting some specific notion of simplicity/low-complexity, and $\lambda > 0$ is the regularization parameter. In the present paper, L and G_λ are imposed on some assumptions covering a large class of estimators. Regularization is now a central theme in many fields including statistics, machine learning and inverse problems. A prominent member covered by (1.2) is the Lasso [9, 12, 13, 17, 23, 45, 59] and its variants such the analysis/fused Lasso [52, 61] or group Lasso [2, 3, 65, 69]. Another example is the nuclear norm minimization for low rank matrix recovery motivated by various applications including robust PCA, phase retrieval, control and computer vision [14, 15, 31, 48]. See [12, 42, 62, 63] for generalizations and comprehensive reviews.

A classical choice of L is the quadratic loss (i.e., $L(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2$ where $\|.\|_2$ denote the Euclidean norm). The associated estimator can be viewed as a Maximum a Posteriori (MAP) estimator with the prior $p_\lambda(\boldsymbol{\theta}) \propto 1/(2\sigma^2)\exp(-G_\lambda(\boldsymbol{\theta}))$. Indeed, since $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I}_n)$, we have that $p(\boldsymbol{y}|\boldsymbol{\theta}) \propto \exp\left(-\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2/(2\sigma^2)\right)$. Then the posteriori is defined by Bayes' rule as

$$p_\lambda(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p_\lambda(\boldsymbol{\theta})}{\int_{\mathbb{R}^M} p(\boldsymbol{y}|\boldsymbol{\theta}')p_\lambda(\boldsymbol{\theta}')d\boldsymbol{\theta}'} \propto \exp\left(-\frac{V(\boldsymbol{\theta})}{2\sigma^2}\right). \tag{1.3}$$

Thus, the minimization of V corresponds to the maximization of the posteriori (1.3).

1.3 Exponential Weighted Aggregation (EWA)

An alternative to the variational estimator (1.2) is the aggregation by exponential weighting which combines all of candidate solutions with the aggregators promoting the prior information. The aggregators are defined via the probability density function

$$\hat{\mu}(\boldsymbol{\theta}) = \frac{\exp\left(-V(\boldsymbol{\theta})/\beta\right)}{\int_\Theta \exp\left(-V(\boldsymbol{\omega})/\beta\right)d\boldsymbol{\omega}}, \tag{1.4}$$

where $\beta > 0$ is called temperature parameter. If all $\boldsymbol{\theta}$ are candidates to estimate the true vector $\boldsymbol{\theta}_0$, then $\Theta = \mathbb{R}^M$. The aggregate is thus defined by

$$\hat{\boldsymbol{\theta}}^{\mathrm{EWA}} = \int_{\mathbb{R}^M} \boldsymbol{\theta}\hat{\mu}(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{1.5}$$

Aggregation by exponential weighting has been widely considered in the statistical and machine learning literatures, see e.g. [18, 19, 21, 22, 27, 34, 38, 43, 49, 68] to name a few.

1.4 The Langevin diffusion

The penalized estimators are computed by solving the optimization problems (1.2). Several types of penalty are well-studied, e.g. ℓ_0, ℓ_1, clipped ℓ_1, transformed ℓ_1, SCAD, FIRM, non-negative garotte and elastic net penalties (see [24, 28, 29, 44, 56, 57, 60, 66, 70]).

The present paper focus on the computation of EWA estimators which is a recent challenge in literature. Computing $\hat{\boldsymbol{\theta}}^{\mathrm{EWA}}$ in (1.5) 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 $\hat{\mu}$ by constructing an appropriate Markov chain whose stationary distribution is $\hat{\mu}$, 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 [22], we here use the Langevin diffusion process; see [50].

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)

$$d\boldsymbol{L}(t) = \frac{1}{2}\boldsymbol{\rho}(\boldsymbol{L}(t))dt + d\boldsymbol{W}(t),\ t > 0,\ \boldsymbol{L}(0) = \boldsymbol{l}_0, \tag{1.6}$$

where $\boldsymbol{\rho} = \nabla \log \mu$, μ 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 (1.6) has a unique strong solution and, $\boldsymbol{L}(t)$ has a stationary distribution with density precisely μ [50, Theorem 2.1]. $\boldsymbol{L}(t)$ is therefore interesting for sampling from μ. In particular, this opens the door to approximating integrals $\int_{\mathbb{R}^M} f(\boldsymbol{\theta})\mu(\boldsymbol{\theta})d\boldsymbol{\theta}$, where $f : \mathbb{R}^M \to \mathbb{R}$, by the average value of a Langevin diffusion, i.e., $\frac{1}{T}\int_0^T f(\boldsymbol{L}(t))dt$ for a large enough T. Under additional assumptions on μ, the expected squared error of the approximation can be controlled [67].

Forward Euler discretization In practice, in simulating the diffusion sample path, we cannot follow exactly the dynamic defined by the SDE (1.6). 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

$$\frac{\delta}{T}\sum_{k=0}^{\lfloor T/\delta \rfloor - 1} \boldsymbol{L}_k, \tag{1.7}$$

where $\lfloor T/\delta \rfloor$ denotes the interger part of T/δ. It is then natural to approximate $\widehat{\boldsymbol{\theta}}$ by applying this discretization strategy to the Langevin diffusion with μ as the target density. However, quantitative consistency guarantees of this discretization require μ (hence $\boldsymbol{\rho}$) to be sufficiently smooth. For a complete review about sampling by Langevin diffusion from smooth and log-concave densities, we refer the studies in [20]. To cope with non-smooth densities, several works have proposed to replace $\log \mu$ with a smoothed version (typically involving the Moreau-Yosida regularization/envelope, see Definition 2.2) [22, 25, 26, 46]. In [26, 46] for instance, the authors proposed proximal-type algorithms to sample from possibly non-smooth log-concave densities μ using the forward Euler discretization and the Moreau-Yosida regularization. In [46][1], $-\log \mu$ is replaced with its Moreau envelope, while in [26], it is assumed that $-\log \mu = F + H$, F is convex Lipschitz continuously differentiable, and H is a proper closed convex function replaced by its Moreau envelope. In both these works, convexity plays a crucial role to get quantitative convergence guarantees. Proximal steps within MCMC methods have been recently proposed for some simple (convex) signal processing problems [16], though without any guarantees.

1.5 Contributions

Our main contributions are summarized as follows.

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

- We aim to enlarge the family of μ covered by [22, 25, 26, 46] by relaxing some conditions. Especially, in our study, μ is not differentiable, nor log-concave. Two algorithms are proposed with some proven theoretical guarantees.

- We apply our algorithms to several penalties in literature: ℓ_1, ℓ_∞, SCAD and FIRM. All of them are considered in a very general forms (group-analysis) as Fused Group Lasso.

- In numerical aspect, our algorithms are applied to compute EWA estimators with these penalties in some classical problems: Compressed Sensing, Inpainting and Deconvolution.

1.6 Paper organization

Some preliminaries, definitions and notations are introduced in Section 2. Section 3 presents some assumptions on μ to exploit the required properties for Moreau-Yosida smoothing on SDE (1.6). The well-posedness of the smoothed SDE and the convergence of the associated Riemann sum (1.7) to $\hat{\boldsymbol{\theta}}^{\mathrm{EWA}}$ are proven in Section 4. Section 5 considers the previous theoretical analysis for prox-regular functions. From these analysis, two algorithms are proposed in Section 6 and applied in Section 7 to establish the EWA estimators with several penalties. The numerical experiments are described in Section 8. The proofs of all results are collected in Section 9.

2 Notations and Preliminaries

Before proceeding, let us introduce some notations and definitions.

Vectors and matrices For a d-dimensional Euclidean space $\mathbb{R}^d$, we endow it with its usual inner product $\langle ., . \rangle$ and associated norm $\|.\|_2$. $\boldsymbol{I}_d$ is the identity matrix on $\mathbb{R}^d$. For $p \geq 1$, $\|.\|_p$ will denote the ℓ_p norm of a vector with the usual adaptation for $p = +\infty$.

Let $\boldsymbol{M} \in \mathbb{R}^{p\times p}$ symmetric positive definite. For a p-dimensional Euclidean space $\mathbb{R}^p$, we endow it with the inner product $\langle ., . \rangle_{\boldsymbol{M}}$ defined as

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\boldsymbol{M}} = \left\langle \boldsymbol{M}^{1/2}\boldsymbol{u}, \boldsymbol{M}^{1/2}\boldsymbol{v} \right\rangle = \langle \boldsymbol{M}\boldsymbol{u}, \boldsymbol{v} \rangle, \ \forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^p,$$

and $\|.\|_{\boldsymbol{M}}$ its associated norm. For a matrix $\boldsymbol{M}$ we denote $\sigma_{\min}(\boldsymbol{M})$ its smallest singular value and $|||\boldsymbol{M}|||$ its spectral norm, i.e., its largest singular value. We have the following equivalence between $\|.\|_{\boldsymbol{M}}$ and ℓ_2 norm

$$\sigma_{\min}(\boldsymbol{M})\|\boldsymbol{v}\|_2^2 \leq \|\boldsymbol{v}\|_{\boldsymbol{M}}^2 \leq |||\boldsymbol{M}|||\|\boldsymbol{v}\|_2^2.$$

Let $\boldsymbol{v} \in \mathbb{R}^p$ and $\mathcal{A} \subset \{1, \ldots, p\}$. We denote $\boldsymbol{v}_{\mathcal{A}}$ the sub-vector whose entries are those of $\boldsymbol{v}$ restricted to the indices in $\mathcal{A}$. For any matrix $\boldsymbol{M}$, $\boldsymbol{M}^T$ denotes its transpose. For a $x \in \mathbb{R}$ we denote $\lfloor x \rfloor$ the stands of integer part of x.

Sets For a set Ω, denote $|\Omega|$ its cardinality, I_Ω its characteristic function, i.e., 1 if the argument is in Ω and 0 otherwise, and ι_Ω its the indicator function, i.e., 0 if the argument is in Ω and $+\infty$ otherwise.

Functions For a function $V : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$, the effective domain of V is defined by $\mathrm{dom}(V) = \{\boldsymbol{x} \in \mathbb{R}^p : V(\boldsymbol{x}) < +\infty\}$ and V is proper if $V(\boldsymbol{x}) > -\infty$ for all $\boldsymbol{x}$ and $\mathrm{dom}(V) \neq 0$ as is the case when it is finite-valued.

For a differentiable function V, ∇V is its (Euclidean) gradient. Define $C^{1+}(\mathbb{R}^p)$ (resp. $C^{11}(\mathbb{R}^p)$) the set of differentiable functions in $\mathbb{R}^p$ whose gradient is locally (resp. globally) Lipschitz continuous. We define also $\widetilde{C^{1+}}(\mathbb{R}^p)$ as

$$\widetilde{C^{1+}}(\mathbb{R}^p) \stackrel{\text{def}}{=} \Big\{ V \in C^{1+}(\mathbb{R}^p) \text{ s.t. } \exists K > 0, \forall \boldsymbol{x} \in \mathrm{dom}(V), \langle \boldsymbol{x}, \nabla V(\boldsymbol{x}) \rangle \leq K(1 + \big\|\boldsymbol{x}\big\|_2^2) \Big\}.$$

The following lemma shows that $C^{11}(\mathbb{R}^p) \subset \widetilde{C^{1+}}(\mathbb{R}^p)$.

Lemma 2.1. *Assume that $V : \Omega \subset \mathbb{R}^p \to \mathbb{R}^p$ is Lipschitz continuous, then there exists $K > 0$ such that*

$$\langle V(\boldsymbol{x}), \boldsymbol{x} \rangle \leq K(1 + \big\|\boldsymbol{x}\big\|_2^2), \ \forall \boldsymbol{x} \in \Omega.$$

Let us consider also some definitions and preperties of variational analysis. A more comprehensive account on variational analysis in finite-dimensional Eudlidean spaces can be found in [51].

Definition 2.1 (Subdifferential)**.** *Given a point $\boldsymbol{x} \in \mathbb{R}^p$ where a function $V : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is finite, the subdifferential of V at $\boldsymbol{x}$ is defined as*

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

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

$$V(\boldsymbol{w}) \geq V(\boldsymbol{x}) + \langle \boldsymbol{v}, \boldsymbol{w} - \boldsymbol{x} \rangle + o(\big\|\boldsymbol{w} - \boldsymbol{x}\big\|_2).$$

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

Let us note that $\partial V(\boldsymbol{x})$ and $\partial^F V(\boldsymbol{x})$ are closed, with $\partial^F V(\boldsymbol{x})$ convex and $\partial^F V(\boldsymbol{x}) \subset \partial V(\boldsymbol{x})$ [51, Theorem 8.6]. A proper lsc convex function is subdifferentially regular.

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

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

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

$\mathrm{prox}_{\gamma V}^{\boldsymbol{M}}$ 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 ${}^{\boldsymbol{M},\gamma}V(\boldsymbol{x}) > -\infty$ for some $\boldsymbol{x}$. The supremum of the set of all such γ is the threshold of prox-boundedness for V.

Definition 2.3 (Hypomonotone and monotone operators)**.** *A set-valued operator $S : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ is hypomonotone of modulus $r > 0$ if*

$$\langle \boldsymbol{x}' - \boldsymbol{x}, \boldsymbol{\eta}' - \boldsymbol{\eta} \rangle \geq -r\|\boldsymbol{x}' - \boldsymbol{x}\|_2^2, \quad \forall \boldsymbol{x}', \boldsymbol{x} \in \mathbb{R}^p \text{ and } \boldsymbol{\eta}' \in S(\boldsymbol{x}'), \boldsymbol{\eta} \in S(\boldsymbol{x}).$$

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

Definition 2.4 (Prox-regularity)**.** *Let $V : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$, given a point $\bar{\boldsymbol{x}} \in \mathrm{dom}(V)$. V is prox-regular at $\bar{\boldsymbol{x}}$ for $\bar{\boldsymbol{v}}$, with $\bar{\boldsymbol{v}} \in \partial V(\bar{\boldsymbol{x}})$ if V is locally lsc at $\bar{\boldsymbol{x}}$, $\exists \epsilon > 0$ and $r > 0$ such that*

$$V(\boldsymbol{x}') > V(\boldsymbol{x}) + (\boldsymbol{x}' - \boldsymbol{x})^T \boldsymbol{v} - \frac{r}{2}\|\boldsymbol{x}' - \boldsymbol{x}\|_2^2,$$

when $\|\boldsymbol{x}' - \bar{\boldsymbol{x}}\|_2 < \epsilon$ and $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2 < \epsilon$ with $\boldsymbol{x}' \neq \boldsymbol{x}$ and $\|V(\boldsymbol{x}) - V(\bar{\boldsymbol{x}})\|_2 < \epsilon$ while $\|\boldsymbol{v} - \bar{\boldsymbol{v}}\|_2 < \epsilon$ with $\boldsymbol{v} \in \partial V(\boldsymbol{x})$.

When this holds for all $\bar{\boldsymbol{v}} \in \partial V(\bar{\boldsymbol{x}})$, V is said prox-regular at $\bar{\boldsymbol{x}}$. When V is prox-regular at every $\boldsymbol{x} \in \mathrm{dom}(V)$, V is said prox-regular.

Roughly speaking, a lsc function V is prox-regular at $\bar{\boldsymbol{x}} \in \mathrm{dom}(V)$ if it has a "local quadratic support" at $\bar{\boldsymbol{x}}$ for any $(\boldsymbol{x}, \boldsymbol{v}) \in \mathrm{gph}(\partial V)$ near enough to $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{v}})$ and with $V(\boldsymbol{x})$ near enough to $V(\bar{\boldsymbol{x}})$.

Let us mention some examples of prox-regular functions.

- Proper lsc convex functions.
- Proper lsc lower-C^2 functions (i.e., $V + \frac{1}{2r}\|.\|_2^2$ is convex, $r > 0$).
- Strongly amenable functions (i.e., $V = g \circ H$, $H : \mathbb{R}^p \to \mathbb{R}^q \in C^2(\mathbb{R}^p)$ and $g : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ proper lsc convex).

Furthermore, a set $\mathcal{C} \subset \mathbb{R}^p$ is prox-regular iff for any $\boldsymbol{x} \in \mathbb{R}^p$ and for any $\gamma > 0$,

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

is single valued and continuous. That equivalents,

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

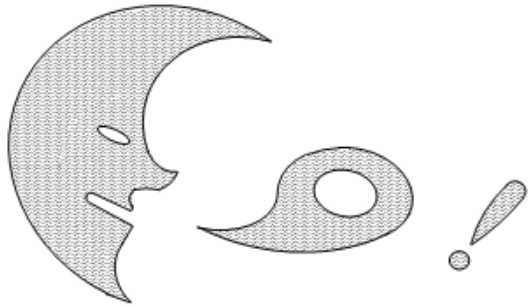

Figure 1: A nonconvex set that is everywhere prox-regular [51, Figure 13.4].

Lemma 2.2 ([47, Theorem 3.2]). *When $V : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is locally lsc at $\bar{\boldsymbol{x}} \in \mathbb{R}^p$, the following are equivalent*

(i) *V is prox-regular at $\bar{\boldsymbol{x}}$ for $\bar{\boldsymbol{v}} \in \partial V(\bar{\boldsymbol{x}})$.*

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

$$V(\boldsymbol{x}) \geq V(\bar{\boldsymbol{x}}) + \langle \bar{\boldsymbol{v}}, \boldsymbol{x} - \bar{\boldsymbol{x}} \rangle - \frac{r}{2}\left\|\boldsymbol{x} - \bar{\boldsymbol{x}}\right\|_2^2, \quad \forall \boldsymbol{x} \in \mathcal{B}(\bar{\boldsymbol{x}}, \epsilon).$$

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

$$\mathrm{T}^V_{\epsilon,\bar{\boldsymbol{x}},\bar{\boldsymbol{v}}}(\boldsymbol{x}) = \begin{cases} \left\{\boldsymbol{v} \in \partial V(\boldsymbol{x}) | \left\|\boldsymbol{v} - \bar{\boldsymbol{v}}\right\|_2 < \epsilon\right\} & \text{if } \left\|\boldsymbol{x} - \bar{\boldsymbol{x}}\right\|_2 < \epsilon \text{ and } \left\|V(\boldsymbol{x}) - V(\bar{\boldsymbol{x}})\right\|_2 < \epsilon, \\ \emptyset \quad \text{otherwise}, \end{cases}$$

such that $\mathrm{T}^V_{\epsilon,\bar{\boldsymbol{x}},\bar{\boldsymbol{v}}} + r\boldsymbol{I}_p$ is montone.

3 Key properties of Moreau-Yosida smoothing

For ease of notation, we denote with the same symbol the measure and its density with respect to the Lebesgue measure. In our framework, the target distribution μ is defined as

$$\mu(\boldsymbol{\theta}) \propto \exp\left(-\Big(F(\boldsymbol{\theta}) + H \circ \widetilde{\boldsymbol{D}}(\boldsymbol{\theta})\Big)\right), \tag{3.1}$$

where $F \in \widetilde{C^{1+}}(\mathbb{R}^M)$, $\widetilde{\boldsymbol{D}} \in \mathbb{R}^{p \times M}$ and $H : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$. Moreover, H is assumed neither differentiable nor convex. To overcome these difficulties, we involve arguments from variational analysis [51]. Namely, we will smooth H by a Moreau envelope and state the following assumptions to exploit the key properties. To avoid trivialities, from now on, we assume that $\mathrm{Argmin}(H) \neq \emptyset$.

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

(H.2) $\mathrm{prox}^{\boldsymbol{M}}_{\gamma H}$ is single valued.

Let us mention some key properties of Moreau-Yosida smoothing in the two following lemmas.

Lemma 3.1. *Let $\boldsymbol{M} \in \mathbb{R}^{p \times p}$ symmetric positive definite, assume that (H.1) hold.*

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

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

(ii) *${}^{\boldsymbol{M},\gamma}H(\theta)$ is finite and depends continuously on $(\boldsymbol{x}, \gamma) \in \mathbb{R}^p \times (0, +\infty)$, and $\left({}^{\boldsymbol{M},\gamma}H(\boldsymbol{x})\right)_{\gamma \in (0,+\infty)}$ is a decreasing net. More precisely,*

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

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

Lemma 3.2. *Let $\boldsymbol{M} \in \mathbb{R}^{p\times p}$ symmetric positive definite, assume that (H.1) and (H.2) hold. Then* $\operatorname{prox}_{\gamma H}^{\boldsymbol{M}}$ *is continuous on $(\boldsymbol{x}, \gamma) \in \mathbb{R}^p \times (0, +\infty)$, and ${}^{\boldsymbol{M},\gamma}H \in C^1(\mathbb{R}^p)$ with gradient*

$$\nabla {}^{\boldsymbol{M},\gamma}H = \gamma^{-1}\boldsymbol{M}\big(\boldsymbol{I}_p - \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}\big).$$

Therefore, for any $\boldsymbol{\theta} \in \mathbb{R}^M$, we have

$$\nabla({}^{\boldsymbol{M},\gamma}H) \circ \widetilde{\boldsymbol{D}}(\boldsymbol{\theta}) = \widetilde{\boldsymbol{D}}^T \nabla {}^{\boldsymbol{M},\gamma}H(\widetilde{\boldsymbol{D}}\boldsymbol{\theta}) = \gamma^{-1}\widetilde{\boldsymbol{D}}^T\boldsymbol{M}\Big(\widetilde{\boldsymbol{D}}\boldsymbol{\theta} - \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\widetilde{\boldsymbol{D}}\boldsymbol{\theta})\Big).$$

In plain words, Lemma 3.2 tells us that the action of the operator $\operatorname{prox}_{\gamma H}^{\boldsymbol{M}}$ is equivalent to a gradient descent on the Moreau envelope of H in the metric $\boldsymbol{M}$ with step-size γ.

Lemma 3.1 (resp. Lemma 3.2) coincides to [27, Lemma 7.1, Lemma 7.2] (resp. [27, Lemma 7.3]) by replacing $\boldsymbol{M}_\gamma$ by $\boldsymbol{M}$.

4 Theoretical guarantees

Let us define the following SDE

$$d\boldsymbol{L}(t) = -\frac{1}{2}\nabla\Big(F + ({}^{\boldsymbol{M},\gamma}H) \circ \widetilde{\boldsymbol{D}}\Big)(\boldsymbol{L}(t))dt + d\boldsymbol{W}(t),\ t > 0. \tag{4.1}$$

For (4.1) to be well-posed, we need also the following assumptions.

(H.3) $\operatorname{prox}_{\gamma H}^{\boldsymbol{M}}$ is locally Lipschitz continuous.

(H.4) For any $\boldsymbol{\theta} \in \mathbb{R}^M$, there exists $K > 0$ such that $\Big\langle \widetilde{\boldsymbol{D}}\boldsymbol{\theta}, \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\widetilde{\boldsymbol{D}}\boldsymbol{\theta})\Big\rangle_{\boldsymbol{M}} \leq K(1 + \|\boldsymbol{\theta}\|_2^2)$.

Denote $D = -\frac{1}{2}\nabla\Big(F + ({}^{\boldsymbol{M},\gamma}H) \circ \widetilde{\boldsymbol{D}}\Big)$, D is called drift coefficient. We introduce the following proposition.

Proposition 4.1. *Assume that (H.1), (H.2), (H.3) and (H.4) hold. Then,*

$$\langle D(\boldsymbol{\theta}), \boldsymbol{\theta}\rangle \leq K(1 + \|\boldsymbol{\theta}\|_2^2),\ \textit{for some } K > 0, \tag{4.2}$$

and

$$D \textit{ is locally Lipschitz continuous.} \tag{4.3}$$

The following proposition guarantees the well-posedness of the SDE (4.1).

Proposition 4.2. *Assume that $D(\boldsymbol{\theta})$ satistifes conditions (4.2) and (4.3). Then, for every initial point $\boldsymbol{L}(0)$ such that $\mathbb{E}\left[\|\boldsymbol{L}(0)\|_2^2\right] < \infty$,*

(i) there exists a unique solution to the SDE (4.1) which is strongly Markovian, and the diffusion is non-explosive, i.e., $\mathbb{E}\left[\|\boldsymbol{L}(t)\|_2^2\right] < \infty$ for all $t > 0$,

(ii) $\boldsymbol{L}$ admits an (unique) invariant measure μ_γ having a density

$$\boldsymbol{\theta} \mapsto \exp\Big(-\Big(F(\boldsymbol{\theta}) + ({}^{\boldsymbol{M},\gamma}H) \circ \widetilde{\boldsymbol{D}}(\boldsymbol{\theta})\Big)\Big)/Z_\gamma$$

where $Z_\gamma = \int_{\mathbb{R}^M} \exp\Big(-\Big(F(\boldsymbol{\theta}') + ({}^{\boldsymbol{M},\gamma}H) \circ \widetilde{\boldsymbol{D}}(\boldsymbol{\theta}')\Big)\Big)d\boldsymbol{\theta}'$.

In Proposition 4.2, Assertion (i) is an application of [67, Theorem 3.6, Chapter II] in our framework. Assertion (ii) is a consequence of [50, Theorem 2.1]. The following proposition answers the natural question on the behaviour of $\mu_\gamma - \mu$ as a function of γ.

Proposition 4.3. *Assume that* (H.1) *hold. Then, μ_γ converges to μ in total variation as $\gamma \to 0$.*

Inserting the identities of Lemma 3.2 into (4.1), we get the SDE

$$d\boldsymbol{L}(t) = -\frac{1}{2}\Big(\nabla F + \gamma^{-1}\widetilde{\boldsymbol{D}}^T \boldsymbol{M}\big(\boldsymbol{I}_p - \mathrm{prox}_{\gamma H}^{\boldsymbol{M}}\big) \circ \widetilde{\boldsymbol{D}}\Big)(\boldsymbol{L}(t))dt + d\boldsymbol{W}(t),\ \boldsymbol{L}(0) = \boldsymbol{l}_0,\ t > 0. \tag{4.4}$$

Consider now the forward Euler discretization of (4.4) with step-size $\delta > 0$, which can be rearranged as

$$\boldsymbol{L}_{k+1} = \boldsymbol{L}_k - \frac{\delta}{2}\nabla F(\boldsymbol{L}_k) - \frac{\delta}{2\gamma}\widetilde{\boldsymbol{D}}^T \boldsymbol{M}\Big(\widetilde{\boldsymbol{D}}\boldsymbol{L}_k - \mathrm{prox}_{\gamma H}^{\boldsymbol{M}}(\widetilde{\boldsymbol{D}}\boldsymbol{L}_k)\Big) + \sqrt{\delta}\boldsymbol{Z}_k,\ t > 0,\ \boldsymbol{L}_0 = \boldsymbol{l}_0. \tag{4.5}$$

Observe that by Lemma 3.2, this is also equivalent to a gradient descent on the Moreau envelope of H in the metric $\boldsymbol{M}$ with step-size δ.

From (4.5), an Euler approximate solution is defined as

$$\boldsymbol{L}^\delta(t) \stackrel{\text{def}}{=} \boldsymbol{L}_0 - \frac{1}{2}\int_0^t \Big(\nabla F(\overline{\boldsymbol{L}}(s)) - \gamma^{-1}\widetilde{\boldsymbol{D}}^T \boldsymbol{M}\Big(\widetilde{\boldsymbol{D}}\overline{\boldsymbol{L}}(s) - \mathrm{prox}_{\gamma H}^{\boldsymbol{M}}(\widetilde{\boldsymbol{D}}\overline{\boldsymbol{L}}(s))\Big)\Big)ds + \int_0^t d\boldsymbol{W}(s),$$

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

Mean square convergence of the pathwise approximation (4.5) and of its first-order moment can be established as follows.

Theorem 4.1. *Assume that $D(\boldsymbol{\theta})$ satistifes conditions* (4.2) *and* (4.3)*, and $\mathbb{E}\left[\left\|\boldsymbol{L}(0)\right\|_2^p\right] < \infty$ for any $p \geq 2$. Then*

$$\left\|\mathbb{E}\left[\boldsymbol{L}^\delta(T)\right] - \mathbb{E}\left[\boldsymbol{L}(T)\right]\right\|_2 \leq \mathbb{E}\left[\sup_{0\leq t\leq T}\left\|\boldsymbol{L}^\delta(t) - \boldsymbol{L}(t)\right\|_2\right] \underset{\delta\to 0}{\longrightarrow} 0. \tag{4.6}$$

Alternative version Assume now that the metric matrix depends also to γ (we denote then $\boldsymbol{M}_\gamma$) with $\boldsymbol{M}_\gamma \underset{\gamma\to 0}{\to} \boldsymbol{I}_p$, and (H.1), (H.2), (H.3) and (H.4) hold. One can consider an alternative version of SDE (4.1), i.e.,

$$d\boldsymbol{L}(t) = -\frac{1}{2}\nabla\Big(F + ({}^{\boldsymbol{M},\gamma}H) \circ \widetilde{\boldsymbol{D}}\Big) \circ {\boldsymbol{M}_\gamma}^{-1/2}(\boldsymbol{L}(t))dt + {\boldsymbol{M}_\gamma}^{1/2}d\boldsymbol{W}(t),\ t > 0. \tag{4.7}$$

Denote the drift coefficient of SDE (4.7) by D_1, we get that

$$\langle D_1(\boldsymbol{\theta}), \boldsymbol{\theta}\rangle = \langle D(\boldsymbol{u}), \boldsymbol{u}\rangle,$$

where $\boldsymbol{u} = {\boldsymbol{M}_\gamma}^{-1/2}\boldsymbol{\theta}$. Then D_1 satisfies also (4.2) and (4.3). From Proposition 4.2, the diffusion $\boldsymbol{L}$ is unique, non explosive and admits an unique invariant measure μ_γ having density

$$\boldsymbol{\theta} \mapsto \exp\left(-\Big(F + ({}^{\boldsymbol{M},\gamma}H) \circ \widetilde{\boldsymbol{D}}\Big) \circ {\boldsymbol{M}_\gamma}^{-1/2}(\boldsymbol{\theta})\right)/Z_\gamma$$

where $Z_\gamma = \sqrt{\det(\boldsymbol{M}_\gamma)}\int_{\mathbb{R}^p}\exp\left(-\Big(F + ({}^{\boldsymbol{M},\gamma}H) \circ \widetilde{\boldsymbol{D}}\Big)(\boldsymbol{u})\right)d\boldsymbol{u}$. Since $\det(\boldsymbol{M}_\gamma) \underset{\gamma\to 0}{\to} 1$, apply the reasoning in the proof of Proposition 4.3, we get also that $\mu_\gamma \to \mu$ as $\gamma \to 0$.

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

$$d\boldsymbol{U}(t) = -\frac{1}{2}{\boldsymbol{M}_\gamma}^{-1}\nabla\Big(F + ({}^{\boldsymbol{M},\gamma}H) \circ \widetilde{\boldsymbol{D}}\Big)(\boldsymbol{U}(t))dt + d\boldsymbol{W}(t),\ t > 0. \tag{4.8}$$

Denote the drift coefficient of SDE (4.8) by D_2, we get that

$$D = \boldsymbol{M}_\gamma D_2.$$

Thus D_2 satisfies the counterpart of the conditions (4.2) and (4.3) in the metric associated to $\boldsymbol{M}_\gamma$. Assume that $\mathbb{E}\left[\left\|L(0)\right\|^p_{\boldsymbol{M}_\gamma}\right] < \infty$ for any $p \geq 2$, the mean square convergence (4.6) can be establish via the metric associated to $\boldsymbol{M}_\gamma$, i.e.,

$$\begin{aligned}\left\|\mathbb{E}\left[\boldsymbol{L}^\delta(T) - \mathbb{E}\left[\boldsymbol{L}(T)\right]\right]\right\|_2 &\leq \frac{1}{\sqrt{\sigma_{\min}(\boldsymbol{M}_\gamma)}}\left\|\mathbb{E}\left[\boldsymbol{L}^\delta(T) - \mathbb{E}\left[\boldsymbol{L}(T)\right]\right]\right\|_{\boldsymbol{M}_\gamma} \\ &\leq \frac{1}{\sqrt{\sigma_{\min}(\boldsymbol{M}_\gamma)}}\mathbb{E}\left[\sup_{0\leq t\leq T}\left\|\boldsymbol{L}^\delta(t) - \boldsymbol{L}(t)\right\|_{\boldsymbol{M}_\gamma}\right] \underset{\delta\to 0}{\longrightarrow} 0.\end{aligned}$$

5 Prox-regularity functions

Let us now discuss about the assumptions imposed in the previous sections. **(H.1)** and **(H.2)** are theoretical assumptions which furnish the properties of Moreau-Yosida smoothing and establish the relation between $\nabla^{\boldsymbol{M},\gamma}H$ and $\operatorname{prox}^{\boldsymbol{M}}_{\gamma H}$ yielding $\nabla^{\boldsymbol{M},\gamma}H$ calculable and exploitable. **(H.3)** and **(H.4)** are the numerical assumptions implying the guarantees of existence, unicity and convergence of LMC diffusions.

This section focus on the prox-regularity which is a classical family of functions in variational analysis. Let us consider a prox-regular function satisfying **(H.1)**. Owing to the following lemma, such type of functions covers also **(H.2)** and **(H.3)**.

Lemma 5.1. *Let $\boldsymbol{M} \in \mathbb{R}^{p\times p}$ symmetric positive definite and γ small enough, assume that H is prox-regular and satisfies* **(H.1)**. *Then for any $\bar{\boldsymbol{x}} \in \mathbb{R}^p$, there exists a neighbourhood $\mathcal{N}_{\bar{\boldsymbol{x}}}$ of $\bar{\boldsymbol{x}}$ on which $\operatorname{prox}^{\boldsymbol{M}}_{\gamma H}$ is single-valued and Lipschitz continuous.*

Corollary 5.1. *Observe that a function H with ∂H r-hypomonotone satisfies the globalized counterpart of Lemma 2.2-(ii). Then such type of functions is prox-regular owing to Lemma 2.2. With H satisfies* **(H.1)**, *the works in [27] have proven the globalized Lipschitz continuous of $\operatorname{prox}^{\boldsymbol{M}}_{\gamma H}$ (see [27, Lemma 7.4]) implying directly* **(H.4)** *according to Lemma 2.1. Futhermore, [27, Theorem 7.1] has specified that the convergence rate in (4.6) is of order $\delta/2$.*

6 Forward-Backward type algorithms

Let us now deal with the main goal: Computing the estimators EWA in (1.5) by sampling $\hat{\mu}$ defined as

$$\hat{\mu}(\boldsymbol{\theta}) \propto \exp\left(-\frac{L(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) + G_\lambda(\boldsymbol{\theta})}{\beta}\right),$$

where $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ is a general loss and $G_\lambda : \mathbb{R}^M \to \mathbb{R}$ is the penalty. In our framework, G_λ has an analysis form, i.e., $G_\lambda = J_\lambda \circ \boldsymbol{D}$ where $\boldsymbol{D} \in \mathbb{R}^{P\times M}$ the analysis operator and $J_\lambda : \mathbb{R}^P \to \mathbb{R}$. In the remaining of the present paper, we set $L_\beta = L(\boldsymbol{X}., \boldsymbol{y})/\beta$ and $J_{\beta,\lambda} = J_\lambda/\beta$.

Assume that $L_\beta \in \widetilde{C^{1+}}(\mathbb{R}^M)$, $J_{\beta,\lambda}$ satisfies **(H.1)**, and $\mathrm{prox}_{\gamma J_{\beta,\lambda}}$ satisfies **(H.2)** and **(H.3)**. Considering $\hat{\mu}$ as the form (3.1), we introduce in the following two algorithms: Forward-Backward LMC Algorithm (with $F \equiv 0$, $H = L_\beta + J_{\beta,\lambda} \circ \boldsymbol{D}$ and $\widetilde{\boldsymbol{D}} = \boldsymbol{I}_M$) and Semi Forward-Backward LMC Algorithm (with $F = L_\beta$, $H = J_{\beta,\lambda}$ and $\widetilde{\boldsymbol{D}} = \boldsymbol{D}$).

6.1 Forward-backward LMC (FBLMC) Algorithm

Consider $H = L_\beta + J_{\beta,\lambda} \circ \boldsymbol{D}$, $F \equiv 0$ and $\widetilde{\boldsymbol{D}} = \boldsymbol{I}_M$ ($p = M$). Since **(H.1)**, **(H.2)** and **(H.3)** are related to $J_{\beta,\lambda}$, a natural idea is to establish the relations between H and $J_{\beta,\lambda}$. In the case of quadratic loss, i.e.,

$$L_\beta(\boldsymbol{\theta}) = \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2}{\beta},$$

this idea is carried out in the following lemma.

Lemma 6.1 ([27, Lemma 7.1]). *Assume that $\gamma \leq \beta/(2\|\boldsymbol{X}\|^2)$ and **(H.1)** holds for H, define $\boldsymbol{M}_\gamma \overset{def}{=} \boldsymbol{I}_M - (2\gamma/\beta)\boldsymbol{X}^T\boldsymbol{X}$, which is symmetric positive definite. Then*

$$\mathrm{prox}_{\gamma H}^{\boldsymbol{M}_\gamma} = \mathrm{prox}_{\gamma J_{\beta,\lambda} \circ \boldsymbol{D}} \circ (\boldsymbol{I}_M - \gamma \nabla L_\beta). \tag{6.1}$$

Assume that $\boldsymbol{D}$ is invertible, operating a simple change of variables $\boldsymbol{u} = \boldsymbol{D}\boldsymbol{\theta}$ and replace L_β with $L_\beta \circ \boldsymbol{D}^{-1}$ and $J_{\beta,\lambda} \circ \boldsymbol{D}$ with $J_{\beta,\lambda}$, Relation (6.1) becomes

$$\mathrm{prox}_{\gamma H}^{\boldsymbol{M}_\gamma} = \mathrm{prox}_{\gamma J_{\beta,\lambda}} \circ (\boldsymbol{I}_M - \gamma \nabla L_\beta). \tag{6.2}$$

From (6.2), one can see that $\mathrm{prox}_{\gamma H}^{\boldsymbol{M}_\gamma}$ satisfies **(H.2)** and **(H.3)**. Again, owing to (6.2), let us impose the following assumption which equivalent to the fact that $\mathrm{prox}_{\gamma H}^{\boldsymbol{M}_\gamma}$ satisfies **(H.4)**.

(H.5) For some $K > 0$, $\left\langle \mathrm{prox}_{\gamma J_{\beta,\lambda}} \circ (\boldsymbol{I}_M - \gamma \nabla L_\beta)(\boldsymbol{D}\boldsymbol{\theta}), \boldsymbol{D}\boldsymbol{\theta} \right\rangle_{\boldsymbol{M}_\gamma} \leq K(1 + \|\boldsymbol{\theta}\|_2^2)$.

From lemmas 3.2 and 6.1, we get

$$\nabla^{\boldsymbol{M}_\gamma,\gamma} H = \gamma^{-1}\boldsymbol{M}_\gamma\Big(\boldsymbol{I}_M - \mathrm{prox}_{\gamma H}^{\boldsymbol{M}_\gamma}\Big) = \gamma^{-1}\boldsymbol{M}_\gamma\Big(\boldsymbol{I}_M - \mathrm{prox}_{\gamma J_{\beta,\lambda}}(\boldsymbol{I}_M - \gamma \nabla L_\beta)\Big).$$

Then, the Euler discretization of SDE (4.1) is defined as

$$\boldsymbol{L}_{k+1} = \boldsymbol{L}_k - \frac{\delta}{2\gamma}\boldsymbol{M}_\gamma\Big(\boldsymbol{L}_k - \mathrm{prox}_{\gamma J_{\beta,\lambda}}(\boldsymbol{L}_k - \gamma \nabla L_\beta(\boldsymbol{L}_k))\Big) + \sqrt{\delta}\boldsymbol{Z}_k,\ t > 0,\ \boldsymbol{L}_0 = \boldsymbol{l}_0, \tag{6.3}$$

Thus, $\hat{\boldsymbol{\theta}}$ is approximated by $\boldsymbol{D}^{-1}(\delta/T \sum_{k=1}^{\lfloor T/\delta \rfloor - 1} \boldsymbol{L}_k)$.

Remark 6.1. *In fact, the "true" Forward-Backward algorithm is established from the SDE (4.7) with the Euler discretization defined as*

$$\boldsymbol{L}_{k+1} = \boldsymbol{L}_k - \frac{\delta}{2\gamma}\Big(\boldsymbol{L}_k - \mathrm{prox}_{\gamma J_{\beta,\lambda}}(\boldsymbol{L}_k - \gamma \nabla L_\beta(\boldsymbol{L}_k))\Big) + \sqrt{\delta}\boldsymbol{Z}_k,\ t > 0,\ \boldsymbol{L}_0 = \boldsymbol{l}_0. \tag{6.4}$$

Observe that, compared to (6.3), there is not the multiplication with matrix $\boldsymbol{M}_\gamma$ that reduces the complexity of algorithm.

6.2 Semi forward-backward LMC (Semi FBLMC) Algorithm

Consider $F = L_\beta$, $H = J_{\beta,\lambda}$, $\widetilde{\boldsymbol{D}} = \boldsymbol{D}$ ($p = P$) and $\boldsymbol{M} = \boldsymbol{I}_P$. Since $J_{\beta,\lambda}$ (resp. $\mathrm{prox}_{\gamma J_{\beta,\lambda}}$) satisfies **(H.1)** (resp. **(H.2)**), from Lemma 3.2 we get that

$$\nabla({}^{\gamma}H) \circ \widetilde{\boldsymbol{D}}(\boldsymbol{\theta}) = \widetilde{\boldsymbol{D}}^T \nabla^{\gamma} H(\widetilde{\boldsymbol{D}}\boldsymbol{\theta}) = \gamma^{-1}\widetilde{\boldsymbol{D}}^T(\widetilde{\boldsymbol{D}}\boldsymbol{\theta} - \mathrm{prox}_{\gamma H}(\widetilde{\boldsymbol{D}}\boldsymbol{\theta})) = \gamma^{-1}\boldsymbol{D}^T(\boldsymbol{D}\boldsymbol{\theta} - \mathrm{prox}_{\gamma J_{\beta,\lambda}}(\boldsymbol{D}\boldsymbol{\theta})).$$

Moreover $\mathrm{prox}_{\gamma J_{\beta,\lambda}}$ satisfies **(H.3)**. By imposing **(H.4)** on $\mathrm{prox}_{\gamma J_{\beta,\lambda}}$, the theoretical properties on SDE (4.1) are validated in the metric $\boldsymbol{M} = \boldsymbol{I}_P$.

We define the Euler discretization of SDE (4.1) as

$$\boldsymbol{L}_{k+1} = \boldsymbol{L}_k - \frac{\delta}{2}\nabla L_\beta(\boldsymbol{L}_k) - \frac{\delta}{2\gamma}\boldsymbol{D}^T\Big(\boldsymbol{D}\boldsymbol{L}_k - \mathrm{prox}_{\gamma J_{\beta,\lambda}}(\boldsymbol{D}\boldsymbol{L}_k)\Big) + \sqrt{\delta}\boldsymbol{Z}_k,\ t > 0,\ \boldsymbol{L}_0 = \boldsymbol{l}_0. \tag{6.5}$$

7 Applications

We exemplify now our algorithms for some popular penalties $G_\lambda = \beta J_{\beta,\lambda} \circ \boldsymbol{D}$ in literature where $J_{\beta,\lambda}$ statisfies Assumption **(H.1)**. The main goal is to calculate $\mathrm{prox}_{\gamma J_{\beta,\lambda}}$ and verify the assumptions **(H.2)** to **(H.5)** on $\mathrm{prox}_{\gamma J_{\beta,\lambda}}$. After that, the computation of our algorithms follows automatically via the Euler schemes (6.4) and (6.5). The computation of $\mathrm{prox}_{\gamma J_{\beta,\lambda}}$ with a general $J_{\beta,\lambda}$ in high dimensional framework is crucial. We focus on a class of penalties where $J_{\beta,\lambda}$ is defined as

$$J_{\beta,\lambda}(\boldsymbol{u}) = \textstyle\sum_{l=1}^{L} w_{\beta,\lambda}\big(\|\boldsymbol{u}_{\mathcal{G}_l}\|_2\big),$$

for some partitions set $(\mathcal{G}_l)_{l\in\{1,\dots,L\}}$ of $\{1,\dots,P\}$, i.e., $\mathcal{G}_1 \oplus \cdots \oplus \mathcal{G}_L = \{1,\dots,P\}$, and $w_{\beta,\lambda} : \mathbb{R}^+ \to \mathbb{R}$. In other words, $J_{\beta,\lambda}$ is separated into the sum of L independent functions, each depends only on the ℓ_2 norm of the group. Let us set the following assumptions for $w_{\beta,\lambda}$.

(H.6) $w_{\beta,\lambda}$ is non decreasing functions on $(0, +\infty)$.

(H.7) $w_{\beta,\lambda}$ is continuously differentiable on $(0, +\infty)$ and the problem $\min_{t\in[0,+\infty)}\{t + \gamma w_{\beta,\lambda}{}'(t)\}$ has a unique solution at 0 for a given γ.

The proximal operator of $J_{\beta,\lambda}$ can be separated by group by the following lemma.

Lemma 7.1. *Assume that* **(H.6)** *hold. For any* $\boldsymbol{u} \in \mathbb{R}^P$ *and* $\gamma > 0$, *we have*

$$\mathrm{prox}_{\gamma J_{\beta,\lambda}}(\boldsymbol{u}) = \begin{pmatrix} \mathrm{prox}_{\gamma w_{\beta,\lambda}}(\|\boldsymbol{u}_{\mathcal{G}_1}\|_2)\frac{\boldsymbol{u}_{\mathcal{G}_1}}{\|\boldsymbol{u}_{\mathcal{G}_1}\|_2} \\ \vdots \\ \mathrm{prox}_{\gamma w_{\beta,\lambda}}\big(\|\boldsymbol{u}_{\mathcal{G}_L}\|_2\big)\frac{\boldsymbol{u}_{\mathcal{G}_L}}{\|\boldsymbol{u}_{\mathcal{G}_L}\|_2} \end{pmatrix}.$$

The computation of $\mathrm{prox}_{\gamma w_{\beta,\lambda}}$ is treated in the following lemma.

Lemma 7.2. *Assume that* **(H.6)** *and* **(H.7)** *hold for some* $\gamma > 0$. *Then,* $\mathrm{prox}_{\gamma w_{\beta,\lambda}}$ *are the single-valued continuous mappings, and satisfy, for* $t \in [0, +\infty)$,

$$\mathrm{prox}_{\gamma w_{\beta,\lambda}}(t) = \begin{cases} 0 & \text{if } t \le \gamma w_{\beta,\lambda}{}'(0^+), \\ t - \gamma w_{\beta,\lambda}{}'\Big(\mathrm{prox}_{\gamma w_{\beta,\lambda}}(t)\Big) & \text{if } t > \gamma w_{\beta,\lambda}{}'(0^+). \end{cases} \tag{7.1}$$

Since $\text{prox}_{\gamma w_{\beta,\lambda}}$ are the single-valued continuous mappings, according to Lemma 7.1, one can see that $\text{prox}_{\gamma J_{\beta,\lambda}}$ satisfies **(H.2)** and **(H.3)**. Let us check **(H.4)** and **(H.5)** in the proof of the following lemma.

Lemma 7.3. *Assume that* **(H.6)** *and* **(H.7)** *hold for some* $\gamma > 0$*, then*

(i) $\text{prox}_{\gamma J_{\beta,\lambda}}$ *satisfies* **(H.4)** *associated with* $\widetilde{\boldsymbol{D}} = \boldsymbol{D}$*,*

(ii) $\text{prox}_{\gamma J_{\beta,\lambda}}$ *satisfies* **(H.5)** *for any* $\boldsymbol{D} \in \mathbb{R}^{M\times M}$*.*

Let us discuss some popular penalties with $w_{\beta,\lambda}$ satisfying **(H.6)** and **(H.7)** for some $\gamma > 0$. These penalties are considered on a very general form taking into account the effects of groups and analysis operators $\boldsymbol{D}$. To retrieve its classical counterpart, it suffices to set $\boldsymbol{D} = \boldsymbol{I}_M$ and the size of groups to 1.

ℓ_1 penalty The ℓ_1 penalty is the most popular penalty in literature which is introduced by the works in [60] and defined as follow

$$G_\lambda(\boldsymbol{\theta}) = \lambda \textstyle\sum_{l=1}^{L} \left\| [\boldsymbol{D}\boldsymbol{\theta}]_{\mathcal{G}_l} \right\|_2.$$

Then $w_{\beta,\lambda}(t) = \lambda t/\beta, \forall t \geq 0$, satisfies **(H.6)** and **(H.7)** for any $\gamma > 0$, and the proximal operator corresponds to a soft thresholding, i.e.,

$$\text{prox}_{\gamma w_{\beta,\lambda}}(t) = (t-\alpha)_+, \quad \forall t \geq 0,$$

where $\alpha = \gamma\lambda/\beta$.

FIRM penalty The FIRM penalty is given by $G_\lambda(\boldsymbol{\theta}) = \sum_{l=1}^{L} \beta w_{\beta,\lambda}\left(\left\| [\boldsymbol{D}\boldsymbol{\theta}]_{\mathcal{G}_l} \right\|_2\right)$ where

$$w_{\beta,\lambda}(t) = \begin{cases} \frac{\lambda}{\beta}\left(t - \frac{t^2}{2\mu}\right) & \text{if } t \leq \mu, \\ \frac{\lambda\mu}{2\beta} & \text{if } t > \mu, \end{cases} \tag{7.2}$$

for any $t \geq 0$. See [66] for a comprehensible review. Since $w_{\beta,\lambda}{}'(t) = \frac{\lambda}{\beta}\left(1 - \frac{t}{\mu}\right)_+$, $w_{\beta,\lambda}$ satisfies **(H.6)** and **(H.7)** for any $\gamma < \beta\mu/\lambda$. The operator $\text{prox}_{\gamma w_{\beta,\lambda}}$ can be constructed from [66, Definition II.3]. Its formula is defined as

$$\text{prox}_{\gamma w_{\beta,\lambda}}(t) = \begin{cases} 0 & \text{if } t \leq \alpha, \\ \frac{\mu}{\mu-\alpha}(t-\alpha) & \text{if } \alpha < t \leq \mu, \\ t & \text{if } t > \mu, \end{cases} \tag{7.3}$$

where $\alpha = \gamma\lambda/\beta$, for any $t \geq 0$. The formula (7.3) can also be found using Lemma 7.2. Observe that the thresholding (7.3) is a generalization of both hard (see [66, Definition II.2]) and soft thresholding. Indeed, (7.3) coincides to a soft (resp. hard) thresholding when $\mu \to \infty$ (resp. $\mu \to \lambda$). That is the interest of FIRM penalty.

SCAD penalty The SCAD penalty is proposed in [30] with $G_\lambda(\boldsymbol{\theta}) = \sum_{l=1}^{L} \beta w_{\beta,\lambda}(\left\| [\boldsymbol{D}\boldsymbol{\theta}]_{\mathcal{G}_l} \right\|_2)$ where

$$w_{\beta,\lambda}(t) = \begin{cases} \frac{\lambda}{\beta}t & \text{if } t \leq \lambda, \\ -\frac{t^2 - 2a\lambda t + \lambda^2}{2\beta(a-1)} & \text{if } \lambda < t \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2\beta} & \text{if } t > a\lambda, \end{cases} \tag{7.4}$$

for any $t \geq 0$, with $a > 2$ and $\lambda > 0$. The following lemma provides the validity of $w_{\beta,\lambda}$ and the formula of its proximal operator.

Lemma 7.4. *Let $w_{\beta,\lambda}$ defined in (7.4). For any $\gamma < (a-1)\beta$,*

(i) $w_{\beta,\lambda}$ satisfies (H.6) and (H.7),

(ii) let $\kappa = \gamma/\beta$, for any $t \geq 0$,

$$\mathrm{prox}_{\gamma w_{\beta,\lambda}}(t) = \begin{cases} (t-\kappa\lambda)_+ & \text{if } t \leq (\kappa+1)\lambda, \\ \frac{(a-1)t-ka\lambda}{a-1-\kappa} & \text{if } (\kappa+1)\lambda < t \leq a\lambda, \\ t & \text{if } t > a\lambda. \end{cases} \tag{7.5}$$

Since $a > 2$, one can set $\kappa = 1$. In this case, (7.5) becomes the classical solution of (1.2) to SCAD penalty which is detailed in [30, Equation (2.8)].

ℓ_∞ penalty The ℓ_∞ penalty has found applications in several fields [36, 40, 58] which is suitable to estimate the signal expected to be flat. It is defined as

$$G_\lambda(\boldsymbol{\theta}) = \lambda \max_{l \in \{1,\dots,L\}} \left\{ \left\| [\boldsymbol{D}\boldsymbol{\theta}]_{\mathcal{G}_l} \right\|_2 \right\}. \tag{7.6}$$

Since $J_{\beta,\lambda}$ can not be separated into the sum of independent functions by group, Lemma 7.1 is now not applicable. However, its proximal operator can be described via the projection in the ℓ_1 unit ball, i.e.,

$$\mathrm{prox}_{\gamma J_{\beta,\lambda}}(\boldsymbol{u}) = \boldsymbol{u} - \underset{\left\{\boldsymbol{x}: \alpha \sum_l \left\| \boldsymbol{x}_{\mathcal{G}_l} \right\|_2 \leq 1\right\}}{\mathrm{proj}} (\boldsymbol{u}), \tag{7.7}$$

where $\alpha = \gamma\lambda/\beta$. One can see that $J_{\beta,\lambda}$ satisfies (H.1), and $\mathrm{prox}_{\gamma J_{\beta,\lambda}}$ satisfies (H.2) and (H.3). We report the verification of (H.4) and (H.5) in the proof of the following lemma.

Lemma 7.5. *Let $J_{\beta,\lambda} = (\lambda/\beta) \max_{l \in \{1,\dots,L\}} \left\{ \left\| \boldsymbol{u}_{\mathcal{G}_l} \right\|_2 \right\}$ for any $\beta > 0$ and $\lambda > 0$, then*

(i) $\mathrm{prox}_{\gamma J_{\beta,\lambda}}$ satisfies (H.4) associated with $\widetilde{\boldsymbol{D}} = \boldsymbol{D}$,

(ii) $\mathrm{prox}_{\gamma J_{\beta,\lambda}}$ satisfies (H.5) for any $\boldsymbol{D} \in \mathbb{R}^{M \times M}$.

8 Numerical experiments

In this section, some numerical experiments are conducted to illustrate and validate our algorithms.

8.1 Problem statement

Let us consider the following linear regression

$$\boldsymbol{y} = f(\boldsymbol{\theta}_0) + \boldsymbol{\epsilon}, \tag{8.1}$$

where $\boldsymbol{\theta}_0$ is a 2-D image which is a matrix in $\mathbb{R}^{128 \times 128}$ (a close up of the image Cameraman.tif, see Figure 2-(a)), $f : \mathbb{R}^{128 \times 128} \to \mathbb{R}^n$, $n \leq 128^2$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$ (the noise level σ is chosen according to the simulated $\boldsymbol{\theta}_0$). Let us denote vec the vectorization operator, then $\mathrm{vec}(\boldsymbol{\theta}_0) \in \mathbb{R}^M$ with $M = 128^2$. From (8.1), we obtain the model (1.1) with $\mathrm{vec}(\boldsymbol{\theta}_0)$ is the unknown regression vector and $\boldsymbol{X} = [\boldsymbol{X}_1, \dots, \boldsymbol{X}_M]$, $\boldsymbol{X}_k = f(\boldsymbol{E}_k)$ where $[\boldsymbol{E}_k]_{i,j} = 1$ when $k = 128(i-1)+j$ and $[\boldsymbol{E}_k]_{i,j} = 0$ otherwise. The goal is estimating $\boldsymbol{\theta}_0$ by computing the EWA estimators via the functions proposed in Section 7. Three numerical problems are considered: Compressed sensing, inpainting and deconvolution whose regression function described in what follows.

Compressed sensing This problem consists in reconstructing the signal via undetermined linear systems, i.e.,

$$f(\boldsymbol{\theta}) = \boldsymbol{X}\,\mathrm{vec}(\boldsymbol{\theta}),$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times M}$. In our experiments, $\boldsymbol{X}$ 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 and $n = 9M/16$.

Inpainting The goal is to recover a masked image (see Figure 2-(b)). Let $\mathcal{M} \subset \{1, \ldots, M\}$ containing the index of masked pixels, the regression function is defined by

$$f(\boldsymbol{\theta}) = [\mathrm{vec}(\boldsymbol{\theta})]_{j,j\in\{1,\ldots,M\}\setminus\mathcal{M}}.$$

In our numerical experiments, we mask 20% of image, thus $n = M - \lfloor 20\%M \rfloor$. The masked positions are chosen randomly from the uniform distribution.

Deconvolution The deconvolution problem (see Figure 2-(c)) is introduced in [41, Section 13.3]. Let $\boldsymbol{p} \in \mathbb{R}^M$ defined as

$$p_i = \exp\left(-\frac{v_i^2}{2\lambda^2}\right),\ i \in \{1, \ldots, M\},$$

where $\boldsymbol{v} = [-M/2, -M/2 + 1, \ldots, M/2]^T \in \mathbb{R}^M$ and $\lambda > 0$. In our experiments, we set $\lambda = 1$.

The regression function corresponds to the convolution operator, i.e.,

$$[f(\boldsymbol{\theta})]_i = \textstyle\sum_{j=1}^{M} [\mathrm{vec}(\boldsymbol{\theta})]_j u(j - i),\ i \in \{1, \ldots, M\},$$

where $u(-k) = u(k) = p_{k+1}$, for any $k \in \{0, \ldots, M - 1\}$. Observe that $n = M$.

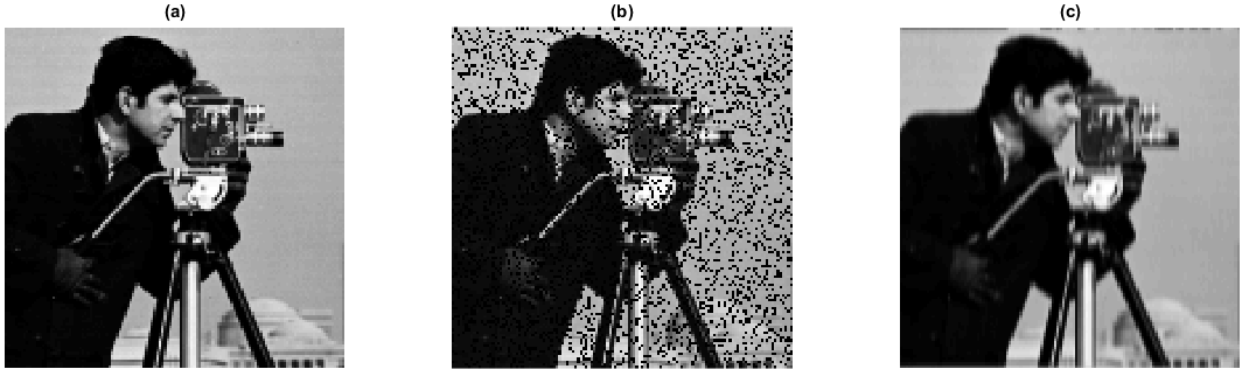

Figure 2: (a): Original image. (b): Masked image (20 percent). (c): Convolved image.

8.2 Analysis operator

Since (8.1) is ill-posed, we need an analysis operator $D : \mathbb{R}^M \to \mathbb{R}^P$ and a partitions set of $\{1, \ldots, P\}$ which gives the sparsity for $\boldsymbol{\theta}_0$. Indeed, since the targeted image is piecewise constant, a popular prior is so called isotropic total variation [53]. Let us now clarify that.

Let $D_c : \mathbb{R}^{\sqrt{M}\times\sqrt{M}} \to \mathbb{R}^{\sqrt{M}\times\sqrt{M}}$ and $D_r : \mathbb{R}^{\sqrt{M}\times\sqrt{M}} \to \mathbb{R}^{\sqrt{M}\times\sqrt{M}}$ the finite difference operators along, respectively, the columns and rows, with appropriate boundary conditions. We define D_1 as

$$D_1 : \boldsymbol{\theta} \in \mathbb{R}^{\sqrt{M}\times\sqrt{M}} \mapsto (D_r(\boldsymbol{\theta}), D_c(\boldsymbol{\theta})) \in \mathbb{R}^{\sqrt{M}\times\sqrt{M}} \times \mathbb{R}^{\sqrt{M}\times\sqrt{M}}.$$

Moreover, the image $\boldsymbol{\theta}_0$ is assumed to be formed by the blocks which overlap. We define the analysis operator as $D = D_2 \circ D_1$ where D_2 separates the overlapping blocks. Namely,

$$D_2 : (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathbb{R}^{\sqrt{M}\times\sqrt{M}} \times \mathbb{R}^{\sqrt{M}\times\sqrt{M}} \to (B(\boldsymbol{\theta}_1), B(\boldsymbol{\theta}_2)) \in \mathbb{R}^L \times \mathbb{R}^L,$$

where B is defined as

$$B(\boldsymbol{\theta}) = (\boldsymbol{B}_i \operatorname{vec}(\boldsymbol{\theta}))_{i\in I} \in \mathbb{R}^L = \prod_{i\in I} \mathbb{R}^{L_i},$$

with $\boldsymbol{B}_i : \mathbb{R}^M \to \mathbb{R}^{L_i}$ are the localization operators whose index belong to the set I which is countable, and $L = \sum_{i\in I} L_i$.

By defining the set of groups by

$$\mathcal{G} = \bigcup_{i\in\{1,\dots,L\}} \{(i,1),(i,2)\},$$

one immediately realizes that measuring sparsity of the above vectorized form is equivalent to group sparsity of $D(\boldsymbol{\theta}_0)$ with groups of size 2 along the third dimension.

8.3 Numerical results

EWA with ℓ_1, FIRM and SCAD penalties Since the analysis operator $\boldsymbol{D}$ is not invertible, we use Semi FBLMC algorithm. Numerical results for ℓ_1 and SCAD penalties are respectively shown in figures 3 and 5.

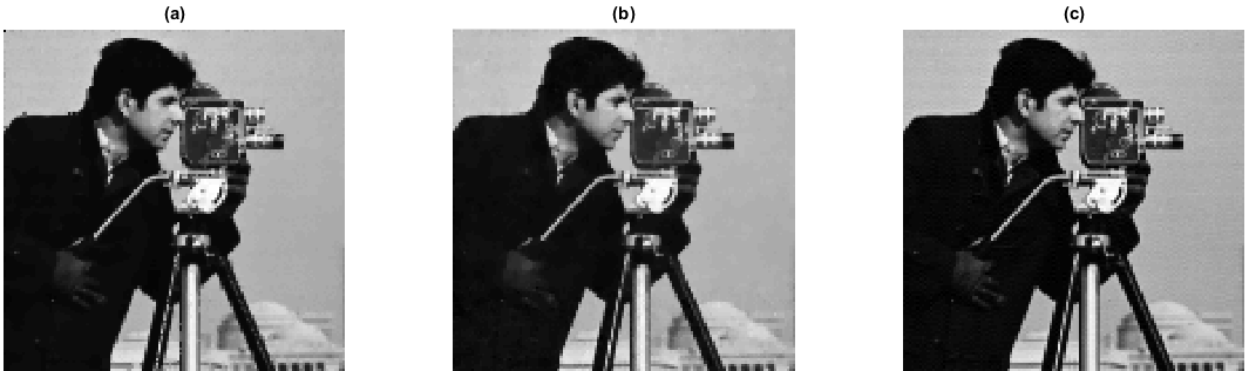

Figure 3: (a): Inpainting with EWA - ℓ_1 penalty. (b): Compressed Sensing with EWA - ℓ_1 penalty. (c) Deconvolution with EWA - ℓ_1 penalty.

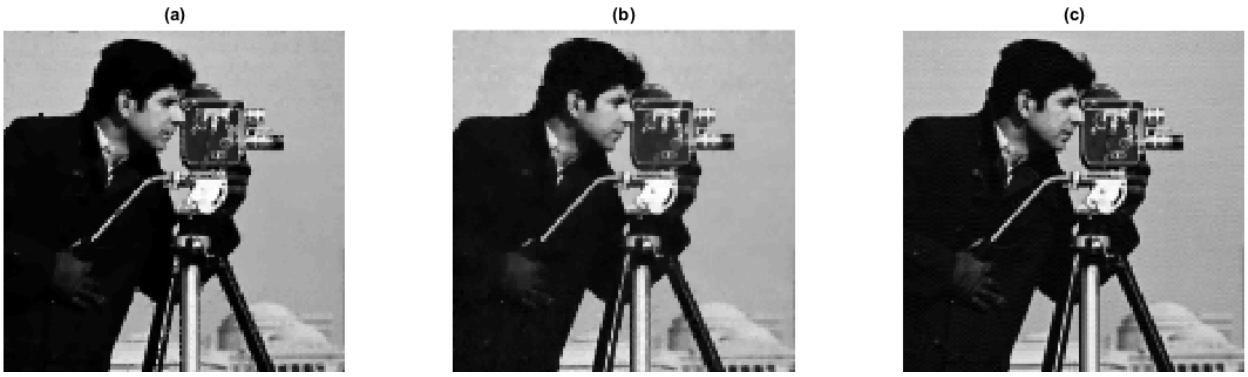

Figure 4: (a): Inpainting with EWA - FIRM penalty. (b): Compressed Sensing with EWA - FIRM penalty. (c) Deconvolution with EWA - FIRM penalty.

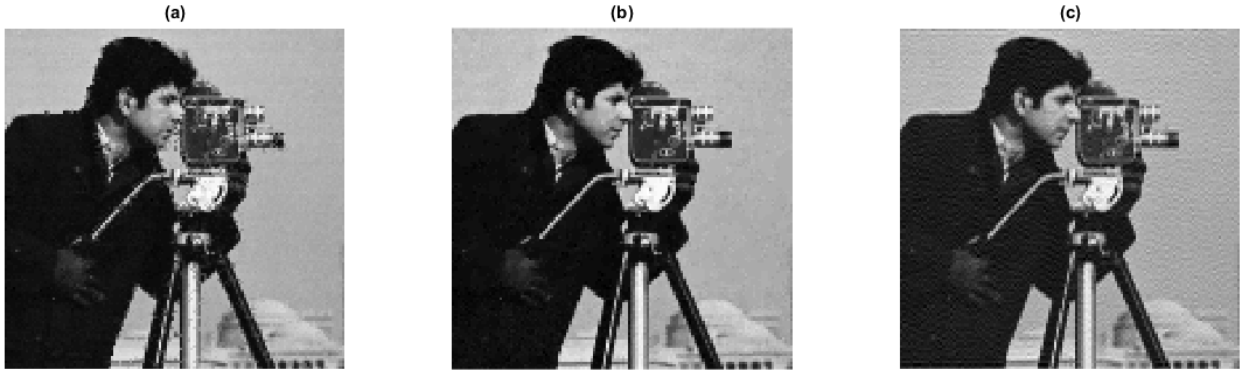

Figure 5: (a): Inpainting with EWA - SCAD penalty. (b): Compressed Sensing with EWA - SCAD penalty. (c) Deconvolution with EWA - SCAD penalty.

EWA with ℓ_∞ penalty ℓ_∞ penalty is a anti-sparsity penalty which is suitable to estimate a flat signal. Thus, let us create a signal whose coordinates are valued in $\{-1, 1\}$. However, we set $\boldsymbol{D} = \boldsymbol{I}_M$ and do not consider the effect of groups, i.e., the size of groups is 1. In this case, we can use FBLMC algorithm. The EWA estimator is implemented in a non-overfitting Compressed Sensing problem, i.e., $n > M$. This condition is classic in literature which guarantees the performance of an estimator with ℓ_∞ penalty. Numerical results are shown in Figure 6.

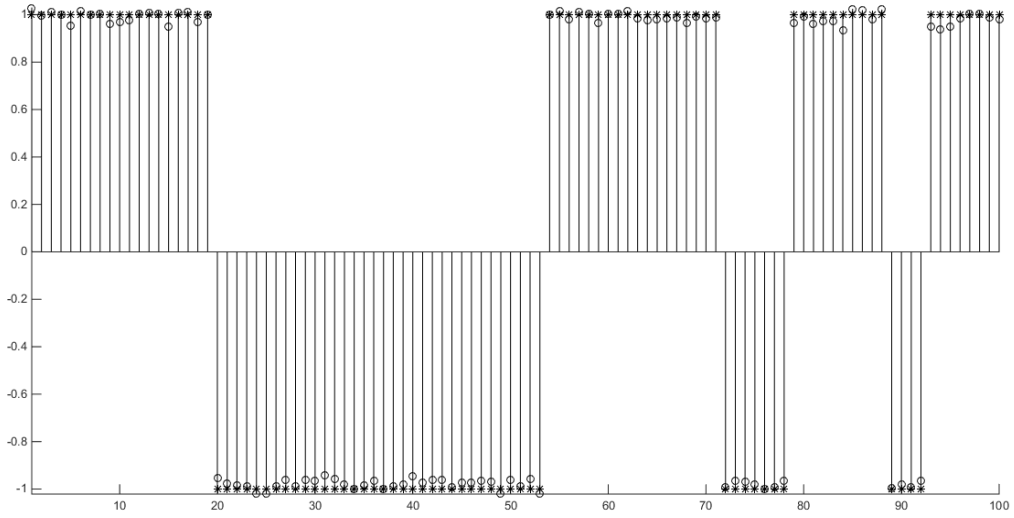

Figure 6: Compressed Sensing with EWA - ℓ_∞ penalty.

9 Proofs

9.1 Proof of Lemma 2.1

Let $\boldsymbol{x}^* \in \Omega$, by Young's inequality and the Lipschitz continuous of V, we obtain

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

for some $\widetilde{K} > 0$, with $K \geq \max\left\{2\widetilde{K} + 1/2, 2\widetilde{K}\|\boldsymbol{x}^*\|_2^2 + \|V(\boldsymbol{x}^*)\|_2^2\right\}$. That concludes the proof of Lemma 2.1. □

9.2 Proof of Proposition 4.1

In view of Lemma 3.2, the drift coefficient becomes

$$D(\boldsymbol{\theta}) = -\frac{1}{2}\nabla(F + ({}^{\boldsymbol{M},\gamma}H) \circ \widetilde{\boldsymbol{D}})(\boldsymbol{\theta}) = -\frac{1}{2}\nabla F(\boldsymbol{\theta}) - \frac{1}{2\gamma}\widetilde{\boldsymbol{D}}^T \boldsymbol{M}\widetilde{\boldsymbol{D}}\boldsymbol{\theta} + \frac{1}{2\gamma}\widetilde{\boldsymbol{D}}^T \boldsymbol{M}\mathrm{prox}_{\gamma H}^{\boldsymbol{M}}(\widetilde{\boldsymbol{D}}\boldsymbol{\theta}).$$

Since $F \in \widetilde{C^{1+}}$ and **(H.4)** hold, there exist $K_1 > 0$ and $K_2 > 0$ such that

$$\begin{aligned}\langle D(\boldsymbol{\theta}), \boldsymbol{\theta}\rangle &= -\frac{1}{2}\langle \nabla F(\boldsymbol{\theta}), \boldsymbol{\theta}\rangle - \frac{1}{2\gamma}\big\|\widetilde{\boldsymbol{D}}\boldsymbol{\theta}\big\|_{\boldsymbol{M}}^2 + \frac{1}{2}\Big\langle \mathrm{prox}_{\gamma H}^{\boldsymbol{M}}(\widetilde{\boldsymbol{D}}\boldsymbol{\theta}), \widetilde{\boldsymbol{D}}\boldsymbol{\theta}\Big\rangle_{\boldsymbol{M}} \\ &\leq K_1(1 + \big\|\boldsymbol{\theta}\big\|_2^2) + \Big|\Big|\Big|\widetilde{\boldsymbol{D}}\Big|\Big|\Big|^2 |||\boldsymbol{M}|||/(2\gamma)\big\|\boldsymbol{\theta}\big\|_2^2 + K_2(1 + \big\|\boldsymbol{\theta}\big\|_2^2) \\ &\leq K(1 + \big\|\boldsymbol{\theta}\big\|_2^2).\end{aligned}$$

where $K \geq K_1 + K_2 + \Big|\Big|\Big|\widetilde{\boldsymbol{D}}\Big|\Big|\Big|^2 |||\boldsymbol{M}|||/(2\gamma)$. Moreover, the local Lipschitz continuous of D follows from **(H.3)** and the local Lipschitz continuous of ∇F. This ends the proof of Proposition 4.1. □

9.3 Proof of Proposition 4.3

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

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

where

$$\mu_\gamma(\boldsymbol{\theta}) = \exp\Big(-\Big(F(\boldsymbol{\theta}) + ({}^{\boldsymbol{M},\gamma}H) \circ \widetilde{\boldsymbol{D}}(\boldsymbol{\theta})\Big)\Big)/Z_\gamma \text{ and } \mu(\boldsymbol{\theta}) = \exp\Big(-\Big(F(\boldsymbol{\theta}) + H \circ \widetilde{\boldsymbol{D}}(\boldsymbol{\theta})\Big)\Big)/Z,$$

and $Z = \int_{\mathbb{R}^M} \exp\Big(-(F(\boldsymbol{\theta}') + H \circ \widetilde{\boldsymbol{D}}(\boldsymbol{\theta}'))\Big) d\boldsymbol{\theta}'$. In view of Lemma 3.1(ii), applying the monotone convergence theorem, we conclude that $Z_\gamma \to Z$. This together with Lemma 3.1(ii) yield that μ_γ converges to μ pointwise. We conclude using Scheffé(-Riesz) theorem [37, 55]. □

9.4 Proof of Theorem 4.1

From (4.2), owing to [67, Theorem 4.1, Chapter II], we get that the p-th moments of $\boldsymbol{L}(t)$ are finite (i.e., $\mathbb{E}\left[\big\|\boldsymbol{L}(t)\big\|_2^p\right] < \infty$) for any $p \geq 2$ and $t \geq 0$. The same property holds for $\boldsymbol{L}^\delta$ because $\boldsymbol{L}^\delta$ is the continuous-time extension to the discrete time chain of $\boldsymbol{L}$. According to the local Lipschitz continuous of D, we conclude the proof of Theorem 4.1 using [35, Theorem 2.2] and Jensen's inequality. □

9.5 Proof of Lemma 5.1

The proof of Lemma 5.1 is based on the one of [51, Proposition 13.37] and generalizes to the proximal mapping in metric $\boldsymbol{M}$ for any $\boldsymbol{M} \in \mathbb{R}^{p\times p}$ symmetric positive definite.

Without loss of generality, we perform the proof on a neighbourhood of $\bar{\boldsymbol{x}}$ where H is lsc. Let $\bar{\boldsymbol{x}} \in \mathbb{R}^p$, $\bar{\boldsymbol{v}} \in \partial H(\bar{\boldsymbol{x}})$, since H is prox-regular at $\bar{\boldsymbol{x}}$ for $\bar{\boldsymbol{v}}$ and H is prox-bounded, owing to [5, Lemma 4.1], there exist $\epsilon > 0$ and $\lambda_0 > 0$ such that

$$\begin{aligned}H(\boldsymbol{x}') &> H(\boldsymbol{x}) + \big\langle \boldsymbol{v}, \boldsymbol{x}' - \boldsymbol{x}\big\rangle - \frac{1}{2\lambda_0}\big\|\boldsymbol{x}' - \boldsymbol{x}\big\|_2^2 \\ &> H(\boldsymbol{x}) + \big\langle \boldsymbol{v}, \boldsymbol{x}' - \boldsymbol{x}\big\rangle - \frac{1}{2\lambda_0\sigma_{\min}(\boldsymbol{M})}\big\|\boldsymbol{x}' - \boldsymbol{x}\big\|_{\boldsymbol{M}}^2, \end{aligned} \tag{9.1}$$

for any $\boldsymbol{x}' \neq \boldsymbol{x}$ and $(\boldsymbol{x}, \boldsymbol{v}) \in \mathrm{gph}\, \mathrm{T}^H_{\epsilon,\bar{\boldsymbol{x}},\bar{\boldsymbol{v}}}$. Let $\gamma_0 = \lambda_0\sigma_{\min}(\boldsymbol{M})$, $\gamma \in (0, \gamma_0)$ and $\boldsymbol{u} = \boldsymbol{x} + \gamma\boldsymbol{M}^{-1}\boldsymbol{v}$, (9.1) becomes

$$H(\boldsymbol{x}') + \frac{1}{2\gamma}\big\|\boldsymbol{x}' - \boldsymbol{u}\big\|_{\boldsymbol{M}}^2 > H(\boldsymbol{x}) + \frac{1}{2\gamma}\big\|\boldsymbol{x} - \boldsymbol{u}\big\|_{\boldsymbol{M}}^2.$$

Therefore, $\operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{u}) = \boldsymbol{x}$ where $(\boldsymbol{x}, \boldsymbol{v}) \in \operatorname{gph} \mathrm{T}_{\epsilon,\bar{\boldsymbol{x}},\bar{\boldsymbol{v}}}^{H}$. That yields $\operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\bar{\boldsymbol{x}} + \gamma \boldsymbol{M}^{-1}\bar{\boldsymbol{v}}) = \bar{\boldsymbol{x}}$.

Since H is lsc, proper and prox-bounded, from [51, Theorem 1.17(c)] (see also [51, Theorem 1.25]), we have

$$\boldsymbol{x} \in \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{u}), \boldsymbol{u} \to \bar{\boldsymbol{x}} + \gamma \boldsymbol{M}^{-1}\bar{\boldsymbol{v}} \implies \begin{cases} \boldsymbol{x} \to \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\bar{\boldsymbol{x}} + \gamma \boldsymbol{M}^{-1}\bar{\boldsymbol{v}}) = \bar{\boldsymbol{x}}, \\ H(\boldsymbol{x}) = {}^{\boldsymbol{M},\gamma}H(\boldsymbol{u}) - \frac{1}{2\gamma}\left\|\boldsymbol{x} - \boldsymbol{u}\right\|_2^2 \to H(\bar{\boldsymbol{x}}). \end{cases} \tag{9.2}$$

For any $\boldsymbol{x} \in \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{u})$, by Fermat rules we get

$$\boldsymbol{v} = \frac{\boldsymbol{M}}{\gamma}(\boldsymbol{u} - \boldsymbol{x}) \in \partial H(\boldsymbol{x}). \tag{9.3}$$

For any $\gamma \in (0, \gamma_0)$, owing to (9.2) and (9.3), there exists $\mathcal{N}_{\gamma,\bar{\boldsymbol{x}},\bar{\boldsymbol{v}}}$ a neighbourhood of $\bar{\boldsymbol{x}} + \gamma \boldsymbol{M}^{-1}\bar{\boldsymbol{v}}$ such that for any $\boldsymbol{u} \in \mathcal{N}_{\gamma,\bar{\boldsymbol{x}},\bar{\boldsymbol{v}}}$, $\left\|\boldsymbol{x} - \bar{\boldsymbol{x}}\right\|_2 \leq \epsilon$, $\left\|H(\boldsymbol{x}) - H(\bar{\boldsymbol{x}})\right\|_2 \leq \epsilon$ and $\left\|\boldsymbol{v} - \bar{\boldsymbol{v}}\right\|_2 \leq \epsilon$. We get then

$$\operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{u}) = \boldsymbol{x} \implies \boldsymbol{v} = \frac{\boldsymbol{M}}{\gamma}(\boldsymbol{u} - \boldsymbol{x}) \in \mathrm{T}_{\epsilon,\bar{\boldsymbol{x}},\bar{\boldsymbol{v}}}^{H}(\boldsymbol{x}).$$

So that

$$\operatorname{prox}_{\gamma H}^{\boldsymbol{M}} = (\boldsymbol{M} + \gamma \mathrm{T}_{\epsilon,\bar{\boldsymbol{x}},\bar{\boldsymbol{v}}}^{H})^{-1} \circ \boldsymbol{M} = (\boldsymbol{M} + \delta^{-1}S)^{-1} \circ (\gamma\delta)^{-1}\boldsymbol{M},$$

where $\delta = 1/\gamma - 1/\gamma_0$, $S = \mathrm{T}_{\epsilon,\bar{\boldsymbol{x}},\bar{\boldsymbol{v}}}^{H} + 1/\gamma_0 \boldsymbol{M}$. From (9.1), S is maximal monotone, the latter operator is well defined as a single valued operator (see [4, Proposition 3.22 (ii)(d)]). Let $\boldsymbol{p} = \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{x})$ and $\boldsymbol{p}' = \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{x}')$. It then follows that

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

and monotonicity of S yields

$$\left\langle \boldsymbol{p}' - \boldsymbol{p}, \boldsymbol{M}(\boldsymbol{x}' - \boldsymbol{x}) \right\rangle \geq \gamma\delta\left\|\boldsymbol{p}' - \boldsymbol{p}\right\|_{\boldsymbol{M}}^2 \geq \gamma\delta\sigma_{\min}(\boldsymbol{M})\left\|\boldsymbol{p}' - \boldsymbol{p}\right\|_2^2.$$

Using Cauchy-Schwarz's inequality, we obtain

$$\left\|\boldsymbol{p}' - \boldsymbol{p}\right\|_2 \leq K\left\|\boldsymbol{x}' - \boldsymbol{x}\right\|_2,$$

where $K^{-1} = \gamma\delta\sigma_{\min}(\boldsymbol{M})/|||\boldsymbol{M}||| = \big(1 - \gamma/\gamma_0\big)\sigma_{\min}(\boldsymbol{M})/|||\boldsymbol{M}|||$.

Let us note that when γ decrease, Inequality (9.1) can be hold for a larger ϵ that enlarges $\mathcal{N}_{\gamma,\bar{\boldsymbol{x}},\bar{\boldsymbol{v}}}$ and $\bar{\boldsymbol{x}} + \gamma \boldsymbol{M}^{-1}\bar{\boldsymbol{v}}$ concentrate to $\bar{\boldsymbol{x}}$ for any $\bar{\boldsymbol{v}}$. Thus, when γ is small enough, there exists a neighbourhood $\bar{\boldsymbol{x}}$ that includes in $\mathcal{N}_{\gamma,\bar{\boldsymbol{x}},\bar{\boldsymbol{v}}}$ for any $\bar{\boldsymbol{v}} \in \partial H(\bar{\boldsymbol{x}})$. That concludes the proof of Lemma 5.1. □

9.6 Proof of Lemma 7.3

Before proceeding, let us discuss about the term $\operatorname{prox}_{\gamma w_{\beta,\lambda}}$. In view of **(H.6)**, ${w_{\beta,\lambda}}'$ is positive of $(0, +\infty)$. According to Lemma 7.2 we get that, for any $t \geq 0$, $\operatorname{prox}_{\gamma w_{\beta,\lambda}}(t) = 0$ if $t \leq \gamma {w_{\beta,\lambda}}'(0)$ and $\operatorname{prox}_{\gamma w_{\beta,\lambda}}(t) = t - \gamma {w_{\beta,\lambda}}'(\operatorname{prox}_{\gamma w_{\beta,\lambda}}(t)) \leq t$ otherwise. That yields the following bound on $\operatorname{prox}_{\gamma w_{\beta,\lambda}}$ which is useful for next steps of the proof

$$0 \leq \operatorname{prox}_{\gamma w_{\beta,\lambda}}(t) \leq t, \quad \forall t > 0. \tag{9.4}$$

(i) Set $\boldsymbol{u} = \boldsymbol{D\theta}$, from Lemma 7.1 and (9.4), we get that

$$\left\langle \mathrm{prox}_{\gamma J_{\beta,\lambda}}(\boldsymbol{u}), \boldsymbol{u} \right\rangle = \sum_{l=1}^{L} \left\langle [\mathrm{prox}_{\gamma J_{\beta,\lambda}}(\boldsymbol{u})]_{\mathcal{G}_l}, \boldsymbol{u}_{\mathcal{G}_l} \right\rangle = \sum_{l=1}^{L} \frac{\mathrm{prox}_{\gamma w_{\beta,\lambda}}(\|\boldsymbol{u}_{\mathcal{G}_l}\|_2)}{\|\boldsymbol{u}_{\mathcal{G}_l}\|_2} \|\boldsymbol{u}_{\mathcal{G}_l}\|_2^2 \leq \|\boldsymbol{u}\|_2^2.$$

(ii) Set $\boldsymbol{u} = \boldsymbol{D\theta}$, $\boldsymbol{v} = 2\gamma \boldsymbol{X}^T \boldsymbol{y}/\beta$ and $\boldsymbol{t_u} = (\boldsymbol{I}_M - \gamma \nabla L_\beta) \circ (\boldsymbol{D\theta}) = \boldsymbol{M}_\gamma \boldsymbol{u} + \boldsymbol{v}$, by Young's inequality, we obtain that

$$\left\langle \mathrm{prox}_{\gamma J_{\beta,\lambda}}(\boldsymbol{t_u}), \boldsymbol{u} \right\rangle_{\boldsymbol{M}_\gamma} = \left\langle \boldsymbol{M}_\gamma \mathrm{prox}_{\gamma J_{\beta,\lambda}}(\boldsymbol{t_u}), \boldsymbol{u} \right\rangle \leq \frac{1}{2} |||\boldsymbol{M}_\gamma|||^2 \|\mathrm{prox}_{\gamma J_{\beta,\lambda}}(\boldsymbol{t_u})\|_2^2 + \frac{1}{2}\|\boldsymbol{u}\|_2^2.$$

Moreover, owing to Lemma 7.1 and (9.4), we get that

$$\begin{aligned}
\|\mathrm{prox}_{\gamma J_{\beta,\lambda}}(\boldsymbol{t_u})\|_2^2 = \| \sum_{l=1}^{L} \frac{\mathrm{prox}_{\gamma J_{\beta,\lambda}}(\|[\boldsymbol{t_u}]_{\mathcal{G}_l}\|_2)}{\|[\boldsymbol{t_u}]_{\mathcal{G}_l}\|_2} [\boldsymbol{t_u}]_{\mathcal{G}_l} \|_2^2 &\leq \left(\sum_{l=1}^{L} |\mathrm{prox}_{\gamma J_{\beta,\lambda}}(\|[\boldsymbol{t_u}]_{\mathcal{G}_l}\|_2)| \right)^2 \\
&\leq \left(\sum_{l=1}^{L} \|[\boldsymbol{t_u}]_{\mathcal{G}_l}\|_2 \right)^2 \\
&\leq L \|\boldsymbol{t_u}\|_2^2 \\
&\leq 2L \Big(|||\boldsymbol{M}_\gamma|||^2 \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2 \Big).
\end{aligned}$$

According to the fact that $\|\boldsymbol{u}\|_2^2 = \|\boldsymbol{D\theta}\|_2^2 \leq |||\boldsymbol{D}|||^2 \|\boldsymbol{\theta}\|_2^2$, we conclude the proof of the assertions (i), (ii) and also Lemma 7.3. □

9.7 Proof of Lemma 7.4

(i) Observe that $w_{\beta,\lambda}$ is continuously differentiable with

$$\gamma w_{\beta,\lambda}{}'(t) = \kappa\lambda \Big(I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \Big).$$

Since $w_{\beta,\lambda}{}'$ is positive on $(0, +\infty)$, $w_{\beta,\lambda}$ satisfies **(H.6)**. Let $u(t) = t + \gamma w_{\beta,\lambda}{}'(t)$, we obtain that

- $u(0) = \kappa\lambda$,
- if $0 < t \leq \lambda$, $u(t) = t + \kappa\lambda > \kappa\lambda$,
- if $\lambda < t \leq a\lambda$, since $a > 2$, $u(t) = t + \frac{\kappa(a\lambda - t)}{a-1} = \frac{a-1+\kappa}{a-1} t + \kappa a\lambda > t + \kappa a\lambda > \lambda + \kappa a\lambda > \kappa\lambda$,
- if $t > a\lambda$, since $a - 1 > \kappa$, $u(t) = t > a\lambda > \kappa\lambda$.

Thus, $t = 0$ is the unique mimimum in $[0, +\infty)$ of $t + p'_\lambda(t)$. According to the continuous differentiability of $w_{\beta,\lambda}$, $w_{\beta,\lambda}$ satisfies **(H.7)**.

(ii) For the sake of simplified notation, we denote $p = \mathrm{prox}_{\gamma w_{\beta,\lambda}}(t)$. Owing to Lemma 7.2, we obtain that

$$p = \begin{cases} 0 & \text{if } t \leq \kappa\lambda, \\ t - \kappa\lambda \Big(I(p \leq \lambda) + \frac{(a\lambda - p)_+}{(a-1)\lambda} I(p > \lambda) \Big) & \text{otherwise.} \end{cases} \tag{9.5}$$

From (9.5), we get the following assertions when $t > \kappa\lambda$,

- if $p \leq \lambda$, $p = t - \kappa\lambda$ and $t = p + \kappa\lambda \leq (\kappa+1)\lambda$,
- if $\lambda < p \leq a\lambda$, $p = t - \kappa(a\lambda - p)/(a-1)$ implies that $p = (((a-1)t - \kappa a\lambda)/(a-1-\kappa))$. Since $\lambda < p \leq a\lambda$, $\kappa < a-1$ and $a > 2$, we get also that

$$(1+\kappa)\lambda < t = \frac{a-1-\kappa}{a-1}p + \frac{\kappa a\lambda}{a-1} \leq a\lambda,$$

- if $p > a\lambda$, $p = t$ implies that $t > a\lambda$.

That concludes the proof of (ii), Lemma 7.4. □

9.8 Proof of Lemma 7.5

(i) Set $\boldsymbol{u} = \boldsymbol{D\theta}$ and $\boldsymbol{p_u} = \mathrm{proj}_{\left\{\boldsymbol{x}:\alpha\sum_l \left\|\boldsymbol{x}_{\mathcal{G}_l}\right\|_2 \leq 1\right\}}(\boldsymbol{u})$. Owing to (7.7) and Young's inequality, we obtain that

$$\left\langle \boldsymbol{u}, \mathrm{prox}_{\gamma J_{\beta,\lambda}}(\boldsymbol{u}) \right\rangle = \langle \boldsymbol{u}, \boldsymbol{u} - \boldsymbol{p_u} \rangle \leq \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{u}\|_2\|\boldsymbol{p_u}\|_2 \leq \frac{3}{2}\|\boldsymbol{u}\|_2^2 + \frac{1}{2}\|\boldsymbol{p_u}\|_2^2 \leq \frac{3}{2}\|\boldsymbol{u}\|_2^2 + \frac{1}{2\alpha^2}.$$

(ii) Set $\boldsymbol{u} = \boldsymbol{D\theta}$, $\boldsymbol{v} = 2\gamma\boldsymbol{X}^T\boldsymbol{y}/\beta$, $\boldsymbol{t_u} = (\boldsymbol{I}_M - \gamma\nabla L_\beta) \circ (\boldsymbol{D\theta}) = \boldsymbol{M}_\gamma\boldsymbol{u} + \boldsymbol{v}$ and $\boldsymbol{p_{t_u}} = \mathrm{proj}_{\left\{\boldsymbol{x}:\alpha\sum_l \left\|\boldsymbol{x}_{\mathcal{G}_l}\right\|_2 \leq 1\right\}}(\boldsymbol{t_u})$. By Young's inequality, we obtain that

$$\left\langle \mathrm{prox}_{\gamma J_{\beta,\lambda}}(\boldsymbol{t_u}), \boldsymbol{u} \right\rangle_{\boldsymbol{M}_\gamma} = \left\langle \boldsymbol{M}_\gamma \mathrm{prox}_{\gamma J_{\beta,\lambda}}(\boldsymbol{t_u}), \boldsymbol{u} \right\rangle \leq \frac{1}{2}|||\boldsymbol{M}_\gamma|||^2 \|\mathrm{prox}_{\gamma J_{\beta,\lambda}}(\boldsymbol{t_u})\|_2^2 + \frac{1}{2}\|\boldsymbol{u}\|_2^2.$$

Moreover, owing to (7.7), we get that

$$\|\mathrm{prox}_{\gamma J_{\beta,\lambda}}(\boldsymbol{t_u})\|_2^2 = \|\boldsymbol{t_u} - \boldsymbol{p_{t_u}}\|_2^2 \leq 2\|\boldsymbol{t_u}\|_2^2 + 2\|\boldsymbol{p_{t_u}}\|_2^2 \leq 4|||\boldsymbol{M}_\gamma|||^2\|\boldsymbol{u}\|_2^2 + \left(4\|\boldsymbol{v}\|_2^2 + \frac{2}{\alpha^2}\right).$$

According to the fact that $\|\boldsymbol{u}\|_2^2 = \|\boldsymbol{D\theta}\|_2^2 \leq |||\boldsymbol{D}|||^2\|\boldsymbol{\theta}\|_2^2$, we conclude the proof of the assertions (i), (ii) and also Lemma 7.5. □

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

References

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

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

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

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

[5] F. Bernard and L. Thibault. Prox-regular functions in hilbert spaces. *Journal of Mathematical Analysis and Applications*, 303(1):1 – 14, 2005.

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

[7] 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.

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

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

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

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

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

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

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

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

[16] 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.

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

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

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

[20] A. S. Dalalyan. Theoretical guarantees for approximate sampling from a smooth and log-concave density. to appear in JRSS B 1412.7392, arXiv, December 2014.

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

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

[23] D. Donoho. For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.

[24] D. L. Donoho, I. M. Johnstone, J. C. Hoch, and A. S. Stern. Maximum entropy and the nearly black object (disc: p67-81). *Journal of the Royal Statistical Society, Series B: Methodological*, 54:41–67, 1992.

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

[28] J. Fan. Comments on "wavelets in statistics: A review" by a. antoniadis. *Journal of the Italian Statistical Society*, 6(2):131, 1997.

[29] J. Fan and R. Li. Variable selection via penalized likelihood. *Journal of American Statistical Association*, 96:1348–1360, 1999.

[30] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties, 2001.

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

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

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

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

[35] 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.

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

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

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

[41] S. Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, 3rd edition, 2008.

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

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

[44] M. Nikolova. Local strong homogeneity of a regularized estimator. *SIAM Journal on Applied Mathematics*, 61(2):633–658, 2000.

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

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

[47] R. A. Poliquin and R. T. Rockafellar. *Prox-regular functions in variational analysis*, 1996.

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

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

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

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

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

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

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

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

[56] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 03 1978.

[57] I. Selesnick. Penalty and shrinkage functions for sparse signal processing. *Connexions*, 11 2012.

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

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

[60] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

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

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

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

[64] 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.

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

[66] J. Woodworth and R. Chartrand. Compressed sensing recovery via nonconvex shrinkage penalties. *CoRR*, abs/1504.02923, 2015.

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

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

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

[70] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.