Sampling from non-smooth distribution through Langevin diffusion

Duy Tung Luu, Jalal M. Fadili, Christophe Chesneau

Sampling from non-smooth distributions through Langevin diffusion

Tung Duy Luu* Jalal Fadili* Christophe Chesneau†

Abstract

In this paper, we propose proximal splitting-type algorithms for sampling from distributions whose densities are not necessarily smooth nor log-concave. Our approach brings together tools from, on the one hand, variational analysis and non-smooth optimization, and on the other hand, stochastic diffusion equations, and in particular the Langevin diffusion. We establish in particular consistency guarantees of our algorithms seen as discretization schemes in this context. These algorithms are then applied to compute the exponentially weighted aggregates for regression problems involving non-smooth priors encouraging some notion of simplicity/complexity. Some popular priors are detailed and implemented on some numerical experiments.

1 Introduction

1.1 Problem statement

We consider the linear regression problem

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

where $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of observations, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the design matrix, $\boldsymbol{\xi}$ is the vector of errors, and $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ is the unknown regression vector we wish to estimate. $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ can be seen as the sensing or degradation operator in inverse problems raising in, e.g., signal and image processing, or the design matrix for a regression problem in statistics and machine learning. Generally, problem (1.1) is either under-determined ($p < n$), or determined ($p = n$) but $\boldsymbol{X}$ is ill-conditioned. In both cases, (1.1) is ill-posed.

The idea of aggregating elements in a dictionary has been introduced in machine learning to combine different techniques (see [43, 66]) with some procedures such as bagging [12], boosting [34, 58] and random forests [1, 8–10, 13, 36]. 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 $\boldsymbol{\theta}_0$ generally conforms to some notions of sparsity/low-complexity. Namely, it has either a simple structure or a small intrinsic dimension. 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).

*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.

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

$$\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} \in \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\mathrm{Argmin}} \Big\{ V(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) + J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \Big\}, \tag{1.2}$$

where $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a general loss function assumed to be differentiable, $J_{\boldsymbol{\lambda}} : \mathbb{R}^p \to \mathbb{R}$ is the regularizing penalty promoting some specific notion of simplicity/low-complexity which depends on a vector of parameters $\boldsymbol{\lambda}$. 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 [11, 14, 15, 19, 27, 47, 61] and its variants such the analysis/fused Lasso [56, 62] or group Lasso [3, 4, 21, 67, 71]. 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 [16, 17, 33, 52]. See [14, 45, 64, 65] for generalizations and comprehensive reviews.

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

$$\widehat{\mu}(\boldsymbol{\theta}) = \frac{\exp\left(-V(\boldsymbol{\theta})/\beta\right)}{\int_{\Theta} \exp\left(-V(\boldsymbol{\omega})/\beta\right) d\boldsymbol{\omega}}, \tag{1.3}$$

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

$$\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} = \int_{\mathbb{R}^p} \boldsymbol{\theta}\widehat{\mu}(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{1.4}$$

Aggregation by exponential weighting has been widely considered in the statistical and machine learning literatures, see e.g. [22, 23, 25, 26, 30, 37, 42, 46, 53, 70] to name a few.

1.4 The Langevin diffusion

In this paper, we focus on the computation of EWA. Computing $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ in (1.4) 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 $\widehat{\mu}$ by constructing an appropriate Markov chain whose stationary distribution is $\widehat{\mu}$, 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 [26], we here use the Langevin diffusion process; see [54].

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

$$d\boldsymbol{L}(t) = \frac{1}{2}\boldsymbol{\rho}(\boldsymbol{L}(t))dt + d\boldsymbol{W}(t), \ t > 0, \ \boldsymbol{L}(0) = \boldsymbol{l}_0, \tag{1.5}$$

where $\boldsymbol{\rho} = \nabla \log \mu$, μ is everywhere non-zero and suitably smooth target density function on $\mathbb{R}^p$, $\boldsymbol{W}$ is a p-dimensional Brownian process and $\boldsymbol{l}_0 \in \mathbb{R}^p$ is the initial value. Under mild assumptions, the SDE (1.5) has a unique strong solution and, $\boldsymbol{L}(t)$ has a stationary distribution with density precisely μ [54, Theorem 2.1]. $\boldsymbol{L}(t)$ is therefore interesting for sampling from μ. In particular, this opens the door to approximating integrals $\int_{\mathbb{R}^p} f(\boldsymbol{\theta})\mu(\boldsymbol{\theta})d\boldsymbol{\theta}$, where $f : \mathbb{R}^p \to \mathbb{R}$, 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 [69].

Forward Euler discretization In practice, in simulating the diffusion sample path, we cannot follow exactly the dynamic defined by the SDE (1.5). 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 iid $\sim \mathcal{N}(0, \mathbf{I}_p)$. The average value $\frac{1}{T}\int_0^T \boldsymbol{L}(t)dt$ can then be naturally approximated via the Riemann sum

$$\frac{\delta}{T}\sum_{k=0}^{\lfloor T/\delta \rfloor - 1} \boldsymbol{L}_k, \tag{1.6}$$

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 comprehensive review of sampling by Langevin diffusion from smooth and log-concave densities, we refer the reader to e.g. [24]. 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) [26, 28, 29, 48]. In [29, 48] 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 [48][1], $-\log \mu$ is replaced with its Moreau envelope, while in [29], it is assumed that $-\log \mu = L + H$, L is convex Lipschitz continuously differentiable, and H is a proper closed convex function replaced by its Moreau envelope. In both these works, convexity plays a crucial role to get quantitative convergence guarantees. Proximal steps within MCMC methods have been recently proposed for some simple (convex) signal processing problems [18], though without any guarantees.

1.5 Contributions

Our main contributions are summarized as follows.

- We aim to enlarge the family of μ covered by [26, 28, 29, 48] by relaxing some underlying conditions. Especially, in our study, μ is structured as $\widehat{\mu}$ in (1.3), and it is not necessarily differentiable nor log-concave.
- We propose two algorithms based on forward-backward proximal splitting for which we prove theoretical consistency guarantees.

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

- These algorithms are applied to compute EWA estimators with several popular penalties in the literature, and illustrated on some numerical problems.

1.6 Paper organization

Some preliminaries, definitions and notations are introduced in Section 2. Section 3 establishes key properties of a Moreau-Yosida regularized version of μ under mild assumptions of the latter. In turn we will consider the SDE (1.5) with such the smoothed density. Well-posedness of this SDE and consistency guarantees for its discrete approximations are proven in Section 4. Section 5 provides a large class of functions, namely prox-regular functions, for which the previous theoretical analysis applies. From this analysis, two algorithms are derived in Section 6 and applied in Section 7 to compute the EWA estimator 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 F or a d-dimensional Euclidean space $\mathbb{R}^d$, we endow it with its usual inner product $\langle \cdot, \cdot \rangle$ and associated norm $\|\cdot\|_2$. $\mathbf{I}_d$ is the identity matrix on $\mathbb{R}^d$. For $r \geq 1$, $\|\cdot\|_r$ will denote the ℓ_r norm of a vector with the usual adaptation for $r = +\infty$.

Let $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ symmetric positive definite, we denote $\langle \cdot, \cdot \rangle_{\boldsymbol{M}} = \langle \cdot, \boldsymbol{M} \cdot \rangle$ and $\|\cdot\|_{\boldsymbol{M}}$ its associated norm. For a matrix $\boldsymbol{M}$, we denote $\sigma_{\min}(\boldsymbol{M})$ its smallest singular value and $|||\boldsymbol{M}|||$ its spectral norm. Of course, $\|\cdot\|_{\boldsymbol{M}}$ and $\|\cdot\|_2$ are equivalent.

Let $\boldsymbol{x} \in \mathbb{R}^d$ and the subset of indices $\mathcal{I} \subset \{1, \ldots, d\}$. We denote $\boldsymbol{x}_{\mathcal{I}}$ the subvector whose entries are those of $\boldsymbol{x}$ indexed by $\mathcal{I}$. For any matrix $\boldsymbol{M}$, $\boldsymbol{M}^\top$ denotes its transpose.

Sets For a set $\mathcal{C}$, denote $I_{\mathcal{C}}$ its characteristic function, i.e., 1 if the argument is in $\mathcal{C}$ and 0 otherwise, and $\iota_{\mathcal{C}}$ its its indicator function, i.e., 0 if the argument is in $\mathcal{C}$ and $+\infty$ otherwise. For an index set $\mathcal{I}$, $\left|\mathcal{I}\right|$ is its cardinality.

Functions We will denote $(\cdot)_+ = \max(\cdot, 0)$ the positive part of a real number. For a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$, its effective domain is $\mathrm{dom}(f) = \left\{\boldsymbol{x} \in \mathbb{R}^d \ : \ f(\boldsymbol{x}) < +\infty\right\}$ and f is proper if $f(\boldsymbol{x}) > -\infty$ for all $\boldsymbol{x}$ and $\mathrm{dom}(f) \neq \emptyset$ as is the case when it is finite-valued. A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$ is lower semi continuous (lsc) at $\boldsymbol{x}_0$ if $\liminf_{\boldsymbol{x} \to \boldsymbol{x}_0} f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0)$.

For a differentiable function f, ∇f is its (Euclidean) gradient. Define $C^{1,+}(\mathbb{R}^d)$ (resp. $C^{1,1}(\mathbb{R}^d)$) the set of differentiable functions in $\mathbb{R}^d$ whose gradient is locally (resp. globally) Lipschitz continuous. We also define $\widetilde{C^{1,+}}(\mathbb{R}^d) \stackrel{\text{def}}{=} \left\{f \in C^{1,+}(\mathbb{R}^d) \ : \ \exists K > 0, \forall \boldsymbol{x} \in \mathbb{R}^d, \langle \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle \leq K(1 + \|\boldsymbol{x}\|_2^2)\right\}$. The following lemma shows that $C^{1,1}(\mathbb{R}^d) \subset \widetilde{C^{1,+}}(\mathbb{R}^d)$.

Lemma 2.1. *Assume that $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous, then there exists $K > 0$ such that*

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

Let us also consider some definitions and properties of variational analysis. A more comprehensive account on variational analysis in finite-dimensional Eudlidean spaces can be found in [55].

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

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

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

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

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

Let us note that $\partial f(\boldsymbol{x})$ and $\partial^F f(\boldsymbol{x})$ are closed, with $\partial^F f(\boldsymbol{x})$ convex and $\partial^F f(\boldsymbol{x}) \subset \partial f(\boldsymbol{x})$ [55, Theorem 8.6]. In particular, if f is a proper lsc convex function, $\partial^F f(\boldsymbol{x}) = \partial f(\boldsymbol{x})$ and f is subdifferentially regular at any point $\boldsymbol{x}$ where $\partial f(\boldsymbol{x}) \neq \emptyset$.

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

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

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

Operators For a set-valued operator $S : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$, its graph is $\operatorname{gph}(S) = \big\{(\boldsymbol{x}, \boldsymbol{v}) \ : \ \boldsymbol{v} \in S(\boldsymbol{x})\big\}$.

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

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

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

3 Moreau-Yosida regularization

In our framework, the target distribution μ is defined as

$$\mu(\boldsymbol{\theta}) \propto \exp\Big(-\Big(L(\boldsymbol{\theta}) + H \circ \boldsymbol{D}^\top(\boldsymbol{\theta})\Big)\Big), \tag{3.1}$$

where $L \in \widetilde{C^{1,+}}(\mathbb{R}^p)$, $\boldsymbol{D} \in \mathbb{R}^{p \times q}$ and $H : \mathbb{R}^q \to \mathbb{R}$. Moreover, H is assumed neither differentiable nor convex. To overcome these difficulties, we invoke arguments from variational analysis [55]. Namely, we will replace H by its Moreau envelope and state the following assumptions to exploit some key properties of the latter. To avoid trivialities, from now on, we assume that $\operatorname{Argmin}(H) \neq \emptyset$.

(H.1) $H : \mathbb{R}^q \to \mathbb{R}$ is lsc and bounded from below.

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

Let us start with some key properties of the Moreau envelope.

Lemma 3.1. *Let $\boldsymbol{M} \in \mathbb{R}^{q\times q}$ depending on $\gamma \in]0,\gamma_0[$ with $\gamma_0 > 0$, we denote it $\boldsymbol{M}_\gamma$, such that $\boldsymbol{M}_\gamma$ is symmetric positive definite for any $\gamma \in]0,\gamma_0[$, and $\gamma \mapsto \|\boldsymbol{\theta}\|_{\boldsymbol{M}_\gamma}$, $\forall \boldsymbol{\theta} \in \mathbb{R}^q$, is a decreasing mapping on $]0,\gamma_0[$. Assume that (H.1) holds.*

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

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

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

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

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 $\boldsymbol{M}_\gamma \in \mathbb{R}^{q\times q}$ symmetric positive definite, assume that (H.1) and (H.2) hold. Then $\operatorname{prox}_{\gamma H}^{\boldsymbol{M}_\gamma}$ is continuous on $(\boldsymbol{x},\gamma) \in \mathbb{R}^q\times]0,\gamma_0[$, and ${}^{\boldsymbol{M}_\gamma,\gamma}H \in C^1(\mathbb{R}^q)$ with gradient*

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

In plain words, Lemma 3.2 tells us that under (H.1)-(H.2), the Moreau envelope is a smooth function, hence the name Moreau-Yosida regularization. Moreover, the action of the operator $\operatorname{prox}_{\gamma H}^{\boldsymbol{M}_\gamma}$ is equivalent to a gradient descent on the Moreau envelope of H in the metric $\boldsymbol{M}_\gamma$ with step-size γ.

Remark 3.1. *When the metric matrix does not depend on γ, Lemmas 3.1 and 3.2 hold with $\gamma_0 = +\infty$.*

Let us now define the smoothed density

$$\mu_\gamma(\boldsymbol{\theta}) = \frac{\exp\left(-\left(L(\boldsymbol{\theta}) + ({}^{\boldsymbol{M},\gamma}H)\circ \boldsymbol{D}^\top(\boldsymbol{\theta})\right)\right)}{Z_\gamma}, \tag{3.2}$$

where

$$Z_\gamma = \int_{\mathbb{R}^p} \exp\left(-\left(L(\boldsymbol{\theta}') + ({}^{\boldsymbol{M},\gamma}H)\circ \boldsymbol{D}^\top(\boldsymbol{\theta}')\right)\right) d\boldsymbol{\theta}'.$$

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

Proposition 3.1. *Assume that (H.1) holds. Then, μ_γ converges to μ in total variation as $\gamma \to 0$.*

4 Langevin diffusion with Moreau-Yosida regularization

Let us define the following SDE with the Moreau-Yosida regularized version of H

$$\begin{aligned} d\boldsymbol{L}(t) &= \boldsymbol{\psi}(\boldsymbol{L}(t))dt + d\boldsymbol{W}(t),\ t>0, \\ \text{where} \quad \boldsymbol{\psi} : \boldsymbol{\theta} \in \mathbb{R}^p &\mapsto -\frac{1}{2}\nabla\big(L + ({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top\big)(\boldsymbol{\theta}), \end{aligned} \tag{4.1}$$

$\boldsymbol{\psi}$ is the drift coefficient.

Recall that (**H.1**) and (**H.2**) were mild assumptions required to establish key properties of Moreau-Yosida regularization, which in turn allow to compute $\nabla^{\boldsymbol{M},\gamma}H$ by exploiting its the relation between $\nabla^{\boldsymbol{M},\gamma}H$ and $\mathrm{prox}_{\gamma H}^{\boldsymbol{M}}$ as stated in Lemma 3.2. Now, to guarantee well-posed (existence and uniqueness) and discretization consistency of the SDE (4.1), we will also need the following assumptions.

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

(**H.4**) There exists $C>0$ such that $\left\langle \boldsymbol{D}^\top\boldsymbol{\theta}, \mathrm{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{D}^\top\boldsymbol{\theta})\right\rangle_{\boldsymbol{M}} \leq C(1+\|\boldsymbol{\theta}\|_2^2), \forall \boldsymbol{\theta}\in\mathbb{R}^p$.

4.1 Well-posedness

We start with the following characterization of the drift $\boldsymbol{\psi}$.

Proposition 4.1. *Assume that* (**H.1**)-(**H.4**) *hold. Then,*

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

and

$$\boldsymbol{\psi} \text{ is locally Lipschitz continuous.} \tag{4.3}$$

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

Proposition 4.2. *Assume that* (4.2) *and* (4.3) *hold. 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 having the density* μ_γ *in* (3.2).

4.2 Discretization

Approach 1 Inserting the identities of Lemma 3.2 into (4.1), we get the SDE

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

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

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

Note that by Lemma 3.2, and without the stochastic term $\sqrt{\delta}\boldsymbol{Z}_k$, (4.5) amounts to a relaxed form of gradient descent on L and the Moreau envelope of H in the metric $\boldsymbol{M}$ with step-size δ.

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

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

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* (4.2) *and* (4.3) *hold and* $\mathbb{E}\left[\|\boldsymbol{L}(0)\|_2^r\right] < \infty$ *for any* $r \geq 2$. *Then*

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

The convergence rate is of order $\delta^{1/2}$ *when* $\operatorname{prox}_{\gamma H}^{\boldsymbol{M}}$ *is globally Lipschitz continuous.*

Approach 2 Assume now that the metric also depends on $\gamma \in]0, \gamma_0[$ with $\gamma_0 > 0$, and we emphasize this by denoting it $\boldsymbol{M}_{\gamma}$ such that $\boldsymbol{M}_{\gamma}$ is symmetric positive definite for any $\gamma \in]0, \gamma_0[$, $\gamma \to \|\boldsymbol{\theta}\|_{\boldsymbol{M}_{\gamma}}$, for any $\boldsymbol{\theta} \in \mathbb{R}^q$, is a decreasing mapping on $]0, \gamma_0[$, and $\boldsymbol{M}_{\gamma} \underset{\gamma\to 0}{\to} \mathbf{I}_q$ (such a choice is motivated by the scheme described in Section 6.1). One can consider an alternative version of the SDE (4.1), i.e.,

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

Denote the drift coefficient of (4.7) by $\boldsymbol{\phi}$, we get that

$$\langle\boldsymbol{\phi}(\boldsymbol{\theta}), \boldsymbol{\theta}\rangle = \langle\boldsymbol{\psi}(\boldsymbol{u}), \boldsymbol{u}\rangle,$$

where $\boldsymbol{u} = {\boldsymbol{M}_{\gamma}}^{-1/2}\boldsymbol{\theta}$. Therefore, it is easily seen that $\boldsymbol{\phi}$ also satisfies (4.2) and (4.3) under assumptions (**H.1**)-(**H.4**). In turn, Proposition 4.2 applies to (4.7) the diffusion $\boldsymbol{L}$ is unique, non explosive and admits an unique invariant measure μ_{γ} having density

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

where $Z_{\gamma} = \sqrt{\det(\boldsymbol{M}_{\gamma})}\int_{\mathbb{R}^p}\exp\left(-\big(L + ({}^{\boldsymbol{M}_{\gamma},\gamma}H)\circ\boldsymbol{D}^{\top}\big)(\boldsymbol{u})\right)d\boldsymbol{u}$. Since $\det(\boldsymbol{M}_{\gamma}) \underset{\gamma\to 0}{\to} 1$, applying the reasoning in the proof of Proposition 3.1, we also deduce that μ_{γ} converges to μ in total variation as $\gamma \to 0$.

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

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

In an analogous way to (4.5), the forward Euler discretization of (4.8) has a deterministic part which is a relaxed gradient descent in the metric ${\boldsymbol{M}_{\gamma}}^{-1}$. In turn, mean square convergence of the Euler discretizations of both (4.7) and (4.8) and of their first-order moments can be established exactly in the same way as in Theorem 4.1. We omit the details here for the sake of brevity.

5 Prox-regular penalties

We now present a large class of penalties, namely prox-regular functions, which satisfy the key assumptions **(H.2)** and **(H.3)**.

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

Definition 5.1 (Prox-regularity)**.** *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, given a point $\bar{\boldsymbol{x}} \in \mathrm{dom}(f)$. f is prox-regular at $\bar{\boldsymbol{x}}$ for $\bar{\boldsymbol{v}}$, with $\bar{\boldsymbol{v}} \in \partial f(\bar{\boldsymbol{x}})$ if f is locally lsc at $\bar{\boldsymbol{x}}$, there exist $\epsilon > 0$ and $r > 0$ such that*

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

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

Example 5.1. *The class of prox-regular functions is large enough to include many of those used in statistics. For instance, here examples where prox-regularity is fullfilled (see [55, Chapter 13, Section F] and [51]):*

(i) Proper lsc convex functions.

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

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

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

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

is single valued and continuous, or equivalently, to

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

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

Lemma 5.1 ([50, Theorem 3.2])**.** *When $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is locally lsc at $\bar{\boldsymbol{x}} \in \mathbb{R}^d$, the following are equivalent*

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

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

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

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

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

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

Let us consider a prox-regular function satisfying **(H.1)**. Owing to the following lemma, such type of functions also fullfills **(H.2)** and **(H.3)**.

Lemma 5.2. *Let $\boldsymbol{M} \in \mathbb{R}^{p\times p}$ symmetric positive definite and γ small enough, assume that $H : \mathbb{R}^p \to \mathbb{R}$ is prox-regular and satisfies **(H.1)**. Then $\mathrm{prox}^{\boldsymbol{M}}_{\gamma H}$ is single-valued and locally Lipschitz continuous.*

Lower-C^2 (or semi-convex) functions, see Example 5.1-(ii), satisfy the global counterpart of Lemma 5.1-(ii). For a lower-C^2 penalty H satisfying **(H.1)**, the following lemma shows that $\mathrm{prox}^{\boldsymbol{M}}_{\gamma H}$ is globally Lipschitz continuous with a proper choice of γ which in turn implies directly **(H.4)** according to Lemma 2.1.

Lemma 5.3. *Assume that H is lower-C^2 (with constant r) satisfying **(H.1)** and $\gamma \in]0, r\sigma_{\min}(\boldsymbol{M})[$, $\mathrm{prox}^{\boldsymbol{M}}_{\gamma H}$ is single-valued and Lipschitz continuous with constant $\frac{\|\boldsymbol{M}\|}{\sigma_{\min}(\boldsymbol{M})}\left(1 - \frac{\gamma}{r\sigma_{\min}(\boldsymbol{M})}\right)^{-1}$.*

When $\mathrm{prox}^{\boldsymbol{M}}_{\gamma H}$ is globally Lipschitz continuous, the optimal convergence rate in (4.6) is of order $\delta^{1/2}$ in view of Theorem 4.1.

6 Forward-Backward type LMC algorithms

Let us now deal with our main goal: computing the EWA estimator in (1.4) by sampling from $\widehat{\mu}$. Remind that

$$\widehat{\mu}(\boldsymbol{\theta}) \propto \exp\left(-\frac{F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) + J_{\boldsymbol{\lambda}}(\boldsymbol{\theta})}{\beta}\right),$$

where $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a general loss and $J_{\boldsymbol{\lambda}} : \mathbb{R}^p \to \mathbb{R}$ is the penalty. Assume that $F(\boldsymbol{X}\cdot, \boldsymbol{y}) \in \widetilde{C^{1,+}}(\mathbb{R}^p)$ and the penalty takes the form $J_{\boldsymbol{\lambda}} = W_{\boldsymbol{\lambda}} \circ \boldsymbol{D}^\top$. Let us impose the following assumptions on $W_{\boldsymbol{\lambda}}$.

(H.1') $W_{\boldsymbol{\lambda}} : \mathbb{R}^q \to \mathbb{R}$ is lsc and bounded from below.

(H.2') $\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}}$ is single valued.

(H.3') $\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}}$ is locally Lipschitz continuous.

To lighten notation, we will write $F_\beta \stackrel{\text{def}}{=} F(\boldsymbol{X}\cdot, \boldsymbol{y})/\beta$. This section aims to describe our Forward-Backward type Langevin Monte-Carlo (LMC) algorithms to implement (1.4). These algorithms are based on wise specializations of the results reported in Section 4.

6.1 Forward-backward LMC (FBLMC)

In (3.1), we set $\boldsymbol{D} = \mathbf{I}_p$ (hence $J_{\boldsymbol{\lambda}} = W_{\boldsymbol{\lambda}}$), $L \equiv 0$, and $H = F_\beta + J_{\boldsymbol{\lambda}}/\beta$, where F is a quadratic loss, i.e., $F_\beta(\boldsymbol{\theta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2/\beta$. Observe that H satisfies (H.1) by assumption (H.1'). To check (H.2)-(H.4), we need to design a metric in which $\mathrm{prox}_{\gamma H}^{\boldsymbol{M}}$ is expressed as a function of $\mathrm{prox}_{\gamma J_{\boldsymbol{\lambda}}/\beta}$. This idea is formalized in the following lemma.

Lemma 6.1. *Assume that* $0 < \gamma < \beta/(2|||\boldsymbol{X}|||^2)$ *and* (H.1') *holds. Define* $\boldsymbol{M}_\gamma \overset{def}{=} \mathbf{I}_p - (2\gamma/\beta)\boldsymbol{X}^\top\boldsymbol{X}$*, which is symmetric positive definite. Then*

$$\mathrm{prox}_{\gamma H}^{\boldsymbol{M}_\gamma} = \mathrm{prox}_{\gamma J_{\boldsymbol{\lambda}}/\beta} \circ (\mathbf{I}_p - \gamma\nabla F_\beta). \tag{6.1}$$

In view of Lemma 6.1, (H.2') and (H.3'), it is immediate to check that (H.2) and (H.3) are satisfied. It remains now to verify (H.4) which is fulfilled by imposing the following assumption on $W_{\boldsymbol{\lambda}}$ (or $J_{\boldsymbol{\lambda}}$).

(H.4'-FB) There exists $C'_{\mathrm{FB}} > 0$ such that

$$\Big\langle \mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta} \circ (\mathbf{I}_p - \gamma\nabla F_\beta)(\boldsymbol{\theta}), \boldsymbol{\theta} \Big\rangle_{\boldsymbol{M}_\gamma} \leq C'_{\mathrm{FB}}(1 + \|\boldsymbol{\theta}\|_2^2), \quad \forall \boldsymbol{\theta} \in \mathbb{R}^p.$$

By Lemma 2.1, a sufficient condition for (H.4'-FB) to hold is that the proximal mapping of $W_{\boldsymbol{\lambda}}$ is Lipschitz continuous.

From Lemmas 3.2 and 6.1, we get

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

With this expression at hand, the forward Euler discretization of the SDE (4.1), specialized to the current case, reads

$$\boldsymbol{L}_{k+1} = \boldsymbol{L}_k - \frac{\delta}{2\gamma}\boldsymbol{M}_\gamma\Big(\boldsymbol{L}_k - \mathrm{prox}_{\gamma J_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{L}_k - \gamma\nabla F_\beta(\boldsymbol{L}_k))\Big) + \sqrt{\delta}\boldsymbol{Z}_k,\ t > 0,\ \boldsymbol{L}_0 = \boldsymbol{l}_0, \tag{6.2}$$

Similarly, the forward Euler discretization of the SDE (4.8) is given by

$$\boldsymbol{U}_{k+1} = (1 - \tfrac{\delta}{2\gamma})\boldsymbol{U}_k + \frac{\delta}{2\gamma}\mathrm{prox}_{\gamma J_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{U}_k - \gamma\nabla F_\beta(\boldsymbol{U}_k)) + \sqrt{\delta}\boldsymbol{Z}_k,\ t > 0,\ \boldsymbol{U}_0 = \boldsymbol{l}_0. \tag{6.3}$$

The familiar reader may have recognized that the deterministic part of (6.3) is nothing but the relaxed form of the so-called Forward-Backward proximal splitting algorithm [6]. This terminology reflects that there is a forward Euler discretization on F_β and a Euler backward discretization on $J_{\boldsymbol{\lambda}}$.

6.2 Semi-Forward-Backward LMC (Semi-FBLMC)

The main limitation of (6.2) is that the proximal mapping of $J_{\boldsymbol{\lambda}}$ must be easy to compute. This may not be true even if the proximal mapping of $W_{\boldsymbol{\lambda}}$ is accessible as, for for example, when $\boldsymbol{D}$ does not have orthogonal rows [6]. Our goal is to circumvent this difficulty.

Toward this goal, in (3.1), consider now $L = F_\beta$, $H = W_{\boldsymbol{\lambda}}/\beta$ and $\boldsymbol{M} = \mathbf{I}_q$. Owing to (H.1')-(H.3'), one can check that (H.1)-(H.3) are fulfilled. Assumption (H.4) is verified by imposing the following on $W_{\boldsymbol{\lambda}}$.

(H.4'-SFB) There exists $C'_{\mathrm{SFB}} > 0$ such that $\left\langle \mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{u}), \boldsymbol{u} \right\rangle \leq C'_{\mathrm{SFB}}(1 + \|\boldsymbol{u}\|_2^2)$, $\forall \boldsymbol{u} \in \mathbb{R}^q$.

From Lemma 3.2, we obtain

$$\nabla\Big((^{\gamma}H) \circ \boldsymbol{D}^\top\Big)(\boldsymbol{\theta}) = \gamma^{-1}\boldsymbol{D}(\boldsymbol{D}^\top\boldsymbol{\theta} - \mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{D}^\top\boldsymbol{\theta})).$$

Thus, the forward Euler discretization of SDE (4.1) now reads

$$\boldsymbol{L}_{k+1} = \boldsymbol{L}_k - \frac{\delta}{2}\nabla F_\beta(\boldsymbol{L}_k) - \frac{\delta}{2\gamma}\boldsymbol{D}\Big(\boldsymbol{D}^\top\boldsymbol{L}_k - \mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{D}^\top\boldsymbol{L}_k)\Big) + \sqrt{\delta}\boldsymbol{Z}_k,\ t > 0,\ \boldsymbol{L}_0 = \boldsymbol{l}_0. \quad (6.4)$$

In the case where $\boldsymbol{D} = \mathbf{I}_p$, F_β and $W_{\boldsymbol{\lambda}}$ are convex, we recover the scheme studied in [29].

7 Applications to penalties in statistics

In this section, we exemplify our LMC sampling algorithms for some popular penalties in the statistical and machine learning literature. Our goal is by no means to be exhaustive, but rather to be illustrative and show the versatility of our framework. For each penalty, we aim at checking that assumptions **(H.1')**-**(H.3')**, **(H.4'-FB)** and **(H.4'-SFB)** hold, and to compute $\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}$. In turn, this allows to apply our algorithms (6.3) and (6.4) to compute EWA with such penalties.

7.1 Analysis group-separable penalties

We first focus on a class of penalties where $J_{\boldsymbol{\lambda}}$ is analysis group-separable, i.e.,

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = W_{\boldsymbol{\lambda}}(\boldsymbol{D}^\top\boldsymbol{\theta}) \quad \text{where} \quad W_{\boldsymbol{\lambda}}(\boldsymbol{u}) = \textstyle\sum_{l=1}^{L} w_{\boldsymbol{\lambda}}\big(\|\boldsymbol{u}_{\mathcal{G}_l}\|_2\big), \quad (7.1)$$

for $w_{\boldsymbol{\lambda}} : \mathbb{R}^+ \to \mathbb{R}$, and some uniform partition $(\mathcal{G}_l)_{l\in\{1,\dots,L\}}$ of $\{1,\dots,q\}$, i.e., $\cup_{l=1}^{L}\mathcal{G}_l = \{1,\dots,q\}$ and $\mathcal{G}_l \cap \mathcal{G}_{l'}$, $\forall l \neq l'$.

Remark 7.1. *It is worth mentioning that separability of $W_{\boldsymbol{\lambda}}$ does not entail that of $J_{\boldsymbol{\lambda}}$. In fact, overlapping groups can be easily taken intro account as any overlapping-group penalty can be written as the composition of $W_{\boldsymbol{\lambda}}$ with a linear operator, say $\boldsymbol{B}$, such that $\boldsymbol{B}^\top\boldsymbol{B}$ is diagonal, and $\boldsymbol{B}$ acts as a group extractor, see [20, 49].*

A first consequence of separability is that $\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}$ is also separable under the following mild assumptions on $w_{\boldsymbol{\lambda}}$.

(W.1) $w_{\boldsymbol{\lambda}}$ is bounded from below on $]0, +\infty[$.

(W.2) $w_{\boldsymbol{\lambda}}$ is non-decreasing functions on $]0, +\infty[$.

Lemma 7.1. *Assume that Assumptions (W.1) and (W.2) hold, and $w_{\boldsymbol{\lambda}}$ is continuous on $]0, +\infty[$. For any $\boldsymbol{u} \in \mathbb{R}^q$ and $\gamma > 0$, we have*

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

Our aim is now to design a family of penalties that will allow to establish (H.1')-(H.3'), (H.4'-FB) and (H.4'-SFB), while involving a form of shrinkage which is ubiquitous in low-complexity regularization. Inspired by the work of [2], we make the following assumptions on $w_{\boldsymbol{\lambda}}$.

(W.3) $w_{\boldsymbol{\lambda}}$ is continuously differentiable on $]0,+\infty[$ and the problem $\min_{t\in[0,+\infty[}\{t+\frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}{}'(t)\}$ has a unique solution at 0 for a given γ.

Under these assumptions, $\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}$ has a indeed convenient shrinkage-type form.

Lemma 7.2 ([2, Theorem 1])**.** *Assume that (W.2) and (W.3) hold for some $\gamma>0$. Then, $\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}$ are the single-valued continuous mappings, and satisfy, for $t\in[0,+\infty[$,*

$$\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t)=\begin{cases}0 & \text{if } t\leq\frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}{}'(0^+),\\ t-\frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}{}'\Big(\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t)\Big) & \text{if } t>\frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}{}'(0^+).\end{cases}\tag{7.2}$$

Let us turn to check our assumptions. (H.1')-(H.3') are fulfilled thanks to (W.1)-(W.3). It remains to check (H.4'-FB) and (H.4'-SFB). This is the subject of the following lemma.

Lemma 7.3. *Assume that (W.2) and (W.3) hold for some $\gamma>0$, then (H.4'-FB) and (H.4'-SFB) also hold.*

We now discuss some popular penalties $w_{\boldsymbol{\lambda}}$ that satisfy (W.1)-(W.3) for some $\gamma>0$.

7.2 Examples

ℓ_1 penalty Take $w_{\boldsymbol{\lambda}}: t\in\mathbb{R}^+\mapsto\lambda t$. This entails the analysis group Lasso penalty

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

Clearly, $w_{\boldsymbol{\lambda}}$ is a continuous positive convex function which verifies (W.1)-(W.3) for any $\gamma>0$, and its proximal mapping corresponds to soft-thresholding, i.e.,

$$\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t)=(t-\gamma\lambda/\beta)_+,\quad\forall t\geq 0.$$

FIRM penalty The FIRM penalty is given by [35]

$$w_{\boldsymbol{\lambda}}(t)=\begin{cases}\lambda\Big(t-\frac{t^2}{2\mu}\Big) & \text{if } 0\leq t\leq\mu,\\ \frac{\lambda\mu}{2} & \text{if } t>\mu.\end{cases}\tag{7.3}$$

which entails the corresponding analysis group FIRM penalty $J_{\boldsymbol{\lambda}}$. Since $w_{\boldsymbol{\lambda}}{}'(t)=\lambda\big(1-\frac{t}{\mu}\big)_+\geq 0$, $w_{\boldsymbol{\lambda}}$ is non-decreasing and bounded from below by $w_{\boldsymbol{\lambda}}(0)=0$ on $]0,+\infty[$. Thus, $w_{\boldsymbol{\lambda}}$ satisfies (W.1) and (W.2). Assumption (W.3) also holds for any $\gamma<\beta\mu/\lambda$. The operator $\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}$ can be constructed from [68, Definition II.3]. Its formula is defined as

$$\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t)=\begin{cases}0 & \text{if } 0\leq t\leq\alpha,\\ \frac{\mu}{\mu-\alpha}(t-\alpha) & \text{if } \alpha<t\leq\mu,\\ t & \text{if } t>\mu,\end{cases}\tag{7.4}$$

where $\alpha=\gamma\lambda/\beta$. The formula (7.4) can also be found using Lemma 7.2. Observe that the FIRM shrinkage (7.4) interpolates between hard- (see [68, Definition II.2]) and soft-thresholding. In particular, (7.4) coincides with soft-thresholding when $\mu\to\infty$.

SCAD penalty The SCAD penalty, proposed in [32] is parameterized by $\boldsymbol{\lambda} = (\lambda, a) \in]0, +\infty[\times]2, +\infty[$ as

$$w_{\boldsymbol{\lambda}}(t) = \begin{cases} \lambda t & \text{if } 0 \leq t \leq \lambda, \\ -\frac{t^2 - 2a\lambda t + \lambda^2}{2(a-1)} & \text{if } \lambda < t \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } t > a\lambda, \end{cases} \tag{7.5}$$

The following lemma establishes the validity of $w_{\boldsymbol{\lambda}}$ and computes $\operatorname{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}$.

Lemma 7.4. *Let $w_{\boldsymbol{\lambda}}$ defined in (7.5), and $\kappa = \gamma/\beta$. For any $\gamma < (a-1)\beta$,*

(i) $w_{\boldsymbol{\lambda}}$ satisfies (W.1) - (W.3),

(ii) The proximal mapping of the SCAD penalty is given by the shrinkage

$$\operatorname{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t) = \begin{cases} (t - \kappa\lambda)_+ & \text{if } 0 \leq t \leq (\kappa+1)\lambda, \\ \frac{(a-1)t - ka\lambda}{a-1-\kappa} & \text{if } (\kappa+1)\lambda < t \leq a\lambda, \\ t & \text{if } t > a\lambda. \end{cases} \tag{7.6}$$

Since $a > 2$, one can set $\kappa = 1$. In this case, (7.6) specializes to [32, Equation (2.8)].

ℓ_∞ penalty The ℓ_∞ norm penalty is convex and continuous but is not separable, unlike the previous ones. It is a suitable prior to promote flat vectors, and has found applications in several fields [39, 44, 60]. It entails the following penalty $W_{\boldsymbol{\lambda}}$:

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = W_{\boldsymbol{\lambda}}(\boldsymbol{D}^\top \boldsymbol{\theta}) \quad \text{where} \quad W_{\boldsymbol{\lambda}}(\boldsymbol{u}) = \lambda \max_{l \in \{1,\dots,L\}} \left\{ \left\| [\boldsymbol{u}]_{\mathcal{G}_l} \right\|_2 \right\}, \tag{7.7}$$

where $\boldsymbol{\lambda} = \lambda > 0$. Since $W_{\boldsymbol{\lambda}}$ is not separable, Lemma 7.1 is not applicable. Nevertheless, the proximal mapping of $W_{\boldsymbol{\lambda}}$ can still be obtained easily from the projector on the ℓ_1 unit ball, i.e.,

$$\operatorname{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{u}) = \boldsymbol{u} - \mathrm{P}_{\left\{ \boldsymbol{x} \,:\, \sum_l \left\| \boldsymbol{x}_{\mathcal{G}_l} \right\|_2 \leq \frac{\beta}{\lambda\gamma} \right\}}(\boldsymbol{u}). \tag{7.8}$$

This projector can be obtained from [31, Proposition 2] (see also references therein). One can see that **(H.1')**-**(H.3')** hold. We report the verification of **(H.4'-FB)** and **(H.4'-SFB)** in the proof of the following lemma.

Lemma 7.5. *Let $W_{\boldsymbol{\lambda}}$ in (7.7). Then **(H.4'-FB)** and **(H.4'-SFB)** hold.*

8 Numerical experiments

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

8.1 Image processing experiments

Let $\boldsymbol{\theta}_0$ is a 2-D image which is a matrix in $\mathbb{R}^{128\times128}$. Let us denote $\operatorname{vec} : \mathbb{R}^{\sqrt{p}\times\sqrt{p}} \to \mathbb{R}^p$ the vectorization operator, i.e. the operator which stacks the columns of its arguments. We then consider the following linear regression problem

$$\boldsymbol{y} = \boldsymbol{X} \operatorname{vec}(\boldsymbol{\theta}_0) + \boldsymbol{\xi}. \tag{8.1}$$

Here $p = 128^2$ and $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. The noise level σ is chosen according to the simulated $\boldsymbol{\theta}_0$ through the signal-to-noise ratio SNR, i.e. $\sigma = n^{-1/2} \|\boldsymbol{X}\boldsymbol{\theta}_0\|_2 / 10^{\mathrm{SNR}/10}$. In our experiments, we take $\mathrm{SNR} = 5$.

The goal is estimating $\boldsymbol{\theta}_0$ by computing the EWA estimators via the penalties proposed in Section 7. Three types of problems are considered: compressed sensing, inpainting and deconvolution whose regression function described in what follows.

- Compressed sensing: in this case $\boldsymbol{X}$ is drawn from a random ensemble. In our experiments, $\boldsymbol{X}$ is drawn uniformly at random from the Rademacher ensemble, i.e., its entries are iid Rademacher random variables. We also set $n = 9p/16$.

- Inpainting In this case, $\boldsymbol{X}$ acts as a masking operator. Let $\mathcal{M} \subset \{1, \dots, p\}$ be the set indexing masked pixels. Thus
$$\boldsymbol{X} \operatorname{vec}(\boldsymbol{\theta}_0) = \big(\operatorname{vec}(\boldsymbol{\theta})_{j \in \{1,\dots,p\} \setminus \mathcal{M}}\big).$$
In our numerical experiments, we mask out 20% of the pixels, and thus $n = p - \lfloor 20\% p \rfloor$ where $\lfloor p \rfloor$ the stands of integer part of p. The masked positions are chosen randomly from the uniform distribution.

- Deconvolution In this case $\boldsymbol{X}$ is the convolution operator with a Gaussian kernel with periodic boundary conditions, such that $\boldsymbol{X}$ is diagonalized in the discrete Fourier basis. In this experiment, the standard deviation of the kernel is set to 1.

Assuming that the targeted image is piecewise smooth, a popular prior is the so-called isotropic total variation [57]. To cas this into our framework, define $\boldsymbol{D}_c : \mathbb{R}^{\sqrt{p} \times \sqrt{p}} \to \mathbb{R}^{\sqrt{p} \times \sqrt{p}}$ and $\boldsymbol{D}_r : \mathbb{R}^{\sqrt{p} \times \sqrt{p}} \to \mathbb{R}^{\sqrt{p} \times \sqrt{p}}$ the finite difference operators along, respectively, the columns and rows of an image, with Neumann boundary conditions. We define $\boldsymbol{D}_{\mathrm{TV}}$ as

$$\boldsymbol{D}_{\mathrm{TV}} \colon \boldsymbol{\theta} \in \mathbb{R}^{\sqrt{p} \times \sqrt{p}} \mapsto \operatorname{vec}\Big(\big(\operatorname{vec}(\boldsymbol{D}_r(\boldsymbol{\theta})), \operatorname{vec}(\boldsymbol{D}_c(\boldsymbol{\theta}))\big)^\top \Big)^\top \in \mathbb{R}^{2p}.$$

With this notation, our prior penalty $J_{\boldsymbol{\lambda}}$ reads

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \sum_{l=1}^{p} w_{\boldsymbol{\lambda}}\Big(\sqrt{\operatorname{vec}(\boldsymbol{D}_r(\boldsymbol{\theta}))_l^2 + \operatorname{vec}(\boldsymbol{D}_c(\boldsymbol{\theta}))_l^2} \Big) = W_{\boldsymbol{\lambda}}(\boldsymbol{D}_{\mathrm{TV}}\boldsymbol{\theta}), \tag{8.2}$$

which clearly has the form (7.1) with p blocks of equal size 2.

To estimate $\boldsymbol{\theta}_0$ from (8.1), we employ the EWA estimator (1.4) with $F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{X} \operatorname{vec}(\boldsymbol{\theta})\|_2^2$ and $J_{\boldsymbol{\lambda}}$ in (8.2). For each problem instance (compressed sensing, inpainting or deconvolution), we tested $w_{\boldsymbol{\lambda}}$ as the ℓ_1, SCAD and FIRM penalties. The corresponding EWA estimators are denoted respectively EWA-ℓ_1, EWA-SCAD and EWA-FIRM. Because of the presence of the analysis operator $\boldsymbol{D}_{\mathrm{TV}}$, which is not unitary, we applied Semi-FBLMC scheme (6.4) to compute EWA with $\beta = 1/(pn)$, $\gamma = \beta$, and $\delta = \{5\beta/10^3, 5\beta/10^2, 5\beta/10^6\}$ respectively associated to inpainting, deconvolution and compressed sensing problems. The results are depicted in Figure 1.

8.2 Signal processing experiments

Here we consider reconstructing a piecewise flat 1D signal from compressed sensing measurements using EWA. For this, we create a $p = 128$ sample signal whose coordinates are valued in $\{-1, 1\}$ and compute the observations (8.1) where $\boldsymbol{X}$ is drawn from the Rademacher ensemble with $n > p$ [2]. We set $F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2$, $J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_\infty$, i.e. $\boldsymbol{D} = \mathbf{I}_p$ and the size of groups is 1. We can then use the FBLMC scheme (6.3), where we choose $\beta = 1/(pn)$, $\gamma = \beta$, and $\delta = 5/10^2$. The results are shown in Figure 2.

[2] The overdetermined regime is known to yield good performance for the ℓ_∞ penalty [63].

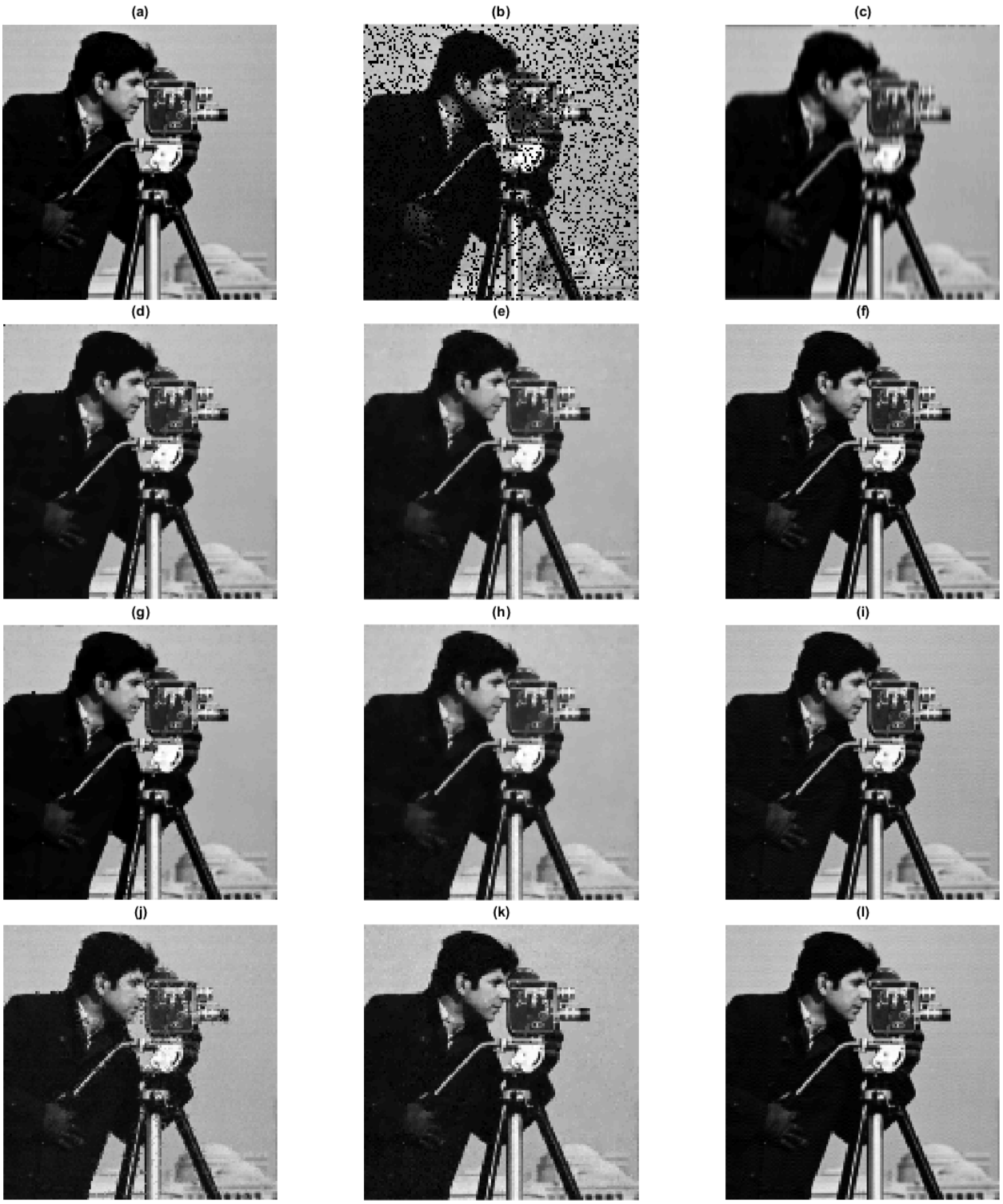

Figure 1: (a): Original image. (b,c) Observed masked and blurry images. (d, e, f): EWA-ℓ_1 estimated images from masked image, compressed sensing measurements, and blurry image. (g, h, i): EWA-FIRM estimated images from masked image, compressed sensing measurements, and blurry image. (j, k, l): EWA-SCAD estimated images from masked image, compressed sensing measurements, and blurry image.

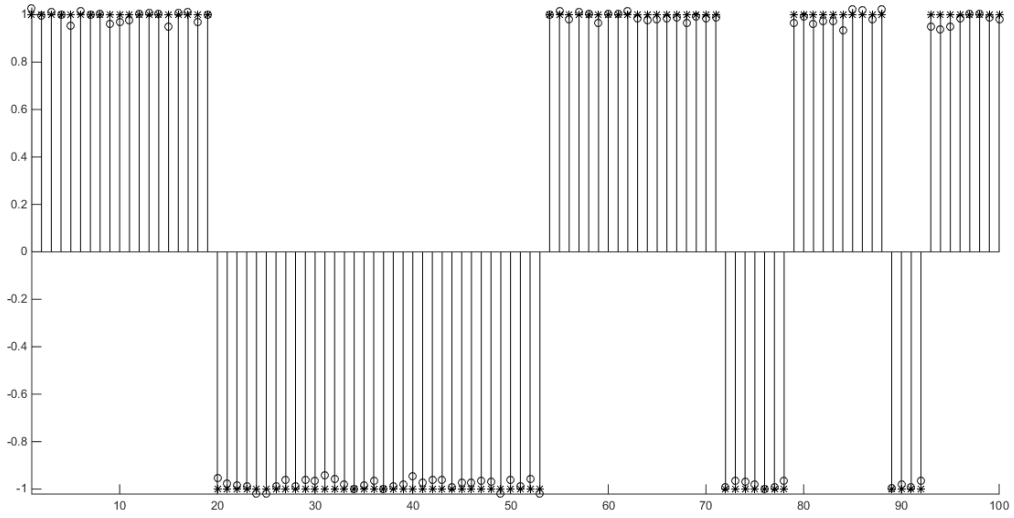

Figure 2: Compressed sensing with EWA using the ℓ_∞ penalty. $'*'$ is the original signal and $'\circ'$ is the the estimated one.

9 Proofs

Proof of Lemma 2.1 Let $\boldsymbol{x}^* \in \mathcal{C}$, a bounded subset of $\mathbb{R}^d$. Using Young and Jensen inequalities as well as $\widetilde{K}$-Lipschitz continuity of $\boldsymbol{f}$, we obtain

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

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

Proof of Lemma 3.1

(i) In view of **(H.1)**, H is prox-bounded by [55, Exercise 1.24] for any $\gamma \in]0, \gamma_0[$, and then for any $\boldsymbol{x}$, $\frac{1}{2\gamma}\left\|\boldsymbol{x} - \cdot\right\|_{\boldsymbol{M}_\gamma}^2 + H$ is proper lsc and level-bounded uniformly in $(\boldsymbol{x}, \gamma) \in \mathbb{R}^q \times]0, \gamma_0[$. This entails that the set of minimizers of this function, i.e. $\operatorname{prox}_{\gamma H}^{\boldsymbol{M}_\gamma}(\boldsymbol{x})$, is a non-empty compact set. For the last claim, suppose that $\boldsymbol{x} \in \operatorname{Argmin}(H) \neq \emptyset$ and bounded but $\boldsymbol{x} \notin \operatorname{prox}_{\gamma H}^{\boldsymbol{M}_\gamma}(\boldsymbol{x})$. Thus, for any $\boldsymbol{p} \in \operatorname{prox}_{\gamma H}^{\boldsymbol{M}_\gamma}(\boldsymbol{x})$, we have $\boldsymbol{p} \neq \boldsymbol{x}$ and

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

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

(ii) The continuity and finiteness properties follow from [55, Theorem 1.17(c)] (see also[55, Theorem 1.25]), where we use (**H.1**). For the second claim, we have $\forall \boldsymbol{x} \in \mathbb{R}^q$

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

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

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

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

□

Proof of Lemma 3.2 By virtue of Lemma 3.1-(i) and (**H.2**), $\operatorname{prox}_{\gamma H}^{\boldsymbol{M}_\gamma}$ is clearly non-empty and single valued. The continuity property follows from [55, Theorem 1.17(b)] (see also [55, Theorem 1.25]) and single-valuedness. By Lemma 3.1-(ii), ${}^{\boldsymbol{M}_\gamma,\gamma}H(\boldsymbol{\theta})$ is finite. Since (**H.1**) holds, H is prox-bounded with threshold ∞ by [55, Exercise 1.24]. Invoking [55, Example 10.32], we get that $-{}^{\boldsymbol{M}_\gamma,\gamma}H$ is locally Lipschitz continuous, subdifferentially regular and

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

Combining this with [55, Theorem 9.18] applied to $-{}^{\boldsymbol{M}_\gamma,\gamma}H$, we obtain that ${}^{\boldsymbol{M}_\gamma,\gamma}H$ is differentiable and its gradient is precisely as given. □

Proof of Proposition 3.1 Recall that we denote with the same symbol the measure and its density with respect to the Lebesgue measure. Thus

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

where

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

and $Z = \int_{\mathbb{R}^M} \exp\big(-(L(\boldsymbol{\theta}') + H \circ \boldsymbol{D}^\top(\boldsymbol{\theta}'))\big) d\boldsymbol{\theta}'$. In view of Lemma 3.1(ii), applying the monotone convergence theorem, we conclude that $Z_\gamma \to Z$ when $\gamma \to 0$. This together with Lemma 3.1(ii) again yield that μ_γ converges to μ pointwise. We conclude using Scheffé(-Riesz) theorem [41, 59]. □

Proof of Proposition 4.1 In view of Lemma 3.2, the drift term reads

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

Since $L \in \widetilde{C^{1,+}}(\mathbb{R}^p)$ and **(H.4)** holds, there exist $K_1 > 0$ and $K_2 > 0$ such that

$$\begin{aligned}\langle \boldsymbol{\psi}(\boldsymbol{\theta}), \boldsymbol{\theta}\rangle &= -\frac{1}{2}\langle \nabla L(\boldsymbol{\theta}), \boldsymbol{\theta}\rangle - \frac{1}{2\gamma}\left\|\boldsymbol{D}^\top\boldsymbol{\theta}\right\|_{\boldsymbol{M}}^2 + \frac{1}{2}\left\langle \mathrm{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{D}^\top\boldsymbol{\theta}), \boldsymbol{D}^\top\boldsymbol{\theta}\right\rangle_{\boldsymbol{M}} \\ &\leq K_1(1 + \|\boldsymbol{\theta}\|_2^2) + |||\boldsymbol{D}|||^2|||\boldsymbol{M}|||/(2\gamma)\,\|\boldsymbol{\theta}\|_2^2 + K_2(1 + \|\boldsymbol{\theta}\|_2^2) \\ &\leq K(1 + \|\boldsymbol{\theta}\|_2^2),\end{aligned}$$

where $K \geq K_1 + K_2 + |||\boldsymbol{D}|||^2|||\boldsymbol{M}|||/(2\gamma)$. Moreover, under **(H.3)**, $({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top$ is locally Lipschitz continuous by virtue of Lemma 3.2, which applies thanks to assumptions **(H.1)**-**(H.2)**. Clearly $({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top \in \widetilde{C^{1,+}}(\mathbb{R}^p)$. Since $\widetilde{C^{1,+}}(\mathbb{R}^p)$ is closed under addition, we conclude the proof. □

Proof of Proposition 4.2 Claim (i) follows by combining Proposition 4.1 and [69, Theorem 3.6, Chapter II]. Claim (ii) is a consequence of Proposition 4.1 and [54, Theorem 2.1]. □

Proof of Theorem 4.1 Owing to Proposition 4.1 and [69, Theorem 4.1, Chapter II], we get that the r-th moments of $\boldsymbol{L}(t)$ are bounded for any $r \geq 2$ and $t \geq 0$. A similar reasoning also entails that the r-th moments of the continuous-time extension $\boldsymbol{L}^\delta$ are also bounded. Moreover, according to Proposition 4.1, the drift $\boldsymbol{\psi}$ is locally Lipschitz continuous. The claim then follows from [38, Theorem 2.2] and Jensen's inequality. In the globally Lipschitz continuous case, we get the claimed rate by putting together Lemma 2.1, Jensen's inequality and [69, Theorem 7.3, Chapter II] or [40, Theorem 10.2.2 and Remark 10.2.3]. □

Proof of Lemma 5.2 The proof of Lemma 5.3 is based on the one of [55, Proposition 13.37] and generalizes to the proximal mapping in metric $\boldsymbol{M}$ for any $\boldsymbol{M} \in \mathbb{R}^{p\times p}$ symmetric positive definite.

Without loss of generality, we prove the claim on a neighbourhood of $\bar{\boldsymbol{x}}$ 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 [7, Lemma 4.1], there exist $\epsilon > 0$ and $\lambda_0 > 0$ such that

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

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

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

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

Since H is lsc, proper and prox-bounded, from [55, Theorem 1.17(c)] (see also [55, Theorem 1.25]), we have

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

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 $\boldsymbol{u} \in \mathcal{N}_{\gamma,\bar{\boldsymbol{x}},\bar{\boldsymbol{v}}}$, $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2 \leq \epsilon$, $\|H(\boldsymbol{x}) - H(\bar{\boldsymbol{x}})\|_2 \leq \epsilon$ and $\|\boldsymbol{v} - \bar{\boldsymbol{v}}\|_2 \leq \epsilon$. We get then

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

So that

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

where $\delta = 1/\gamma - 1/\gamma_0$, $S = \mathrm{T}^{H}_{\epsilon,\bar{\boldsymbol{x}},\bar{\boldsymbol{v}}} + 1/\gamma_0 \boldsymbol{M}$. From (9.1), S is maximal monotone, the latter operator is well defined as a single valued operator (see [5, 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

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

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 5.3. □

Proof of Lemma 5.3 From [55, Example 12.28(b)], ∂H is hypomonotone of modulus $\frac{1}{r}$. In turn $S = \partial H + \frac{1}{\gamma_0}\boldsymbol{M} = \partial\Big(H + \frac{1}{2\gamma_0}\|\cdot\|_{\boldsymbol{M}}^2\Big)$ is monotone with $\gamma_0 = r\sigma_{\min}(\boldsymbol{M})$, or equivalently that $H + \frac{1}{2\gamma_0}\|\cdot\|_{\boldsymbol{M}}^2$ is convex [55, Example 12.28(b)]. Let $\delta = \frac{1}{\gamma} - \frac{1}{\gamma_0}$ and $W(\boldsymbol{w}, \boldsymbol{\theta}) = H(\boldsymbol{w}) + \frac{r'}{2}\|\boldsymbol{w} - \boldsymbol{\theta}\|_{\boldsymbol{M}}^2$. Thus

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

$W(\cdot, \boldsymbol{\theta})$ is a convex function on $\mathbb{R}^p$ and $\delta > 0$ as $\gamma < \gamma_0$. Altogether, this entails that $W(\cdot, \boldsymbol{\theta}) + \frac{\delta}{2}\|\cdot - \boldsymbol{\theta}\|_{\boldsymbol{M}}^2$ is strongly convex uniformly in $\boldsymbol{\theta}$ and γ complying with $\gamma < \gamma_0$. It then follows that $\operatorname{prox}_{\gamma H}^{\boldsymbol{M}}$ is single-valued. We have

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

By Fermat's rule, we then get

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

and the latter operator is well-defined as a single-valued operator since S is maximal monotone; see [5, Proposition 3.22 (ii)(d)]. Let $\boldsymbol{p} = \mathrm{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{\theta})$ and $\boldsymbol{p}' = \mathrm{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{\theta}')$. It then follows that

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

and monotonicity of S yields

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

Using Cauchy-Schwartz inequality, we then obtain

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

where $\kappa^{-1} = \frac{\gamma\delta\sigma_{\min}(\boldsymbol{M})}{\|\boldsymbol{M}\|} = \frac{\sigma_{\min}(\boldsymbol{M})}{\|\boldsymbol{M}\|}\Big(1 - \frac{\gamma}{\gamma_0}\Big) = \frac{\sigma_{\min}(\boldsymbol{M})}{\|\boldsymbol{M}\|}\Big(1 - \frac{\gamma}{r\sigma_{\min}(\boldsymbol{M})}\Big)$. That concludes the proof of Lemma 5.3. □

Proof of Lemma 6.1 We have

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

By the Pythagoras relation, we then get

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

We conclude the proof of Lemma 6.1. □

Proof of Lemma 7.1 This is a probably known result, for which we provide a simple proof. Since $W_{\boldsymbol{\lambda}}$ is separable and $w_{\boldsymbol{\lambda}}$ is continuous and lower-bounded, we have

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

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

$$\Big[\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{u})\Big]_{\mathcal{G}_l} = \underset{\boldsymbol{v}\in\mathbb{R}^G}{\mathrm{Argmin}}\ \frac{1}{2}\left\|\boldsymbol{v} - \boldsymbol{u}_{\mathcal{G}_l}\right\|_2^2 + \frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}(\|\boldsymbol{v}\|_2). \tag{9.4}$$

If $\boldsymbol{u}_{\mathcal{G}_l} = 0$, then as $w_{\boldsymbol{\lambda}}$ is an increasing function, $\left[\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{u})\right]_{\mathcal{G}_l} = 0$. For $\boldsymbol{u}_{\mathcal{G}_l} \neq 0$, by isotropy of problem (9.4), we can write

$$\min_{\boldsymbol{v}\in\mathbb{R}^G} \frac{1}{2}\left\|\boldsymbol{v}-\boldsymbol{u}_{\mathcal{G}_l}\right\|_2^2 + \frac{\gamma}{\beta} w_{\boldsymbol{\lambda}}(\|\boldsymbol{v}\|_2) = \min_{t\geq 0} \frac{\gamma}{\beta} w_{\boldsymbol{\lambda}}(t) + \left(\min_{\|\boldsymbol{v}\|_2=t} \frac{1}{2}\left\|\boldsymbol{v}-\boldsymbol{u}_{\mathcal{G}_l}\right\|_2^2\right). \tag{9.5}$$

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 (9.5) and rearranging the terms, (9.4) becomes

$$\left[\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{u})\right]_{\mathcal{G}_l} = \frac{\boldsymbol{u}_{\mathcal{G}_l}}{\left\|\boldsymbol{u}_{\mathcal{G}_l}\right\|_2} \operatorname*{Argmin}_{t\geq 0} \frac{1}{2}\left(t-\left\|\boldsymbol{u}_{\mathcal{G}_l}\right\|_2\right)^2 + \frac{\gamma}{\beta} w_{\boldsymbol{\lambda}}(t) = \frac{\boldsymbol{u}_{\mathcal{G}_l}}{\left\|\boldsymbol{u}_{\mathcal{G}_l}\right\|_2} \mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(\left\|\boldsymbol{u}_{\mathcal{G}_l}\right\|_2),$$

where we used even-symmetry of $w_{\boldsymbol{\lambda}}$. □

Proof of Lemma 7.3 Before proceeding, let us discuss about the term $\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}$. In view of **(W.2)**, ${w_{\boldsymbol{\lambda}}}'/\beta$ is positive on $]0,+\infty[$. According to Lemma 7.2 we get that, for any $t \geq 0$, $\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t) = 0$ if $t \leq \frac{\gamma}{\beta}{w_{\boldsymbol{\lambda}}}'(0)$ and $\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t) = t - \frac{\gamma}{\beta}{w_{\boldsymbol{\lambda}}}'(\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t)) \leq t$ otherwise. Hencem for any $t \geq 0$,

$$0 \leq \mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t) \leq t, \quad \forall t \geq 0. \tag{9.6}$$

Set $\boldsymbol{u} = \boldsymbol{D}^\top\boldsymbol{\theta}$, from Lemma 7.1 and (9.6), we get that

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

According to the fact that $\|\boldsymbol{u}\|_2^2 = \left\|\boldsymbol{D}^\top\boldsymbol{\theta}\right\|_2^2 \leq |||\boldsymbol{D}|||^2 \|\boldsymbol{\theta}\|_2^2$, Assumption **(H.4'-SFB)** holds.

Set $\boldsymbol{v} = 2\gamma\boldsymbol{X}^\top\boldsymbol{y}/\beta$ and $\boldsymbol{t}_{\boldsymbol{\theta}} = \boldsymbol{\theta} - \gamma\nabla F_\beta(\boldsymbol{\theta}) = \boldsymbol{M}_\gamma\boldsymbol{\theta} + \boldsymbol{v}$, by Young's inequality, we obtain that

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

Moreover, owing to Lemma 7.1 and (9.6), we get that

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

Thus, Assumption **(H.4'-FB)** holds and we conclude the proof of Lemma 7.3. □

Proof of Lemma 7.4

(i) Observe that $w_{\boldsymbol{\lambda}}$ is continuously differentiable on $]0,+\infty[$ with

$$w_{\boldsymbol{\lambda}}{}'(t) = \kappa\lambda\left(I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda)\right) \geq 0,$$

$w_{\boldsymbol{\lambda}}$ is then non decreasing and bounded from below by $w_{\boldsymbol{\lambda}}(0) = 0$ on $]0,+\infty[$. Thus, $w_{\boldsymbol{\lambda}}$ satisfies **(W.1)** and **(W.2)**. Let us check **(W.3)**. Let $u(t) = t + \kappa w_{\boldsymbol{\lambda}}{}'(t)$, we obtain that

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

Thus, $t = 0$ is the unique mimimum in $[0,+\infty[$ of $t + p'_{\lambda}(t)$. In other words, $w_{\boldsymbol{\lambda}}$ satisfies **(W.3)**.

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

$$p = \begin{cases} 0 & \text{if } t \leq \kappa\lambda, \\ t - \kappa\lambda\left(I(p \leq \lambda) + \frac{(a\lambda - p)_+}{(a-1)\lambda} I(p > \lambda)\right) & \text{otherwise.} \end{cases} \tag{9.7}$$

From (9.7), we get the following assertions when $t > \kappa\lambda$,

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

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

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

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

Proof of Lemma 7.5 Set $\boldsymbol{u} = \boldsymbol{D}^\top \boldsymbol{\theta}$, $\alpha = \gamma\lambda/\beta$ and $\boldsymbol{p}_{\boldsymbol{u}} = \mathrm{P}_{\left\{\boldsymbol{x} \,:\, \alpha \sum_l \|\boldsymbol{x}_{\mathcal{G}_l}\|_2 \leq 1\right\}}(\boldsymbol{u})$. Owing to (7.8) and Young's inequality, we obtain that

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

According to the fact that $\|\boldsymbol{u}\|_2^2 = \left\|\boldsymbol{D}^\top \boldsymbol{\theta}\right\|_2^2 \leq |||\boldsymbol{D}|||^2 \|\boldsymbol{\theta}\|_2^2$, **(H.4'-SFB)** holds.

Set $\boldsymbol{v} = 2\gamma \boldsymbol{X}^\top \boldsymbol{y}/\beta$, $\boldsymbol{t}_{\boldsymbol{\theta}} = \boldsymbol{\theta} - \gamma \nabla F_\beta(\boldsymbol{\theta}) = \boldsymbol{M}_\gamma \boldsymbol{\theta} + \boldsymbol{v}$ and $\boldsymbol{p}_{\boldsymbol{t}_{\boldsymbol{\theta}}} = \mathrm{P}_{\left\{\boldsymbol{x} \,:\, \alpha \sum_l \|\boldsymbol{x}_{\mathcal{G}_l}\|_2 \leq 1\right\}}(\boldsymbol{t}_{\boldsymbol{\theta}})$. By Young's inequality, we obtain that

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

Moreover, owing to (7.8), we get that

$$\left\|\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{t}_{\boldsymbol{\theta}})\right\|_2^2 = \left\|\boldsymbol{t}_{\boldsymbol{\theta}} - \boldsymbol{p}_{\boldsymbol{t}_{\boldsymbol{\theta}}}\right\|_2^2 \leq 2\|\boldsymbol{t}_{\boldsymbol{\theta}}\|_2^2 + 2\left\|\boldsymbol{p}_{\boldsymbol{t}_{\boldsymbol{\theta}}}\right\|_2^2 \leq 4|||\boldsymbol{M}_\gamma|||^2 \|\boldsymbol{\theta}\|_2^2 + \left(4\|\boldsymbol{v}\|_2^2 + \frac{2}{\alpha^2}\right).$$

Thus, Assumption **(H.4'-FB)** holds and we conclude the proof of Lemma 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] A. Antoniadis and J. Fan. Regularization of Wavelet Approximations. *Journal of the American Statistical Association*, 96:939–967, 2001.

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

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

[5] 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.

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

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

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

[9] 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.

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

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

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

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

[14] 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.

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

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

[17] 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.

[18] 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.

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

[20] X. Chen, Q. Lin, S. Kim, J. G. Carbonell, and E. P. Xing. An efficient proximal-gradient method for general structured sparse learning. *Preprint arXiv:1005.4717*, 2010.

[21] C. Chesneau and M. Hebiri. Some theoretical results on the grouped variables lasso. *Mathematical Methods of Statistics*, 17(4):317–326, 2008.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] A. Durmus and E. Moulines. Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm. Preprint hal-01176132, July 2015.

[29] A. Durmus, E. Moulines, and M. Pereyra. Sampling from convex non continuously differentiable functions, when Moreau meets Langevin. Preprint hal-01267115, Feb. 2016.

[30] 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.

[31] J. Fadili and G. Peyré. Total variation projection with first order schemes. *IEEE Transactions on Image Processing*, 20(3):657–669, 2011.

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

[33] 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.

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

[35] H.-Y. Gao and A. Bruce. Waveshrink with firm shrinkage. *Statist. Sinica*, 7:855– 874, 1997.

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

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

[38] 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.

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

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

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

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

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

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

[45] 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.

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

[47] 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.

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

[49] G. Peyré, J. Fadili, and C. Chesneau. Group sparsity with overlapping partition functions. In *EUSIPCO*, Barcelona, Spain, Aug. 2011.

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

[51] R. A. Poliquin, R. T. Rockafellar, and L. Thibault. Local differentiability of distance functions. *Transactions of the American mathematical Society*, 352:5231–5249, 2000.

[52] 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.

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

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

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

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

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

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

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

[60] 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.

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

[62] 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.

[63] S. Vaiter, M. Golbabaee, M. J. Fadili, and G. Peyré. Model selection with low complexity priors. *Information and Inference: A Journal of the IMA (IMAIAI)*, 2015.

[64] 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.

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

[66] 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.

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

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

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

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

[71] 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.