Sampling Nonsmooth Log-Concave Densities: A Comparative Study of Primal-Dual Based Proposal Distributions

Juliette Chevallier, Gersende Fort

Sampling Nonsmooth Log-Concave Densities: A Comparative Study of Primal-Dual Based Proposal Distributions

Juliette Chevallier
IMT, UMR 5219, INSA Toulouse
Université de Toulouse, France
juliette.chevallier@math.univ-toulouse.fr

Gersende Fort
IMT, UMR 5219, CNRS
Université de Toulouse, France
gersende.fort@math.univ-toulouse.fr

***Abstract*—Sampling from a real-valued distribution, whose density is nonsmooth and log-concave, is a computational issue that often arises in Machine Learning and Statistics. Langevin-based Hastings-Metropolis methods were proposed: they extend the Unadjusted Langevin Algorithm by using proximal methods to define a smoothed version of the density of interest. We consider the case when these extensions do not apply: the involved proximal operators do not have closed forms and the density is defined on a subset of the real numbers. We derive new Gaussian proposal mechanisms in a Metropolis Adjusted Langevin Algorithm, which use first-order information about the density function. We numerically compare these strategies and discuss the benefits of a change of geometry. The gain in using partial updates instead of global parameter updates is also illustrated.**

***Index Terms*—Monte Carlo Sampling, Langevin-based algorithms, Proximal and Subgradient methods, Epidemiological model.**

I. Introduction

Context. Sampling from a distribution on $\mathbb{R}^d$ is a computational question that often arises in Machine Learning and Statistics, including Statistical Signal and Image Processing, among scientific domains. This paper considers the case when the distribution possesses a density π, with respect to the Lebesgue measure, of the form

$$-\log \pi(\theta) = \begin{cases} f(\theta) + g(\theta) + h(\mathsf{A}\theta) & \text{for} \quad \theta \in \mathcal{D}, \\ +\infty & \text{otherwise}, \end{cases} \tag{1}$$

where $\mathcal{D}$ is a measurable convex subset of $\mathbb{R}^d$, A is a $d' \times d$ matrix, and f, g, h are measurable convex functions finite on $\mathcal{D}$. The appeal of these three functions allows for various regularity properties: the function $f : \mathcal{O} \to \mathbb{R}$ is assumed to be continuously differentiable on an open subset $\mathcal{O}$ including $\mathcal{D}$; the functions $g : \mathbb{R}^d \to \mathbb{R}$ and $h : \mathbb{R}^{d'} \to \mathbb{R}$ are lower semicontinuous, their proximal operators exist and have a closed form expression; nevertheless, the matrix A is such that the proximal operator of the function $h(\mathsf{A}\cdot) : \theta \mapsto h(\mathsf{A}\theta)$ is intractable. Such a nonsmooth convex composite negative log-density arises in many problems: let us cite aggregation of estimators by exponential weighting in PAC-Bayesian learning [1]–[3] and Bayesian inverse problems in signal and image processing [4]–[12] as few examples.

Related work. Efficient Monte Carlo sampling takes benefit of optimization methods (see [13] for a survey). The `Metropolis Adjusted Langevin Algorithm (MALA)`, introduced in [14], exploits the discretization of a diffusion process targeting π in order to design an efficient proposal mechanism in a Hastings-Metropolis scheme. Given a current point $\theta^k \in \mathcal{D}$, sample

$$\theta^{k+1/2} \sim \mu(\theta^k) + \sqrt{2\gamma}\,\mathcal{N}_d(0, \mathsf{I}_d), \tag{2}$$

where the *drift* term μ is defined by

$$\mu(\theta) := \theta + \gamma \nabla \log \pi(\theta^k); \tag{3}$$

Work supported by ANR-23-CE48-0009 "*OptiMoCSI*".

$\gamma > 0$, I_d is the $d \times d$ identity matrix, $\mathcal{N}_d(m, \mathsf{C})$ denotes a Gaussian distribution on $\mathbb{R}^d$ with expectation m and covariance matrix C, and $\nabla \ell$ is the gradient of the function ℓ. The candidate $\theta^{k+1/2}$ is then accepted or rejected through an acceptance-rejection (AR) step, which yields θ^{k+1}. The AR step implies that $\theta^{k+1} \in \mathcal{D}$. `MALA` can be seen as a sampler using first-order informations on $\log \pi$ when π satisfies Eq. (1) with $g = h = 0$.

The `Unadjusted Langevin Algorithm (ULA)` was then proposed and studied: the acceptance-rejection step of `MALA` is removed providing a Monte Carlo approximation $\{\theta^k, k \geq 0\}$ targeting a distribution which is no more π [15], [16]. Since $\theta^{k+1} = \theta^{k+1/2}$ in `ULA`, it samples via the Gaussian distribution Eq. (2), so that the Monte Carlo approximation can not be restricted to a subset $\mathcal{D}$ of $\mathbb{R}^d$: `ULA` addresses the case $\mathcal{D} = \mathbb{R}^d$.

When π is nonsmooth but $h = 0$, Langevin-based samplers were proposed by combining a smoothing technique using the Moreau-Yoshida envelope and proximal operators (see `MYULA` in [9] and proximal Markov Chain Monte Carlo (MCMC) samplers in [9], [17]–[20] possibly in a Riemannian geometry [21]). All of these contributions assume that $\mathcal{D} = \mathbb{R}^d$. When $\mathcal{D} \subsetneq \mathbb{R}^d$ (and $h = 0$), Langevin-based Monte Carlo samplers using projections or reflections on the boundaries of $\mathcal{D}$ were proposed for specific topologies of $\mathcal{D}$ (see e.g. [22], [23]).

Outline, Goals and contributions. The originality of our setting Eq. (1) is to consider both *(i)* that a nonsmooth component of $-\log \pi$ is a convex function combined with a linear operator A, and *(ii)* a domain $\mathcal{D} \subsetneq \mathbb{R}^d$ exists. Since we consider the case when the proximal operator of $h(\mathsf{A}\cdot)$ does not have a closed form expression, the Proximal-based MCMC samplers proposed in the literature do not apply except the samplers in [12], [19]. Nevertheless, [19] are `ULA`-type samplers thus addressing the case $\mathcal{D} = \mathbb{R}^d$.

This paper completes the methodological contributions in [12] and the numerical comparisons in [24]. It is structured as follows:

First, in Section II, we generalize the gradient step in Eq. (3) to the case $\log \pi$ is nonsmooth and of the form Eq. (1), and propose new drift terms by using explicit first-order information on $\log \pi$. When the matrix A is full column rank, other strategies are derived. The domain $\mathcal{D}$ is not considered in the proposal mechanism: despite the AR step being known to slow down the convergence of `MALA`, we consider an unrestricted Gaussian proposal of the form Eq. (2) but combined with an AR step which forces θ^{k+1} to be in $\mathcal{D}$. Such a strategy avoids complex sampling of the candidate $\theta^{k+1/2}$ to ensure $\theta^{k+1/2} \in \mathcal{D}$ especially when the topology of $\mathcal{D}$ is difficult to handle.

Second, in Section III, we numerically compare these new drift terms $\mu(\cdot)$, and discuss the interest of first-order strategies with respect to a `Random Walk` drift term $\mu(\theta) = \theta$. We also consider the benefit of a change of geometry, which is a natural idea in situations when A is full column rank; and compare global updates of the full vector

θ to sequential updates such as a *one-at-a-time* strategy. Finally, we comment how adaption mechanisms for an automated choice of design parameters drastically improve the samplers.

Notations. Hereafter, $\mathsf{A}^\top$ denotes the transpose of the matrix A, and $\mathsf{A}^{-\top} := (\mathsf{A}^{-1})^\top$. $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^p$ and $\|\cdot\|_1$ is the L_1-norm. $\nabla f(\theta)$ is the gradient at θ of a continuously differentiable function f. ∂g is the subdifferential operator of a proper function $g : \mathbb{R}^p \to (-\infty, +\infty]$.

II. Primal Dual based proposal distributions

In this work, we investigate a density π satisfying Eq. (1), such that:

A) $\mathcal{D}$ is a measurable convex subset of $\mathbb{R}^d$;

B) there exists an open neighborhood $\mathcal{O}$ of $\mathcal{D}$ such that f is continuously differentiable on $\mathcal{O}$;

C) $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ and $h : \mathbb{R}^{d'} \to \mathbb{R} \cup \{+\infty\}$ are lower semicontinuous convex functions; for all $\gamma > 0$, $\theta \in \mathbb{R}^d$ and $\theta' \in \mathbb{R}^{d'}$: $\mathrm{Prox}_{\gamma g}(\theta)$ and $\mathrm{Prox}_{\gamma h}(\theta')$ have a closed form expression;

D) A is a $d' \times d$ matrix. For all $\gamma > 0$ and $\theta \in \mathbb{R}^d$, $\mathrm{Prox}_{\gamma h(\mathsf{A}\cdot)}(\theta)$ is intractable.

Remember that the proximal operator of a proper lower semicontinuous convex function $\ell : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is defined by

$$\mathrm{Prox}_\ell(\cdot) := \mathrm{argmin}_{\theta \in \mathbb{R}^p} \left(\ell(\theta) + \frac{1}{2}\|\theta - \cdot\|^2 \right);$$

for all ϑ, the minimizer of $\theta \mapsto \ell(\theta) + \|\theta - \vartheta\|^2/2$ exists and is unique (see [25, Definition 12.23]). When $\mathsf{A}\mathsf{A}^\top$ is an invertible diagonal matrix and for specific functions h (see e.g. [12, Lemma 4]), the assumption C implies that $\mathrm{Prox}_{\gamma h(\mathsf{A}\cdot)}(\theta)$ has a closed form expression.

In this section, we design Hastings-Metropolis (HM) samplers, with a focus on the drift term $\mu(\cdot)$. One iteration of HM is described by Alg. 1. Define

$$\alpha(\theta, \theta') := \min\left(1; \frac{\pi(\theta')}{\pi(\theta)} \frac{q(\mu(\theta'), \theta)}{q(\mu(\theta), \theta')}\right), \tag{4}$$

where $y \mapsto q(x, y)$ denotes the density of a Gaussian distribution centered at x, with covariance matrix I_d. The rate of ergodicity of a Markov chain is related to how fast, starting from a *small set*, the chain returns back to this small set (see e.g. [26, Theorem 15.0.1]). Since a MCMC sampler produces a Markov chain with a given invariant distribution, we are interested in drift terms that can push the chain towards the modes of π when the chain is visiting the tails of π. For these reasons, the drift terms below are chosen as one step of a convex optimization procedure for minimizing $-\log \pi$. It holds ([25, Theorem 16.3]): $\theta^\star$ is a minimizer of $-\log \pi$ if and only if (iff) $0 \in \partial(f + g + h(\mathsf{A}\cdot))(\theta^\star)$, or equivalently under the stated assumptions: for any $\gamma > 0$,

$$\begin{cases} 0 = \gamma\, \nabla f(\theta^\star) + \gamma\, u^\star + \gamma\, \mathsf{A}^\top s^\star; \\ u^\star \in \partial g(\theta^\star), s^\star \in \partial h(\mathsf{A}\theta^\star). \end{cases} \tag{5}$$

From Eq. (5), a first strategy is to define

$$\mu(\theta) := \theta - \gamma \nabla f(\theta) - \gamma G(\theta) - \gamma \mathsf{A}^\top H(\mathsf{A}\theta), \tag{6}$$

where $G(\theta) \in \partial g(\theta)$, $H(\tau) \in \partial h(\tau)$ and G, H are measurable point-to-point maps. For example, when g is continuously differentiable, $G(\theta) = \nabla g(\theta)$; when $h(\cdot) = \|\cdot\|_1$, $H(\tau) = \mathrm{sign}(\tau)$ with the convention $\mathrm{sign}(0) = 0$. When $\partial g(\theta)$ or $\partial h(\tau)$ is not a singleton, it is far simpler to have a deterministic selection of one subgradient, which here, is denoted by $G(\theta)$ and $H(\tau)$ respectively. A random selection of one of the subgradients is possible, but in that case, the candidate $\theta^{k+1/2}$ is no more Gaussian (see Eq. (2)) and the expression of its probability distribution has to be derived in closed form for the AR step (see Eq. (4)); this method is not investigated in this paper. Eq. (6) is a *full sub-gradient* strategy, named hereafter, `FSG`. Using again Eq. (5) and the property: $p = \mathrm{Prox}_\ell(\tau)$ iff $\tau - p \in \partial \ell(p)$ (see e.g. [25, Proposition 16.44]), another strategy for the drift term is

$$\mu(\theta) = \mathrm{Prox}_{\gamma g}\left(\theta - \gamma \nabla f(\theta) - \gamma \mathsf{A}^\top H(\mathsf{A}\theta)\right). \tag{7}$$

Again, $H(\tau) \in \partial h(\tau)$ and H is a one-to-one map. Eq. (7) combines a subgradient of h and an *implicit* gradient of g through the proximal operator of γg; hereafter, it is named `Prox-SG`. Following the same idea, another strategy could involve a subgradient of g and the proximal operator of γh: this would yield an update for $\mathsf{A}\theta$ but, without additional assumptions on A, it can not be used for an update of θ.

When A is invertible. When A is invertible, a change of geometry is possible. We have indeed

Lemma. *Set $\tilde{\theta} := \mathsf{A}\theta$. If $\theta \sim \pi$ then $\tilde{\theta} \sim \tilde{\pi}$, where $\tilde{\pi}$ is proportional to $\pi(\mathsf{A}^{-1}\cdot)$. If $\{\tilde{\theta}^k, k \geq 0\}$ is a Markov chain with unique invariant distribution $\tilde{\pi}$, then the chain $\{\mathsf{A}^{-1}\tilde{\theta}^k, k \geq 0\}$ is a Markov chain with unique invariant distribution π.*

As a corollary, iterating Alg. 2 produces a Markov chain $\{\theta^k, k \geq 0\}$ with unique invariant distribution π. The candidate $\tilde{\theta}^{k+1/2}$ in Alg. 2 has the same distribution as $\mathsf{A}\theta^{k+1/2}$ where

$$\theta^{k+1/2} \sim \mathsf{A}^{-1}\tilde{\mu}(\mathsf{A}\theta^k) + \sqrt{2\gamma}\mathcal{N}_d(0, \mathsf{A}^{-1}\mathsf{A}^{-\top}). \tag{8}$$

It shows that, transcribed in the original θ-space, the change of geometry has two effects on the proposal mechanism: it yields *(i)* a new drift term $\mathsf{A}^{-1}\tilde{\mu}(\mathsf{A}\cdot)$ where $\tilde{\mu}$ is chosen to be efficient for sampling $\tilde{\pi}$, and *(ii)* a covariance matrix of the Gaussian proposal which is not the identity matrix. Since $-\ln \tilde{\pi}$ is, up to an additive constant, equal to $\tilde{\theta} \mapsto f(\mathsf{A}^{-1}\tilde{\theta}) + g(\mathsf{A}^{-1}\tilde{\theta}) + h(\tilde{\theta})$ on the set $\{\tilde{\theta} : \mathsf{A}^{-1}\tilde{\theta} \in \mathcal{D}\}$ and equal to $+\infty$ otherwise, we define two strategies for a drift term $\tilde{\mu}$.

The first one is, again, a *full sub-gradient* approach

$$\tilde{\mu}(\tilde{\theta}) := \tilde{\theta} - \gamma \mathsf{A}^{-\top} \nabla f(\mathsf{A}^{-1}\tilde{\theta}) - \gamma \mathsf{A}^{-\top} G(\mathsf{A}^{-1}\tilde{\theta}) - \gamma H(\tilde{\theta}), \tag{9}$$

where G, H are point-to-point maps such that $G(\theta) \in \partial g(\theta)$ and $H(\tilde{\theta}) \in \partial h(\tilde{\theta})$. Hereafter, it is named `inv-FSG`. The second one uses the sub-gradient of $g(\mathsf{A}^{-1}\cdot)$ and the proximal operator of γh:

$$\tilde{\mu}(\tilde{\theta}) := \mathsf{A}^{-1}\mathrm{Prox}_{\gamma h}\left(\tilde{\theta} - \gamma \mathsf{A}^{-\top}\Big(\nabla f(\mathsf{A}^{-1}\tilde{\theta}) + G(\mathsf{A}^{-1}\tilde{\theta})\Big)\right). \tag{10}$$

Again, G is a one-to-one map and $G(\theta) \in \partial g(\theta)$. When $g = 0$, such a HM sampler was proposed in [12], under the name `PGdual`. Hereafter, for consistency purposes, it is named `SG-Prox`.

Algorithm 1 one iteration of a HM targeting π

Input: θ^k
Sample $\theta^{k+1/2} \sim \mu(\theta^k) + \sqrt{2\gamma}\mathcal{N}_d(0, \mathsf{I}_d)$
(AR step) Set $\theta^{k+1} = \theta^{k+1/2}$ with probability $\alpha(\theta^k, \theta^{k+1/2})$ and $\theta^{k+1} = \theta^k$ otherwise.
Output: θ^{k+1}.

Algorithm 2 one iteration via change of geometry

Input: θ^k
Set $\tilde{\theta}^k := \mathsf{A}\theta^k$
Sample $\tilde{\theta}^{k+1/2} \sim \tilde{\mu}(\tilde{\theta}^k) + \sqrt{2\gamma}\mathcal{N}_d(0, \mathsf{I}_d)$
Run an AR step with target distribution $\tilde{\pi}$ and obtain $\tilde{\theta}^{k+1}$.
Output: $\theta^{k+1} := \mathsf{A}^{-1}\tilde{\theta}^{k+1}$.

When A is full column rank. Write $h(\mathsf{A}\theta) = \bar{h}(\bar{\mathsf{A}}\theta)$ where $\bar{\mathsf{A}}$ is a $d \times d$ invertible matrix with rows $\#(d - d' + 1)$ to $\#d$ equal to A, and $\bar{h}(x_1, \cdots, x_d) := h(x_{d-d'+1}, \cdots, x_d)$. Then apply the case "$\mathsf{A}$ invertible" with $\mathsf{A} \leftarrow \bar{\mathsf{A}}$ and $h \leftarrow \bar{h}$. Such an idea is detailed in [12, Section III-C].

When π is known up to a normalizing constant. All the drift terms depend on f via its gradient, so f can be defined up to an additive constant. This remark combined with the expression of the AR ratio (see Eq. (4)), shows that the above HM samplers apply even when π is known up to a multiplicative constant.

III. A COMPARATIVE STUDY

We consider a density π given by a penalized Poisson likelihood introduced to model counts $\mathsf{Z}_1, \cdots, \mathsf{Z}_T$ in an epidemiological model (see e.g. [12, Section II-C]). Let D be a $T \times T$ matrix, and $\Phi_1, \cdots, \Phi_T$ be non-negative numbers; set $\mathbf{\Phi}$ the $T \times T$ diagonal matrix with diagonal entries $\{\Phi_s, 1 \leq s \leq T\}$. On a set $\mathcal{D}$, $-\ln \pi(\theta)$ is equal to

$$\sum_{t=1}^{T} (\Phi_t(\mathsf{R}_t + \mathsf{O}_t) - \mathsf{Z}_t \ln(\mathsf{R}_t + \mathsf{O}_t)) + \lambda_{\mathsf{R}} \|\mathsf{D}\mathbf{R} + \delta\|_1 + \lambda_{\mathsf{O}} \|\mathbf{\Phi}\mathbf{O}\|_1$$

and to $+\infty$ otherwise. Here, $\theta := (\mathsf{R}_1, \cdots, \mathsf{R}_T, \mathsf{O}_1, \cdots, \mathsf{O}_T)$, $\mathbf{R} := (\mathsf{R}_1, \cdots, \mathsf{R}_T) \in \mathbb{R}^T$, $\mathbf{O} := (\mathsf{O}_1, \cdots, \mathsf{O}_T) \in \mathbb{R}^T$. This model aims to learn a time-varying reproduction number $t \mapsto \mathsf{R}_t$ from highly corrupted data $\mathsf{Z}_1, \cdots, \mathsf{Z}_T$ and averaged past counts $\Phi_1, \cdots, \Phi_T$ (see Fig. 1). The O_t's model errors in the counts, and D is the discrete-time 2nd order derivative matrix: the penalty terms favor sparse errors and slowly varying reproduction numbers. δ corresponds to initial values of the 2nd order derivative of $\mathbf{R}$. Finally, the set $\mathcal{D} \subseteq (\mathbb{R}_+)^T \times \mathbb{R}^T$ ensures that the Poisson intensity and the reproduction numbers are non-negative (a Poisson with null parameter is a Dirac mass at zero):

$$\mathcal{D} := \bigcap_{t:\mathsf{Z}_t > 0} \{(\mathbf{R}, \mathbf{O}) : \mathsf{R}_t + \mathsf{O}_t > 0\} \cap \bigcap_{t:\mathsf{Z}_t = 0} \{(\mathbf{R}, \mathbf{O}) : \mathsf{R}_t + \mathsf{O}_t \geq 0\}.$$

Note that the density π is of the form Eq. (1) with $d = 2T$; $f = 0$; $g\colon \theta \mapsto \sum_{t=1}^{T} (\Phi_t(\mathsf{R}_t + \mathsf{O}_t) - \mathsf{Z}_t \ln(\mathsf{R}_t + \mathsf{O}_t)) + \iota_{\mathcal{D}}(\theta)$, where $\iota_{\mathcal{D}}$ is the characteristic function of the convex set $\mathcal{D}$; A is the block-diagonal $d \times d$ matrix with block entries D and $\lambda_{\mathsf{O}}\mathbf{\Phi}/\lambda_{\mathsf{R}}$; $h(\cdot) := \lambda_{\mathsf{R}} \|\cdot + \bar{\delta}\|_1$, where $\bar{\delta}$ concatenates $\delta \in \mathbb{R}^T$ and the null vector $0 \in \mathbb{R}^T$. The functions g and h are convex, finite on $\mathcal{D}$ and lower semicontinuous. The proximal operators have closed form expressions: for example, the component $\#t$ of $\mathrm{Prox}_{\gamma h}(\theta)$ is $-\bar{\delta}_t + \max(0, |\theta_t + \bar{\delta}_t| - \gamma\lambda_{\mathsf{R}}) \operatorname{sign}(\theta_t + \bar{\delta}_t)$; for g, see e.g. [11, Proposition 3] for a similar computation. We have $\partial g(\theta) = \{\nabla g(\theta)\}$ and $\partial h(\theta) = \{\lambda_{\mathsf{R}} \operatorname{sign}(\theta + \bar{\delta})\}$ for $\theta \neq 0$; for $\theta = 0$, we choose $H(0) = 0$. Finally, $\mathsf{D}\mathsf{D}^\top$ is not a diagonal matrix and $\mathrm{Prox}_{\gamma h(\mathsf{A}\cdot)}(\theta)$ is intractable.

COVID-19 data. The data $\mathsf{Z}_1, \cdots, \mathsf{Z}_T$ are the daily new infection counts of COVID-19, in France, from 2022/02/20 to 2022/04/28; $T = 68$. They are provided by the Johns Hopkins University repository[1]. Due to poor reporting of the counts during the weekends and public holidays, there are regularly abnormally low counts followed by abnormally high ones.

Algorithms setup. All the samplers are run for 10^7 iterations and start from a poor initialization $\mathbf{R}^0 := (1, \cdots, 1) \in \mathbb{R}^T$ and $\mathbf{O}^0 := (0, \cdots, 0) \in \mathbb{R}^T$. They all depend on a step size γ. This design parameter is adapted during $5 \cdot 10^6$ iterations and then, it is fixed to its current value. The adaption mechanism drives the empirical mean acceptance rate to 25%. The initial value of γ is chosen so that each sampler accepts a candidate before the $5 \cdot 10^4$-th iteration.
In Fig. 2 to 4, the curves are averaged curves obtained from ten independent runs of the samplers.

[1] https://coronavirus.jhu.edu/

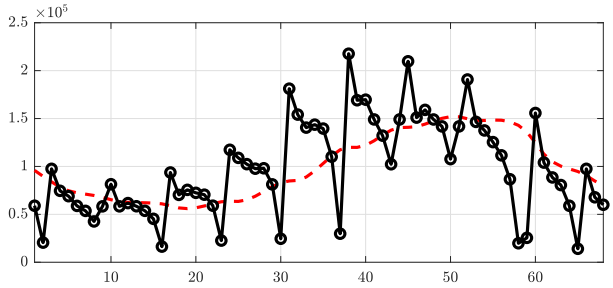

Fig. 1: The data $\mathsf{Z}_1, \cdots, Z_T$ (black curve) and the averaged past counts $\Phi_1, \cdots, \Phi_T$ (dashed line curve)

Analyzes. $\theta \mapsto \log \pi(\theta)$ has a unique maximizer $\theta^\star := (\mathbf{R}^\star, \mathbf{O}^\star)$ (see [12, Proposition 1]) which can be approximated by a primal-dual algorithm (see [11]). In order to visualize how fast the sampler moves towards the mode of π when started from a poor initialization, we display the $\log \pi$ criterion $k \mapsto (\log \pi(\theta^\star) - \log \pi(\theta^k))/\log \pi(\theta^\star)$ (see Fig. 2, top row). We also display the `MAP-R` criterion: $k \mapsto \|\mathbf{R}^k - \mathbf{R}^\star\|/\|\mathbf{R}^\star\|$ (see Fig. 2, bottom row). The asymptotic efficiency of the samplers is compared via a mean auto-correlation function (ACF) criterion: on Fig. 4, we report $\tau \mapsto (1/T) \sum_{t=1}^{T} |\mathrm{ACF}_t^U(\tau)|$ where $\mathrm{ACF}_t^U(\cdot)$ is the ACF of the component $\#t$ of the vector U; we consider in turn $U = \mathbf{R}$ and $U = \mathbf{O}$. A rapid decay to zero of this criterion is expected iff the Markov chain has good mixing properties (see e.g. [26, Section 17.5]). Finally, we report the value of the step size γ at the end of the adaption period (see Fig. 3).

RW, FSG and Prox-SG. We run the HM samplers (see Alg. 1) with the drift term `FSG` (see Eq. (6)) and the drift term `Prox-SG` (see Eq. (7)). For comparison, we also run a `Random Walk (RW)` algorithm which corresponds to the drift term $\mu(\theta) = \theta$; see solid red, solid green and solid black curves respectively, on Fig. 2 left to Fig. 4 left. `FSG` and `Prox-SG` are almost equivalent, and better than `RW`: there is a gain in using first-order information on $\log \pi$. From the `ACF` criterion, there is a slight advantage to `FSG`. The step size γ is the same for the $2T$ components of θ: it is very small (about 10^{-11} for the three samplers, see Fig. 3 left) thus explaining the poor decay of the ACF (see Fig. 4 left).

Change of geometry. We compare `inv-FSG` (see Eq. (9)), `SG-Prox` (see Eq. (10)) and `inv-RW` which corresponds to $\tilde{\mu}(\tilde{\theta}) = \tilde{\theta}$; see solid red, solid blue and solid black curves respectively, on Fig. 2 right to Fig. 4 right. From Eq. (8), `inv-RW` and `RW` have the same drift term in the θ-space but differ through the covariance matrix. Comparing `RW` and `inv-RW` shows that the covariance matrix in the Gaussian proposal mechanism drastically modifies the behavior (see Fig. 2 and Fig 4). All the methods are improved by this change of geometry: we observe a faster convergence to the mode of π on Fig. 2, and a faster decay to zero of the ACF on Fig. 4 even if the step size γ remains very small, less than 10^{-12}, see Fig 3.

Global update or sequential update. Until now, the vector θ is updated globally; it is named a `no-Gibbs` approach. This yields very small values of the step size γ in order to accept 25% of the candidates, which in turn implies a high correlation between successive points $\theta^k, \theta^{k+1}, \cdots$, and therefore very poor mixing

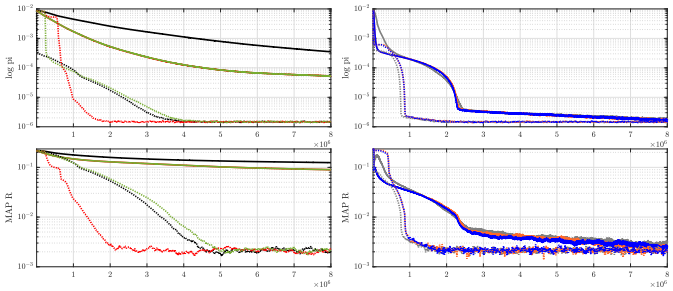

Fig. 2: $\log \pi$ (top) and `MAP-R` (bottom) criteria, for algorithms run in the original space (left, RW, FSG, Prox-SG) or algorithms after a change of geometry (right, inv-RW, inv-FSG, SG-Prox). `No Gibbs` samplers are in solid lines and `one-at-a-time Gibbs` samplers are in dotted lines.

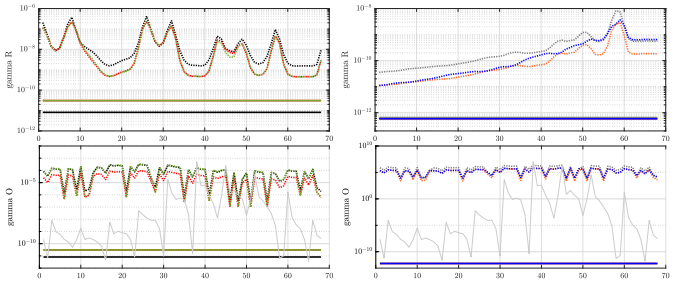

Fig. 3: The step size γ, for algorithms run in the original space (left, RW, FSG, Prox-SG) or algorithms after a change of geometry (right, inv-RW, inv-FSG, SG-Prox). `No Gibbs` samplers are in solid lines and `one-at-a-time Gibbs` samplers are in dotted lines. Bottom row: the data $\mathsf{Z}_1, \cdots, \mathsf{Z}_T$ are displayed in light gray on an independent y-scale.

properties of the chain. We investigate the benefit of Gibbs strategies i.e. sequential updates of θ; here, we consider a `one-at-a-time Gibbs` sampler where each component of $\mathbf{R}$ and then each component of $\mathbf{O}$ are updated successively: for each of the $2T$ components of θ, a candidate is proposed and accepted conditionally to the current value of the $(2T-1)$ other components. The acceptance ratio uses the conditional distribution of θ_t given $\theta_1, \cdots, \theta_{t-1}, \theta_{t+1}, \cdots, \theta_{2T}$. This distribution (on $\mathbb{R}$) is again of the form (1) so that the drift terms `FSG`, `Prox-SG`, `inv-FSG`, and `SG-Prox` can be derived; details are omitted here but are given in [27]. A key observation is that for this `one-at-a-time` approach, there is one step size per component of θ, whose values at the end of the adaption period are displayed in dotted lines on Fig. 3.

The six algorithms learn a step size which turns out to vary strongly along the components (up to a factor 10^3); this variation is almost the same for the `FSG`, `Prox-SG` and `RW` on one side (see Fig. 3 left, dotted curves) and `inv-FSG`, `SG-Prox`, and `inv-RW` on the other side (see Fig. 3 right, dotted curves). On Fig. 3 bottom, we display the data $\mathsf{Z}_1, \cdots, \mathsf{Z}_T$ (light gray curve) in order to detect similarities between the step sizes for the $\mathbf{O}$-block and the data; there is no clear correlation, and the behavior of the step sizes will have to be explored. Compared to the `no Gibbs` strategy, the step sizes are at least 100 times larger (Fig. 3, top left), and even 10^{10} times larger (Fig. 3, bottom right). As a consequence, the $\log \pi$, `MAP-R`, and ACF criteria decay more rapidly to zero with a `one-at-a-time Gibbs` strategy, than with a `no Gibbs` one (see Fig 2 and Fig 4, dotted curves and solid curves).

The three `one-at-a-time Gibbs` strategies look equivalent

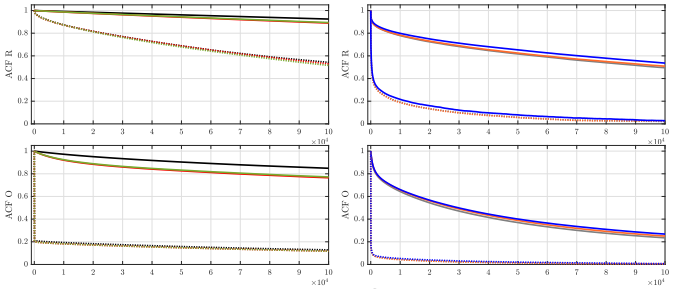

Fig. 4: `ACF` criteria for $\mathbf{R}$ (top) and $\mathbf{O}$ (bottom), for algorithms run in the original space (left, RW, FSG, Prox-SG) or algorithms after a change of geometry (right, inv-RW, inv-FSG, SG-Prox). `No Gibbs` samplers are in solid lines and `one-at-a-time Gibbs` samplers are in dotted lines.

when combined with a change of geometry (see Fig. 2 and Fig. 4 right, dotted curves), and otherwise, `FSG` is preferable since it moves more rapidly from a low-density region to a high-density region (see Fig. 2 left, dotted curves).

Note however that the computational cost of a `one-at-a-time Gibbs` sampler is higher than the `no Gibbs` one; partial updates with intermediate size of the blocks could be considered to balance the computational cost and the gain in efficiency (see [27]).

Conclusion. These numerical analyzes illustrate that the covariance structure of the Gaussian proposal plays an important role in the efficiency of the `Random Walk` sampler. When A is not full column rank, `FSG` is the best strategy; when A is full column rank, there is a gain in using a change of geometry, for both `Random Walk` and the HM samplers with a drift term using first-order information on $\log \pi$. The crux of the dimension d of the sampling space exists, even with a change of geometry: partial sequential updates or even one-at-a-time updates are preferable to a global update. Based on these analyzes, we recommend `FSG`-type approaches, which are among the most efficient samplers in both the original θ-space and with a change of geometry, when compared to `Random Walk` and to methods using (sub-)gradients and proximal operators.

IV. Conclusions and Perspectives

We compared different strategies for the definition of the drift term in a HM sampler with a Gaussian proposal, that uses first-order information on $\log \pi$.

A next step is to test adaptive methods for learning a covariance matrix of the Gaussian proposal, following for example the seminal paper of [28]. Another direction of research is to use `fully proximal` drift terms, by adapting the primal-dual optimization method `PD3O` of [29]: such an approach will rely on data augmentation and will necessitate the definition of an adequate target distribution on an extended (θ, s)-space whose marginal in θ is π (see e.g. [30] for a similar idea).

Finally, `MALA` is known to be poor for heavy-tailed distributions, except when combined with preconditioning strategies [31]: the benefit of primal-dual based methods with state-dependent preconditioners will be investigated.

Matlab codes implementing the proposed samplers will be shared publicly at the time of publication.

REFERENCES

[1] A. Dalalyan and A. Tsybakov, "Sparse regression learning by aggregation and Langevin Monte-Carlo," *J. Comput. Syst. Sci.*, vol. 78, no. 5, pp. 1423–1443, 2012.

[2] B. Guedj and P. Alquier, "PAC-Bayesian estimation and prediction in sparse additive models," *ElectroN. J. Stat.*, vol. 7, no. none, pp. 264 – 291, 2013.

[3] T. Luu, J. Fadili, and C. Chesneau, "Pac-bayesian risk bounds for group-analysis sparse regression by exponential weighting," *J. Multivar. Anal.*, vol. 171, pp. 209–233, 2019.

[4] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors," *IEEE Trans. Inf. Theory*, vol. 45, no. 3, pp. 909–919, 1999.

[5] N. Dobigeon, A. O. Hero, and J.-Y. Tourneret, "Hierarchical Bayesian Sparse Image Reconstruction With Application to MRFM," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 2059–2070, 2009.

[6] L. Chaâri, J.-C. Pesquet, J.-Y. Tourneret, P. Ciuciu, and A. Benazza-Benyahia, "A hierarchical Bayesian model for frame representation," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4086–4089.

[7] F. Lucka, "Fast Markov chain Monte Carlo sampling for sparse Bayesian inference in high-dimensional inverse problems using L1-type priors," *Inverse Probl.*, vol. 28, no. 12, p. 125012, nov 2012.

[8] F. Costa, H. Batatia, L. Chaari, and J.-Y. Tourneret, "Sparse EEG Source Localization Using Bernoulli Laplacian Priors," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 12, pp. 2888–2898, 2015.

[9] M. Pereyra, "Proximal Markov chain Monte Carlo algorithms," *Stat. Comput.*, vol. 26, pp. 745–760, 2016.

[10] L. Chaari, J.-Y. Tourneret, C. Chaux, and H. Batatia, "A Hamiltonian Monte Carlo Method for Non-Smooth Energy Sampling," *IEEE Trans. Signal Process.*, vol. 64, no. 21, pp. 5585–5594, 2016.

[11] B. Pascal, P. Abry, N. Pustelnik, R. S. G, R. Gribonval, and P. Flandrin, "Nonsmooth convex optimization to estimate the Covid-19 reproduction number space-time evolution with robustness against low quality data," *IEEE Trans. Signal Process.*, vol. 70, pp. 2859–2868, 2022.

[12] G. Fort, B. Pascal, P. Abry, and N. Pustelnik, "Covid19 reproduction number: Credibility intervals by blockwise proximal monte carlo samplers," *IEEE Trans. Signal Process.*, vol. 71, pp. 888–900, 2023.

[13] M. Pereyra, P. Schniter, E. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A. Hero, and S. McLaughlin, "A survey of stochastic simulation and optimization methods in signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 2, pp. 224–241, 2015.

[14] G. Roberts and R. Tweedie, "Exponential Convergence of Langevin Distributions and Their Discrete Approximations," *Bernoulli*, vol. 2, no. 4, pp. 341–363, 1996.

[15] A. Durmus and E. Moulines, "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm," *Ann. Appl. Probab.*, vol. 27, no. 3, pp. 1551 – 1587, 2017.

[16] A. Dalalyan, "Theoretical guarantees for approximate sampling from smooth and log-concave densities," *J. R. Stat. Soc. Ser. B Methodol.*, vol. 79, no. 3, pp. 651–676, 2017.

[17] A. Durmus, E. Moulines, and M. Pereyra, "Efficient Bayesian Computation by Proximal Markov Chain Monte Carlo: When Langevin Meets Moreau," *SIAM J. Imaging Sci.*, vol. 11, no. 1, pp. 473–506, 2018.

[18] M. Pereyra, L. V. Mieles, and K. C. Zygalakis, "Accelerating Proximal Markov Chain Monte Carlo by Using an Explicit Stabilized Method," *SIAM J. Imaging Sci.*, vol. 13, no. 2, pp. 905–935, 2020.

[19] T. Luu, J. Fadili, and C. Chesneau, "Sampling from Non-smooth Distributions Through Langevin Diffusion," *Methodology and Computing in Applied Probability*, vol. 23, pp. 1173–1201, 2021.

[20] A. Durmus, E. Moulines, and M. Pereyra, "A Proximal Markov Chain Monte Carlo Method for Bayesian Inference in Imaging Inverse Problems: When Langevin Meets Moreau," *SIAM Review*, vol. 64, no. 4, pp. 991–1028, 2022.

[21] T. T.-K. Lau and H. Liu, "Bregman proximal Langevin Monte Carlo via Bregman-moreau envelopes," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162, 2022, pp. 12 049–12 077.

[22] S. Bubeck, R. Eldan, and J. Lehec, "Sampling from a Log-Concave Distribution with Projected Langevin Monte Carlo," *Discrete Comput. Geom.*, vol. 59, pp. 757–783, 2018.

[23] S. Melidonis, P. Dobson, Y. Altmann, M. Pereyra, and K. Zygalakis, "Efficient bayesian computation for low-photon imaging problems," *SIAM J. Imaging Sci.*, vol. 16, no. 3, pp. 1195–1234, 2023.

[24] P. Abry, G. Fort, B. Pascal, and N. Pustelnik, "Proximal-langevin samplers for nonsmooth composite posteriors: Application to the estimation of covid19 reproduction number," in *2023 31th European Signal Processing Conference (EUSIPCO)*, 2023.

[25] H. Bauschke and P.-L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. New York: Springer, 2019.

[26] S. Meyn and R. Tweedie, *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.

[27] J. Chevallier and G. Fort, "Monte carlo sampling for nonsmooth log-concave densities: Application to pandemic monitoring," work in progress, Tech. Rep., 2024.

[28] H. Haario, E. Saksman, and J. Tamminen, "An adaptive Metropolis algorithm," *Bernoulli*, vol. 7, no. 2, pp. 223 – 242, 2001.

[29] Y. Ming, " A New Primal–Dual Algorithm for Minimizing the Sum of Three Functions with a Linear Operator," *J. Sci. Comput.*, vol. 76, pp. 1698–1717, 2018.

[30] M. Vono, N. Dobigeon, and P. Chainais, "Split-and-Augmented Gibbs Sampler—Application to Large-Scale Inference Problems," *IEEE Trans. Signal Process.*, vol. 67, no. 6, pp. 1648–1661, 2019.

[31] G. Fort and G. O. Roberts, "Subgeometric ergodicity of strong Markov processes," *Ann. Appl. Probab.*, vol. 15, no. 2, pp. 1565 – 1589, 2005.