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 involving Markov chain Monte Carlo schemes. 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 the exponential families of distributions with smooth energy function. In this paper, we address the problem of using Hamiltonian dynamics to sample from probability distributions having non-differentiable energy functions such as ℓ_1. Such distributions are being more and more used in sparse signal and image recovery applications. The proposed technique uses a modified leapfrog transform involving a proximal step. The resulting non-smooth Hamiltonian Monte Carlo (ns-HMC) 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, most of the recent applications such as remote sensing [2] and medical image reconstruction [3, 4] generate large and intractable data volumes that have to be processed either independently or jointly. To handle such inverse problems, Bayesian techniques have demonstrated their usefulness especially when the model hyperparameters are difficult to fix. Such techniques generally rely on a maximum a posteriori (MAP) estimation built upon the signal/image likelihood and priors. Since efficient priors generally have a complicated form, analytical expressions of the MAP estimators are often difficult to obtain. For this reason, many Bayesian techniques resort to 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-

L. Chaari, J.-Y. Tourneret and Hadj 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.

C. Chaux is with LATP and CNRS UMR 7353, Aix-Marseille University, 39 rue F. Joliot-Curie, 13453 Marseille Cedex 13, France. E-mail: caroline.chaux@latp.univ-mrs.fr.

adjusted Langevin algorithm (MALA) of [7, 8], or the high-dimensional Gaussian sampling methods of [9, 10]. To handle log-concave smooth probability distributions, a Hamiltonian Monte Carlo (HMC) sampling technique has recently been proposed in [11, 12]. This technique uses the analogy with the kinetic energy conservation in physics to design efficient proposals that better follow the target distribution's geometry. HMC has recently been investigated in a number of works dealing with multi-dimensional sampling problems for various applications [13, 14], which demonstrates the efficiency of such sampling schemes. 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 adapted to log-concave probability distributions with smooth energy functions since the gradient of these functions has to be calculated. This constraint represents a real limitation in some applications where sparsity is one of the main processing ingredients especially for large data. Indeed, sparsity promoting probability distributions generally have a non-differentiable energy function such as the Laplace or the generalized Gaussian (GG) [15] distributions which involve ℓ_1 and ℓ_p energy functions, respectively. Such 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 [16–18]. Using the HMC technique in this context is therefore not possible.

This paper presents a modified HMC scheme that makes it possible to sample from log-concave probability distributions with non-differentiable energy functions. The so called non-smooth HMC (ns-HMC) sampling scheme relies on a modified leapfrog transform [11, 12] that circumvents the non-differentiability of the target energy function. The modified leapfrog transform relies on the sub-differential and proximity operator concepts [19]. The proposed scheme is validated on a sampling example where samples are drawn from a GG distribution with different shape parameters. It is also illustrated on a signal recovery problem where a sparse regularization scheme is applied to recover a high-dimensional signal.

The rest 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 log-concave distributions. Finally, some conclusions and perspectives are drawn in Section V.

II. PROBLEM FORMULATION

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

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

where $E_\theta(x)$ is the energy function. As stated above, in this paper we concentrate on sampling from the class of log-concave probability densities, which means that the energy function E_θ 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 particular values of p such as $p = 2, 4, \ldots$, the energy function $E_\theta(x) = \frac{|x|^p}{\gamma}$ is not differentiable (where $\theta = (\gamma, p)$). 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 scheme.

III. NON-SMOOTH SAMPLING

A. Hamiltonian Monte Carlo methods

HMC methods [11, 12, 14] are powerful tools that use the principle of Hamiltonian dynamics. These methods have been originally proposed by analogy to kinetic energy evolution. They usually provide better mixing and convergence properties than standard schemes. Let us assume that $E_\theta(x)$ represents a *potential* energy function. If we associate to x a *momentum* vector $q \in \mathbb{R}^N$, the energy function H_θ for the Hamiltonian dynamics is the combination of the potential energy $E_\theta(x)$ and a *kinetic* energy $K(q)$, i.e.,

$$H_\theta(x, q) = E_\theta(x) + K(q). \tag{3}$$

The energy function H_θ is called the *Hamiltonian* and completely describes the considered system. For simplicity reasons, a quadratic kinetic energy corresponding to a unitary diagonal covariance matrix is usually assumed so that $K(q) = \frac{1}{2}q^\mathsf{T}q$. The Hamiltonian's motion equations determine the evolution of the state as a function of time t [12]

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

where ∂ denotes the partial derivative operator. These equations define a transformation $\mathcal{F}_s$ that maps the state of the system at time t to the state at time $t + s$. The distribution of the Hamiltonian dynamics energy defined in (3) is therefore given by

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

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

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

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

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

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

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

where H_θ is the energy function 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 x and 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 following definition of the sub-differential and a useful example.

Definition III.1 *[19, p. 223] Let φ be in $\Gamma_0(\mathbb{R})$, the class of lower semi-continuous convex functions from $\mathbb{R}$ to $]-\infty, +\infty]$. The sub-differential of φ is the set $\partial\varphi(x) = \{\rho \in \mathbb{R} | \varphi(\eta) \geq \varphi(x) + \langle\rho|\eta - x\rangle \, \forall \eta \in \mathbb{R}\}$, where $\langle\cdot|\cdot\rangle$ defines the standard scalar product. Every element $\rho \in \partial\varphi(x)$ is a sub-gradient of φ at point x. If φ is differentiable, the sub-differential reduces to its gradient: $\partial\varphi(x) = \{\nabla\varphi(x)\}$.*

Example III.1 *Let φ be defined as*

$$\varphi : \; \mathbb{R} \longmapsto \mathbb{R}$$
$$x \longrightarrow |x|. \tag{10}$$

The sub-differential of φ at x is defined by

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

If in addition we consider a scalar $\lambda \in \mathbb{R}_+$ and we call $\varphi_\lambda(\cdot) = \lambda\varphi(\cdot)$, then we have $\partial\varphi_\lambda(x) = \lambda\partial\varphi(x)$ for every $x \in \mathbb{R}$ [19, Prop. 16.5].

For distributions with smooth energy, we propose to 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 $\rho \in \partial E_{\boldsymbol{\theta}}$ 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.

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, and 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 [20] such as the *Clarke* generalization [21], and by conducting calculations similar to [12, Chapter 5, p. 118]. To facilitate the reading of the paper, the volume

---

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

- Initialize with some $\boldsymbol{x}^{(0)}$.
- 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 $\boldsymbol{q}^{(r,\frac{1}{2}\epsilon)} = \boldsymbol{q}^{(r,0)} - \frac{\epsilon}{2}\rho(\boldsymbol{x}^{(r,0)})$;
  - Compute $\boldsymbol{x}^{(r,\epsilon)} = \boldsymbol{x}^{(r,0)} + \epsilon\boldsymbol{q}^{(r,\frac{1}{2}\epsilon)}$;
  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

---

preservation issue is addressed in Appendix A.

2) Scheme 2 - proximal based approach:

Since the calculation of the sub differential 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 [22–24], and more recently in [25] where stochastic proximal algorithms have been investigated. Let us first recall the following definition.

Definition III.2 *[19, Definition 12.23][26] 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 $\mathrm{prox}_\varphi(x)$.*

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

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

Many other examples and interesting properties that make this tool very powerful and commonly used in the recent optimization literature are given in [27]. One of these properties in which we are interested here is given in the following property.

Property 1 *[28, 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 $x - \widehat{x} \in \partial\varphi(\widehat{x})$. Using the proximity operator definition hereabove, it turns out that $\widehat{x} = \mathrm{prox}_\varphi(x)$.*

By modifying the discretization scheme T_s, we propose the

following l-th leapfrog discretization scheme denoted by T_s''

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

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

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

If $E_\theta(x)$ 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 $\text{prox}_{E_\theta}(x)$ instead of x. As for scheme 1, the proposed candidates are given by $q^* = q^{(r,L_f)}$ and $x^* = x^{(r,\epsilon L_f)}$ 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. As well as for Algorithm 1, and due to the presence of the MH acceptance rule, the elements $x^{(r)}$ generated by this algorithm are asymptotically distributed according to the target distribution $f(x;\theta)$ defined in (1).

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

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

---

3) Discussion:

Fig. 1 illustrates the use of the adopted discretization schemes associated with algorithms 1 and 2 in approximating 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 sub-differential can be analytically calculated and is given by

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

where $\partial\varphi$ is defined in Example III.1 and Id is the identity operator. The two proposed algorithms can therefore be
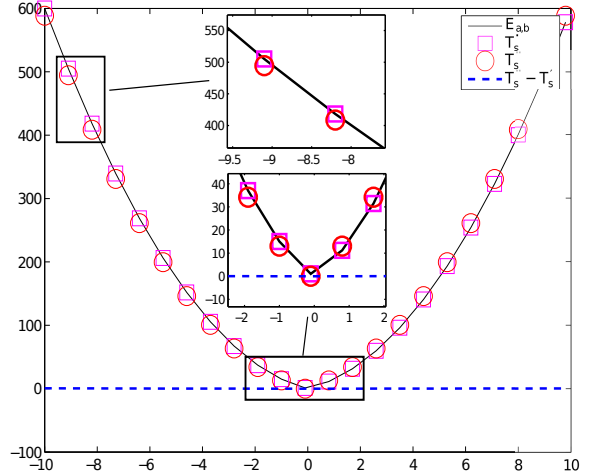

Fig. 1. The potential energy $E_{a,b}$ (solid black line) in (19) ($a = 10, b = 5$) and its discretizations using the modified leapfrog schemes T_s' (squares) and T_s'' (circles), as well as the difference between the two discretizations $T_s'' - T_s'$ (dashed blue line).

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, $x \simeq x - \text{prox}_{E_{a,b}} x$ for $x \in [-\varepsilon, \varepsilon]$. In particular, for the considered example, we have $x = x - \text{prox}_{E_{a,b}} x$ 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.
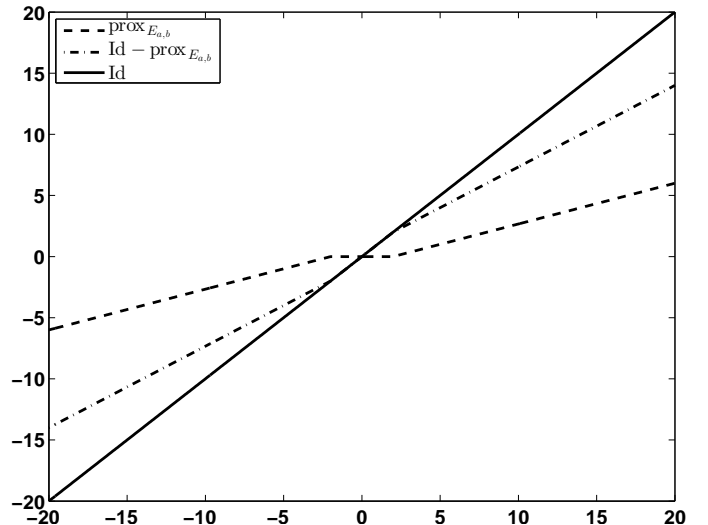

Fig. 2. The proximity operator $\text{prox}_{E_{a,b}}$, the identity function (Id) and the difference $\text{Id} - \text{prox}_{E_{a,b}}$ for $a = b = 2$.

Since Algorithm 2 is more general than Algorithm 1 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 discretization scheme for our experiments.

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 different values of the shape and scale parameters of a GG distribution (p and λ). Chains generated using the proposed ns-HMC sampling scheme are compared to the ones obtained with a random walk Metropolis-Hastings (rw-MH) scheme. 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 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 displays the mean square error (MSE) between the target GG pdf and the histogram of the generated samples with respect to the number of sampled coefficients. This figure shows slightly faster convergence for the proposed ns-HMC scheme compared to the rw-MH algorithm.

To further investigate the sampling efficiency, Fig. 4 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 fast convergence of the ns-HMC scheme. In fact, 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. 4 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: multivariate sampling

In this experiment, sampling is performed according to a multivariate GG distribution. First, sampling using rw-MH
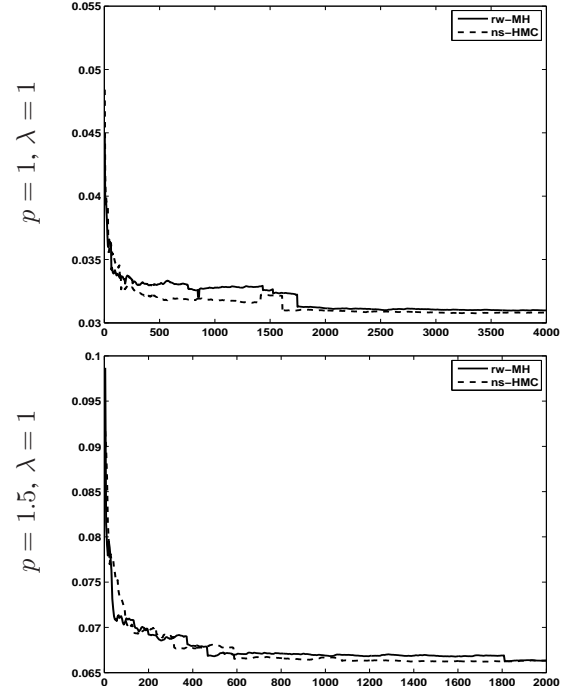

Fig. 3. MSEs between the target 1D GG pdf and the histogram of the generated samples using the rw-MH and ns-HMC algorithms for two different combinations of p and λ.
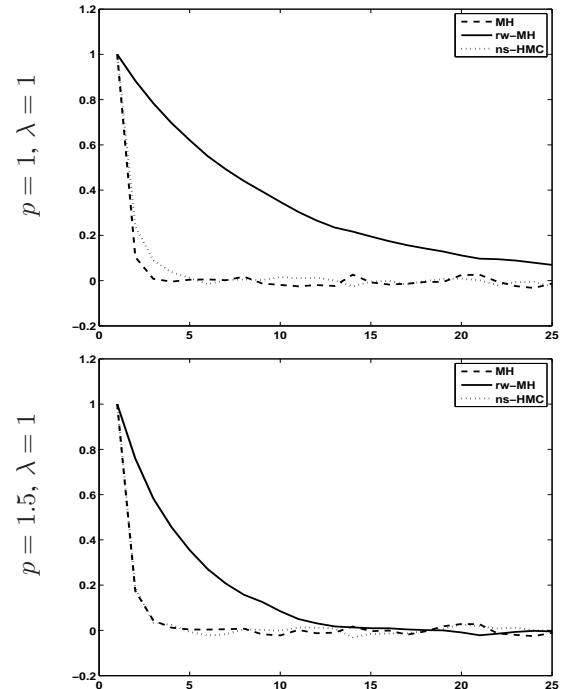

Fig. 4. ACFs of sampled chains using Metropolis-Hastings (MH) and random walk Metropolis-Hastings (rw-MH) algorithms, in addition to the proposed method (ns-HMC) for two values of (p, λ).
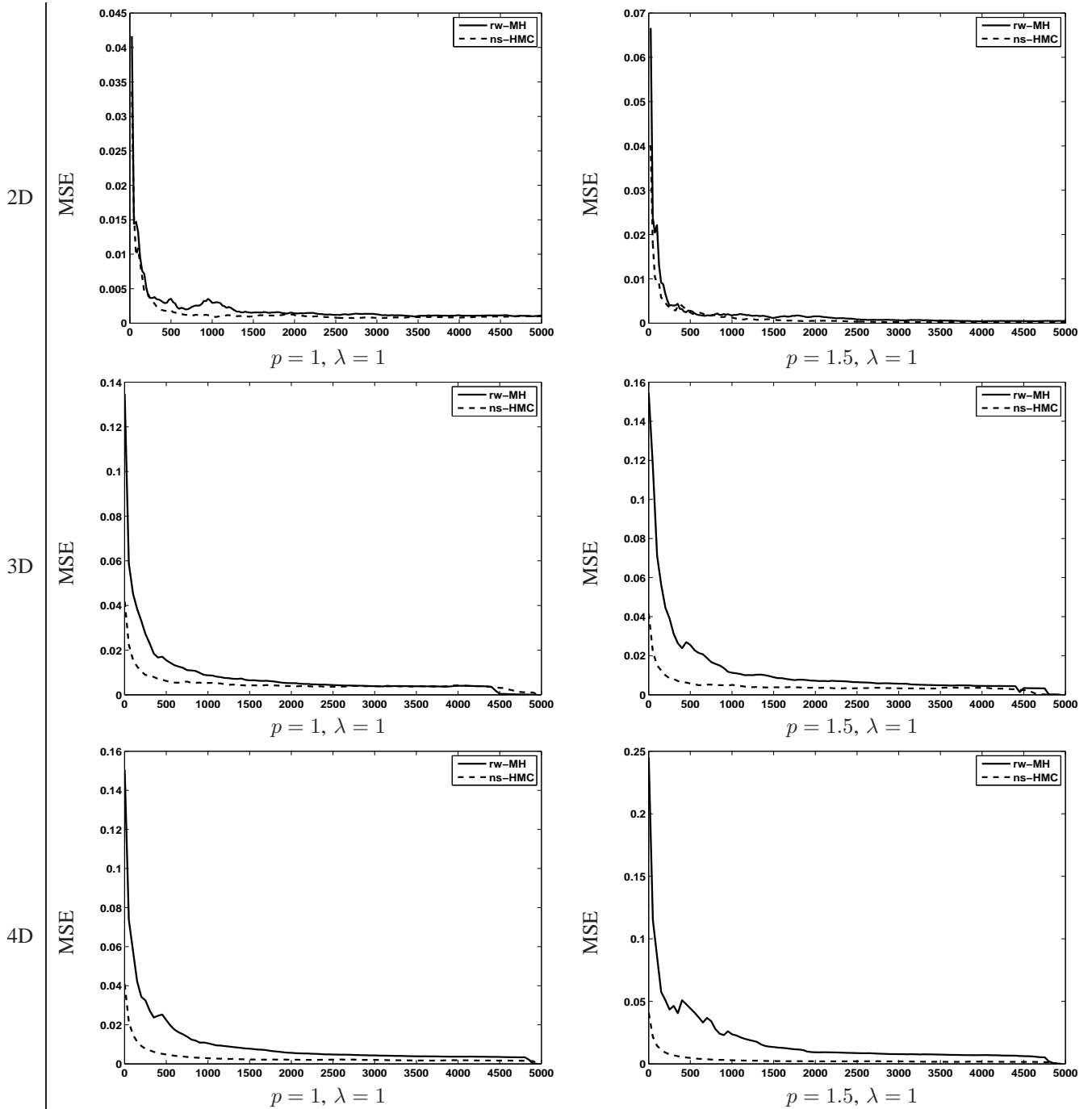
Fig. 5. MSEs between the target GG pdf and the histogram of the generated samples for the rw-MH and ns-HMC algorithms.

is performed with a large number of iterations (long burn-in period) so that the target distribution is guaranteed to be reached. The histogram of the obtained samples is calculated and corresponds to our ground truth. Then, samplings using the proposed ns-HMC method and the rw-MH scheme are performed on the same dataset. The MSE between the obtained histograms and the ground truth are computed versus the number of iterations of the sampler. Fig. 5 displays this MSE for the multivariate GG sampling and for different values of the scale and shape parameters (λ and p). Note that simulations have been performed for the 2D, 3D and 4D cases. The reported MSE values show that the target distribution is

more rapidly approached using ns-HMC compared to rw-MH. Specifically, for the 2D case, the ns-HMC scheme converges after about 500 iterations for $p = 1$ (resp. 1000 iterations for $p = 1.5$), while the rw-MH sampling needs about 3500 iterations for $p = 1$ and $p = 2$. When looking at the same curves in Fig. 5 for the 3D and 4D cases, it is worth noticing that the gap between the two methods in terms of convergence speed increases with the problem dimensionality. This effect is confirmed for the shape and scale parameters. This corroborates the usefulness of the proposed ns-HMC scheme especially for large data sampling where the convergence speed of the standard MH or rw-MH algorithms is altered

by the size of the data.

C. Experiment 3: 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

$$\boldsymbol{y} = F^{-1}\boldsymbol{x} + \boldsymbol{n} \tag{21}$$

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 [29, 30]

$$f(\boldsymbol{x}; \lambda) = \left(\frac{1}{2\lambda}\right)^N \exp\left(-\frac{||\boldsymbol{x}||_1}{\lambda}\right) \tag{22}$$

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 [31, 32]

$$f(\lambda|a, b) = \mathcal{IG}(\lambda|a, b) = \frac{b^a}{\Gamma(a)} \lambda^{-a-1} \exp\left(-\frac{b}{\lambda}\right) \tag{23}$$

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 ($\sigma_n^2 \sim \frac{1}{\sigma_n^2}\mathbb{1}_{\mathbb{R}+}(\sigma_n^2)$), 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

$$f(\boldsymbol{x}|\boldsymbol{y}, \sigma_n^2, \lambda) \propto \exp\left[-U(\boldsymbol{x})\right] \tag{24}$$

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

$$\mathrm{prox}_U(\boldsymbol{x}) = \mathrm{prox}_{||\cdot||_1/(1+\alpha)}\left(\frac{\boldsymbol{x} + F\boldsymbol{y}}{1+\alpha}\right) \tag{25}$$

where $\alpha = \frac{1}{\sigma_n^2}$ and $\mathrm{prox}_{||\cdot||_1/(1+\alpha)}$ can easily be calculated using standard properties of the proximity operators [19, 23, 33]. See Appendix B for more details.

Regarding the noise variance and the prior hyperparameter, straightforward calculations lead to the following conditional distributions which are easy to sample

$$\sigma_n^2|\boldsymbol{x}, \boldsymbol{y} \sim \mathcal{IG}\left(\sigma_n^2|N/2, ||\boldsymbol{y} - F^{-1}\boldsymbol{x}||^2/2\right) \tag{26}$$

$$\lambda|\boldsymbol{x}, a, b \sim \mathcal{IG}\left(\lambda|a + N, b + ||\boldsymbol{x}||_1\right) \tag{27}$$

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 perior, i.e., after convergence of the Gibbs sampler in Algorithm 3.

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 (26);
  - Sample $\lambda^{(r)}$ according to (27);
  - 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 $\widehat{\boldsymbol{z}} = F^{-1}\widehat{\boldsymbol{x}}$.

An example of denoised image using Algorithm 3 is displayed in Fig. 6[bottom-left]. For the sake of comparison, a denoised image using the Wiener filter is displayed in 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 [34] (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 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 $L_f = 10$.

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. Validation results showed that the proposed technique provides faster convergence and interesting decorrelation properties for the sampled chains when compared to more standard methods

Reference

Noisy
SNR = 5.68 dB, SSIM = 0.699

Algorithm 3
SNR = 20.48 dB, SSIM = 0.985
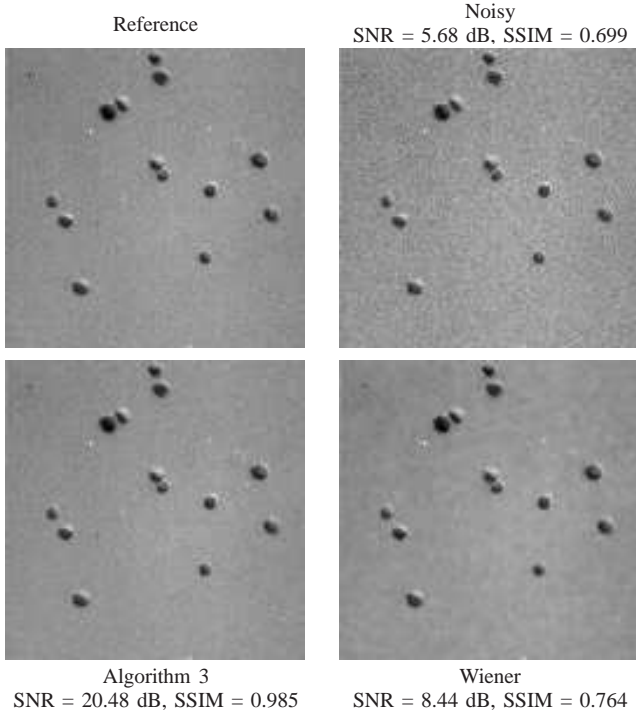
Wiener
SNR = 8.44 dB, SSIM = 0.764

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

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.

APPENDIX

A. Volume preservation

This appendix studies the volume preservation condition both for the Hamiltonian dynamics and the proposed discretization schemes. Volume preservation is a key property of sampling algorithms involving an acceptance-rejection step such as the Metropolis within Gibbs algorithm since it allows simpler acceptance probability to be obtained.

1) Volume preservation for the Hamiltonian dynamics:
Akin to [12], we consider here volume preservation for Hamiltonian dynamics for the one dimensional case. The multi-dimensional case can then be handled through simple generalizations. 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 $t+\delta$. For δ small enough, $\mathcal{F}_\delta$ can be approximated by [12]

$$\mathcal{F}_\delta(q,x) = \begin{bmatrix} q \\ x \end{bmatrix} + \delta \begin{bmatrix} dq/dt \\ dx/dt \end{bmatrix} + O(\delta^2) \quad (28)$$

where $O(\delta^2)$ involves terms of order δ^2 or higher. After replacing the time derivatives of (4) in (28), and accounting for the fact that the Hamiltonian may be non-differentiable with respect to x, the generalized Jacobian matrix can be written as

$$\mathcal{J}_\delta = \begin{bmatrix} 1 + \delta \frac{\partial^2 H_\theta}{\partial q \partial x} & \delta \frac{\partial^2 H_\theta}{\partial q^2} \\ -\delta \frac{\partial^2 H_\theta}{\partial q^2} & 1 - \delta \frac{\partial^2 H_\theta}{\partial x \partial q} \end{bmatrix} + O(\delta^2) \quad (29)$$

where ∂ denotes the sub-gradient and $\frac{\partial^2 H_\theta}{\partial q \partial x}$ 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

$$\det(\mathcal{J}_\delta) = 1 + \delta \frac{\partial^2 H_\theta}{\partial q \partial x} - \delta \frac{\partial^2 H_\theta}{\partial x \partial q} + O(\delta^2)$$
$$= 1 + O(\delta^2) \quad (30)$$

where $\det(A)$ is the determinant of the matrix A. Following the construction proposed in [12], it turns out that for some time interval s that is not close to zero, $\det(\mathcal{J}_s) = 1$, which means that the transformation $\mathcal{F}_s$ ensures volume preservation.

2) Volume preservation for the proposed discretization schemes:
It is woth noticing here that the two modified leapfrog discretization schemes T'_s and T''_s defined respectively in (12)-(14) and (16)-(18), as the original leapfrog scheme defined in (6)-(8), preserve volume since they are shear transformations. The interested reader can refer to [12] or [35, page 121] for more details.

B. Proximity operator calculation for the experiment of Section IV-C

The energy function considered in this appendix is the one involved in the conditional distribution of the wavelet coefficients in (24), i.e.,

$$U(\boldsymbol{x}) = \frac{\alpha}{2}\|\boldsymbol{y} - F^{-1}\boldsymbol{x}\|_2^2 + \varphi(\boldsymbol{x}) \quad (31)$$

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 [23, 26], we can write

$$\text{prox}_U(\boldsymbol{x}) = \boldsymbol{p} \Leftrightarrow \boldsymbol{x} - \boldsymbol{p} \in \partial U(\boldsymbol{p})$$
$$\Leftrightarrow \boldsymbol{x} - \boldsymbol{p} \in \partial \varphi(\boldsymbol{p}) + \alpha \boldsymbol{p} - \alpha F \boldsymbol{y}$$
$$\Leftrightarrow \boldsymbol{x} + \alpha F \boldsymbol{y} - (\alpha+1)\boldsymbol{p} \in \partial \varphi(\boldsymbol{p})$$
$$\Leftrightarrow \frac{\boldsymbol{x} + \alpha F \boldsymbol{y}}{\alpha+1} - \boldsymbol{p} \in \partial \varphi/(\alpha+1)(\boldsymbol{p})$$
$$\Leftrightarrow \boldsymbol{p} = \text{prox}_{\varphi/(\alpha+1)}\left(\frac{\boldsymbol{x} + F\boldsymbol{y}}{\alpha+1}\right) \quad (32)$$

which proves the expresssion of the proximity operator given in (25).

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. Pereyra, "Proximal markov chain monte carlo algorithms," *Arxiv*, 2013, http://arxiv.org/abs/1306.0187.

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

[10] 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.

[11] 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.

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

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

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

[23] 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.

[24] 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.

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

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

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

[28] 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.

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

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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