Noname manuscript No.
(will be inserted by the editor)

Sampling from non-smooth distributions through Langevin diffusion

Tung Duy Luu · Jalal Fadili · Christophe Chesneau

the date of receipt and acceptance should be inserted later

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 penalties that are commonly used to promote some notion of simplicity/complexity. Some popular penalties are detailed and implemented on some numerical experiments.

Keywords Langevin diffusion · Monte-Carlo · Non-smooth distributions · Proximal splitting · Exponentially Weighted aggregation

1 Introduction

1.1 Problem statement

We consider the linear regression problem

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_0 + \boldsymbol{\zeta}, \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of observations, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the design matrix, $\boldsymbol{\zeta}$ 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) is either under-determined ($p > n$), or determined ($p = n$) but $\boldsymbol{X}$ is ill-conditioned. In both cases, (1) is ill-posed.

The idea of aggregating elements in a dictionary has been introduced in machine learning to combine different techniques (see (Vovk, 1990; Littlestone and Warmuth, 1994)) with some procedures such as bagging (Breiman, 1996), boosting (Freund, 1995; Schapire, 1990) and random forests (Amit and Geman, 1997; Breiman, 2001; Biau et al, 2008; Biau and Devroye, 2010; Genuer, 2010; Biau, 2012). In the recent years, there has been a flurry of research on the use of low-complexity regularization/penalties (among which sparsity and low-rank are the most popular) in various areas including statistics and machine learning in high dimension.The idea is to promote vectors $\boldsymbol{\theta}_0$ that conform to some notion of simplicity. 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, two families of estimators have been considered in this context: Penalized Estimators and Exponentially Weighted Aggregates (EWA).

Tung Duy Luu and Jalal Fadili
Normandie University, ENSICAEN, UNICAEN, CNRS, GREYC, France,
E-mail: duy-tung.luu@ensicaen.fr, Jalal.Fadili@greyc.ensicaen.fr

Christophe Chesneau
Normandie University, UNICAEN, CNRS, LMNO, France. E-mail: christophe.chesneau@unicaen.fr

1.2 Variational/Penalized Estimators

This class of estimators are obtained by solving the optimization problem

$$\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}} \in \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\mathrm{Argmin}} \Big\{ V(\boldsymbol{\theta}) \stackrel{\text{def}}{=} F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) + J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \Big\}, \tag{2}$$

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 (2) is the Lasso (Chen et al, 1999; Tibshirani, 1996; Osborne et al, 2000; Donoho, 2006; Candès and Plan, 2009; Bickel et al, 2009; Bühlmann and van de Geer, 2011) and its variants such the analysis/fused Lasso (Rudin et al, 1992a; Tibshirani et al, 2005) or group Lasso (Bakin, 1999; Yuan and Lin, 2006; Bach, 2008; Wei and Huang, 2010; Chesneau and Hebiri, 2008). 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 (Recht et al, 2010; Candès and Recht, 2009; Fazel et al, 2001; Candès et al, 2013). See (Negahban et al, 2012; Bühlmann and van de Geer, 2011; van de Geer, 2014; Vaiter et al, 2015b) for generalizations and comprehensive reviews.

1.3 Exponential Weighted Aggregation (EWA)

An alternative to the variational estimator (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 a probability distribution supported on $\Theta \subset \mathbb{R}^p$, having the density with respect to the Lebesgue measure

$$\widehat{\mu}(\boldsymbol{\theta}) = \frac{\exp\left(-V(\boldsymbol{\theta})/\beta\right)}{\int_{\Theta} \exp\left(-V(\boldsymbol{\xi})/\beta\right) d\boldsymbol{\xi}}, \tag{3}$$

where $\beta > 0$ is the temperature parameter, and V defined in (2) is supposed to be a measurable function such that $\int_{\Theta} \exp\left(-V(\boldsymbol{\xi})/\beta\right) d\boldsymbol{\xi} < +\infty$. Typically V should grow sufficiently fast for the latter to hold (this will be made precise later). 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

$$\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}} = \int_{\mathbb{R}^p} \boldsymbol{\theta} \widehat{\mu}(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{4}$$

Aggregation by exponential weighting has been widely considered in the statistical and machine learning literatures, see e.g. (Dalalyan and Tsybakov, 2007, 2008, 2009, 2012; Nemirovski, 2000; Yang, 2004; Rigollet and Tsybakov, 2007; Lecué, 2007; Guedj and Alquier, 2013; Duy Luu et al, 2016) 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 (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 (Dalalyan and Tsybakov, 2012), we here use the Langevin diffusion process; see (Roberts and Tweedie, 1996). Note that there are other Monte Carlo approaches to compute estimators such as $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$, see for example the recent survey paper (Pereyra et al, 2016) and references therein.

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)

$$d\boldsymbol{L}(t) = \frac{1}{2}\boldsymbol{\rho}(\boldsymbol{L}(t))dt + d\boldsymbol{W}(t),\ t > 0,\ \boldsymbol{L}(0) = \boldsymbol{l}_0, \tag{5}$$

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 (5) has a unique strong solution and, $\boldsymbol{L}(t)$ has a stationary distribution with density precisely μ (Roberts and Tweedie, 1996, 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 (Xuerong, 2007).

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

$$\frac{\delta}{T}\sum_{k=0}^{\lfloor T/\delta \rfloor - 1} \boldsymbol{L}_k, \tag{6}$$

where $\lfloor T/\delta \rfloor$ denotes the integer part of T/δ. It is then natural to approximate $\widehat{\boldsymbol{\theta}}_n^{\mathrm{EWA}}$ 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. (Dalalyan, 2014).

1.5 Contributions and relation to prior work

Our main goal in this paper is to propose a provably consistent estimator of EWA by efficiently sampling from a distribution with the density $\widehat{\mu}$ in (3), where V is given in (2), and the latter is not necessarily smooth nor convex. In (Pereyra, 2016; Durmus et al, 2016), the authors proposed proximal-type algorithms to sample from non-smooth log-concave densities μ using the forward Euler discretization applied to a smooth version of μ involving the Moreau-Yosida regularization/envelope; see Definition 2. In (Pereyra, 2016), $-\log\mu$ is replaced with its Moreau envelope. However, the author applied it to problems where $-\log\mu = L + H$ assuming the Moreau envelope of this sum is available. 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. He then suggested an approximation reminiscent of the forward-backward splitting strategy that we propose, albeit without consistency guarantees. In (Durmus et al, 2016), 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. The authors then derived non-asymptotic bounds on the mixing time of the Markov chain $\boldsymbol{L}_k$ in total variation with a markedly different dependence of these bounds on the dimension. Proximal steps within MCMC methods have been proposed for some simple (convex) signal processing problems (Chaari et al, 2014), though without any guarantees.

In all these works, however, convexity is of paramount importance, for instance to get non-asymptotic bounds with polynomial dependence on the dimension. We here propose to cope with both the lack of smoothness and convexity at the same time, which allows to cover distributions that are beyond the current state of the art as covered in (Dalalyan and Tsybakov, 2012; Pereyra, 2016; Durmus and Moulines, 2015; Durmus et al, 2016). One of our key tools is the Moreau-Yosida regularization/envelope, but extended to the non-convex setting, which necessitate to invoke arguments from variational analysis. We first show in Proposition 1 that under mild assumptions, the smoothed distribution is well-defined and converges in total variation to the distribution μ. We also show in Proposition 3 that the Langevin diffusion based on the smoothed density is well-posed. We then turn to discretizing such an SDE. We describe two approaches in Section 4 that yield two fast and easy to implement algorithms that are reminiscent of the forward-backward proximal splitting popular in non-smooth optimization. For these algorithms, we prove in Theorem 1 theoretical consistency guarantees by showing convergence of the ergodic average to the EWA. However, given that we do not assume not even log-concavity, proving non-asymptotic bounds on convergence of the distribution of $\boldsymbol{L}_k$ to its stationary distribution as in (Durmus et al, 2016), is far more challenging. We believe it is an important direction to pursue that we leave to a future work. We finally exemplify our proposed algorithms to compute EWA estimators with several popular penalties in the literature, and illustrate their performance 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 (5) with such a 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 For a d-dimensional Euclidean space $\mathbb{R}^d$, we endow it with its usual inner product $\langle \cdot, \cdot \rangle$ and associated norm $\|\cdot\|_2$. $\mathbf{I}_d$ is the identity matrix on $\mathbb{R}^d$. For $r \geq 1$, $\|\cdot\|_r$ 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 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 $\liminf_{\boldsymbol{x} \to \boldsymbol{x}_0} f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0)$. A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is level-coercive if it is bounded below on bounded sets and satisfies

$$\liminf_{\|\boldsymbol{x}\|_2 \to +\infty} \frac{f(\boldsymbol{x})}{\|\boldsymbol{x}\|_2} > 0.$$

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 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 (Rockafellar and Wets, 1998).

Definition 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

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

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

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

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 $\partial^F f(\boldsymbol{x}) \subset \partial f(\boldsymbol{x})$ (Rockafellar and Wets, 1998, 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 (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

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

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

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

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

Definition 3 (Hypomonotone and monotone operators) A set-valued operator $S : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is hypomonotone of modulus $r > 0$ if

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

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

3 Moreau-Yosida regularization

In our framework, the target density μ is defined as

$$\mu(\boldsymbol{\theta}) = Z^{-1} \exp\Big(-\Big(L(\boldsymbol{\theta}) + H \circ \boldsymbol{D}^\top(\boldsymbol{\theta})\Big)\Big), \tag{7}$$

where $L \in \widetilde{C^{1,+}}(\mathbb{R}^p)$, $\boldsymbol{D} \in \mathbb{R}^{p \times q}$ and $H : \mathbb{R}^q \to \mathbb{R}$, and $Z = \int_{\mathbb{R}^p} \exp\Big(-(L(\boldsymbol{\xi}) + H \circ \boldsymbol{D}^\top(\boldsymbol{\xi}))\Big) d\boldsymbol{\xi}$ is the partition function.

Moreover, H is assumed neither differentiable nor convex. To overcome these difficulties, we invoke arguments from variational analysis (Rockafellar and Wets, 1998). 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 $\mathrm{Argmin}(H) \neq \emptyset$.

(**H.1**) $H : \mathbb{R}^q \to \mathbb{R}$ is lsc and bounded from below.
(**H.2**) $\mathrm{prox}_{\gamma H}^{\boldsymbol{M}}$ is single valued.
(**H.3**) Either one of the following holds:
- (a) L is bounded below, H is level-coercive and $\boldsymbol{D}$ is surjective.
- (b) $L + H \circ \boldsymbol{D}^\top$ is level-coercive and H is Lipschitz continuous.

As we will see shortly in Proposition 1, assumption (**H.3**) is crucial to ensure that $Z < +\infty$ and that the densities μ and μ_γ (see (8)) are well-defined.

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

Lemma 2 *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 $\gamma \mapsto \big\|\boldsymbol{\theta}\big\|_{\boldsymbol{M}_\gamma}$, $\forall \boldsymbol{\theta} \in \mathbb{R}^q$, is a decreasing mapping on $]0, \gamma_0[$. Assume that* (**H.1**) *holds.*

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

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

(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 $\Big({}^{\boldsymbol{M}_\gamma, \gamma}H(\boldsymbol{x})\Big)_{\gamma \in]0, \gamma_0[}$ 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 *Let $\boldsymbol{M}_\gamma \in \mathbb{R}^{q \times q}$ symmetric positive definite, assume that* (**H.1**) *and* (**H.2**) *hold. Then $\mathrm{prox}_{\gamma H}^{\boldsymbol{M}_\gamma}$ 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*

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

In plain words, Lemma 3 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 $\mathrm{prox}_{\gamma H}^{\boldsymbol{M}_\gamma}$ is equivalent to a gradient descent on the Moreau envelope of H in the metric $\boldsymbol{M}_\gamma$ with step-size γ.

Remark 1 When the metric matrix does not depend on γ, Lemmas 2 and 3 hold with $\gamma_0 = +\infty$.

Let us now define the smoothed density

$$\mu_\gamma(\boldsymbol{\theta}) = Z_\gamma^{-1} \exp\Big(-\Big(L(\boldsymbol{\theta}) + ({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top(\boldsymbol{\theta})\Big)\Big), \tag{8}$$

where

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

The following proposition answers the natural question on the behaviour of $\mu_\gamma - \mu$ as a function of γ. Recall that for two probability measures on $\mathbb{R}^d$ which have densities μ and ν with respect to the Lebesgue measure, the total variation between them is defined as

$$\big\|\mu - \nu\big\|_{\mathrm{TV}} = \int_{\mathbb{R}^d} \big|\mu(\boldsymbol{\theta}) - \nu(\boldsymbol{\theta})\big| d\boldsymbol{\theta}.$$

Proposition 1 *Let $\gamma > 0$ and $\boldsymbol{M}$ be symmetric positive definite with $\sigma_{\min}(\boldsymbol{M}) > 0$ uniformly in γ. Assume that* (**H.1**) *and* (**H.3**) *hold. Then, $Z, Z_\gamma < +\infty$ uniformly in γ, and $\big\|\mu_\gamma - \mu\big\|_{\mathrm{TV}} \to 0$ 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

$$\begin{aligned} &d\boldsymbol{L}(t) = \boldsymbol{\psi}(\boldsymbol{L}(t))dt + d\boldsymbol{W}(t),\ t > 0, \\ &\text{where} \quad \boldsymbol{\psi} : \boldsymbol{\theta} \in \mathbb{R}^p \mapsto -\frac{1}{2}\nabla\big(L + ({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top\big)(\boldsymbol{\theta}), \end{aligned} \tag{9}$$

$\boldsymbol{\psi}$ 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 computation of $\nabla^{\boldsymbol{M},\gamma}H$ by exploiting the relation between $\nabla^{\boldsymbol{M},\gamma}H$ and $\mathrm{prox}_{\gamma H}^{\boldsymbol{M}}$ as stated in Lemma 3. Now, to guarantee well-posedness (existence and uniqueness) and discretization consistency of the SDE (9), we will also need the following assumptions.

(**H.4**) $\mathrm{prox}_{\gamma H}^{\boldsymbol{M}}$ is locally Lipschitz continuous.

(**H.5**) There exists $C > 0$ such that $\Big\langle \boldsymbol{D}^\top\boldsymbol{\theta}, \mathrm{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{D}^\top\boldsymbol{\theta})\Big\rangle_{\boldsymbol{M}} \leq C(1 + \|\boldsymbol{\theta}\|_2^2), \forall \boldsymbol{\theta} \in \mathbb{R}^p$.

4.1 Well-posedness

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

Proposition 2 *Assume that* (**H.1**)*,* (**H.2**)*,* (**H.4**) *and* (**H.5**) *hold. Then,*

$$\langle \boldsymbol{\psi}(\boldsymbol{\theta}), \boldsymbol{\theta} \rangle \leq K(1 + \|\boldsymbol{\theta}\|_2^2), \text{ for some } K > 0, \tag{10}$$

and

$$\boldsymbol{\psi} \text{ is locally Lipschitz continuous.} \tag{11}$$

The following proposition guarantees the well-posedness of the SDE (9).

Proposition 3 *Assume that* (**H.1**)-(**H.5**) *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 (9) *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 (8)*.*

4.2 Discretization

4.2.1 Approach 1

Inserting the identities of Lemma 3 into (9), we get the SDE

$$d\boldsymbol{L}(t) = -\frac{1}{2}\Big(\nabla L + \gamma^{-1}\boldsymbol{D}\boldsymbol{M}\Big(\mathbf{I}_q - \mathrm{prox}_{\gamma H}^{\boldsymbol{M}}\Big) \circ \boldsymbol{D}^\top\Big)(\boldsymbol{L}(t))dt + d\boldsymbol{W}(t),\ \boldsymbol{L}(0) = \boldsymbol{l}_0,\ t > 0. \tag{12}$$

Consider now the forward Euler discretization of (12) with step-size $\delta > 0$, which can be rearranged as

$$\boldsymbol{L}_{k+1} = \boldsymbol{L}_k - \frac{\delta}{2}\nabla L(\boldsymbol{L}_k) - \frac{\delta}{2\gamma}\boldsymbol{D}\boldsymbol{M}\Big(\boldsymbol{D}^\top \boldsymbol{L}_k - \mathrm{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{D}^\top \boldsymbol{L}_k)\Big) + \sqrt{\delta}\boldsymbol{Z}_k,\ t > 0,\ \boldsymbol{L}_0 = \boldsymbol{l}_0. \tag{13}$$

Note that by Lemma 3, and without the stochastic term $\sqrt{\delta}\boldsymbol{Z}_k$, (13) 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 (13), an Euler approximate solution is defined as

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

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 (13) and of its first-order moment can be established as follows.

Theorem 1 *Assume that* (**H.1**)-(**H.5**) *hold and* $\mathbb{E}\left[\|\boldsymbol{L}(0)\|_2^r\right] < \infty$ *for any* $r \geq 2$. *Then*

$$\Big\|\mathbb{E}\big[\boldsymbol{L}^\delta(T)\big] - \mathbb{E}\left[\boldsymbol{L}(T)\right]\Big\|_2 \leq \mathbb{E}\left[\sup_{0\leq t\leq T}\Big\|\boldsymbol{L}^\delta(t) - \boldsymbol{L}(t)\Big\|_2\right] \underset{\delta\to 0}{\longrightarrow} 0. \tag{14}$$

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

4.2.2 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 $\boldsymbol{M}_\gamma$. We assume that $\boldsymbol{M}_\gamma$ is symmetric positive definite for any $\gamma \in]0, \gamma_0[$ with $\sigma_{\min}(\boldsymbol{M}) > 0$ uniformly in γ, that for each $\boldsymbol{\theta} \in \mathbb{R}^q$, the mapping $\gamma \mapsto \|\boldsymbol{\theta}\|_{\boldsymbol{M}_\gamma}$ is decreasing on $]0, \gamma_0[$, and that $\boldsymbol{M}_\gamma \underset{\gamma\to 0}{\to} \mathbf{I}_q$ (such a choice is motivated by the scheme described in Section 6.1). One can consider an alternative version of the SDE (9), i.e.,

$$d\boldsymbol{L}(t) = -\frac{1}{2}\nabla\Big(\big(L + ({}^{\boldsymbol{M}_\gamma,\gamma}H) \circ \boldsymbol{D}^\top\big) \circ {\boldsymbol{M}_\gamma}^{-1/2}\Big)(\boldsymbol{L}(t))dt + {\boldsymbol{M}_\gamma}^{1/2}d\boldsymbol{W}(t),\ t > 0. \tag{15}$$

Denote the drift coefficient of (15) 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 (10) and (11) under assumptions (**H.1**), (**H.2**), (**H.4**) and (**H.5**). Moreover, arguing exactly as in the proof of the first part of Proposition 1, (**H.3**) and that $\det(\boldsymbol{M}_\gamma) \underset{\gamma\to 0}{\to} 1$ allow to show that

$$\boldsymbol{\theta} \mapsto Z_\gamma^{-1}\exp\Big(-\Big(L + ({}^{\boldsymbol{M}_\gamma,\gamma}H) \circ \boldsymbol{D}^\top\Big) \circ {\boldsymbol{M}_\gamma}^{-1/2}(\boldsymbol{\theta})\Big), \tag{16}$$

is a well-defined uniformly in γ, where $Z_\gamma = \sqrt{\det(\boldsymbol{M}_\gamma)}\int_{\mathbb{R}^p}\exp\Big(-\Big(L + ({}^{\boldsymbol{M}_\gamma,\gamma}H) \circ \boldsymbol{D}^\top\Big)(\boldsymbol{\xi})\Big)d\boldsymbol{\xi} < +\infty$. In turn, Proposition 3 applies to (15) to show that the diffusion $\boldsymbol{L}$ is unique, non explosive and admits an unique invariant measure whose density is precisely (16). In addition, applying again the same reasoning as in the proof of the last part of Proposition 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

$$d\boldsymbol{U}(t) = -\frac{1}{2}{\boldsymbol{M}_\gamma}^{-1}\nabla\Big(L + ({}^{\boldsymbol{M}_\gamma,\gamma}H) \circ \boldsymbol{D}^\top\Big)(\boldsymbol{U}(t))dt + d\boldsymbol{W}(t),\ t > 0. \tag{17}$$

In an analogous way to (13), the forward Euler discretization of (17) 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 (15) and (17) and of their first-order moments can be established exactly in the same way as in Theorem 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.4**).

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 4 (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 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 (Rockafellar and Wets, 1998, Chapter 13, Section F) and (Poliquin et al, 2000)):

(i) Proper lsc convex functions.
(ii) Proper lsc lower-C^2 (or semi-convex) functions, i.e., f is such that $f + \frac{1}{2r}\left\|\cdot\right\|_2^2$ 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 4 ((Poliquin and Rockafellar, 1996, 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

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

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.4**).

Lemma 5 *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 1-(ii), satisfy the global counterpart of Lemma 4-(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.5**) according to Lemma 1.

Lemma 6 *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 $\frac{|||\boldsymbol{M}|||}{\sigma_{\min}(\boldsymbol{M})}\left(1 - \frac{\gamma}{r\sigma_{\min}(\boldsymbol{M})}\right)^{-1}$. In turn,* (14) *holds with the optimal rate $\delta^{1/2}$.*

6 Forward-Backward type LMC algorithms

Let us now deal with our main goal: computing the EWA estimator in (4) by sampling from $\widehat{\mu}$. Recall that

$$\widehat{\mu}(\boldsymbol{\theta}) = Z^{-1} \exp\left(-\frac{F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) + J_{\boldsymbol{\lambda}}(\boldsymbol{\theta})}{\beta}\right),$$

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, and $Z = \int_{\mathbb{R}^p} \exp\left(-\frac{F(\boldsymbol{X}\boldsymbol{\xi}, \boldsymbol{y}) + J_{\boldsymbol{\lambda}}(\boldsymbol{\xi})}{\beta}\right) d\boldsymbol{\xi}$. 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'**) Either one of the following holds:
  - (a) F is bounded below, $W_{\boldsymbol{\lambda}}$ is level-coercive and $\boldsymbol{D}$ is surjective.
  - (b) $F(\boldsymbol{X}\cdot, \boldsymbol{y}) + J_{\boldsymbol{\lambda}}$ is level-coercive and $W_{\boldsymbol{\lambda}}$ is Lipschitz continuous.
- (**H.4'**) $\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}}$ is locally Lipschitz continuous.

These assumptions are specializations of those in Section 3 to the density $\widehat{\mu}$. In particular, assumption (**H.3'**) is instrumental to ensure that $\widehat{\mu}$ and its smoothed version are well-defined. Note that (**H.3'**) is also known to ensure existence of the variational/penalized estimator $\widehat{\boldsymbol{\theta}}_n^{\mathrm{PEN}}$.

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

6.1 Forward-backward LMC (FBLMC)

In (7), 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., $F_\beta(\boldsymbol{\theta}) = \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right\|_2^2/\beta$. Observe that H satisfies (**H.1**) owing to assumption (**H.1'**). Also (**H.3'**) implies that (**H.3**) holds. In the particular, (**H.3'**)(a) is true if $W_{\boldsymbol{\lambda}}$ is level-coercive (since F_β is non-negative), which in turn implies (**H.3'**)(b). The converse is, however, not true. To check (**H.2**), (**H.4**) and (**H.5**), 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 7 *Assume that* (**H.1'**) *holds and* $0 < \gamma \leq \beta/(2|||\boldsymbol{X}|||^2)(1-\delta)$ *with* $\delta]0,1[$. *Define* $\boldsymbol{M}_\gamma \overset{\text{def}}{=} \mathbf{I}_p - (2\gamma/\beta)\boldsymbol{X}^\top\boldsymbol{X}$. *Then* $\boldsymbol{M}_\gamma$ *is symmetric positive definite with* $\sigma_{\min}(\boldsymbol{M}_\gamma) \geq \delta > 0$. *Moreover,*

$$\mathrm{prox}_{\gamma H}^{\boldsymbol{M}_\gamma} = \mathrm{prox}_{\gamma J_{\boldsymbol{\lambda}}/\beta} \circ \big(\mathbf{I}_p - \gamma \nabla F_\beta\big). \tag{18}$$

In view of Lemma 18, (**H.2'**) and (**H.4'**), it is immediate to check that (**H.2**) and (**H.4**) are satisfied.

It remains now to verify (**H.5**) which is fulfilled by imposing the following assumption on $W_{\boldsymbol{\lambda}}$ (or $J_{\boldsymbol{\lambda}}$).

(**H.5'-FB**) There exists $C'_{\mathrm{FB}} > 0$ such that

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

By Lemma 1, a sufficient condition for (**H.5'-FB**) to hold is that the proximal mapping of $W_{\boldsymbol{\lambda}}$ is Lipschitz continuous.

From Lemmas 3 and 7, 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 (9), specialized to the current case, reads

$$\boldsymbol{L}_{k+1} = \boldsymbol{L}_k - \frac{\delta}{2\gamma}\boldsymbol{M}_\gamma\Big(\boldsymbol{L}_k - \mathrm{prox}_{\gamma J_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{L}_k - \gamma\nabla F_\beta(\boldsymbol{L}_k))\Big) + \sqrt{\delta}\boldsymbol{Z}_k,\ t > 0,\ \boldsymbol{L}_0 = \boldsymbol{l}_0. \tag{19}$$

Similarly, the forward Euler discretization of the SDE (17) is given by

$$\boldsymbol{U}_{k+1} = (1 - \tfrac{\delta}{2\gamma})\boldsymbol{U}_k + \frac{\delta}{2\gamma}\mathrm{prox}_{\gamma J_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{U}_k - \gamma\nabla F_\beta(\boldsymbol{U}_k)) + \sqrt{\delta}\boldsymbol{Z}_k,\ t > 0,\ \boldsymbol{U}_0 = \boldsymbol{l}_0. \tag{20}$$

The familiar reader may have recognized that the deterministic part of (20) is nothing but the relaxed form of the so-called Forward-Backward proximal splitting algorithm (Bauschke and Combettes, 2011). 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 (19) 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 (Bauschke and Combettes, 2011). Our goal now is to overcome this difficulty.

Toward this goal, in (7), consider now $L = F_\beta$, $H = W_{\boldsymbol{\lambda}}/\beta$ and $\boldsymbol{M} = \mathbf{I}_q$. Owing to (**H.1'**)-(**H.4'**), one can check that (**H.1**)-(**H.4**) are fulfilled. Assumption (**H.5**) is verified by imposing the following assumption on $W_{\boldsymbol{\lambda}}$.

(**H.5'-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, 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 (9) now reads

$$\boldsymbol{L}_{k+1} = \boldsymbol{L}_k - \frac{\delta}{2}\nabla F_\beta(\boldsymbol{L}_k) - \frac{\delta}{2\gamma}\boldsymbol{D}\Big(\boldsymbol{D}^\top\boldsymbol{L}_k - \mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{D}^\top\boldsymbol{L}_k)\Big) + \sqrt{\delta}\boldsymbol{Z}_k,\ t > 0,\ \boldsymbol{L}_0 = \boldsymbol{l}_0. \tag{21}$$

In the case where $\boldsymbol{D} = \mathbf{I}_p$, F_β and $W_{\boldsymbol{\lambda}}$ are convex, we recover the scheme studied in (Durmus et al, 2016).

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.4'**), (**H.5'-FB**) and (**H.5'-SFB**) hold, and to compute $\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}$. In turn, this allows to apply our algorithms (20) and (21) 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.,

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = W_{\boldsymbol{\lambda}}(\boldsymbol{D}^\top\boldsymbol{\theta}) \quad \text{where} \quad W_{\boldsymbol{\lambda}}(\boldsymbol{u}) = \textstyle\sum_{l=1}^L w_{\boldsymbol{\lambda}}\big(\|\boldsymbol{u}_{\mathcal{G}_l}\|_2\big), \tag{22}$$

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

Remark 2 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 (Peyré et al, 2011; Chen et al, 2010).

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 a non-decreasing function on $]0,+\infty[$.

Lemma 8 *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*

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

Our aim is now to design a family of penalties that will allow to establish (**H.1'**)-(**H.4'**), (**H.5'-FB**) and (**H.5'-SFB**), while involving a form of shrinkage which is ubiquitous in low-complexity regularization. Inspired by the work of (Antoniadis and Fan, 2001), 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 indeed a convenient shrinkage-type form.

Lemma 9 ((Antoniadis and Fan, 2001, Theorem 1)) *Assume that* (**W.1**)*,* (**W.2**) *and* (**W.3**) *hold for some* $\gamma > 0$*. Then,* $\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}$ *is a single-valued continuous mapping on* $\mathbb{R}$*, and satisfies, for* $t \in [0, +\infty[$*,*

$$\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t) = \begin{cases} 0 & \text{if } t \leq \frac{\gamma}{\beta} {w_{\boldsymbol{\lambda}}}'(0^+), \\ t - \frac{\gamma}{\beta} {w_{\boldsymbol{\lambda}}}'\Big(\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t)\Big) & \text{if } t > \frac{\gamma}{\beta} {w_{\boldsymbol{\lambda}}}'(0^+). \end{cases} \tag{23}$$

Let us turn to check our assumptions. (**H.1'**), (**H.2'**) and (**H.4'**) are fulfilled thanks to (**W.1**), (**W.2**) and (**W.3**). To comply with (**H.3'**), it is sufficient to impose that:

(**W.4**) Either one of the following holds:
(a) F is bounded below, $w_{\boldsymbol{\lambda}}$ is level-coercive on $]0, +\infty[$, and $\boldsymbol{D}$ is surjective.
(b) $F(\cdot, \boldsymbol{y})$ is level-coercive, $w_{\boldsymbol{\lambda}}$ is level-coercive on $]0, +\infty[$ and $\ker(\boldsymbol{X}) \cap \ker(\boldsymbol{D}^\top) = \{0\}$.

Sufficiency of the first condition is immediate. For the second, the argument is standard. It is easy to see that by level-coercivity, we indeed have the existence of of $a > 0$ and $b \in \mathbb{R}$ such that for all $\boldsymbol{\theta}$ outside $\ker(\boldsymbol{X}) \cap \ker(\boldsymbol{D}^\top)$,

$$F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) + J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \geq a\big\|\boldsymbol{\theta}\big\|_2 + b.$$

It remains to check (**H.5'-FB**) and (**H.5'-SFB**). This is the subject of the following lemma.

Lemma 10 *Assume that* (**W.1**)*,* (**W.2**) *and* (**W.3**) *hold for some* $\gamma > 0$*, then* (**H.5'-FB**) *and* (**H.5'-SFB**) *also hold.*

We now discuss some popular penalties $w_{\boldsymbol{\lambda}}$ that satisfy (**W.1**)-(**W.4**).

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.,

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

The ℓ_1 penalty is obviously level-coercive and thus (**W.4**) is verified if either F is bounded below and $\boldsymbol{D}$ is surjective, or F is level-coercive and $\ker(\boldsymbol{X}) \cap \ker(\boldsymbol{D}^\top) = \{0\}$.

FIRM penalty The FIRM penalty is given by (Gao and Bruce, 1997)

$$w_{\boldsymbol{\lambda}}(t) = \begin{cases} \lambda\Big(t - \frac{t^2}{2\mu}\Big) & \text{if } 0 \leq t \leq \mu, \\ \frac{\lambda\mu}{2} & \text{if } t > \mu. \end{cases} \tag{24}$$

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 (Woodworth and Chartrand, 2015, Definition II.3). Its formula is defined as

$$\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t) = \begin{cases} 0 & \text{if } 0 \leq t \leq \alpha, \\ \frac{\mu}{\mu - \alpha}(t - \alpha) & \text{if } \alpha < t \leq \mu, \\ t & \text{if } t > \mu, \end{cases} \tag{25}$$

where $\alpha = \gamma\lambda/\beta$. The formula (25) can also be found using Lemma 9. Observe that the FIRM shrinkage (25) interpolates between hard- (see (Woodworth and Chartrand, 2015, Definition II.2)) and soft-thresholding. In particular, (25) coincides with soft-thresholding when $\mu \to \infty$.

SCAD penalty The SCAD penalty, proposed in (Fan and Li, 2001) is parameterized by $\boldsymbol{\lambda} = (\lambda, a) \in]0, +\infty[\times]2, +\infty[$ as

$$w_{\boldsymbol{\lambda}}(t) = \begin{cases} \lambda t & \text{if } 0 \leq t \leq \lambda, \\ -\frac{t^2 - 2a\lambda t + \lambda^2}{2(a-1)} & \text{if } \lambda < t \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } t > a\lambda, \end{cases} \tag{26}$$

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

Lemma 11 *Let $w_{\boldsymbol{\lambda}}$ defined in* (26)*, 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

$$\mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t) = \begin{cases} (t - \kappa\lambda)_+ & \text{if } 0 \leq t \leq (\kappa+1)\lambda, \\ \frac{(a-1)t - ka\lambda}{a-1-\kappa} & \text{if } (\kappa+1)\lambda < t \leq a\lambda, \\ t & \text{if } t > a\lambda. \end{cases} \tag{27}$$

Since $a > 2$, one can set $\kappa = 1$. In this case, (27) specializes to (Fan and Li, 2001, 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 (Jégou et al, 2012; Lyubarskii and Vershynin, 2010; Studer et al, 2012). It entails the following penalty $W_{\boldsymbol{\lambda}}$:

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = W_{\boldsymbol{\lambda}}(\boldsymbol{D}^\top \boldsymbol{\theta}) \quad \text{where} \quad W_{\boldsymbol{\lambda}}(\boldsymbol{u}) = \lambda \max_{l \in \{1,\dots,L\}} \left\{ \left\| [\boldsymbol{u}]_{\mathcal{G}_l} \right\|_2 \right\}, \tag{28}$$

where $\boldsymbol{\lambda} = \lambda > 0$. Since $W_{\boldsymbol{\lambda}}$ is not separable, Lemma 8 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.,

$$\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{u}) = \boldsymbol{u} - \mathrm{P}_{\left\{ \boldsymbol{x} \,:\, \sum_l \left\| \boldsymbol{x}_{\mathcal{G}_l} \right\|_2 \leq \frac{\beta}{\lambda\gamma} \right\}}(\boldsymbol{u}). \tag{29}$$

This projector can be obtained from (Fadili and Peyré, 2011, Proposition 2) (see also references therein). One can see that (**H.1'**), (**H.2'**) and (**H.4'**) hold. Since $W_{\boldsymbol{\lambda}}$ is level-coercive, (**H.3**) can be fulfilled under the same assumptions as for the ℓ_1 norm discussed before. We report the verification of (**H.5'-FB**) and (**H.5'-SFB**) in the proof of the following lemma.

Lemma 12 *Let $W_{\boldsymbol{\lambda}}$ in* (28)*. Then* (**H.5'-FB**) *and* (**H.5'-SFB**) *hold.*

8 Numerical experiments

In this section, some numerical experiments are conducted to illustrate and validate our LMC algorithms. Following the philosophy of reproducible research, all the code implementing our sampling algorithms and reproducing the experiments of this paper are made publicly available for download at `https://github.com/luuduytung/LMCToolbox`.

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 $\mathrm{vec} : \mathbb{R}^{\sqrt{p}\times\sqrt{p}} \to \mathbb{R}^p$ the vectorization operator, i.e. the operator which stacks the columns of its arguments. We then consider the following linear regression problem

$$\boldsymbol{y} = \boldsymbol{X}\,\mathrm{vec}(\boldsymbol{\theta}_0) + \boldsymbol{\zeta}. \tag{30}$$

Here $p = 128^2$ and $\boldsymbol{\zeta} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. The noise level σ is chosen according to the simulated $\boldsymbol{\theta}_0$ through the signal-to-noise ratio SNR, i.e. $\sigma = n^{-1/2} \left\| \boldsymbol{X}\boldsymbol{\theta}_0 \right\|_2 / 10^{\mathrm{SNR}/10}$. 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 (Rudin et al, 1992b). 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

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \sum_{l=1}^{p} w_{\boldsymbol{\lambda}}\Big(\sqrt{\operatorname{vec}(\boldsymbol{D}_r(\boldsymbol{\theta}))_l^2 + \operatorname{vec}(\boldsymbol{D}_c(\boldsymbol{\theta}))_l^2} \Big) = W_{\boldsymbol{\lambda}}(\boldsymbol{D}_{\mathrm{TV}} \boldsymbol{\theta}), \tag{31}$$

which clearly has the form (22) with p blocks of equal size 2.

To estimate $\boldsymbol{\theta}_0$ from (30), we employ the EWA estimator (4) with $F(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{X} \operatorname{vec}(\boldsymbol{\theta})\|_2^2$ and $J_{\boldsymbol{\lambda}}$ in (31). For each problem instance (compressed sensing, inpainting or deconvolution), we tested $w_{\boldsymbol{\lambda}}$ as the ℓ_1, SCAD and FIRM penalties. Observe that (**W.4**) holds in this setting for ℓ_1 as soon as $\ker(\boldsymbol{X})$ does not contain constants. 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 (21) to compute EWA with $\beta = 1/(pn)$, $\gamma = \beta$, and $\delta = \big\{5\beta/10^3, 5\beta/10^2, 5\beta/10^6\big\}$ 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 (30) where $\boldsymbol{X}$ is drawn from the Rademacher ensemble with $n > p$ [1]. 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. All required assumptions are again verified in this setting, including (**W.4**) as $J_{\boldsymbol{\lambda}}$ is level-coercive. We can then use the FBLMC scheme (20), where we choose $\beta = 1/(pn)$, $\gamma = \beta$, and $\delta = 5/10^2$. The results are shown in Figure 2.

9 Proofs

Proof of Lemma 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

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

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

[1] The overdetermined regime is known to yield good performance for the ℓ_∞ penalty (Vaiter et al, 2015a).

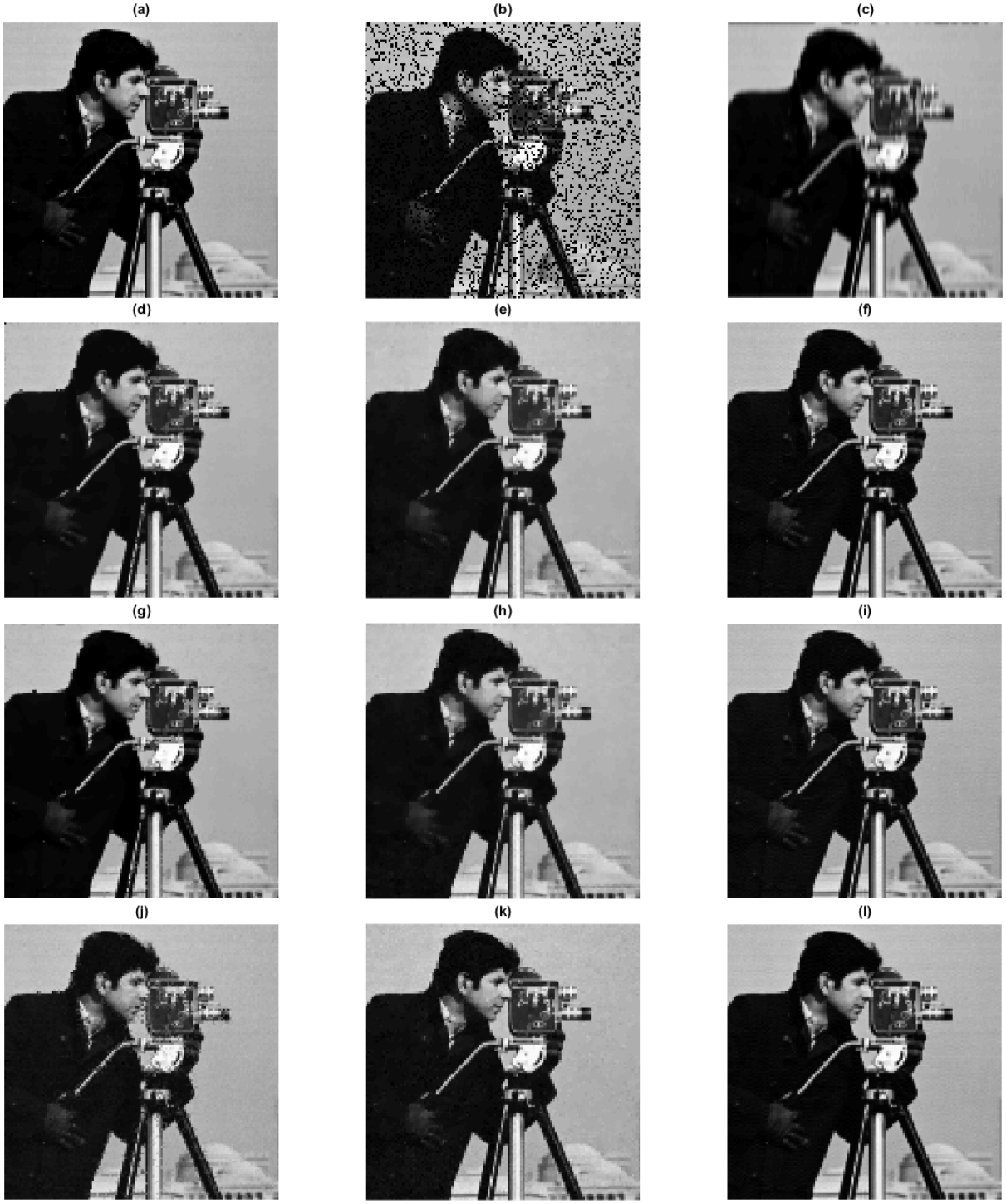

Fig. 1 (a): Original image. (b,c) Observed masked and blurry images. (d, e, f): EWA-ℓ_1 estimated images from masked image, compressed sensing measurements, and blurry image. (g, h, i): EWA-FIRM estimated images from masked image, compressed sensing measurements, and blurry image. (j, k, l): EWA-SCAD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

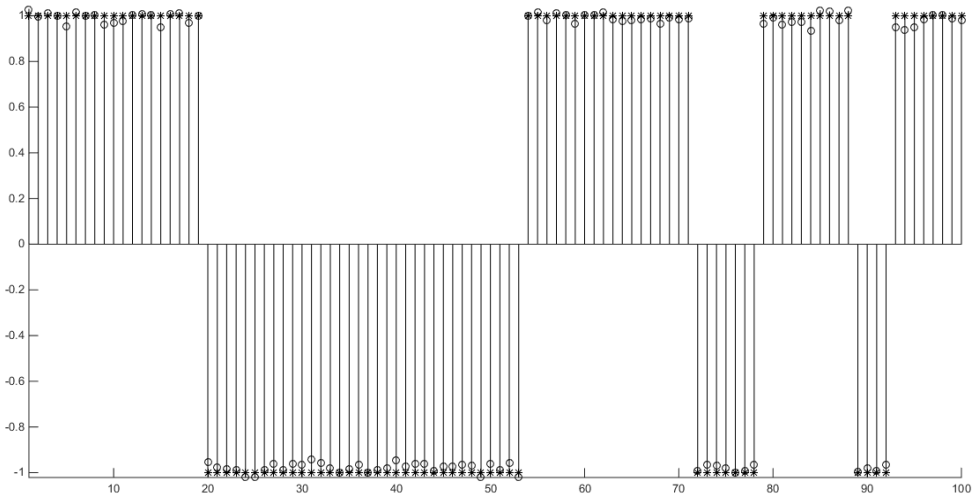

Fig. 2 Compressed sensing with EWA using the ℓ_∞ penalty. $'*'$ is the original signal and $'\circ'$ is the the estimated one.

Proof of Lemma 2

(i) In view of (**H.1**), H is prox-bounded by (Rockafellar and Wets, 1998, Exercise 1.24) for any $\gamma \in]0, \gamma_0[$, and then for any $\boldsymbol{x}$, $\frac{1}{2\gamma}\big\|\boldsymbol{x} - \cdot\big\|^2_{\boldsymbol{M}_\gamma} + H$ 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. $\operatorname{prox}^{\boldsymbol{M}_\gamma}_{\gamma H}(\boldsymbol{x})$, is a non-empty compact set. For the last claim, suppose that $\boldsymbol{x} \in \operatorname{Argmin}(H) \neq \emptyset$ and bounded but $\boldsymbol{x} \notin \operatorname{prox}^{\boldsymbol{M}_\gamma}_{\gamma H}(\boldsymbol{x})$. Thus, for any $\boldsymbol{p} \in \operatorname{prox}^{\boldsymbol{M}_\gamma}_{\gamma H}(\boldsymbol{x})$, we have $\boldsymbol{p} \neq \boldsymbol{x}$ and

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

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

(ii) The continuity and finiteness properties follow from (Rockafellar and Wets, 1998, Theorem 1.17(c)) (see also(Rockafellar and Wets, 1998, 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 $\boldsymbol{p} \in \operatorname{prox}^{\boldsymbol{M}_\gamma}_{\gamma H}(\boldsymbol{x})$. Then, $\forall \delta > \gamma$,

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

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

□

Proof of Lemma 3 By virtue of Lemma 2-(i) and (**H.2**), $\operatorname{prox}^{\boldsymbol{M}_\gamma}_{\gamma H}$ is clearly non-empty and single valued. The continuity property follows from (Rockafellar and Wets, 1998, Theorem 1.17(b)) (see also (Rockafellar and Wets, 1998, Theorem 1.25)) and single-valuedness. By Lemma 2-(ii), ${}^{\boldsymbol{M}_\gamma,\gamma}H(\boldsymbol{\theta})$ is finite. Since (**H.1**) holds, H is prox-bounded with threshold ∞ by (Rockafellar and Wets, 1998, Exercise 1.24). Invoking (Rockafellar and Wets, 1998, Example 10.32), we get that $-{}^{\boldsymbol{M}_\gamma,\gamma}H$ is locally Lipschitz continuous, subdifferentially regular and

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

Combining this with (Rockafellar and Wets, 1998, 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 1 In view of (Rockafellar and Wets, 1998, Theorem 3.26(a)), assumption (**H.3**)(a) entails that there exists $a > 0$ and $b \in \mathbb{R}$ such that for all $\boldsymbol{\theta} \in \mathbb{R}^p$

$$L(\boldsymbol{\theta}) + H \circ \boldsymbol{D}^\top(\boldsymbol{\theta}) \geq a\big\|\boldsymbol{D}^\top\boldsymbol{\theta}\big\|_2 + b \geq a\sigma_{\min}(\boldsymbol{D}^\top)\,\|\boldsymbol{\theta}\|_2 + b,$$

where $\sigma_{\min}(\boldsymbol{D}^\top) > 0$ by injectivity. Thus,

$$Z \leq e^{-b}\int_{\mathbb{R}^p} \exp\Big(-a\sigma_{\min}(\boldsymbol{D}^\top)\,\|\boldsymbol{\theta}\|_2\Big)d\boldsymbol{\theta} < +\infty.$$

In addition,

$$\begin{aligned}
L(\boldsymbol{\theta}) + ({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top(\boldsymbol{\theta}) &= L(\boldsymbol{\theta}) + \min_{\boldsymbol{w}\in\mathbb{R}^q}\left\{\frac{1}{2\gamma}\big\|\boldsymbol{w} - \boldsymbol{D}^\top\boldsymbol{\theta}\big\|_{\boldsymbol{M}}^2 + H(\boldsymbol{w})\right\} \\
&\geq b + \min_{\boldsymbol{w}\in\mathbb{R}^q}\left\{\frac{1}{2\gamma}\big\|\boldsymbol{w} - \boldsymbol{D}^\top\boldsymbol{\theta}\big\|_{\boldsymbol{M}}^2 + a\big\|\boldsymbol{w}\big\|_2\right\} \\
&\geq b + \min_{\boldsymbol{w}\in\mathbb{R}^q}\left\{\frac{\sigma_{\min}(\boldsymbol{M})}{2\gamma}\big\|\boldsymbol{w} - \boldsymbol{D}^\top\boldsymbol{\theta}\big\|_2^2 + a\big\|\boldsymbol{w}\big\|_2\right\}.
\end{aligned}$$

The solution to the above minimization problem is the well-known soft-thresholding operator

$$\boldsymbol{w}^\star = \boldsymbol{D}^\top\boldsymbol{\theta}\left(1 - \frac{a\gamma}{\sigma_{\min}(\boldsymbol{M})\big\|\boldsymbol{D}^\top\boldsymbol{\theta}\big\|_2}\right)_+.$$

Replacing in the above inequality, we get

$$\begin{aligned}
L(\boldsymbol{\theta}) + ({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top(\boldsymbol{\theta}) &\geq b + \begin{cases} \frac{\sigma_{\min}(\boldsymbol{M})}{2\gamma}\big\|\boldsymbol{D}^\top\boldsymbol{\theta}\big\|_2^2 & \big\|\boldsymbol{D}^\top\boldsymbol{\theta}\big\|_2 \leq \frac{a\gamma}{\sigma_{\min}(\boldsymbol{M})} \\ a\big\|\boldsymbol{D}^\top\boldsymbol{\theta}\big\|_2 - \frac{a^2\gamma}{\sigma_{\min}(\boldsymbol{M})} & \text{otherwise.} \end{cases} \\
&\geq b + \begin{cases} \frac{\sigma_{\min}(\boldsymbol{D}^\top)^2\sigma_{\min}(\boldsymbol{M})}{2\gamma}\big\|\boldsymbol{\theta}\big\|_2^2 & \big\|\boldsymbol{D}^\top\boldsymbol{\theta}\big\|_2 \leq \frac{a\gamma}{\sigma_{\min}(\boldsymbol{M})} \\ a\sigma_{\min}(\boldsymbol{D}^\top)\big\|\boldsymbol{\theta}\big\|_2 - \frac{a^2\gamma}{\sigma_{\min}(\boldsymbol{M})} & \text{otherwise.} \end{cases}
\end{aligned}$$

Hence,

$$\liminf_{\|\boldsymbol{\theta}\|_2\to+\infty} \frac{L(\boldsymbol{\theta}) + ({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top(\boldsymbol{\theta})}{\|\boldsymbol{\theta}\|_2} = a\sigma_{\min}(\boldsymbol{D}^\top) > 0,$$

or equivalently, that $L + ({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top$ is level-coercive uniformly in γ and $\boldsymbol{M}$. Arguing as for Z, we then get that $Z_\gamma < +\infty$ uniformly in γ.

Let us consider now the case of assumption (**H.3**)(b). This assumption is equivalent to the existence of $a > 0$ and $b \in \mathbb{R}$ (possibly different from those above) such that, for all $\boldsymbol{\theta}$

$$L(\boldsymbol{\theta}) + H \circ \boldsymbol{D}^\top(\boldsymbol{\theta}) \geq a\big\|\boldsymbol{\theta}\big\|_2 + b.$$

We then have $Z < +\infty$. It remains to show that $Z_\gamma < +\infty$ in this case. As H is β-Lipschitz continuity, we get

$$\begin{aligned}
L(\boldsymbol{\theta}) + ({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top(\boldsymbol{\theta}) &= L(\boldsymbol{\theta}) + \inf_{\boldsymbol{w}\in\mathbb{R}^q}\left\{\frac{1}{2\gamma}\big\|\boldsymbol{w} - \boldsymbol{D}^\top\boldsymbol{\theta}\big\|_{\boldsymbol{M}}^2 + H(\boldsymbol{w})\right\} \\
&\geq L(\boldsymbol{\theta}) + H \circ \boldsymbol{D}^\top(\boldsymbol{\theta}) + \min_{\boldsymbol{w}\in\mathbb{R}^q}\left\{\frac{1}{2\gamma}\big\|\boldsymbol{w} - \boldsymbol{D}^\top\boldsymbol{\theta}\big\|_{\boldsymbol{M}}^2 + (H(\boldsymbol{w}) - H(\boldsymbol{D}^\top\boldsymbol{\theta}))\right\} \\
&\geq L(\boldsymbol{\theta}) + H \circ \boldsymbol{D}^\top(\boldsymbol{\theta}) + \min_{\boldsymbol{w}\in\mathbb{R}^q}\left\{\frac{1}{2\gamma}\big\|\boldsymbol{w} - \boldsymbol{D}^\top\boldsymbol{\theta}\big\|_{\boldsymbol{M}}^2 - \beta\big\|\boldsymbol{w} - \boldsymbol{D}^\top\boldsymbol{\theta}\big\|_2\right\} \\
&\geq L(\boldsymbol{\theta}) + H \circ \boldsymbol{D}^\top(\boldsymbol{\theta}) + \min_{\boldsymbol{w}\in\mathbb{R}^q}\left\{\frac{1}{2\gamma}\big\|\boldsymbol{w} - \boldsymbol{D}^\top\boldsymbol{\theta}\big\|_{\boldsymbol{M}}^2 - \frac{\beta}{\sqrt{\sigma_{\min}(\boldsymbol{M})}}\big\|\boldsymbol{w} - \boldsymbol{D}^\top\boldsymbol{\theta}\big\|_{\boldsymbol{M}}\right\} \\
&= L(\boldsymbol{\theta}) + H \circ \boldsymbol{D}^\top(\boldsymbol{\theta}) + \min_{t\geq 0,\big\|\boldsymbol{w}-\boldsymbol{D}^\top\boldsymbol{\theta}\big\|_{\boldsymbol{M}}=t}\left\{\frac{1}{2\gamma}t^2 - \frac{\beta}{\sqrt{\sigma_{\min}(\boldsymbol{M})}}t\right\} \\
&= L(\boldsymbol{\theta}) + H \circ \boldsymbol{D}^\top(\boldsymbol{\theta}) + \min_{t\geq 0}\left\{\frac{1}{2\gamma}t^2 - \frac{\beta}{\sqrt{\sigma_{\min}(\boldsymbol{M})}}t\right\}.
\end{aligned}$$

The minimization problem has a unique solution $u^\star = \gamma\beta\sigma_{\min}(\boldsymbol{M})^{-1/2}$, and thus

$$L(\boldsymbol{\theta}) + ({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top(\boldsymbol{\theta}) \geq L(\boldsymbol{\theta}) + H \circ \boldsymbol{D}^\top(\boldsymbol{\theta}) - \gamma\beta^2\sigma_{\min}(\boldsymbol{M})/2.$$

Thus level-coercivity of $L + H \circ \boldsymbol{D}^\top$ transfers to $L + ({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top$ whence $Z_\gamma < +\infty$ follows immediately.

Overall, we have shown that both μ and μ_γ are well-defined uniformly in γ under assumption (**H.3**) via the fact that $Z < +\infty$ and there exists $a > 0$ and $b \in \mathbb{R}$ (a is independent of γ and $\boldsymbol{M}$) such that

$$\exp\Big(-(L(\boldsymbol{\theta}) + ({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top(\boldsymbol{\theta}))\Big) \leq e^b \exp\left(-a\big\|\boldsymbol{\theta}\big\|_2\right).$$

This means that the function $e^- \circ (L + ({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top)$ is dominated by an integrable function. This together with the pointwise convergence provided by Lemma 2(ii), allow to apply the dominated convergence theorem to conclude that $Z_\gamma \to Z$ as $\gamma \to 0$. Combining this with Lemma 2(ii) again yields that μ_γ converges to μ pointwise. We conclude using Scheffé(-Riesz) theorem (Scheffe, 1947; Kusolitsch, 2010). □

Proof of Proposition 2 In view of Lemma 3, 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.5**) holds, there exist $K_1 > 0$ and $K_2 > 0$ such that

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

where $K \geq K_1 + K_2 + |||\boldsymbol{D}|||^2|||\boldsymbol{M}|||/(2\gamma)$. Moreover, under (**H.4**), $({}^{\boldsymbol{M},\gamma}H) \circ \boldsymbol{D}^\top$ is locally Lipschitz continuous by virtue of Lemma 3, 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 3 First observe that by Proposition 1, μ_γ is well-defined for any γ under (**H.3**). Claim (i) follows by combining Proposition 2 and (Xuerong, 2007, Theorem 3.6, Chapter II). Claim (ii) is a consequence of Proposition 2 and (Roberts and Tweedie, 1996, Theorem 2.1). □

Proof of Theorem 1 Again, μ_γ is well-defined for any γ thanks to Proposition 1. Thus by virtue of Proposition 2 and (Xuerong, 2007, 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 2, the drift $\boldsymbol{\psi}$ is locally Lipschitz continuous. The claim then follows from (Higham et al, 2003, Theorem 2.2) and Jensen's inequality. In the globally Lipschitz continuous case, we get the claimed rate by putting together Lemma 1, Jensen's inequality and (Xuerong, 2007, Theorem 7.3, Chapter II) or (Kloeden and Platen, 1995, Theorem 10.2.2 and Remark 10.2.3). □

Proof of Lemma 5 The proof of Lemma 5 is based on the one of (Rockafellar and Wets, 1998, 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 (Bernard and Thibault, 2005, Lemma 4.1), there exist $\epsilon > 0$ and $\lambda_0 > 0$ such that

$$\begin{aligned}
H(\boldsymbol{x}') &> H(\boldsymbol{x}) + \big\langle \boldsymbol{v}, \boldsymbol{x}' - \boldsymbol{x}\big\rangle - \frac{1}{2\lambda_0}\big\|\boldsymbol{x}' - \boldsymbol{x}\big\|_2^2 \\
&> H(\boldsymbol{x}) + \big\langle \boldsymbol{v}, \boldsymbol{x}' - \boldsymbol{x}\big\rangle - \frac{1}{2\lambda_0\sigma_{\min}(\boldsymbol{M})}\big\|\boldsymbol{x}' - \boldsymbol{x}\big\|_{\boldsymbol{M}}^2, \qquad (32)
\end{aligned}$$

for any $\boldsymbol{x}' \neq \boldsymbol{x}$ and $(\boldsymbol{x}, \boldsymbol{v}) \in \mathrm{gph}\,\mathrm{T}_{\epsilon,\bar{\boldsymbol{x}},\bar{\boldsymbol{v}}}^H$. Let $\gamma_0 = \lambda_0\sigma_{\min}(\boldsymbol{M})$, $\gamma \in]0, \gamma_0[$ and $\boldsymbol{u} = \boldsymbol{x} + \gamma\boldsymbol{M}^{-1}\boldsymbol{v}$, (32) becomes

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

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

Since H is lsc, proper and prox-bounded, from (Rockafellar and Wets, 1998, Theorem 1.17(c)) (see also (Rockafellar and Wets, 1998, Theorem 1.25)), we have

$$\boldsymbol{x} \in \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{u}), \boldsymbol{u} \to \bar{\boldsymbol{x}} + \gamma \boldsymbol{M}^{-1}\bar{\boldsymbol{v}} \implies \begin{cases} \boldsymbol{x} \to \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\bar{\boldsymbol{x}} + \gamma \boldsymbol{M}^{-1}\bar{\boldsymbol{v}}) = \bar{\boldsymbol{x}}, \\ H(\boldsymbol{x}) = {}^{\boldsymbol{M},\gamma}H(\boldsymbol{u}) - \frac{1}{2\gamma}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 \to H(\bar{\boldsymbol{x}}). \end{cases} \tag{33}$$

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

$$\boldsymbol{v} = \frac{\boldsymbol{M}}{\gamma}(\boldsymbol{u} - \boldsymbol{x}) \in \partial H(\boldsymbol{x}). \tag{34}$$

For any $\gamma \in]0, \gamma_0[$, owing to (33) and (34), 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

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

So that

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

where $\delta = 1/\gamma - 1/\gamma_0$, $S = \mathrm{T}_{\epsilon,\bar{\boldsymbol{x}},\bar{\boldsymbol{v}}}^{H} + 1/\gamma_0 \boldsymbol{M}$. From (32), S is maximal monotone, the latter operator is well defined as a single valued operator (see (Bauschke et al, 2003, Proposition 3.22 (ii)(d))). Let $\boldsymbol{p} = \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{x})$ and $\boldsymbol{p}' = \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{x}')$. It then follows that

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

and monotonicity of S yields

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

Using Cauchy-Schwarz's inequality, we obtain

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

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

Let us note that when γ decrease, Inequality (32) 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. □

Proof of Lemma 6 From (Rockafellar and Wets, 1998, 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 (Rockafellar and Wets, 1998, Example 12.28(b)). Let $\delta = \frac{1}{\gamma} - \frac{1}{\gamma_0}$ and $W(\boldsymbol{w}, \boldsymbol{\theta}) = H(\boldsymbol{w}) + \frac{r'}{2}\left\|\boldsymbol{w} - \boldsymbol{\theta}\right\|_{\boldsymbol{M}}^2$. Thus

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

$W(\cdot, \boldsymbol{\theta})$ is a convex function on $\mathbb{R}^p$ and $\delta > 0$ as $\gamma < \gamma_0$. Altogether, this entails that $W(\cdot, \boldsymbol{\theta}) + \frac{\delta}{2}\left\|\cdot - \boldsymbol{\theta}\right\|_{\boldsymbol{M}}^2$ 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 (Bauschke et al, 2003, Proposition 3.22 (ii)(d)). Let $\boldsymbol{p} = \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{\theta})$ and $\boldsymbol{p}' = \operatorname{prox}_{\gamma H}^{\boldsymbol{M}}(\boldsymbol{\theta}')$. 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 6.

Since $\mathrm{prox}_{\gamma H}^{\boldsymbol{M}}$ is globally Lipschitz continuous, the optimal convergence rate in (14) is of order $\delta^{1/2}$ in view of Theorem 1. □

Proof of Lemma 7 The fact that $\boldsymbol{M}_\gamma$ is symmetric definite positive with a spectrum bounded below by δ is immediate. We now have

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

By the Pythagoras relation, we then get

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

We conclude the proof of Lemma 7. □

Proof of Lemma 8 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

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

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

$$\Big[\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{u})\Big]_{\mathcal{G}_l} = \operatorname*{Argmin}_{\boldsymbol{v}\in\mathbb{R}^G} \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{u}_{\mathcal{G}_l}\|_2^2 + \frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}(\|\boldsymbol{v}\|_2). \tag{35}$$

If $\boldsymbol{u}_{\mathcal{G}_l} = 0$, then as $w_{\boldsymbol{\lambda}}$ is an increasing function, $\Big[\mathrm{prox}_{\gamma W_{\boldsymbol{\lambda}}/\beta}(\boldsymbol{u})\Big]_{\mathcal{G}_l} = 0$. For $\boldsymbol{u}_{\mathcal{G}_l} \neq 0$, by isotropy of problem (35), we can write

$$\min_{\boldsymbol{v}\in\mathbb{R}^G} \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{u}_{\mathcal{G}_l}\|_2^2 + \frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}(\|\boldsymbol{v}\|_2) = \min_{t\geq 0} \frac{\gamma}{\beta}w_{\boldsymbol{\lambda}}(t) + \left(\min_{\|\boldsymbol{v}\|_2 = t} \frac{1}{2}\|\boldsymbol{v} - \boldsymbol{u}_{\mathcal{G}_l}\|_2^2\right). \tag{36}$$

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 $\boldsymbol{v} = t\frac{\boldsymbol{u}_{\mathcal{G}_l}}{\|\boldsymbol{u}_{\mathcal{G}_l}\|_2}$ since $\boldsymbol{u}_{\mathcal{G}_l} \neq 0$. Inserting this into (36) and rearranging the terms, (35) becomes

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

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

Proof of Lemma 10 In view of (**W.2**), ${w_{\boldsymbol{\lambda}}}'/\beta$ is positive on $]0,+\infty[$. According to Lemma 9 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. Hence for any $t \geq 0$,

$$0 \leq \mathrm{prox}_{\gamma w_{\boldsymbol{\lambda}}/\beta}(t) \leq t, \quad \forall t \geq 0. \tag{37}$$

Set $\boldsymbol{u} = \boldsymbol{D}^\top \boldsymbol{\theta}$, from Lemma 8 and (37), we get that

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

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.5'-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

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

Moreover, owing to Lemma 8 and (37), we get that

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

Thus, assumption (**H.5'-FB**) holds which concludes the proof. □

Proof of Lemma 11

(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 9, we obtain that

$$p = \begin{cases} 0 & \text{if } t \leq \kappa\lambda, \\ t - \kappa\lambda\Big(I(p \leq \lambda) + \frac{(a\lambda - p)_+}{(a-1)\lambda} I(p > \lambda)\Big) & \text{otherwise.} \end{cases} \tag{38}$$

From (38), 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 11. □

Proof of Lemma 12 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 (29) 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.5'-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 (29), we get that

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

Thus, Assumption (**H.5'-FB**) holds and we conclude the proof of Lemma 12. □

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

References

Amit Y, Geman D (1997) Shape quantization and recognition with randomized trees. Neural Comput 9(7):1545–1588, DOI 10.1162/neco.1997.9.7.1545, URL http://dx.doi.org/10.1162/neco.1997.9.7.1545

Antoniadis A, Fan J (2001) Regularization of Wavelet Approximations. Journal of the American Statistical Association 96:939–967, URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.8.6694

Bach F (2008) Consistency of the group lasso and multiple kernel learning. Journal of Machine Learning Research 9:1179–1225

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

Bauschke HH, Combettes PL (2011) Convex analysis and monotone operator theory in Hilbert spaces. Springer

Bauschke HH, Borwein JM, Combettes PL (2003) Bregman monotone optimization algorithms. SIAM Journal on Control and Optimization 42(2):596–636

Bernard F, Thibault L (2005) Prox-regular functions in hilbert spaces. Journal of Mathematical Analysis and Applications 303(1):1 – 14, DOI http://dx.doi.org/10.1016/j.jmaa.2004.06.003, URL http://www.sciencedirect.com/science/article/pii/S0022247X04004718

Biau G (2012) Analysis of a random forests model. J Mach Learn Res 13(1):1063–1095, URL http://dl.acm.org/citation.cfm?id=2503308.2343682

Biau G, Devroye L (2010) 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, DOI 10.1016/j.jmva.2010.06.019, URL http://dx.doi.org/10.1016/j.jmva.2010.06.019

Biau G, Devroye L, Lugosi G (2008) Consistency of random forests and other averaging classifiers. J Mach Learn Res 9:2015–2033, URL http://dl.acm.org/citation.cfm?id=1390681.1442799

Bickel PJ, Ritov Y, Tsybakov A (2009) Simultaneous analysis of lasso and Dantzig selector. Annals of Statistics 37(4):1705–1732

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140, DOI 10.1023/A:1018054314350, URL http://dx.doi.org/10.1023/A:1018054314350

Breiman L (2001) Random forests. Mach Learn 45(1):5–32, DOI 10.1023/A:1010933404324, URL http://dx.doi.org/10.1023/A:1010933404324

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

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

Candès EJ, Recht B (2009) Exact matrix completion via convex optimization. Foundations of Computational mathematics 9(6):717–772

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

Chaari L, Tourneret JY, Chaux C, Batatia H (2014) A hamiltonian monte carlo method for non-smooth energy sampling. Tech. Rep. arXiv:1401.3988,

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

Chen X, Lin Q, Kim S, Carbonell JG, Xing EP (2010) An efficient proximal-gradient method for general structured sparse learning. Preprint arXiv:10054717

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

Dalalyan A, Tsybakov A (2009) Pac-bayesian bounds for the expected error of aggregation by exponential weights. Tech. rep., Université Paris 6, CREST and CERTIS, Ecole des Ponts ParisTech, personal communication

Dalalyan A, Tsybakov AB (2008) Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. Mach Learn 72(1-2):39–61, DOI 10.1007/s10994-008-5051-0, URL http://dx.doi.org/10.1007/s10994-008-5051-0

Dalalyan AS (2014) Theoretical guarantees for approximate sampling from a smooth and log-concave density. to appear in JRSS B 1412.7392, arXiv, URL http://arxiv.org/pdf/1412.7392v3.pdf

Dalalyan AS, Tsybakov AB (2007) Aggregation by exponential weighting and sharp oracle inequalities. In: Proceedings of the 20th Annual Conference on Learning Theory, Springer-Verlag, Berlin, Heidelberg, COLT'07, pp 97–111, URL http://dl.acm.org/citation.cfm?id=1768841.1768854

Dalalyan AS, Tsybakov AB (2012) Sparse regression learning by aggregation and langevin monte-carlo. J Comput Syst Sci 78(5):1423–1443, DOI 10.1016/j.jcss.2011.12.023, URL http://dx.doi.org/10.1016/j.jcss.2011.12.023

Donoho D (2006) 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

Durmus A, Moulines E (2015) Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm, URL https://hal.archives-ouvertes.fr/hal-01176132, preprint hal-01176132

Durmus A, Moulines E, Pereyra M (2016) Sampling from convex non continuously differentiable functions, when Moreau meets Langevin, URL https://arxiv.org/abs/1612.07471, arxiv:1612.07471

Duy Luu T, Fadili JM, Chesneau C (2016) PAC-Bayesian risk bounds for group-analysis sparse regression by exponential weighting. Tech. rep., hal-01367742, URL https://hal.archives-ouvertes.fr/hal-01367742

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

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties

Fazel M, Hindi H, Boyd SP (2001) A rank minimization heuristic with application to minimum order system approximation. In: Proceedings of the American Control Conference, IEEE, vol 6, pp 4734–4739

Freund Y (1995) Boosting a weak learning algorithm by majority. Information and Computation 121(2):256 – 285, DOI http://dx.doi.org/10.1006/inco.1995.1136, URL http://www.sciencedirect.com/science/article/pii/S0890540185711364

Gao HY, Bruce A (1997) Waveshrink with firm shrinkage. Statist Sinica 7:855– 874

van de Geer S (2014) Weakly decomposable regularization penalties and structured sparsity. Scandinavian Journal of Statistics 41(1):72–86, DOI 10.1111/sjos.12032, URL http://dx.doi.org/10.1111/sjos.12032

Genuer R (2010) Random Forests: elements of theory, variable selection and applications. Theses, Université Paris Sud - Paris XI, URL https://tel.archives-ouvertes.fr/tel-00550989

Guedj B, Alquier P (2013) Pac-bayesian estimation and prediction in sparse additive models. Electron J Statist 7:264–291, DOI 10.1214/13-EJS771, URL http://dx.doi.org/10.1214/13-EJS771

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

Jégou H, Furon T, Fuchs JJ (2012) Anti-sparse coding for approximate nearest neighbor search. In: IEEE ICASSP, pp 2029–2032

Kloeden PE, Platen E (1995) Numerical solution of stochastic differential equations. Stochastic Modelling and Applied Probability, Springer

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

Lecué G (2007) Simultaneous adaptation to the margin and to complexity in classification. Ann Statist 35(4):1698–1721, DOI 10.1214/009053607000000055, URL http://dx.doi.org/10.1214/009053607000000055

Littlestone N, Warmuth MK (1994) The weighted majority algorithm. Inf Comput 108(2):212–261, DOI 10.1006/inco.1994.1009, URL http://dx.doi.org/10.1006/inco.1994.1009

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

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

Nemirovski A (2000) Topics in non-parametric statistics
Osborne M, Presnell B, Turlach B (2000) A new approach to variable selection in least squares problems. IMA journal of numerical analysis 20(3):389–403
Pereyra M (2016) Proximal markov chain monte carlo algorithms. Statistics and Computing 26(4):745–760
Pereyra M, Schniter P, Chouzenoux E, Pesquet J, Tourneret J, Hero AO, McLaughlin S (2016) Tutorial on stochastic simulation and optimization methods in signal processing. IEEE Sel Topics in Signal Processing 10(2):224–241
Peyré G, Fadili J, Chesneau C (2011) Group sparsity with overlapping partition functions. In: EUSIPCO, Barcelona, Spain
Poliquin RA, Rockafellar RT (1996) Prox-regular functions in variational analysis. Transactions of the American Mathematical Society 348(5):1805–1838
Poliquin RA, Rockafellar RT, Thibault L (2000) Local differentiability of distance functions. Transactions of the American mathematical Society 352:5231–5249
Recht B, Fazel M, Parrilo PA (2010) Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM review 52(3):471–501
Rigollet P, Tsybakov A (2007) Linear and convex aggregation of density estimators. Mathematical Methods of Statistics 16(3):260–280, DOI 10.3103/S1066530707030052, URL http://dx.doi.org/10.3103/S1066530707030052
Roberts GO, Tweedie RL (1996) Exponential Convergence of Langevin Distributions and Their Discrete Approximations. Bernoulli 2(4):341–363, URL http://www.jstor.org/stable/3318418
Rockafellar RT, Wets R (1998) Variational analysis, vol 317. Springer Verlag
Rudin L, Osher S, Fatemi E (1992a) Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena 60(1-4):259–268
Rudin LI, Osher S, Fatemi E (1992b) Nonlinear total variation based noise removal algorithms. Phys D 60(1-4):259–268, DOI 10.1016/0167-2789(92)90242-F, URL http://dx.doi.org/10.1016/0167-2789(92)90242-F
Schapire RE (1990) The strength of weak learnability. Mach Learn 5(2):197–227, DOI 10.1023/A:1022648800760, URL http://dx.doi.org/10.1023/A:1022648800760
Scheffe H (1947) A useful convergence theorem for probability distributions. Ann Math Statist 18(3):434–438
Studer C, Yin W, Baraniuk RG (2012) Signal representations with minimum ℓ_∞-norm. In: 50th Annual Allerton Conference on Communication, Control, and Computing,
Tibshirani R (1996) Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B Methodological 58(1):267–288
Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused Lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67(1):91–108
Vaiter S, Golbabaee M, Fadili MJ, Peyré G (2015a) Model selection with low complexity priors. Information and Inference: A Journal of the IMA (IMAIAI)
Vaiter S, Peyré G, Fadili MJ (2015b) Low complexity regularization of linear inverse problems. In: Pfander G (ed) Sampling Theory, a Renaissance, Applied and Numerical Harmonic Analysis (ANHA), Birkhäuser/Springer
Vovk VG (1990) Aggregating strategies. In: Proceedings of the Third Annual Workshop on Computational Learning Theory, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, COLT '90, pp 371–386, URL http://dl.acm.org/citation.cfm?id=92571.92672
Wei F, Huang J (2010) Consistent group selection in high-dimensional linear regression. Bernoulli 16(4):1369–1384
Woodworth J, Chartrand R (2015) Compressed sensing recovery via nonconvex shrinkage penalties. CoRR abs/1504.02923, URL http://arxiv.org/abs/1504.02923
Xuerong M (2007) Stochastic differential equations and applications. Woodhead Publishing
Yang Y (2004) Aggregating regression procedures to improve performance. Bernoulli 10(1):25–47
Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(1):49–67