A Hamiltonian Monte Carlo Method for Non-Smooth Energy Sampling

Lotfi Chaari *IEEE Member*, Jean-Yves Tourneret, *IEEE Senior member*, Caroline Chaux, *IEEE Senior member*, and Hadj Batatia, *Member, IEEE*

***Abstract*—Efficient sampling from high-dimensional distributions is a challenging issue which is encountered in many large data recovery problems. In this context, sampling using Hamiltonian dynamics is one of the recent techniques that have been proposed to exploit the target distribution geometry. Such schemes have clearly been shown to be efficient for multi-dimensional sampling, but are rather adapted to distributions from the exponential family with smooth energy functions. In this paper, we address the problem of using Hamiltonian dynamics to sample from probability distributions having non-differentiable energy functions such as those based on the ℓ_1 norm. Such distributions are being used intensively in sparse signal and image recovery applications. The technique studied in this paper uses a modified leapfrog transform involving a proximal step. The resulting non-smooth Hamiltonian Monte Carlo method is tested and validated on a number of experiments. Results show its ability to accurately sample according to various multivariate target distributions. The proposed technique is illustrated on synthetic examples and is applied to an image denoising problem.**

***Index Terms*—Sparse sampling, Bayesian methods, MCMC, Hamiltonian, proximity operator, leapfrog.**

I. Introduction

Sparse signal and image recovery is a hot topic which has gained a lot of interest during the last decades, especially after the emergence of the compressed sensing theory [1]. In addition, many recent applications, especially in remote sensing [2] and medical image reconstruction [3, 4], deal with large data volumes that are processed either independently or jointly. To handle such inverse problems, Bayesian techniques have demonstrated their usefulness especially when the model hyperparameters are difficult to be adjusted a priori. These techniques generally rely on a maximum a posteriori (MAP) estimation built upon the signal/image likelihood and priors. Analytical expressions of the MAP estimators are often difficult to obtain due to the complex form of the associated efficient priors. For this reason, many Bayesian estimators are computed using samples generated according to the posterior using Markov chain Monte Carlo (MCMC) sampling techniques [5]. To handle large-dimensional sampling, several techniques have been proposed during the last decades. In addition to the random walk Metropolis Hastings (MH) algorithm [5], one can mention the work in [6] about efficient high-dimensional importance sampling, the Metropolis-adjusted Langevin algorithm (MALA) [7–9], elliptical slice sampling [10] or the high-dimensional Gaussian sampling methods of [11, 12]. To handle log-concave or multi-modal smooth probability distributions, a Hamiltonian Monte Carlo (HMC) sampling technique has recently been proposed in [8, 13, 14]. This technique uses the analogy with the kinetic energy conservation in physics to design efficient proposals that better follow the geometry of the target distribution. HMC has recently been investigated in a number of works dealing with multi-dimensional sampling problems for various applications [15, 16], demonstrating its efficiency. Efficient sampling is obtained using these strategies where the convergence and mixing properties of the simulated chains are improved compared to classical sampling schemes such as the Gibbs and MH algorithms. However, these techniques are only appropriate for probability distributions with smooth energy functions whose gradient can be calculated. This constraint represents a real limitation in applications where sparsity is a key property, especially with large datasets. Indeed, sparsity promoting probability distributions generally have a non-differentiable energy function such as the Laplace or the generalized Gaussian (GG) distributions [17] which involve ℓ_1 and ℓ_p regularizations, respectively. These distributions have been used as priors for the target signals or images in a number of works where inverse problems are handled in a Bayesian framework [18–21]. Sampling from non-smooth posteriors has been considered in a number of signal and image processing problems such as image deblurring [22], magnetic resonance force microscopy reconstruction [23] and electroencephalography signal recovery [24]. However, these works did not use efficient sampling moves based on HMC or MALA, since the definition of these moves for non-differentiable functions is not an easy problem.

This paper introduces a modified HMC algorithm allowing us to sample from possibly non-differentiable energy functions. The objective of this algorithm is therefore to be applicable to both differentiable and non-differentiable energy functions.

The so called non-smooth HMC (ns-HMC) sampling scheme relies on a modified leapfrog transform [13, 14] that circumvents the non-differentiability of the target energy function. The modified leapfrog transform relies on the sub-differential and proximity operator concepts [25]. The proposed scheme is validated on a sampling example where samples are drawn from a GG distribution with different shape

L. Chaari, J.-Y. Tourneret and H. Batatia are with the University of Toulouse, IRIT - INP-ENSEEIHT (UMR 5505), 2 rue Charles Camichel, BP 7122, Toulouse Cedex 7 France. E-mail: firstname.lastname@enseeiht.fr. L. Chaari is also with the MIRACL laboratory (Univ. of Sfax, Tunisia).

C. Chaux is with I2M and CNRS UMR 7373, Aix-Marseille University, 39 rue F. Joliot-Curie, 13453 Marseille Cedex 13, France. E-mail: caroline.chaux@univ-amu.fr.

This work was supported by the CNRS ImagIn project under grant OPTIMISME.

parameters. It is also applied to a signal recovery problem where a sparse regularization scheme is used to recover a high-dimensional signal.

The remainder of the paper is organized as follows. Section II formulates the problem of non-smooth sampling for large data using Hamiltonian dynamics. Section III presents the proposed ns-HMC sampling scheme. This technique is then validated in Section IV to illustrate its efficiency for sampling from non-smooth distributions. Finally, some conclusions and perspectives are drawn in Section V.

II. Problem formulation

Let us consider a signal of interest $\boldsymbol{x} \in \mathbb{R}^N$ and let $f(\boldsymbol{x};\boldsymbol{\theta})$ be its probability density function (pdf) which is parametrized by the vector of parameters $\boldsymbol{\theta}$. In this work, we focus on an exponential family of distributions such that

$$f(\boldsymbol{x};\boldsymbol{\theta}) \propto \exp[-E_{\boldsymbol{\theta}}(\boldsymbol{x})] \tag{1}$$

where $E_{\boldsymbol{\theta}}(\boldsymbol{x})$ is the energy function. Precisely, we concentrate on sampling from the class of log-concave probability densities, where the energy function $E_{\boldsymbol{\theta}}$ is assumed to be convex but not necessarily differentiable. In addition, we will also make the assumption that E_θ belongs to $\Gamma_0(\mathbb{R})$, the class of proper lower semi-continuous convex functions from $\mathbb{R}$ to $]-\infty,+\infty]$. Finally, we will consider probability distributions from which direct sampling is not possible and requires the use of an acceptance-rejection step. Example II.1 presents the case of the GG distribution which satisfies the above mentioned assumptions.

Example II.1 *Let $\gamma > 0$ and $p \geq 1$ two real-positive scalars. The generalized Gaussian distribution* $\mathrm{GG}(x;\gamma,p)$ *is defined by the following probability density function*

$$\mathrm{GG}(x;\gamma,p) = \frac{p}{2\gamma^{1/p}\Gamma(1/p)} \exp\left(-\frac{|x|^p}{\gamma}\right) \tag{2}$$

for $x \in \mathbb{R}$.

Except for even values of p, such as $p = 2, 4, \ldots$, the energy function $E_{\boldsymbol{\theta}}(x) = \frac{|x|^p}{\gamma}$ is not differentiable (where $\boldsymbol{\theta} = (\gamma, p)$). In what follows, we are interested in efficiently drawing samples according to the probability distribution f defined in (1). The following section describes the proposed non-smooth sampling algorithm that can be used for this generation.

III. Non-smooth sampling

A. Hamiltonian Monte Carlo methods

HMC methods [13, 14, 16] are powerful tools that use the principle of Hamiltonian dynamics and energy preservation. The theory of Hamiltonian dynamics is a reformulation of the theory of classical mechanics. It is generally used to model dynamic physical systems [26]. Let us consider the one-dimensional case to explain the principle of HMC methods. A dynamic particle of mass m can be characterized by its position x and momentum $q = mv$, where $v = \dfrac{\partial x}{\partial t}$ is the velocity of the particle (∂ denotes the partial derivative). The Hamiltonian models the total energy of this particle, namely the *potential* energy $E(x)$ and the *kinetic* energy $K(v) = \dfrac{1}{2}mv^2$, which can also be expressed as a function of the momentum since $K(q) = \dfrac{1}{2m}q^2$. The Hamiltonian $H(x,q)$ can be expressed as

$$H(x,q) = E(x) + K(q). \tag{3}$$

The Hamiltonian's motion equations determine the evolution of $x(t)$ as a function of time t [14]

$$\begin{aligned} \frac{dq}{dt} &= \frac{\partial H}{\partial x} \\ \frac{dx}{dt} &= -\frac{\partial H}{\partial q}. \end{aligned} \tag{4}$$

These equations define a transformation $\mathcal{F}_s$ that maps the state of the system at time t to the state at time $t+s$.
In the multidimensional case (where the particle has a unitary mass), Hamiltonian dynamics are used to sample from a target distribution $f(\boldsymbol{x};\boldsymbol{\theta})$ departing from a given position $\boldsymbol{x}$ by introducing an auxiliary momentum variable $\boldsymbol{q}$. The pdf of the Hamiltonian dynamics energy defined in (3) is given by

$$\begin{aligned} f_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{q}) &\propto \exp\left[-H(\boldsymbol{x},\boldsymbol{q})\right] \\ &\propto f(\boldsymbol{x};\boldsymbol{\theta}) \exp\left(-\frac{\boldsymbol{q}^\intercal\boldsymbol{q}}{2}\right). \end{aligned} \tag{5}$$

HMC methods iteratively proceed by alternating updates of samples $\boldsymbol{x}$ and $\boldsymbol{q}$ drawn according to the distribution (5). At iteration $\#r$, the HMC algorithm starts with the current values of vectors $\boldsymbol{x}^{(r)}$ and $\boldsymbol{q}^{(r)}$. Two steps have then to be performed. The first updates the momentum vector leading to $\boldsymbol{q}^{(r)}$ by sampling according to the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_N)$, where $\boldsymbol{I}_N$ is the $N \times N$ identity matrix. The second step updates both momentum $\boldsymbol{q}$ and position $\boldsymbol{x}$ by proposing two candidates $\boldsymbol{x}^*$ and $\boldsymbol{q}^*$. These two candidates are generated by simulating the Hamiltonian dynamics, which are discretized using some discretization techniques such as the leapfrog method. For instance, the discretization can be performed using L_f steps of the leapfrog method with a stepsize $\epsilon > 0$. The parameter L_f can either be manually fixed or automatically tuned such as in [27].
The lth leapfrog discretization will be denoted by T_s and can be summarized as follows

$$\boldsymbol{q}^{(r,(l+\frac{1}{2})\epsilon)} = \boldsymbol{q}^{(r,l\epsilon)} - \frac{\epsilon}{2}\frac{\partial E_{\boldsymbol{\theta}}}{\partial \boldsymbol{x}^\intercal}\left(\boldsymbol{x}^{(r,l\epsilon)}\right) \tag{6}$$

$$\boldsymbol{x}^{(r,(l+1)\epsilon)} = \boldsymbol{x}^{(r,l\epsilon)} + \epsilon\boldsymbol{q}^{(r,(l+\frac{1}{2})\epsilon)} \tag{7}$$

$$\boldsymbol{q}^{(r,(l+1)\epsilon)} = \boldsymbol{q}^{(r,(l+\frac{1}{2})\epsilon)} - \frac{\epsilon}{2}\frac{\partial E_{\boldsymbol{\theta}}}{\partial \boldsymbol{x}^\intercal}\left(\boldsymbol{x}^{(r,(l+1)\epsilon)}\right). \tag{8}$$

After the L_f steps, the proposed candidates are given by $\boldsymbol{q}^* = \boldsymbol{q}^{(r,\epsilon L_f)}$ and $\boldsymbol{x}^* = \boldsymbol{x}^{(r,\epsilon L_f)}$. These candidates are then accepted using the standard MH rule, i.e., with the following probability

$$\min\left\{1, \exp\left[H(\boldsymbol{x}^{(r)},\boldsymbol{q}^{(r)}) - H(\boldsymbol{x}^*,\boldsymbol{q}^*)\right]\right\} \tag{9}$$

where H is the Hamiltonian defined in (3).

B. Non-smooth Hamiltonian Monte Carlo schemes

The key step in standard HMC sampling schemes is the approximation of the Hamiltonian dynamics. This approximation allows the random simulation of uncorrelated samples according to a target distribution while exploiting the geometry of its corresponding energy. In this section, we propose two non-smooth Hamiltonian Monte Carlo (ns-HMC) schemes to perform this approximation for non-smooth energy functions. The first scheme is based on the subdifferential operator while the second one is based on proximity operators. For both schemes, the whole algorithm to sample $\boldsymbol{x}$ and $\boldsymbol{q}$ is detailed in Algorithms 1 and 2. These algorithms describe all the necessary steps to sample from a log-concave target distribution.

1) Scheme 1 - subdifferential based approach:
Let us first give the definition of the sub-differential and a useful example.

Definition III.1 *[25, p. 223] Let φ be in $\Gamma_0(\mathbb{R})$. The sub-differential of φ is the set $\partial_s\varphi(x) = \{\rho \in \mathbb{R}|\ \varphi(\eta) \geq \varphi(x) + \langle \rho|\eta - x\rangle\ \forall \eta \in \mathbb{R}\}$, where $\langle\cdot|\cdot\rangle$ defines the standard scalar product. Every element $\rho \in \partial_s\varphi(x)$ is a sub-gradient of φ at point x. If φ is differentiable, the sub-differential reduces to its gradient: $\partial_s\varphi(x) = \{\nabla\varphi(x)\}$.*

Example III.1 *Let φ be defined as*

$$\begin{aligned}\varphi:\ \mathbb{R} &\longmapsto \mathbb{R}\\ x &\longrightarrow |x|.\end{aligned} \tag{10}$$

The sub-differential of φ at x is defined by

$$\partial_s\varphi(x) = \begin{cases}\{\text{sign}(x)\} & \text{if } x \neq 0\\ [-1, 1] & \text{if } x = 0.\end{cases} \tag{11}$$

In addition, if we consider a scalar $\lambda \in \mathbb{R}_+$ and we call $\varphi_\lambda(\cdot) = \lambda\varphi(\cdot)$, then we have $\partial_s\varphi_\lambda(x) = \lambda\partial_s\varphi(x)$ for every $x \in \mathbb{R}$ [25, Prop. 16.5].

For distributions with smooth energy, one can use the leapfrog method whose basic form requires to compute the gradient of the potential energy $E_{\boldsymbol{\theta}}(\boldsymbol{x})$. Since we cannot determine this gradient for non-smooth energy functions, we resort to the following reformulation of the leapfrog scheme by using the concept of sub-differential introduced hereabove

$$\boldsymbol{q}^{(r,(l+\frac{1}{2})\epsilon)} = \boldsymbol{q}^{(r,l\epsilon)} - \frac{\epsilon}{2}\rho\left(\boldsymbol{x}^{(r,l\epsilon)}\right) \tag{12}$$

$$\boldsymbol{x}^{(r,(l+1)\epsilon)} = \boldsymbol{x}^{(r,l\epsilon)} + \epsilon\boldsymbol{q}^{(r,(l+\frac{1}{2})\epsilon)} \tag{13}$$

$$\boldsymbol{q}^{(r,(l+1)\epsilon)} = \boldsymbol{q}^{(r,(l+\frac{1}{2})\epsilon)} - \frac{\epsilon}{2}\rho\left(\boldsymbol{x}^{(r,(l+1)\epsilon)}\right) \tag{14}$$

where $\rho \in \partial_s E_{\boldsymbol{\theta}}$ is sampled uniformly in the sub-differential of $E_{\boldsymbol{\theta}}$. This discretization scheme will be denoted by T_s'. If $E_{\boldsymbol{\theta}}(\boldsymbol{x})$ is differentiable, the mapping T_s' in (12), (13) and (14) exactly matches the conventional HMC mapping T_s in (6), (7) and (8).

As for the standard HMC scheme, the proposed candidates are defined by $\boldsymbol{q}^* = \boldsymbol{q}^{(r,\epsilon L_f)}$ and $\boldsymbol{x}^* = \boldsymbol{x}^{(r,\epsilon L_f)}$ that can be computed after L_f leapfrog steps. These candidates are then accepted based on the standard MH rule defined in (9). The resulting sampling algorithm is summarized in Algorithm 1.

Algorithm 1: Gibbs sampler using Hamiltonian dynamics for non-smooth log-concave probability distributions: Scheme 1.

- Initialize with some $\boldsymbol{x}^{(0,0)}$.
- Set the iteration number $r = 0$, L_f and ϵ;
- **for** $r = 1 \dots S$ **do**
  - Sample $\boldsymbol{q}^{(r,0)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_N)$;
  - Compute $\boldsymbol{q}^{(r,\frac{1}{2}\epsilon)} = \boldsymbol{q}^{(r,0)} - \frac{\epsilon}{2}\rho(\boldsymbol{x}^{(r-1,0)})$;
  - Compute $\boldsymbol{x}^{(r,\epsilon)} = \boldsymbol{x}^{(r-1,0)} + \epsilon\boldsymbol{q}^{(r,\frac{1}{2}\epsilon)}$;
  - **for** $l_f = 1$ *to* $L_f - 1$ **do**
    - Compute $\boldsymbol{q}^{(r,(l_f+\frac{1}{2})\epsilon)} = \boldsymbol{q}^{(r,l_f\epsilon)} - \frac{\epsilon}{2}\rho(\boldsymbol{x}^{(r,l_f\epsilon)})$;
    - Compute $\boldsymbol{x}^{(r,(l_f+1)\epsilon)} = \boldsymbol{x}^{(r,l_f\epsilon)} + \epsilon\boldsymbol{q}^{(r,(l_f+\frac{1}{2})\epsilon)}$;
  - **end**
  - Compute $\boldsymbol{q}^{(r,(L_f+\frac{1}{2})\epsilon)} = \boldsymbol{q}^{(r,L_f\epsilon)} - \frac{\epsilon}{2}\rho(\boldsymbol{x}^{(r,L_f\epsilon)})$;
  - Apply standard MH acceptation/rejection rule by taking $\boldsymbol{q}^* = \boldsymbol{q}^{(r,\epsilon L_f)}$ and $\boldsymbol{x}^* = \boldsymbol{x}^{(r,\epsilon L_f)}$;
- **end**

Note that we do not need to account for any additional term in the acceptance ratio in (9) since volume preservation is ensured by the Metropolis update. Volume preservation is equivalent to having an absolute value of the Jacobian matrix determinant for the mapping T_s equal to one. This is due to the fact that candidates are proposed according to Hamiltonian dynamics. More precisely, volume preservation can be easily demonstrated by using the concept of Jacobian matrix approximation [28] such as the *Clarke* generalization [29], and by conducting calculations similar to [14, Chapter 5, p. 118].

2) Scheme 2 - proximal based approach:
Since the calculation of the subdifferential is not straightforward for some classes of convex functions, a second scheme modifying the leapfrog steps (12), (13) and (14) can be considered by using the concept of *proximity operators*. These operators have been found to be fundamental in a number of recent works in convex optimization [30–32], and more recently in [33] where stochastic proximal algorithms have been investigated. Let us first recall the proximity operator definition.

Definition III.2 *[25, Definition 12.23][34] Let $\varphi \in \Gamma_0(\mathbb{R})$. For every $x \in \mathbb{R}$, the function $\varphi + \|\cdot - x\|^2/2$ reaches its infimum at a unique point referred to as proximity operator and denoted by* $\text{prox}_\varphi(x)$.

Example III.2 *For the function φ defined in Example III.1, the proximity operator is given by*

$$\text{prox}_\varphi(x) = \text{sign}(x)\max\{|x| - 1, 0\}\ \forall x \in \mathbb{R}. \tag{15}$$

Many other examples and interesting properties that make this tool very powerful and commonly used in the recent optimization literature are given in [35]. One of these properties in which we are interested here is the following.

Property 1 *[36, Prop. 3] Let $\varphi \in \Gamma_0(\mathbb{R})$ and $x \in \mathbb{R}$. There exists a unique point $\widehat{x} \in \mathbb{R}$ such that $x - \widehat{x} \in \partial_s \varphi(\widehat{x})$. Using the proximity operator definition hereabove, it turns out that $\widehat{x} = \text{prox}_{\varphi}(x)$.*

By modifying the discretization scheme T_s (Eqs. (6)-(8)), we propose the following l-th leapfrog discretization scheme denoted by T_s''

$$\boldsymbol{q}^{(r,(l+\frac{1}{2})\epsilon)} = \boldsymbol{q}^{(r,l\epsilon)} - \frac{\epsilon}{2}\left[\boldsymbol{x}^{(r,l\epsilon)} - \text{prox}_{E_{\boldsymbol{\theta}}}(\boldsymbol{x}^{(r,l\epsilon)})\right] \quad (16)$$

$$\boldsymbol{x}^{(r,(l+1)\epsilon)} = \boldsymbol{x}^{(r,l\epsilon)} + \epsilon \boldsymbol{q}^{(r,(l+\frac{1}{2})\epsilon)} \quad (17)$$

$$\boldsymbol{q}^{(r,(l+1)\epsilon)} = \boldsymbol{q}^{(r,(l+\frac{1}{2})\epsilon)} - \frac{\epsilon}{2}\times \left[\boldsymbol{x}^{(r,(l+1)\epsilon)} - \text{prox}_{E_{\boldsymbol{\theta}}}(\boldsymbol{x}^{(r,(l+1)\epsilon)})\right]. \quad (18)$$

If $E_{\boldsymbol{\theta}}(\boldsymbol{x})$ is differentiable, the mapping T_s'' in (16), (17) and (18) exactly matches the mapping T_s in (6), (7) and (8). The only difference is that the sub-differential of the mapping T_s'' is evaluated in $\text{prox}_{E_{\boldsymbol{\theta}}}(\boldsymbol{x})$ instead of $\boldsymbol{x}$. As for scheme 1, the proposed candidates are given by $\boldsymbol{q}^* = \boldsymbol{q}^{(r,\epsilon L_f)}$ and $\boldsymbol{x}^* = \boldsymbol{x}^{(r,\epsilon L_f)}$ after L_f leapfrog steps. These candidates are then accepted based on the standard MH rule (9).

The Gibbs sampler resulting from the transformation T_s'' is summarized in Algorithm 2. Similarly to Algorithm 1, and due to the presence of the MH acceptance rule, the elements $\boldsymbol{x}^{(r)}$ generated by this algorithm are asymptotically distributed according to the target distribution $f(\boldsymbol{x};\boldsymbol{\theta})$ defined in (1).

Algorithm 2: Gibbs sampler using Hamiltonian dynamics for non-smooth log-concave probability distributions.

- Initialize with some $\boldsymbol{x}^{(0,0)}$.
- Set the iteration number $r = 0$, L_f and ϵ;
- **for** $r = 1, \ldots, S$ **do**
  - Sample $\boldsymbol{q}^{(r,0)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_N)$;
  - Compute $\boldsymbol{q}^{(r,\frac{1}{2}\epsilon)} = \boldsymbol{q}^{(r,0)} - \frac{\epsilon}{2}\left[\boldsymbol{x}^{(r-1,0)} - \text{prox}_{E_{\boldsymbol{\theta}}}(\boldsymbol{x}^{(r-1,0)})\right]$;
  - Compute $\boldsymbol{x}^{(r,\epsilon)} = \boldsymbol{x}^{(r-1,0)} + \epsilon \boldsymbol{q}^{(r,\frac{1}{2}\epsilon)}$;
  - **for** $l_f = 1$ *to* $L_f - 1$ **do**
    - Compute $\boldsymbol{q}^{(r,(l_f+\frac{1}{2})\epsilon)} = \boldsymbol{q}^{(r,l_f\epsilon)} - \frac{\epsilon}{2}\left[\boldsymbol{x}^{(r,l_f\epsilon)} - \text{prox}_{E_{\boldsymbol{\theta}}}(\boldsymbol{x}^{(r,l_f\epsilon)})\right]$;
    - Compute $\boldsymbol{x}^{(r,(l_f+1)\epsilon)} = \boldsymbol{x}^{(r,l_f\epsilon)} + \epsilon \boldsymbol{q}^{(r,(l_f+\frac{1}{2})\epsilon)}$;
  - **end**
  - Compute $\boldsymbol{q}^{(r,(L_f+\frac{1}{2})\epsilon)} = \boldsymbol{q}^{(r,L_f\epsilon)} - \frac{\epsilon}{2}\left[\boldsymbol{x}^{(r,L_f\epsilon)} - \text{prox}_{E_{\boldsymbol{\theta}}}(\boldsymbol{x}^{(r,L_f\epsilon)})\right]$;
  - Apply standard MH acceptation/rejection rule by taking $\boldsymbol{q}^* = \boldsymbol{q}^{(r,\epsilon L_f)}$ and $\boldsymbol{x}^* = \boldsymbol{x}^{(r,\epsilon L_f)}$;
- **end**

C. Discussions

1) Comparison of the two schemes:
Fig. 1 illustrates the use of the proposed discretization schemes (algorithms 1 and 2) to approximate a Hamiltonian made up of a quadratic kinetic energy and a potential energy having the following form

$$E_{a,b}(x) = a|x| + bx^2 \quad (19)$$

where $(a, b) \in (\mathbb{R}_+^*)^2$. For this potential energy, the subdifferential can be analytically calculated and is given by

$$\partial_s E_{a,b} = a\partial_s\varphi + (2b)\text{Id} \quad (20)$$

where $\partial_s\varphi$ is defined in Example III.1 and Id is the identity operator. Both proposed algorithms can therefore be compared for this example.

Fig. 1 shows that the discretized energy is close to the continuous one for the two mappings T_s' and T_s''. Moreover, the slight difference $(T_s'' - T_s')$ between the two mappings shows that the two discretization schemes perform very similarly close to the critical region of non-differentiability (the interval $[-\varepsilon, \varepsilon]$ with small $\varepsilon \in \mathbb{R}_+$, see the zoom around the origin in Fig. 1). Fig. 2 illustrates the shape of the proximity operator for the considered energy function $E_{a,b}$, as well as the identity function Id and the difference $\text{Id} - \text{prox}_{E_{a,b}}$. This figure clearly shows that, due to the thresholding property of the proximity operator, we have $x \simeq x - \text{prox}_{E_{a,b}} x$ for $x \in [-\varepsilon, \varepsilon]$. In particular, for the considered example, we have $x \simeq x - \text{prox}_{E_{a,b}} x$ for every $x \in [\frac{-a}{b+1}, \frac{a}{b+1}]$. This comparison confirms that the two schemes perform similarly especially close to the non-differentiability point.

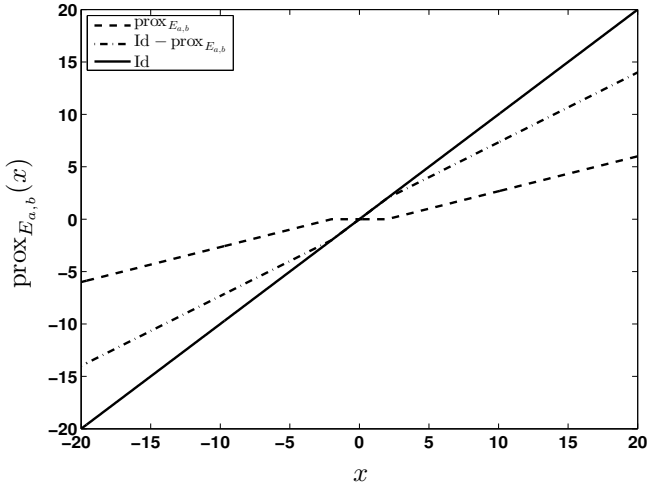

Fig. 2. The proximity operator $\text{prox}_{E_{a,b}}$, the identity function (Id) and the difference $\text{Id} - \text{prox}_{\text{E,b}}$ for $a = b = 2$.

Since Algorithm 2 is more general than Algorithm 1 (the proximity operator is generally unique with a closed-form) and allows us to handle energies for which the sub-differential is not straightforward (while performing well especially close to the critical regions), we will focus on this second discretization scheme for our experiments.

2) Convergence analysis:
The convergence conditions of the proposed sampling scheme are discussed in this section. Since the proposed scheme relies on an MH acceptance step with an infinite support of the proposal distribution (which therefore includes the support of the target distribution), ensuring volume preservation of the discretization scheme suffices to guarantee the convergence of the proposed scheme.
From a geometric point of view, it is worth to note that the two modified leapfrog discretization schemes T_s' and T_s'' defined

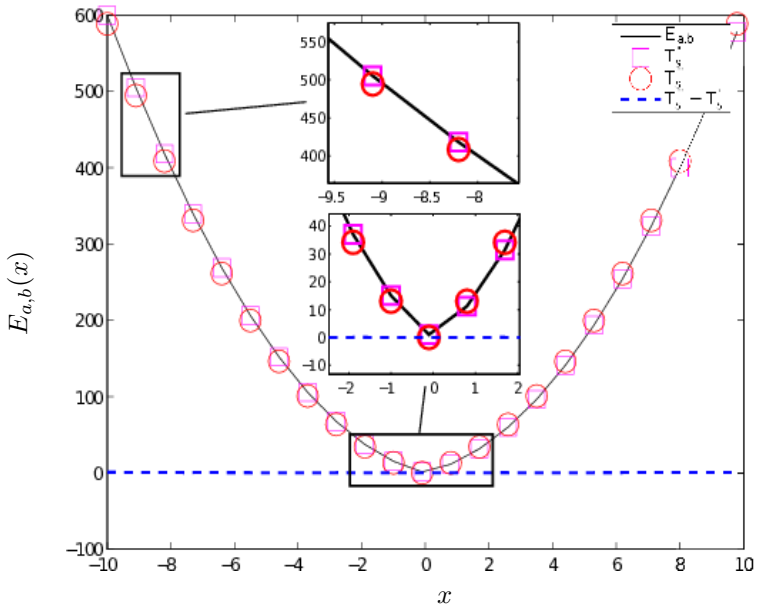

Fig. 1. The potential energy $E_{a,b}$ (solid black line) in (19) ($a = 10, b = 5$) and its discretizations using the modified leapfrog schemes T_s' (squares) and T_s'' (circles), as well as the difference between the two discretizations $T_s'' - T_s'$ (dashed blue line).

respectively in (12)-(14) and (16)-(18), as well as the original leapfrog scheme defined in (6)-(8), preserve volume since they are shear transformations. The interested reader can refer to [14] or [37, page 121] for more details.

Analytically speaking, volume preservation can also be demonstrated by using the generalization of the Jacobian matrix which is defined using the sub-gradient instead of the gradient itself. Let us denote by $\mathcal{F}_\delta$ (see Section III-A) the mapping between the state at time t, denoted by $(x(t), q(t))$, and the state $(x(t+\delta), q(t+\delta))$ at time $t+\delta$. Without loss of generality, we consider here the one-dimensional case since the multi-dimensional case can be handled through simple generalizations. Developments similar to [14] lead to the following form of the generalized Jacobian matrix for the one-dimensional case

$$\mathcal{J}_\delta = \begin{bmatrix} 1+\delta\dfrac{\partial_s^2 H_{\boldsymbol{\theta}}}{\partial_s q \partial_s x} & \delta\dfrac{\partial_s^2 H_{\boldsymbol{\theta}}}{\partial_s q^2} \\ -\delta\dfrac{\partial_s^2 H_{\boldsymbol{\theta}}}{\partial_s q^2} & 1-\delta\dfrac{\partial_s^2 H_{\boldsymbol{\theta}}}{\partial_s x \partial_s q} \end{bmatrix} + O(\delta^2) \qquad (21)$$

where ∂_s denotes the sub-gradient and $\dfrac{\partial_s^2 H_{\boldsymbol{\theta}}}{\partial_s q \partial_s x}$ is an element of the second-order sub-differential with respect to q and x. The determinant of this matrix can therefore be written as

$$\begin{aligned} \det(\mathcal{J}_\delta) &= 1 + \delta\frac{\partial_s^2 H_{\boldsymbol{\theta}}}{\partial_s q \partial_s x} - \delta\frac{\partial_s^2 H_{\boldsymbol{\theta}}}{\partial_s x \partial_s q} + O(\delta^2) \\ &= 1 + O(\delta^2). \end{aligned} \qquad (22)$$

Following the construction proposed in [14], it turns out that for some time interval s that is not close to zero, $\det(\mathcal{J}_s) = 1$. Since the transformation $\mathcal{F}_s$ is reversible (by replacing the stepsize by its opposite), it then preserves volume. Hence, the determinant of the Jacobian matrix does not need to be involved in the MH acceptance probability. Therefore, for the deterministic mapping $\mathcal{F}_s$, the corresponding acceptance probability is given by (9). Under the reversibility condition of the dynamics, the joint density as well as the marginals $f(\boldsymbol{x})$ and $f(\boldsymbol{q})$ are left invariant. Moreover, and as explained in a number of works such as [8], HMC schemes for $f(\boldsymbol{x})$ can be interpreted as a Gibbs sampler with an auxiliary variable $\boldsymbol{q}$. The proposed scheme produces therefore an ergodic and time reversible Markov chain whose stationary distribution is $f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{q})$ with marginal distribution $f(\boldsymbol{x})$.

3) Effectiveness analysis:

We give here a theoretical analysis of the effectiveness of the proposed sampling scheme. Combining (16) and (17) in a single step yields the following update form

$$\boldsymbol{x}^{(r,(l+1)\epsilon)} = \boldsymbol{x}^{(r,l\epsilon)} + \epsilon\boldsymbol{q}^{(r,l\epsilon)} - \frac{\epsilon^2}{2}\left[\boldsymbol{x}^{(r,l\epsilon)} - \text{prox}_{E_{\boldsymbol{\theta}}}(\boldsymbol{x}^{(r,l\epsilon)})\right] \qquad (23)$$

which can also be rewritten as

$$\boldsymbol{x}^{(r,(l+1)\epsilon)} = \frac{\epsilon^2}{2}\left[\frac{(1-\frac{\epsilon^2}{2})}{\frac{\epsilon^2}{2}}\boldsymbol{x}^{(r,l\epsilon)} + \text{prox}_{E_{\boldsymbol{\theta}}}(\boldsymbol{x}^{(r,l\epsilon)})\right] + \epsilon\boldsymbol{q}^{(r,l\epsilon)}. \qquad (24)$$

One can notice that this update scheme is similar to a random walk step with random $\boldsymbol{q}$ updated around the point

$$\boldsymbol{x}^{\#} = \frac{\epsilon^2}{2}\left[\frac{(1-\frac{\epsilon^2}{2})}{\frac{\epsilon^2}{2}}\boldsymbol{x}^{(r,l\epsilon)} + \text{prox}_{E_{\boldsymbol{\theta}}}(\boldsymbol{x}^{(r,l\epsilon)})\right]. \qquad (25)$$

Using the definition of the proximity operator given in Definition III.2, we can write

$$\forall \boldsymbol{y} \in \mathbb{R}^N,\ \mathrm{prox}_{E_{\boldsymbol{\theta}}}(\boldsymbol{x}^{(r,l\epsilon)}) = \arg\inf_{\boldsymbol{y}} E_{\boldsymbol{\theta}}(\boldsymbol{y}) + \|\boldsymbol{y}-\boldsymbol{x}\|^2/2. \tag{26}$$

If the infimum is reached, we can write

$$\forall \boldsymbol{y} \in \mathbb{R}^N,\ \mathrm{prox}_{E_{\boldsymbol{\theta}}}(\boldsymbol{x}^{(r,l\epsilon)}) = \arg\min_{\boldsymbol{y}\in\mathbb{R}^N} E_{\boldsymbol{\theta}}(\boldsymbol{y}) + \|\boldsymbol{y}-\boldsymbol{x}\|^2/2 \tag{27}$$

which can be interpreted as a regularized minimization of the target distribution energy function $E_{\boldsymbol{\theta}}$. The term $\boldsymbol{x}^{\#}$ can therefore be seen as a linear combination of the current point $\boldsymbol{x}$ and the point of minimal energy value (due to the minimization of the energy function). As a consequence, it turns out that the proposed scheme reduces to a random walk that is applied not around the current point as done in the standard random walk Metropolis-Hastings (rw-MH) algorithm, but around a more optimal point which provides a good compromise between the current state and the state of minimal energy (i.e., the maximal value of the target probability density function).

As regards computational costs, the proposed scheme presents the same level of complexity than the standard HMC scheme. Indeed, the modified leapfrog transform relies on the calculation of the proximity operator, *whose cost is not necessarily higher than that of calculating a gradient*. For example, let us consider the function

$$\begin{aligned}\varphi:\ \mathbb{R} &\longmapsto \mathbb{R}\\ x &\longrightarrow \alpha x^2.\end{aligned} \tag{28}$$

The gradient of φ is given by $\nabla\varphi(x) = 2\alpha x$, while the proximity operator is $\mathrm{prox}_{\varphi}(x) = x/(2\alpha+1)$. *For this example, the gradient and proximal operators can be computed with similar complexity.* Other examples of proximity operator calculations are available in [31, 35].

IV. Experimental validation

This section validates the proposed ns-HMC scheme for non-smooth log-concave distributions through three experiments. The two first experiments consider the GG distribution whose energy function is non-differentiable for the values of the shape parameter considered here ($p = 1$ and $p = 1.5$). For the third experiment, a Laplace distribution (GG distribution with $p = 1$) is used for an image denoising problem where the clean image is recovered from noisy measurements using a Bayesian regularization scheme involving a sampling technique based on the proposed ns-HMC algorithm.

A. Experiment 1: 1D sampling

In the first experiment, a 1D sampling is performed for a given configuration of the shape and scale parameters of a GG distribution ($p = \lambda = 1$). Chains generated using the proposed ns-HMC sampling scheme are compared to the ones obtained with an rw-MH scheme. For our ns-HMC, the number of leapfrog steps L_f has been empirically set to 10. The stepsize ϵ of the algorithm has been set to $1/L_f$. Indeed, on the one hand, a too large stepsize leads to a low acceptation ratio. On the other hand, a too small stepsize leads to a slow exploration of the target space and thus decreases the convergence rate of the method. Choosing $\epsilon = 1/L_f$ guarantees a reasonable trajectory length ϵL_f. The rw-MH strategy is used here for comparison since it generally improves the mixing properties of the generated samples when compared to a fixed proposal distribution. Let $x^{(r)}$ be the current sample of the chain and x^* the proposed one. A Gaussian proposal centered on the current sample with unitary variance is used for the rw-MH algorithm, i.e., $x^* \sim \mathcal{N}(x^{(r)}, 1)$. Fig. 3[top] displays the Kullback-Leibler (KL) divergence between the target GG pdf (with $p = 1$ and $\lambda = 1$) and the histogram of the generated samples with respect to the number of sampled coefficients. Note that the different curves have been obtained by averaging the outputs of 50 Monte Carlo (MC) runs. Errorbars indicate the standard deviation around this mean. To further illustrate the sampling efficiency of the ns-HMC algorithm, Fig. 3[bottom] displays the autocorrelation functions (ACFs) of the sampled chains for the same values of (p, λ). This figure clearly shows that samples generated using the ns-HMC scheme are less correlated than those generated using rw-MH, which corroborates the faster convergence of the ns-HMC scheme. Note that the proposed technique does not need any adjustment of the proposal variance contrary to the rw-HM algorithm while giving acceptable level of intra-chain correlation. For the sake of comparison, Fig. 3[bottom] also displays the ACFs of chains sampled using a standard MH algorithm with a centered Gaussian proposal ($x^* \sim \mathcal{N}(0, 1)$). Indeed, it has been reported that rw-MH increases the correlation level within sampled chains [5], while an MH algorithm provides uncorrelated samples. The comparison between the ACFs corresponding to ns-HMC and MH shows that chains sampled using ns-HMC are as less correlated as the standard MH algorithm with $\mathcal{N}(0, 1)$ proposal.

B. Experiment 2: sampling from multivariate distributions

This section studies different algorithms for sampling according to a multivariate GG distribution. The scale and shape parameters of this GG distribution have been adjusted to the values of Experiment 1. Multimensional sampling is performed in each simulation. However, to evaluate the convergence rate of the algorithms, we compute the KL divergence between the marginal one-dimensional GG pdfs and the histograms of the samples generated by the different algorithms. The obtained KL divergence is then averaged over all the dimensions to get the mean value. Over 50 MC runs, Fig. 4 displays the averages of the KL divergences and the corresponding error bars (mean $\pm$ standard deviations) w.r.t the iteration number. In addition to the results provided by the rw-MH algorithm, and for the sake of comparison with other existing algorithms that are adapted to the multidimensional case, the KL divergence curves are also provided for the elliptical slice sampling (ESS) [10] technique.

Fig. 4 shows that the convergence to the marginal distributions is faster with ns-HMC than with rw-MH or ESS. It is also worth noticing that the variance over the 50 MC realizations is lower with ns-HMC. An interesting property of the ns-HMC

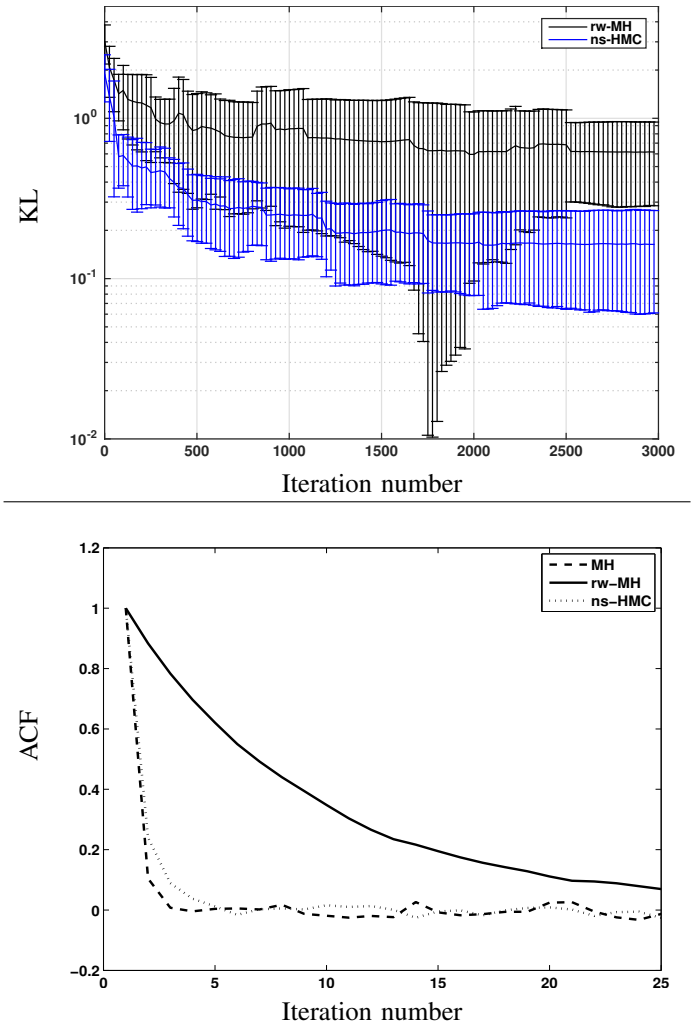

Fig. 3. 1D GG sampling with $p = 1$ and $\lambda = 1$. Top: mean KL divergence (with standard deviation) between the target pdf and the histogram of the generated samples using the MH, rw-MH and ns-HMC algorithms (Logarithmic scale); bottom: ACFs of the sampled chains using the MH and the rw-MH algorithms, in addition to the proposed ns-HMC method.

method is that its convergence rate does not depend on the dimension of the problem, contrary to rw-MH. This stability is due to the fact that HMC exploits the shape of the energy function in contrary to the rw-MH algorithm.
These comparisons confirm the usefulness of the proposed ns-HMC scheme especially in high-dimensional scenarios where the convergence speed of the standard MH, rw-MH or ESS algorithms is altered by the size of the data.

C. Experiment 3: sampling from a Bernoulli-GG distribution

In this experiment, we want to sample a vector $\boldsymbol{x}$ distributed according to a 2D Bernoulli-GG distribution, i.e., such that

$$\forall \boldsymbol{x} \in \mathbb{R}^2,\ f(\boldsymbol{x};\omega,\lambda,p) = \omega\delta(\mathbf{0}) + (1-\omega)GG(\boldsymbol{x};\lambda,p)$$

where $\delta(\cdot)$ is the Dirac delta function. The aim of this example is to investigate the performance of the ns-HMC algorithm when the target energy function has a non-diifferentiable point ($\mathbf{x} = \mathbf{0}$) that is reached with a non-zero probability. All simulations were performed with $\omega = 0.6$, which denotes

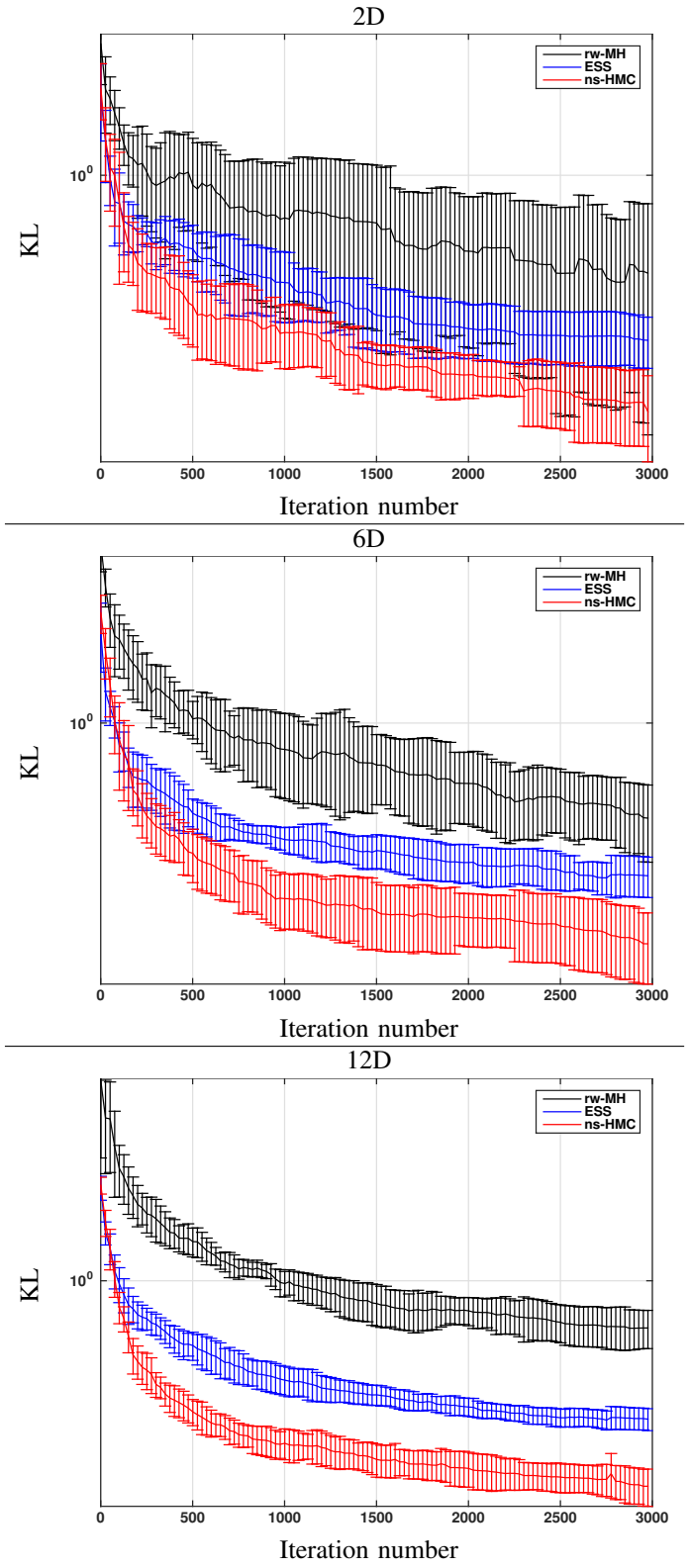

Fig. 4. Mean KL divergence (w.r.t iteration number) between the target GG pdf and the histogram of the generated samples using the rw-MH, ESS and ns-HMC algorithms for multidimensional signals: 2D, 6D and 12D cases (logarithmic scale). The errorbars indicate the estimated standard deviation around the mean values.

the probability of sampling $\boldsymbol{x} = \mathbf{0}$. The GG shape and scale parameters are the same as in experiment # 2. Fig. 5 shows the

average KL distances (computed using 50 MC runs) between the continuous part of the distribution and the histograms of the corresponding generated samples, with the corresponding error bars. The ns-HMC algorithm is clearly converging faster than the rw-MH and ESS methods. Comparing the results of Figs. 4 and 5, we can observe that the gain in terms of convergence rate is faster for a Bernoulli-GG distribution than for a multivariate GG distribution. Indeed, the modified Leapfrog scheme investigated in this paper based on proximal operators allows a faster convergence to be obtained.
An experiment where the target distribution has a infinite number of singularities is detailed in the technical report [38].

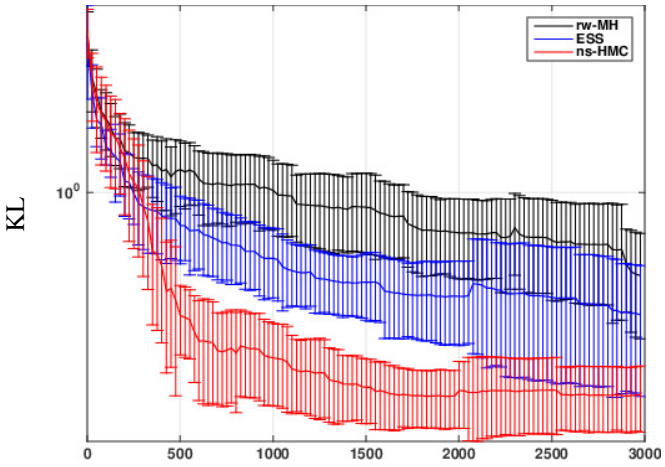

Fig. 5. Mean KL divergence (w.r.t iteration number) between the target GG pdf and the histogram of the generated samples using the rw-MH, ESS and ns-HMC algorithms for a 2D Bernoulli-GG distribution (logarithmic scale). The errorbars indicate the estimated standard deviation around the mean values.

D. Experiment 4: denoising

In this experiment, the performance of the proposed ns-HMC sampling algorithm is analyzed for an image denoising problem. The 2D image of size $N = 128 \times 128$ displayed in Fig. 6[top-left] has been used as a ground truth for this example. An independent identically distributed additive Gaussian noise of variance $\sigma_n^2 = 40$ has been added to this image to obtain the noisy image depicted in Fig. 6[top-right]. The objective of this third experiment is to promote the sparsity of the wavelet coefficients associated with the target image. To this end, we express the image formation model as function of the wavelet coefficients $\boldsymbol{x} \in \mathbb{R}^N$ which are related to the ground truth image $\boldsymbol{z}$ through the relation $\boldsymbol{z} = F^{-1}\boldsymbol{x}$ where $F^{-1} \in \mathbb{R}^{N\times N}$ denotes the dual frame operator. The analysis frame operator thus corresponds to $F \in \mathbb{R}^{N\times N}$ and as orthonormal bases are considered here, the dual frame operator reduces to the inverse operator yielding $F^{-1}F = FF^{-1} = \mathrm{Id}$. The observation model can thus be expressed as

$$\boldsymbol{y} = F^{-1}\boldsymbol{x} + \boldsymbol{n} \tag{29}$$

where $\boldsymbol{y} \in \mathbb{R}^N$ is the observed image, $\boldsymbol{x} \in \mathbb{R}^N$ contains the unknown wavelet coefficients and $\boldsymbol{n} \in \mathbb{R}^N$ is the additive noise. Note that the denoised image $\widehat{\boldsymbol{z}}$ can be easily recovered from the estimated wavelet coefficients $\widehat{\boldsymbol{x}}$ by taking $\widehat{\boldsymbol{z}} = F^{-1}\widehat{\boldsymbol{x}}$.
Based on this model and the Gaussian likelihood assumption, a hierarchical Bayesian model has been built using an independent Laplace prior for the wavelet coefficients [39, 40]

$$f(\boldsymbol{x};\lambda) = \left(\frac{1}{2\lambda}\right)^N \exp\left(-\frac{||\boldsymbol{x}||_1}{\lambda}\right) \tag{30}$$

where λ is an unknown parameter that is estimated within the proposed Bayesian algorithm. More precisely, an inverse gamma prior distribution is assigned to λ [23, 41]

$$f(\lambda|a,b) = \mathcal{IG}(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\lambda^{-a-1}\exp\left(-\frac{b}{\lambda}\right) \tag{31}$$

where $\Gamma(.)$ is the gamma function, and a and b are fixed hyperparameters (in our experiments these hyperparameters have been set to $a = b = 10^{-3}$).
Using a Jeffrey's prior for the noise variance ($\sigma_n^2 \sim \frac{1}{\sigma_n^2}1_{\mathbb{R}^+}(\sigma_n^2)$), the full posterior of this denoising model can be derived. The associated Gibbs sampler generates samples according to the conditional distributions of the posterior. The conditional distribution of the wavelet coefficients $\boldsymbol{x}$ writes

$$f(\boldsymbol{x}|\boldsymbol{y},\sigma_n^2,\lambda) \propto \exp\left[-U(\boldsymbol{x})\right] \tag{32}$$

where the energy function U is defined by $U(\boldsymbol{x}) = \frac{||\boldsymbol{x}||_1}{\lambda} + \frac{||\boldsymbol{y}-F^{-1}\boldsymbol{x}||_2^2}{2\sigma_n^2}$. Sampling according to this distribution is performed using the proposed ns-HMC scheme, which requires the calculation of the proximity operator of its energy function given by

$$\mathrm{prox}_U(\boldsymbol{x}) = \mathrm{prox}_{||\cdot||_1/(1+\alpha)}\left(\frac{\boldsymbol{x}+\alpha F\boldsymbol{y}}{1+\alpha}\right) \tag{33}$$

where $\alpha = \dfrac{1}{\sigma_n^2}$ and $\mathrm{prox}_{||\cdot||_1/(1+\alpha)}$ can easily be calculated using standard properties of the proximity operators [25, 31, 42] (see Appendix A for more details).
Regarding the noise variance and the prior hyperparameter, straightforward calculations lead to the following conditional distributions which are easy to sample

$$\sigma_n^2|\boldsymbol{x},\boldsymbol{y} \sim \mathcal{IG}\Big(\sigma_n^2|N/2, ||\boldsymbol{y}-F^{-1}\boldsymbol{x}||^2/2\Big) \tag{34}$$

$$\lambda|\boldsymbol{x},a,b \sim \mathcal{IG}\Big(\lambda|a+N, b+||\boldsymbol{x}||_1\Big) \tag{35}$$

where $\mathcal{IG}$ is the inverse gamma distribution. The estimation of the denoised image is performed based on the sampled wavelet coefficients after an appropriate burn-in period, i.e., after convergence of the Gibbs sampler in Algorithm 3.

An example of denoised image using Algorithm 3 is displayed in Fig. 6[bottom-left]. This result is compared with the Wiener filter, which is a state-of-the-art denoising method, (Fig. 6[bottom-right]). From a visual point of view, we can easily notice that Algorithm 3 provides a better denoised image compared to the Wiener filter. Quantitatively speaking, the evaluation of the noisy and denoised images is based on both SNR (signal to noise ratio) and SSIM [43] (structural similarity). These values are directly reported in the figure and show the efficiency of the denoising algorithm based on the proposed ns-HMC technique to sample from the conditional

Algorithm 3: Gibbs sampler for image denoising.

- Initialize with some $\boldsymbol{x}^{(0)}$.
for $r = 1, 2 \ldots$ **do**
  - Sample $\sigma_n^{2\,(r)}$ according to (34);
  - Sample $\lambda^{(r)}$ according to (35);
  - Sample $\boldsymbol{x}^{(r)}$ according to its conditional distribution using the proposed ns-HMC scheme;
end
- After convergence, compute the MMSE estimator $\widehat{\boldsymbol{x}}$ and return the estimated image $\widehat{\boldsymbol{z}} = F^{-1}\widehat{\boldsymbol{x}}$.

distribution of the wavelet coefficients $\boldsymbol{x}$. As regards the computational time, only 1000 iterations are necessary for the proposed algorithm involving a burn-in period of 500 iterations, taking around 9 seconds on a 64-bit 2.00GHz i7-3667U architecture with a Matlab implementation. For the ns-HMC step, the second scheme has been used with $L_f = 10$.

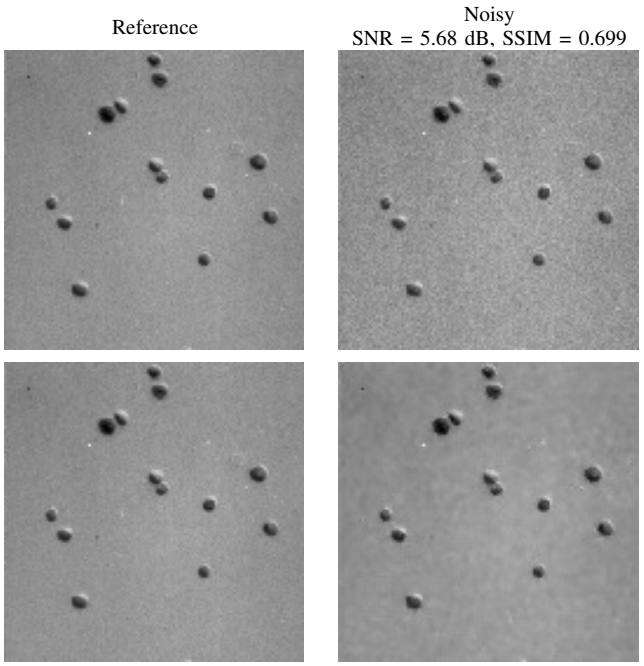

Fig. 6. Reference (top-left), noisy (top-right) and denoised images using Algorithm 3 (bottom-left) and the Wiener filter (bottom-right).

V. Conclusion

This paper proposed a solution to make feasible the use of Hamiltonian dynamics for sampling according to log-concave probability distributions with non-smooth energy functions. The proposed sampling technique relies on some interesting results from convex optimization and Hamiltonian Monte Carlo methods. More precisely, proximity operators were investigated to address the non-differentiability problem of the energy function related to the target distribution. The proposed technique provided faster convergence and interesting decorrelation properties for the sampled chains when compared to more standard methods such as the random walk Metropolis Hastings algorithm. The proposed technique was evaluated on synthetic data and applied to an image denoising problem. Our results showed that the use of proximity operators in a Hamiltonian Monte Carlo method allows faster convergence to the target distribution to be obtained. This conclusion is particularly important for large scale data sampling since the gain in convergence speed increases with the problem dimensionality. In a future work, we will focus on the investigation of this technique for sparse signal recovery where a non-tight linear operator is involved in the observation model. A preliminary attempt on synthetic data has already been performed in [44]. A complete comparison with other multidimensional sampling techniques would also be interesting.

Appendix

A. Proximity operator calculation for the experiment of Section IV-D

The energy function considered in this appendix is the one involved in the conditional distribution of the wavelet coefficients in (32), i.e.,

$$U(\boldsymbol{x}) = \frac{\alpha}{2}||\boldsymbol{y} - F^{-1}\boldsymbol{x}||_2^2 + \varphi(\boldsymbol{x}) \tag{36}$$

where $\alpha = 1/\sigma_n^2$ and $\varphi(\boldsymbol{x}) = \frac{||\boldsymbol{x}||_1}{\lambda}$. In order to use the proposed ns-HMC sampling algorithm, the proximity operator of the function U has to be calculated. Following the standard definition of the proximity operator [31, 34], we can write

$$\begin{aligned}
\text{prox}_U(\boldsymbol{x}) = \boldsymbol{p} \Leftrightarrow & \boldsymbol{x} - \boldsymbol{p} \in \partial U(\boldsymbol{p}) \\
\Leftrightarrow & \boldsymbol{x} - \boldsymbol{p} \in \partial\varphi(\boldsymbol{p}) + \alpha\boldsymbol{p} - \alpha F\boldsymbol{y} \\
\Leftrightarrow & \boldsymbol{x} + \alpha F\boldsymbol{y} - (\alpha+1)\boldsymbol{p} \in \partial\varphi(\boldsymbol{p}) \\
\Leftrightarrow & \frac{\boldsymbol{x} + \alpha F\boldsymbol{y}}{\alpha+1} - \boldsymbol{p} \in \partial\varphi/(\alpha+1)(\boldsymbol{p}) \\
\Leftrightarrow & \boldsymbol{p} = \text{prox}_{\varphi/(\alpha+1)}\left(\frac{\boldsymbol{x} + \alpha F\boldsymbol{y}}{\alpha+1}\right)
\end{aligned} \tag{37}$$

which proves the expresssion of the proximity operator given in (33).

References

[1] D.L Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[2] C. Chaux, P. L. Combettes, J.-C Pesquet, and V. Wajs, "Iterative image deconvolution using overcomplete representations," in *Proc. European Signal Processing Conference (EUSIPCO)*, Florence, Italy, Sep. 4-8 2006.

[3] L. Chaari, P. Ciuciu, S. Mériaux, and J.-C. Pesquet, "Spatio-temporal wavelet regularization for parallel MRI reconstruction: application to functional MRI," *Mag. Reson. Mater. in Phys., Biol. and Med. (MAGMA)*, vol. 27, pp. 509–529, 2014.

[4] L. Boubchir and B. Boashash, "Wavelet denoising based on the MAP estimation using the BKF prior with application to images and EEG signals," *IEEE Trans. Signal Process.*, vol. 61, no. 8, pp. 1880–1894, Apr. 2013.

[5] C. Robert and G. Castella, *Monte Carlo statistical methods*, Springer, New York, 2004.

[6] J.-F. Richard and Wei Zhang, "Efficient high-dimensional importance sampling," *J. Econom.*, vol. 141, no. 2, pp. 1385 – 1411, 2007.

[7] G. O. Roberts and R. L. Tweedie, "Exponential convergence of Langevin distributions and their discrete approximations," *Bernouilli*, vol. 1, no. 4, pp. 341–363, 1996.

[8] M. Girolami and B. Calderhead, "Riemann manifold Langevin and Hamiltonian Monte Carlo methods," *J. R. Statist. Soc. B*, vol. 73, no. 2, pp. 123214, 2011.

[9] M. Pereyra, "Proximal Markov chain Monte Carlo algorithms," *Statistics and Computing*, pp. 1–16, 2015.

[10] I. Murray, R. P. Adams, and D. MacKay, "Elliptical slice sampling," *Journal of Machine Learning Research*, vol. 9, pp. 541–548, 2010.

[11] F. Orieux, O. Fron, and J.-F. Giovannelli, "Sampling high-dimensional Gaussian distributions for general linear inverse problems," *IEEE Signal Process. Lett.*, vol. 19, no. 5, pp. 251–254, May 2012.

[12] C. Gilavert, S. Moussaoui, and J. Idier, "Efficient Gaussian Sampling for Solving Large-Scale Inverse Problems using MCMC Methods," *IEEE Trans. Signal Process.*, vol. 63, no. 1, Aug. 2014.

[13] K. M. Hanson, "Markov chain Monte Carlo posterior sampling with the Hamiltonian method," in *SPIE Medical Imaging: Image Processing, M. Sonka and K. M. Hanson, eds.*, 2001, pp. 456–467.

[14] R. M. Neal, "MCMC using Hamiltonian dynamics," in *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. Jones, and X. L. Meng, Eds., chapter 5. Chapman and Hall/CRC Press, Boston, USA, 2010.

[15] A. Pakman and L. Paninski, "Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians," *http://arxiv.org/abs/1208.4118*, 2013.

[16] Y. Altmann, N. Dobigeon, and J.-Y. Tourneret, "Unsupervised post-nonlinear unmixing of hyperspectral images using a Hamiltonian Monte Carlo algorithm," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2663–2675, 2014.

[17] H. P. Wai and B. D. Jeffs, "Adaptive image restoration using a generalized Gaussian model for unknown noise," *IEEE Trans. Image Process.*, vol. 4, no. 10, pp. 1451–1456, Oct. 1995.

[18] E. P. Simoncelli and E. H. Adelson, "Noise removal via Bayesian wavelet coring," in *IEEE Int. Conf. on Image Process. (ICIP)*, Lausanne, Switzerland, Sep. 16-19 1996, pp. 379–382.

[19] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalized-Gaussian priors," in *IEEE Int. Symp. on Time-Frequency and Time-Scale Analysis*, Pittsburgh, USA, Oct. 1998, pp. 633–636.

[20] L. Chaari, J.-C. Pesquet, J.-Y. Tourneret, Ph. Ciuciu, and A. Benazza-Benyahia, "A hierarchical Bayesian model for frame representation," *IEEE Trans. Signal Process.*, vol. 18, no. 11, pp. 5560–5571, Nov. 2010.

[21] 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," 2015, http://arxiv.org/abs/1505.00273.

[22] F. Lucka, "Fast MCMC sampling for sparse Bayesian inference in high-dimensional inverse problems using L1-type priors," 2012, arXiv:1206.0262v1.

[23] N. Dobigeon, A. O. Hero, and J.-Y. Tourneret, "Hierarchical Bayesian sparse image reconstruction with application to MRFM," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2059–2070, Sept. 2009.

[24] F. Costa, H. Batatia, L. Chaari, and J.-Y. Tourneret, "Sparse EEG source localization using Bernoulli laplacian priors," *IEEE Trans. on Biomed. Eng.*, vol. 62, no. 12, pp. 2888 – 2898, 2015.

[25] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, New York, 2011.

[26] B. J. Alder and T. E. Wainwright, "Studies in molecular dynamics.i. general method," *J. Chem. Phys.*, vol. 31, pp. 459–466, 1959.

[27] Z. Wang, S. Mohamed, and D. Nando, "Adaptive Hamiltonian and Riemann manifold Monte Carlo," in *Proc. Int. Conf. Machine Learning*, Atlanta, USA, May 2013, pp. 1462–1470.

[28] V. Jeyakumar and D. T. Luc, "Approximate Jacobian matrices for nonsmooth continuous maps and C^1 optimization," *SIAM J. Control and Optim.*, vol. 36, no. 5, pp. 1815–1832, 1998.

[29] F. H. Clarke, *Optimization and nonsmooth analysis*, Wiley-Interscience, New York, 1983.

[30] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Model. Simul.*, vol. 4, pp. 1168–1200, 2005.

[31] C. Chaux, P. Combettes, J.-C. Pesquet, and V.R Wajs, "A variational formulation for frame-based inverse problems," *Inv. Prob.*, vol. 23, no. 4, pp. 1495–1518, Aug. 2007.

[32] L. Chaari, J.-C. Pesquet, A. Benazza-Benyahia, and P. Ciuciu, "A wavelet-based regularized reconstruction algorithm for SENSE parallel MRI with applications to neuroimaging," *Med. Image Anal.*, vol. 15, no. 2, pp. 185–201, Nov. 2011.

[33] Y. F. Atchade, G. Fort, and E. Moulines, "On stochastic proximal gradient algorithms," *Arxiv*, 2014, http://arxiv.org/abs/1402.2365.

[34] J.-J. Moreau, "Proximité et dualité dans un espace hilbertien," *Bulletin de la Société Mathématique de France*, vol. 93, pp. 273–299, 1965.

[35] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, H. H. Bauschke, R. Burachik, P. L. Combettes, V. Elser, D. R. Luke, , and H. Wolkowicz, Eds., pp. 185–212. Springer-Verlag, New York, 2011.

[36] J. Douglas and H. H. Rachford, "On the numerical solution of heat conduction problems in two or three space variables," *Trans. Amer. Math. Soc.*, vol. 82, pp. 421–439., 1956.

[37] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, Eds., *Handbook of Markov Chain Monte Carlo*, Chapman & Hall/CRC, Boston, USA, 2011.

[38] L. Chaari, J.-Y. Tourneret, C. Chaux, and H. Batatia, "A Hamiltonian Monte Carlo method for non-smooth energy sampling: Technical report," Tech. Rep., University of Toulouse, 2015, http://lotfi-chaari.net/.

[39] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2008.

[40] M. Seeger, "Bayesian inference and optimal design in the sparse linear model," *J. Mach. Learn. Res.*, vol. 9, pp. 759–813, 2008.

[41] L. Chaari, J.-Y. Tourneret, and H. Batatia, "Sparse Bayesian regularization using Bernoulli-Laplacian priors," in *Proc. European Signal Process. Conf. (EUSIPCO)*, Marrakech, Morocco, September 9-13, 2013.

[42] P. L. Combettes and J.-C. Pesquet, "A proximal decomposition method for solving convex variational inverse problems," *Inv. Prob.*, vol. 24, no. 6, Dec. 2008, 27 p.

[43] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[44] L. Chaari, J.-Y. Tourneret, and C. Chaux, "Sparse signal recovery using a bernouilli generalized gaussian prior," in *European Signal Processing Conference (EUSIPCO)*, Nice, France, August 31- September 4 2015.