Efficient Sampling of Non Log-Concave Posterior Distributions With Mixture of Noises

Pierre Palud, Pierre-Antoine Thouvenin, Pierre Chainais, Emeric Bron, Franck Le Petit

Efficient sampling of non log-concave posterior distributions with mixture of noises

Pierre Palud*, Pierre-Antoine Thouvenin*, Pierre Chainais*, Emeric Bron†, Franck Le Petit†

*Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
†LERMA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne Universités, 92190 Meudon, France

***Abstract*—This paper focuses on a challenging class of inverse problems that is often encountered in applications. The forward model is a complex non-linear black-box, potentially non-injective, whose outputs cover multiple decades in amplitude. Observations are supposed to be simultaneously damaged by additive and multiplicative noises and censorship. As needed in many applications, the aim of this work is to provide uncertainty quantification on top of parameter estimates. The resulting log-likelihood is intractable and potentially non-log-concave. An adapted Bayesian approach is proposed to provide credibility intervals along with point estimates. An MCMC algorithm is proposed to deal with the multimodal posterior distribution, even in a situation where there is no global Lipschitz constant (or it is very large). It combines two kernels, namely an improved version of PMALA [1] and a Multiple Try Metropolis (MTM) kernel [2]. This sampler addresses all the challenges induced by the complex form of the likelihood. The proposed method is illustrated on classical test multimodal distributions as well as on a challenging and realistic inverse problem in astronomy.**

***Index Terms*—Bayesian inference, black-box forward model, inverse problem, Markov Chain Monte Carlo algorithms.**

I. Introduction

Physics and experimental sciences often produce non-linear, potentially non-injective, forward models. Such models are often encoded by expensive black-box functions, e.g., the solution of a large set of partial differential equations in epidemiology [3] or astrophysics [4], [5]. The forward model may also span multiple decades, as in astrophysics [6] where orders of magnitude can be gigantic. As a consequence, when the log-likelihood function is smooth, the Lipschitz constant of the gradient is too large to be numerically useful. Inverse problems that involve such models can lead to a non-log-concave, potentially multimodal likelihood function.

For the sake of simplicity, most observation models consider one source of noise only, while more detailed models may involve multiple noises as well as censored data, due to sensitivity limitations. Such difficult models are often addressed by simplifying the scenario in practice. For instance, one noise is assumed to dominate the others that are neglected, as in medical ultrasound imaging [7] or in synthetic aperture radar [8].

This work addresses a family of inverse problems involving both a non-linear black-box forward model covering multiple decades, and censored observations affected by both an additive and a multiplicative noise. The log-posterior is intractable, admits a Lipschitz continuous gradient with a very large constant (if finite), may be non-concave and potentially multimodal. A Bayesian approach is proposed. The problem is addressed with a Markov Chain Monte Carlo (MCMC) algorithm [9]–[11] to provide point estimates with the corresponding credibility intervals. This uncertainty quantification is particularly critical for applications where no ground truth is available, as in cosmology and astrophysics [4], [5]. An explicit and smooth approximation of the likelihood is proposed. It relies on a model reduction and on controlled approximations of the noise model. Since the Lipschitz constant of the gradient of the log-posterior is assumed to be very large or even infinite, efficient sampling methods relying on smoothness assumptions, such as the Metropolis Adjusted Langevin Algorithm (MALA) [12] or Hamiltonian Monte Carlo (HMC) [13] typically fail to explore the parameter space. To address this issue, a preconditioned MALA (PMALA) kernel [14], [15] exploiting the RMSProp preconditioner from deep-learning [16] is considered. Compared to HMC and MALA, the exploration of the parameter space is based on local second order information of the log-posterior, instead of the global Lipschitz constant of its gradient only. A first version of this kernel, introduced in [1], led to an approximate sampler. We further improve it in this paper to obtain an exact sampler, *i.e.*, asymptotically drawing samples from the distribution of interest. To account for the non-log-concavity and potential multimodality of the posterior, a combination of PMALA with a Multiple Try Metropolis (MTM) kernel [2] is proposed.

The proposed sampler is validated on two classical multimodal examples: a 2D Gaussian mixture model and the sensor localization problem [17]. It is then applied with good performances to a realistic higher dimensional astrophysical problem that combines all the aforementioned challenges. A preliminary version of this work was published in [18], where the main principle of the method was summarized for a simplified illustration.

Section II introduces the Bayesian model, the proposed likelihood approximation and the resulting posterior distribution. Section III introduces the MCMC algorithm used to derive estimators. Section IV demonstrates the performance of the proposed method on the three experiments outlined above. Section V provides conclusions and perspectives.

II. Bayesian model

This section introduces the general Bayesian model considered in this article. A tractable surrogate likelihood with

controlled error is built on a reduced forward model and a likelihood approximation that deals with the mixture of noises. The prior and the resulting posterior are then introduced.

A. Notation

Throughout this paper, scalars are denoted with regular letters, e.g., indices n, d and ℓ, or the corresponding dimensions N, D and L. In the following, N is the number of observations over L channels and D is the number of parameters of the model. Vectors are denoted using bold lowercase letters, e.g., parameters $\boldsymbol{\theta} \in \mathbb{R}^D$ or observations $\boldsymbol{y} = (y_1, \dots, y_\ell, \dots, y_L) \in \mathbb{R}^L$. Matrices are written with bold uppercase letters, e.g., matrices of observations $\boldsymbol{Y} = (\boldsymbol{y}_n)_{n=1}^N \in \mathbb{R}^{N \times L}$. The notation for functions is set accordingly, e.g., the forward model $\boldsymbol{f}(\boldsymbol{\theta}) = (f_\ell(\boldsymbol{\theta}))_{\ell=1}^L$.

B. Problem statement

Individual observations $\boldsymbol{y} = (y_\ell)_{\ell=1}^L$ gather L channels. They are considered to be generated from some parameter $\boldsymbol{\theta} \in \mathbb{R}^D$ and a forward model $\boldsymbol{f} : \mathbb{R}^D \to \mathbb{R}^L$, where the number D of parameters is assumed to remain moderate, e.g. $D \lesssim 10$. The forward model prediction for a channel ℓ is denoted by f_ℓ, so that for any $\boldsymbol{\theta} \in \mathbb{R}^D$, $\boldsymbol{f}(\boldsymbol{\theta}) = (f_\ell(\boldsymbol{\theta}))_{\ell=1}^L$. The forward model $\boldsymbol{f}$ is assumed to be a non-linear black-box function, e.g., the result of a physical experiment or a numerical simulation. It is considered valid on a compact subset $\mathcal{C} = [l_1, u_1] \times \cdots \times [l_D, u_D] \subset \mathbb{R}^D$, with $l_d, u_d \in \mathbb{R}$ for any $d \in [\![1, D]\!]$, that can correspond to a typical domain of validity. To reflect physical considerations on the nature of the data, the function $\boldsymbol{f}$ is further assumed to have positive values that can span multiple decades, as is often the case in astrophysics [6]. Individual observations and parameters are grouped in indexed sets $\boldsymbol{Y} = (\boldsymbol{y}_n)_{n=1}^N$ and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_n)_{n=1}^N$ of size N, such as an image, a time series or more generally a graph, with N potentially very large, e.g., of the order of millions. The sensors are assumed to have a lower limit of sensitivity $\omega \in \mathbb{R}$ below which an observation is censored. Both an additive and multiplicative noise degrade the observations. Such a mixture of noises occurs in astrophysics as well as in medical ultrasound imaging [7] or laser imaging and synthetic aperture radars [8] for instance. Turning to inference, one of the two noises is generally neglected for sake of tractability [7]. However, when the forward model spans several decades, the nature of the dominant noise depends on the amplitude of $f_\ell(\boldsymbol{\theta})$. The resulting observation model is, for $n \in [\![1, N]\!]$ and $\ell \in [\![1, L]\!]$,

$$y_{n,\ell} = \max\left\{\omega,\ \epsilon_{n,\ell}^{(m)} f_\ell(\boldsymbol{\theta}_n) + \epsilon_{n,\ell}^{(a)}\right\}, \tag{1}$$

where $\epsilon_{n,\ell}^{(a)} \sim \mathcal{N}(0, \sigma_a^2)$ is an additive Gaussian white noise, and $\epsilon_{n,\ell}^{(m)} \sim \log\mathcal{N}(-\sigma_m^2/2, \sigma_m^2)$ is a lognormal multiplicative noise such that $\mathbb{E}[\epsilon_{n,\ell}^{(m)}] = 1$. The noise terms $\epsilon_{n,\ell}^{(a)}$ and $\epsilon_{n,\ell}^{(m)}$ are assumed independent with known variances σ_a^2 and σ_m^2, respectively. They are also assumed independent of $f_\ell(\boldsymbol{\theta}_n)$. At low intensities, the additive noise dominates; high intensities are mainly damaged by the multiplicative noise.

C. Likelihood approximation

The likelihood associated to (1) involves a potentially expensive forward model. A model reduction can be used to ensure the computational efficiency of the inference process. The presence of the two sources of noise makes the likelihood intractable so that we propose a parametric surrogate model.

1) Model reduction: The forward model $\boldsymbol{f}$ is assumed to be encoded by an expensive black-box function. Such black-box models may be addressed with a likelihood-free method such as Approximate Bayesian Computation (ABC) [19] that yield approximate samplers of the true posterior distribution. These methods are limited by the cost of numerous evaluations of the black-box function. A cheaper reduced model is preferred when this cost becomes prohibitive [20], [21]. This solution often permits to exactly sample from an approximate posterior. Model reduction largely remains an application specific problem, with only a few generic approaches [22], [23]. Here the forward model f_ℓ is positive and covers several decades, for all $\ell \in [\![1, L]\!]$. We propose to replace it by approximations $\widetilde{P}_\ell$ of its logarithm so that

$$\forall \ell, \quad \tilde{f}_\ell(\boldsymbol{\theta}) = \exp\left[\widetilde{P}_\ell(\boldsymbol{\theta})\right]. \tag{2}$$

The error introduced by replacing $\boldsymbol{f}$ by $\tilde{\boldsymbol{f}}$ should be negligible compared to $\boldsymbol{f}$ and to the noise standard deviations σ_a and σ_m. In the present approach, evaluations of $\widetilde{P}_\ell$ and its gradients should be fast, as each iteration of the proposed MCMC algorithm in Section III will require such operations. This approximation is also required to be twice differentiable to satisfy the requirements of the PMALA kernel involved in III-A. The derivation of such a reduced model is feasible with a wide family of methods including polynomials or neural networks but remain out of the scope of this work. In the following, $\tilde{\boldsymbol{f}}$ is assumed available so that from now on f_ℓ is replaced by $\tilde{f}_\ell$.

2) Modeling the noise mixture: For simplicity, we first consider the uncensored part of the model (1). Since the corresponding likelihood is intractable, most approaches in the literature [7], [8] neglect one source of noise. This strategy obviates the need for handling both the additive and multiplicative noises at intermediate intensities. A slightly different mixture model is addressed in [24] with a hierarchical approach and linear forward models. The proposed approach builds on [25], where the mixture is approximated with a purely additive model. The additive noise $\epsilon_{n,\ell}^{(a)}$ in (1) can be neglected when $\tilde{f}_\ell(\boldsymbol{\theta}_n) \to \infty$, while the multiplicative noise $\epsilon_{n,\ell}^{(m)}$ becomes negligible as $\tilde{f}_\ell(\boldsymbol{\theta}_n) \to 0$. Therefore, for each observation $y_{n,\ell}$, the true likelihood is approximated using three different regimes: low, intermediate and high values of $\tilde{f}_\ell(\boldsymbol{\theta}_n)$. In the low value regime, the true likelihood function $\pi(y_{n,\ell}|\boldsymbol{\theta}_n)$ is approximated by an additive Gaussian approximation $\pi^{(a)}(y_{n,\ell}|\boldsymbol{\theta}_n)$ corresponding to

$$y_{n,\ell} \simeq \tilde{f}_\ell(\boldsymbol{\theta}_n) + e_{n,\ell}^{(a)}, \quad e_{n,\ell}^{(a)} \sim \mathcal{N}(m_{a,n,\ell}, s_{a,n,\ell}^2), \tag{3}$$

where $m_{a,n,\ell}$ and $s^2_{a,n,\ell}$ are obtained by matching the two first moments with model (1), which yields

$$\begin{cases} m_{a,\ell,n} = 0, \\ s^2_{a,\ell,n} = \tilde{f}_\ell(\boldsymbol{\theta}_n)^2(e^{\sigma_m^2}-1) + \sigma_a^2. \end{cases} \tag{4}$$

Conversely, in the high value regime, a multiplicative lognormal approximation $\pi^{(m)}(y_{n,\ell}|\boldsymbol{\theta}_n)$ is used. It reads

$$y_{n,\ell} \simeq e^{(m)}_{n,\ell}\, \tilde{f}_\ell(\boldsymbol{\theta}_n), \quad e^{(m)}_{n,\ell} \sim \log\mathcal{N}(m_{m,n,\ell}, s^2_{m,n,\ell}), \tag{5}$$

where moment matching with (1) yields:

$$\begin{cases} m_{m,\ell,n} = -\frac{1}{2}\left\{\sigma_m^2 + \log\left[1 + \frac{\sigma_a^2}{\tilde{f}_\ell(\boldsymbol{\theta}_n)^2 e^{\sigma_m^2}}\right]\right\}, \\ s^2_{m,\ell,n} = -2\, m_{m,\ell,n} \quad \text{so that } \mathbb{E}[e^{(m)}_{n,\ell}] = 1. \end{cases} \tag{6}$$

For the intermediate regime, for each channel ℓ, we introduce parameters $\boldsymbol{a}_\ell = (a_{\ell,0}, a_{\ell,1}) \in \mathbb{R}^2$. $a_{\ell,0}$ pinpoints the low to intermediate value transition and $a_{\ell,1}$ the intermediate to high value transition. In this intermediate regime, i.e., $a_{\ell,0} \leq \widetilde{P}_\ell(\boldsymbol{\theta}_n) \leq a_{\ell,1}$, we propose to use a geometric average of the two likelihood approximations $\pi^{(a)}(y_{n,\ell}|\boldsymbol{\theta}_n)$ and $\pi^{(m)}(y_{n,\ell}|\boldsymbol{\theta}_n)$ with weights $1-\lambda$ and λ, respectively, see the first term of (8) below. The weight function λ is defined as a twice differentiable sigmoid with values in $[0,1]$:

$$\lambda(\boldsymbol{\theta}_n, \boldsymbol{a}_\ell) = \begin{cases} 0 & \text{if } \widetilde{P}_\ell(\boldsymbol{\theta}_n) \leq a_{\ell,0} \\ 1 & \text{if } \widetilde{P}_\ell(\boldsymbol{\theta}_n) \geq a_{\ell,1} \\ Q\left(\frac{\widetilde{P}_\ell(\boldsymbol{\theta}_n) - \log a_{\ell,0}}{\log a_{\ell,1} - \log a_{\ell,0}}\right) & \text{otherwise} \end{cases}, \tag{7}$$

where Q is a polynomial such that $Q(0)=0$, $Q(0)=1$ and $Q^{'}(0) = Q^{'}(1) = Q^{''}(0) = Q^{''}(1) = 0$ for λ to be $\mathscr{C}^2$. One of the simplest such polynomials is $Q(u) = u^3(6u^2 - 15u + 10)$.

To take censorship into account, let $\mathbf{C} = (c_{n,\ell})_{n,\ell} \in \{0,1\}^{NL}$ be a matrix such that $c_{n,\ell} = 1$ for a censored observation, and $c_{n,\ell} = 0$ otherwise. Let $F^{(a)}(\cdot|\boldsymbol{\theta}_n)$ and $F^{(m)}(\cdot|\boldsymbol{\theta}_n)$ be the cumulative density functions (cdf) of $\pi^{(a)}(\cdot|\boldsymbol{\theta}_n)$ and $\pi^{(m)}(\cdot|\boldsymbol{\theta}_n)$, respectively. The likelihood of censored data involves $F^{(a)}(\omega|\boldsymbol{\theta}_n)$ and $F^{(m)}(\omega|\boldsymbol{\theta}_n)$. The proposed likelihood approximation of model (1) finally reads

$$\begin{aligned} &\tilde{\pi}(y_{n,\ell}|\boldsymbol{\theta}_n, \boldsymbol{a}_\ell) \propto \\ &\left[\pi^{(a)}(y_{n,\ell}|\boldsymbol{\theta}_n)^{1-\lambda(\boldsymbol{\theta}_n,\boldsymbol{a}_\ell)}\, \pi^{(m)}(y_{n,\ell}|\boldsymbol{\theta}_n)^{\lambda(\boldsymbol{\theta}_n,\boldsymbol{a}_\ell)}\right]^{1-c_{n,\ell}} \\ &\times \left[F^{(a)}(\omega|\boldsymbol{\theta}_n)^{1-\lambda(\boldsymbol{\theta}_n,\boldsymbol{a}_\ell)}\, F^{(m)}(\omega|\boldsymbol{\theta}_n)^{\lambda(\boldsymbol{\theta}_n,\boldsymbol{a}_\ell)}\right]^{c_{n,\ell}}. \end{aligned} \tag{8}$$

The accuracy of this likelihood approximation clearly depends on the choice of the parameter $\boldsymbol{a}_\ell$. Appendix A proposes a procedure to adjust it in a relevant manner.

D. Prior and resulting posterior

We will consider applications on multispectral images so that this work combines two penalties to build the prior distribution. The first one favors the spatial regularity of estimations. It is based on a local regularizer $h : \mathbb{R}^N \to \mathbb{R}_+$ applied to each map $\boldsymbol{\Theta}_{\cdot d} = (\theta_{n,d})_{1\leq n\leq N}$, with $d \in [\![1, D]\!]$. The regularizer can be the Euclidean norm of the usual gradient or Laplacian of the component map, with regularization parameter $\tau_d > 0$. The second term encodes the validity of $\boldsymbol{f}$ on a compact set $\mathcal{C} = [l_1, u_1] \times \cdots \times [l_D, u_D]$. Note that the reduced model may be defined out of $\mathcal{C}$ but will not be considered as valid since it was not trained on such points. The most natural approach would be to use the indicator function $\iota_{\mathcal{C}^N}$ of the set $\mathcal{C}^N$, where $\iota_{\mathcal{C}^N}(\boldsymbol{\Theta}) = 0$ if $\boldsymbol{\Theta} \in \mathcal{C}^N$, and $+\infty$ otherwise. Since the PMALA kernel to be introduced in Section III-A requires twice differentiability, the following twice differentiable approximation of $\iota_{\mathcal{C}^N}$, known as the quartic penalty in constrained optimization [26], is used:

$$\tilde{\iota}_{\mathcal{C}^N} : \boldsymbol{\Theta} \mapsto \sum_{n=1}^{N}\sum_{d=1}^{D} \left[\max(0, \theta_{n,d} - u_d, l_d - \theta_{n,d})\right]^4, \tag{9}$$

leading to a smooth uniform distribution. Finally, the resulting prior distribution is given by

$$\pi(\boldsymbol{\Theta}) \propto \exp\left(-\delta\, \tilde{\iota}_{\mathcal{C}^N}(\boldsymbol{\Theta}) - \sum_{d=1}^{D} \tau_d\, h(\boldsymbol{\Theta}_{\cdot d})\right), \tag{10}$$

where $\delta > 0$ is a penalty parameter. The posterior distribution combines NL independent likelihoods (8) and the priors (10).

$$\pi(\boldsymbol{\Theta}|\boldsymbol{Y}) \propto \exp\left[-g(\boldsymbol{\Theta})\right] \tag{11}$$

$$\propto \left[\prod_{n=1}^{N}\prod_{\ell=1}^{L} \tilde{\pi}(y_{n,\ell}|\boldsymbol{\theta}_n, \boldsymbol{a}_\ell)\right] \pi(\boldsymbol{\Theta}). \tag{12}$$

III. Proposed MCMC sampler

MCMC algorithms can provide point estimates along with the associated uncertainty quantification. However, the posterior distribution of complex systems is in general non-log-concave, hence potentially multimodal, which makes the sampling task challenging. In addition, when the forward model spans several decades, the gradient of the negative log-posterior ∇g has a potentially very large Lipschitz constant, if any. To address these two challenges, a new transition kernel is proposed as a combination of two kernels: PMALA [15] and MTM [2]. PMALA tackles the regularity issue to efficiently explore the neighborhood of a local mode, whereas MTM permits jumps between modes.

A. PMALA transition kernel

In absence of an exploitable gradient-Lipschitz regularity of the log-posterior g in (11), a preconditioned MALA equipped with RMSProp [16] is introduced to perform an efficient local exploration of the posterior distribution. To simplify notations, we temporarily use the vector version of $\boldsymbol{\Theta}$ in lexicographic order so that $\boldsymbol{\Theta} \in \mathbb{R}^{ND}$. Metropolis-Hastings (MH) [27], [28] is arguably the most famous MCMC algorithm. At each step t, a candidate $\boldsymbol{\Theta}_c^{(t)}$ is sampled from a proposal distribution $q(\boldsymbol{\Theta}_c^{(t)}|\boldsymbol{\Theta}^{(t-1)})$ that is accepted with probability

$$\rho^{(t)} = 1 \wedge \frac{\pi\left(\boldsymbol{\Theta}_c^{(t)}|\boldsymbol{Y}\right)}{\pi\left(\boldsymbol{\Theta}^{(t-1)}|\boldsymbol{Y}\right)} \frac{q\left(\boldsymbol{\Theta}^{(t-1)}|\boldsymbol{\Theta}_c^{(t)}\right)}{q\left(\boldsymbol{\Theta}_c^{(t)}|\boldsymbol{\Theta}^{(t-1)}\right)}. \tag{13}$$

The random walk proposal is often used but does not scale up well due to its blind nature [10]. Hamiltonian Monte Carlo (HMC) [13] and Metropolis Adjusted Langevin Algorithm (MALA) [12] both exploit gradient information. MALA

is defined as a discretized Langevin diffusion process with an accept-reject step, while HMC relies on Hamiltonian dynamics and auxiliary variables. Both propose larger steps than the random walk with high acceptance probability, which favors scaling up [10]. They both rely on a step size inversely proportional to the Lipschitz constant of ∇g, if it exists. Here the forward model $\tilde{\boldsymbol{f}}$ covers several decades so that this Lipschitz constant is potentially very large or even infinite. Therefore, MALA and HMC will typically fail to efficiently explore the parameter space of a posterior distribution such as (12).

A transition kernel that handles such situations relies on extensions of HMC and MALA to Riemannian manifolds [14]. The MALA extension is favored over its HMC counterpart as MALA yields faster individual iterations and requires less parameter tuning. Moreover, Riemannian manifolds MALA was improved in [15], resulting in the so-called position dependent MALA (PMALA) kernel. It permits exploiting local information geometry thanks to a position dependent preconditioner. We propose to use the RMSProp preconditioner [16] that was initially defined in the deep learning literature for fast neural networks training. In the absence of exploitable Lipschitz constant, it adaptively estimates a local variance of the gradient ∇g by keeping memory of former proposals $\boldsymbol{\Theta}_c^{(t)}$. At each step t, it updates a surrogate gradient variance vector $\boldsymbol{v}^{(t)} \in \mathbb{R}^{ND}$ such that for all $i \in [\![1, ND]\!]$,

$$v_i^{(t)} = \alpha v_i^{(t-1)} + (1-\alpha)\left[\frac{\partial g}{\partial \theta_i}\left(\boldsymbol{\Theta}_c^{(t)}\right)\right]^2 \tag{14}$$

$$= (1-\alpha)\sum_{j=1}^{t} \alpha^{t-j}\left[\frac{\partial g}{\partial \theta_i}\left(\boldsymbol{\Theta}_c^{(t-j)}\right)\right]^2, \tag{15}$$

where $\alpha \in]0,1[$ is an exponential decay rate. Note that the variance vector $\boldsymbol{v}^{(t)}$ relies on candidates $\boldsymbol{\Theta}_c^{(t)}$ instead of iterates $\boldsymbol{\Theta}^{(t)}$: candidates might not be kept in the Markov chain, but they still contain important information about the shape of the distribution. The RMSProp preconditioner is defined as [16]

$$\boldsymbol{G}^{(t)} = \operatorname{diag}\left(\frac{1}{\eta + \sqrt{\boldsymbol{v}^{(t)}}}\right) \in \mathbb{R}^{ND\times ND}, \tag{16}$$

with η a small damping parameter. This preconditioner has already been used in a MCMC context [1] within an approximate sampler. The goal in [1] was to sample from a distribution defined over the parameters of a neural network trained on a large dataset. Accept or reject steps were omitted as they would have required expensive computations on the full dataset. Additionally, the discretization of the Langevin diffusion process equipped with a position-dependent preconditioner comes with an additional drift term [15] that was neglected in [1]. We correct these two approximations to sample exactly from (12). Following [15], the proposal distribution corresponding to PMALA with the RMSProp preconditioner is the Gaussian distribution:

$$q\left(\boldsymbol{\Theta}_c^{(t)}|\boldsymbol{\Theta}^{(t-1)}\right) = \mathcal{N}\left(\boldsymbol{\Theta}_c^{(t)}|\boldsymbol{\mu}^{(t)}, \boldsymbol{\Lambda}^{(t)}\right) \tag{17}$$

with

$$\begin{cases} \boldsymbol{\mu}^{(t)} = \boldsymbol{\Theta}^{(t-1)} - \frac{\epsilon}{2}\boldsymbol{G}^{(t-1)}\nabla g(\boldsymbol{\Theta}^{(t-1)}) + \epsilon\boldsymbol{\gamma}^{(t-1)}, \\ \boldsymbol{\Lambda}^{(t)} = \epsilon\boldsymbol{G}^{(t-1)}, \end{cases} \tag{18}$$

where ϵ is a step size and $\boldsymbol{\gamma}^{(t-1)}$ is the additional drift term due to the position-dependent preconditioner [15]. In full generality, for all $i \in [\![1, ND]\!]$,

$$\gamma_i^{(t-1)} = \frac{1}{2}\sum_{j=1}^{ND}\frac{\partial G_{ij}^{(t-1)}}{\partial \theta_j^{(t-1)}}. \tag{19}$$

However, the RMSProp preconditioner is diagonal so that the sum in (19) reduces to the $j = i$ term only. Note that $\boldsymbol{\gamma}^{(t-1)}$ is defined from a differentiation with respect to iterate $\boldsymbol{\Theta}^{(t-1)}$ while the variance vector $\boldsymbol{v}$ in (14) is defined from candidates. Since all iterates start as candidates, let $j^{(t)}$ be the number of iterations since last accept: $j^{(t)} = \min\left\{j \geq 0 | \boldsymbol{\Theta}^{(t)} = \boldsymbol{\Theta}_c^{(t-j)}\right\}$. The correction terms $\gamma_i^{(t-1)}$ are then given by

$$\gamma_i^{(t-1)} = -\frac{(1-\alpha)\alpha^{j^{(t-1)}}\left(\frac{\partial g}{\partial \theta_i}\cdot\frac{\partial^2 g}{\partial \theta_i^2}\right)\left(\boldsymbol{\Theta}^{(t-1)}\right)}{2\sqrt{v_i^{(t-1)}}\left(\eta + \sqrt{v_i^{(t-1)}}\right)^2}, \tag{20}$$

To compute $q(\boldsymbol{\Theta}^{(t-1)}|\boldsymbol{\Theta}_c^{(t)}) = \mathcal{N}\left(\boldsymbol{\Theta}^{(t-1)}|\boldsymbol{\mu}_c^{(t)}, \boldsymbol{\Lambda}_c^{(t)}\right)$, one needs to update the variance $\boldsymbol{v}^{(t)}$ and preconditioner $\boldsymbol{G}^{(t)}$ and evaluate the candidate additional drift term $\boldsymbol{\gamma}_c^{(t)}$. By definition, $j^{(t)} = 0$ for candidates, so for all $i \in [\![1, ND]\!]$,

$$\gamma_{c,i}^{(t)} = -\frac{(1-\alpha)\left(\frac{\partial g}{\partial \theta_i}\cdot\frac{\partial^2 g}{\partial \theta_i^2}\right)\left(\boldsymbol{\Theta}^{(t-1)}\right)}{2\sqrt{v_i^{(t)}}\left(\eta + \sqrt{v_i^{(t)}}\right)^2}. \tag{21}$$

The parameters $\boldsymbol{\mu}_c^{(t)}, \boldsymbol{\Lambda}_c^{(t)}$ are thus given by

$$\begin{cases} \boldsymbol{\mu}_c^{(t)} = \boldsymbol{\Theta}_c^{(t)} - \frac{\epsilon}{2}\boldsymbol{G}^{(t)}\nabla g\left(\boldsymbol{\Theta}_c^{(t)}\right) + \epsilon\boldsymbol{\gamma}_c^{(t)}, \\ \boldsymbol{\Lambda}_c^{(t)} = \epsilon\boldsymbol{G}^{(t)}. \end{cases} \tag{22}$$

Algorithm 1 describes the proposed PMALA kernel with RMSProp preconditioner. It relies on three scalar parameters: a damping parameter η, an exponential decay rate α and a step size ϵ. The first two are generally set to $\eta = 10^{-5}$ and $\alpha = 0.99$ [1]. The step size is chosen empirically. MALA achieves optimal convergence rates with an acceptance rate equal to 0.574 when the components of $\boldsymbol{\Theta}$ are independent [9]. Despite the interdependencies in the posterior distribution, we also set ϵ to obtain an average acceptance rate close to 0.574, which yields good results in practice.

B. MTM transition kernel

The non-log-concavity and potential multimodality of the posterior (12) is the second major difficulty to be addressed. After a discussion on state-of-the-art methods, we will propose a MTM kernel. In practice, samplers such as MH, MALA, HMC or even PMALA fail to explore the full distribution when modes are far away: they get stuck in one. Alternative MCMC algorithms dedicated to multimodal distributions have been proposed in the literature. Tempering-based samplers,

Algorithm 1: PMALA kernel $\mathcal{K}_1$ at step t

Input: $\boldsymbol{\Theta}^{(t-1)}$, $\boldsymbol{v}^{(t-1)}$, $j^{(t-1)}$
Output: $\boldsymbol{\Theta}^{(t)}$, $\boldsymbol{v}^{(t)}$, $j^{(t)}$

// Propose candidate
$\boldsymbol{G}^{(t-1)}$ and $\boldsymbol{\gamma}^{(t-1)}$ // using (16), (20)
$\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\Lambda}^{(t)}$ // using (18)
$\boldsymbol{\Theta}_c^{(t)} \sim \mathcal{N}(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Lambda}^{(t)})$
// Accept or reject
$\boldsymbol{v}^{(t)}$, $\boldsymbol{G}^{(t)}$ and $\boldsymbol{\gamma}_c^{(t)}$ // using (14), (16), (21)
$\boldsymbol{\mu}_c^{(t)}$, $\boldsymbol{\Lambda}_c^{(t)}$ and $\rho^{(t)}$ // using (22), (13)
Draw $\zeta \sim \text{Unif}(0,1)$
if $\zeta \leq \rho^{(t)}$ **then** $\boldsymbol{\Theta}^{(t)} = \boldsymbol{\Theta}_c^{(t)}$, $j^{(t)} = 0$
else $\boldsymbol{\Theta}^{(t)} = \boldsymbol{\Theta}^{(t-1)}$, $j^{(t)} = j^{(t-1)} + 1$

Algorithm 2: MTM kernel $\mathcal{K}_2$ at step t

Input: $\boldsymbol{\Theta}^{(t-1)}$
Output: $\boldsymbol{\Theta}^{(t)}$
for $n = 1$ **to** N **do**
  // Propose candidates, select one
  $\boldsymbol{\theta}_n^{(k)} \sim q\left(\boldsymbol{\theta}_n \middle| \boldsymbol{\Theta}_{\setminus n}^{\left(t-1+\frac{n-1}{N}\right)}\right)$ for $k = 1$ to K
  $w\left(\boldsymbol{\theta}_n^{(k)}\right)$ for $k = 1$ to K // using (23)
  w_k for $k = 1$ to K // using (24)
  $i \sim \text{Cat}(w_1, \cdots, w_K)$
  // Accept or reject
  $r_n^{(t)}$ // using (25)
  Draw $\zeta \sim \text{Unif}(0,1)$
  if $\zeta \leq r_n^{(t)}$ **then**
    $\boldsymbol{\theta}_n^{\left(t-1+\frac{n}{N}\right)} = \boldsymbol{\theta}_n^{(i)}$, $\boldsymbol{\Theta}_{\setminus n}^{\left(t-1+\frac{n}{N}\right)} = \boldsymbol{\Theta}_{\setminus n}^{\left(t-1+\frac{n-1}{N}\right)}$
  else $\boldsymbol{\Theta}^{\left(t-1+\frac{n}{N}\right)} = \boldsymbol{\Theta}^{\left(t-1+\frac{n-1}{N}\right)}$

e.g., the Equi-Energy Sampler [29] and the Adaptive Parallel Tempering Algorithm [30], run parallel interacting Markov chains at different temperatures. High temperature chains can navigate between modes and only one chain at low temperature is actually used for estimations. Other methods consider an augmented distribution with a latent mode index and sample it with two kernels: a local kernel explores around a mode and a jump kernel permits jumps between modes. Such methods include Darting MC (DMC) [31], Jumping Adaptative Multimodal Sampler (JAMS) [32], Regeneration Darting MC (RDMC) [33] and Wormhole HMC (WHMC) [34]. WHMC is a particular case of the Riemannian Manifold Hamiltonian Monte Carlo algorithm [14]. The metric of the corresponding manifold combines the standard Euclidean distance and a wormhole metric that shortens the distances between already identified modes, which simplifies transitions from one to another. DMC and JAMS require a prior identification of the distribution modes by some optimization methods. RDMC and WHMC allow running optimization methods in parallel to the sampler and update the distribution and the sampler parameters at random regeneration times [35]. A more complete review of samplers dedicated to multimodal distributions is available in [32]. Most of these methods are computationally very expensive or rely on optimization methods to identify the modes. When $\tilde{\boldsymbol{f}}$ covers multiple decades and is non-linear, the posterior (12) has potentially many modes with only a few of significant weight in the distribution. The identification of relevant modes with standard optimization methods is difficult.

We propose to use a Multiple-Try Metropolis (MTM) [36] kernel that can escape a local mode and explore other ones without any knowledge about the number, positions or variances of the modes. Instead of sampling the whole vector $\boldsymbol{\Theta} \in \mathbb{R}^{ND}$ at once, it uses a Gibbs sampler to decompose it into N individual $\boldsymbol{\theta}_n$. For each conditional distribution, it harnesses an Independent Multiple-Try Metropolis (I-MTM) approach [2], [36], [37]. This method generates $K \geq 1$ candidates $(\boldsymbol{\theta}_n^{(k)})_{k=1}^K$ independently of $\boldsymbol{\theta}_n^{(t-1)}$. This divide-to-conquer approach permits considering N conditional distributions $\pi\left(\boldsymbol{\theta}_n | \boldsymbol{y}_n, \boldsymbol{\Theta}_{\setminus n}^{(t-1)}\right)$ of small dimension, where $\boldsymbol{\Theta}_{\setminus n}^{(t-1)} = \left(\boldsymbol{\theta}_1^{(t-1)}, \ldots, \boldsymbol{\theta}_{n-1}^{(t-1)}, \boldsymbol{\theta}_{n+1}^{(t-1)}, \ldots, \boldsymbol{\theta}_N^{(t-1)}\right)$. Candidates are sampled from a proposal distribution $q\left(\boldsymbol{\theta}_n | \boldsymbol{\Theta}_{\setminus n}^{(t-1)}\right)$ that should be permissive enough to generate candidates in all modes of $\pi\left(\boldsymbol{\theta}_n | \boldsymbol{y}_n, \boldsymbol{\Theta}_{\setminus n}^{(t-1)}\right)$. Then, using an importance weight function w

$$w\left(\boldsymbol{\theta}_n^{(k)}\right) = \frac{\pi\left(\boldsymbol{\theta}_n^{(k)} | \boldsymbol{y}_n, \boldsymbol{\Theta}_{\setminus n}^{(t-1)}\right)}{q\left(\boldsymbol{\theta}_n^{(k)} | \boldsymbol{\Theta}_{\setminus n}^{(t-1)}\right)}, \tag{23}$$

one candidate is selected using a categorical distribution with selection probability w_k for candidate k

$$w_k = \frac{w\left(\boldsymbol{\theta}_n^{(k)}\right)}{\sum_{j=1}^K w\left(\boldsymbol{\theta}_n^{(j)}\right)}. \tag{24}$$

The MH step is then performed with the selected candidate i and the generalized acceptance probability [2], [36]

$$r^{(t)} = 1 \wedge \frac{w\left(\boldsymbol{\theta}_n^{(i)}\right) + \sum_{j=1, j\neq i}^K w\left(\boldsymbol{\theta}_n^{(j)}\right)}{w\left(\boldsymbol{\theta}_n^{(t-1)}\right) + \sum_{j=1, j\neq i}^K w\left(\boldsymbol{\theta}_n^{(j)}\right)}. \tag{25}$$

Algorithm 2 summarizes the MTM sampler. Note that due to the Gibbs approach, it updates one component at a time and returns the result of all updates. A succession of intermediate updates $\left(\boldsymbol{\Theta}^{(t-1+n/N)}\right)_{n=1}^N$ is therefore introduced. This transition kernel relies on the choice of the proposal distribution q and on the number of candidates K generated at each step. This parameter is chosen as a trade-off between computational intensity and average acceptance probability: the higher K, the higher the acceptance probability, the mixing capability but also the computational cost.

In image inverse problems, many common spatial priors are based on a local operator such as the image gradient or Laplacian. In such cases, many components $\boldsymbol{\theta}_n$ are conditionally independent. They can be sampled in parallel using a Chromatic Gibbs sampler [38], which can significantly speed up computations.

C. Proposed sampler and implementation details

To combine a good local exploration of modes as well as jumps between modes, the proposed kernel mixes the PMALA

and MTM transition kernels above. At every step t, the MTM kernel is selected with probability p, and the PMALA kernel with probability $1-p$. Since the MTM kernel divides the parameter space in N D-dimensional subspaces, the PMALA global integer $j^{(t)} \geq 0$ is replaced by a vector $\boldsymbol{j}^{(t)} \in \mathbb{N}^N$, where $j_n^{(t)}$ counts the number of steps since last acceptance for component $\boldsymbol{\theta}_n$. When a component $\boldsymbol{\theta}_n$ is accepted by the MTM kernel, the counter j_n is reset to 0 and the variance component $\boldsymbol{v}_n \in \mathbb{R}^D$ is updated as in (14) with $\frac{\partial g}{\partial \boldsymbol{\theta}_n}\big(\boldsymbol{\Theta}^{(t)}\big)$.

Algorithm 3 reports the complete proposed sampler. Similarly to RDMC and WHMC, the proposed sampler mixes a kernel dedicated to local exploration – PMALA – and another to jump between modes – MTM. The decomposition of the parameter space into N D-dimensional subspaces makes the sampler much simpler than previous approaches. It will perform well in structured problems that allow such decomposition, e.g., images and graphs, and poorly in high-dimensional problems that do not, e.g., Gaussian Mixtures over the full space.

Regarding theoretical properties, the PMALA kernel satisfies the detailed balance property – from [9, theorem 7.2] – and produces ergodic Markov chains – from [9, corollary 7.5]). The proposed MTM kernel is a Metropolis-within-Gibbs algorithm with propositions independent to the current location and with multiple candidates K. In the particular case where $K = 1$, it satisfies the detailed balance property and produces uniformly ergodic Markov Chains – from [39, theorem 7]. Using $K > 1$ candidates in a Multiple-Try Metropolis framework maintains detailed balance and ergodicity [36]. As a mixture of kernels having the same stationary distribution, the proposed kernel also admits the posterior as a stationary distribution – from [9, chapter 10]. As the MTM kernel produces uniformly ergodic Markov chains, so does the proposed mixture kernel – from [9, proposition 10.20]). These results of convergence towards the posterior are mostly asymptotic and also hold for simpler algorithms such as Random Walk Metropolis-Hastings [40]. A comparative theoretical study of non-asymptotic properties that could demonstrate a faster convergence of the proposed sampler is beyond the scope of this paper. However, empirical results show that the proposed sampler yields state-of-the-art performance on multimodal distributions in low-dimensional settings, or higher dimensional applications with relevant low-dimensional conditional distributions.

IV. Numerical experiments

The performance of the proposed method is evaluated on three examples of increasing complexity: (i) sampling a 2D-Gaussian mixture model with unknown modes, (ii) the sensor localization problem [17], and (iii) a synthetic astrophysical inverse problem inspired from [4]. The first two examples focus on the ability of the sampler to efficiently explore multimodal distributions. The astrophysical inverse problem combines all the challenges addressed in Sections II and III: censorship, mixture of noises, forward model spanning multiple decades and multiple local minima.

Algorithm 3: Proposed sampler: PMALA and MTM

Input: number of iterations T, starting point $\boldsymbol{\Theta}^{(0)}$
Output: Markov chain $\{\boldsymbol{\Theta}^{(t)}\}_{t=1}^{T}$

Initialize $v_{nd}^{(0)} = \big[\frac{\partial g}{\partial \theta_{nd}}\big(\boldsymbol{\Theta}^{(0)}\big)\big]^2$ for all n and d
Initialize $\boldsymbol{j}^{(0)} = \mathbf{0}_N$
for $t = 1$ **to** T **do**
  Draw $\zeta \sim \text{Unif}(0,1)$
  if $\zeta > p$ **then** `// PMALA kernel (Algo. 1)`
    $\boldsymbol{\Theta}^{(t)}, \boldsymbol{v}^{(t)}, \boldsymbol{j}^{(t)} = \mathcal{K}_1\big(\boldsymbol{\Theta}^{(t-1)}, \boldsymbol{v}^{(t-1)}, \boldsymbol{j}^{(t-1)}\big)$
  else `// MTM kernel (Algo. 2)`
    $\boldsymbol{\Theta}^{(t)} = \mathcal{K}_2\big(\boldsymbol{\Theta}^{(t-1)}\big)$
    `// Update PMALA parameters`
    for $n = 1$ **to** N **do**
      if *candidate for* $\boldsymbol{\theta}_n$ *was accepted* **then** $\forall d$,
        $v_{nd}^{(t)} = \alpha v_{nd}^{(t-1)} + (1-\alpha)\big[\frac{\partial g}{\partial \theta_{nd}}(\boldsymbol{\Theta}^{(t)})\big]^2$,
        $j_n^{(t)} = 0$, $\boldsymbol{v}_{\backslash n}^{(t)} = \boldsymbol{v}_{\backslash n}^{(t-1)}$, $\boldsymbol{j}_{\backslash n}^{(t)} = \boldsymbol{j}_{\backslash n}^{(t-1)}$

A. Gaussian mixture model

A two-dimensional Gaussian mixture model (GMM) restricted to the square $\mathcal{C} = [-15, 15]^2$ is considered. This simple multimodal distribution, shown on Fig. 1 (top left), is set to contain 15 modes $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. It will demonstrate the ability of the proposed sampler to jump between modes. For simplicity, all the modes have an equal weight in the mixture

$$\pi(\boldsymbol{\Theta}) \propto \left[\sum_{i=1}^{15} \mathcal{N}\left(\boldsymbol{\Theta} | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)\right] \exp\left(-\delta\, \tilde{\iota}_{\mathcal{C}}(\boldsymbol{\Theta})\right), \tag{26}$$

with $\delta = 10^4$. No natural structure decomposition exists for a GMM since each observation consists of $N = 1$ point only in dimension $D = 2$. A Markov chain composed of 10 000 samples is considered, including 100 burn-in samples. To illustrate the role of each of the two kernels in the proposed sampler, two different values are considered for the probability of selecting the MTM kernel: $p = 0.1$ or $p = 0.9$. The number of candidates of the MTM kernel is set to $K = 50$, and the proposal distribution q is the smooth uniform prior on $\mathcal{C}$ (see appendix B). The MTM candidates weights $w\big(\boldsymbol{\theta}_n^{(k)}\big)$ in (23) are then equal to the likelihood term, i.e., the sum of Gaussian pdfs. The default values $\alpha = 0.99$ and $\eta = 10^{-5}$ are considered for the exponential decay and damping factor of the PMALA kernel [1], and its step size is set to $\epsilon = 0.5$. The proposed approach is compared to the state-of-the-art Wormhole Hamiltonian Monte Carlo sampler (WHMC) [34], using the same number of samples. Note that WHMC needs the prior knowledge of mode positions $(\boldsymbol{\mu}_i)_{1 \leq i \leq 15}$, while the proposed kernel does not.

Fig. 1 shows the 2D histograms obtained with the three samplers: WHMC and the proposed ones with either $p = 0.1$ or $p = 0.9$. The three Markov chains efficiently explore all the modes, and their local dispersion obeys the covariance structures equally well. Table I compares their effective sample sizes (ESS) [9] and biases. When $p = 0.9$, the proposed

TABLE I
SAMPLERS COMPARISON ON 2D-GMM.

MCMC sampler	Bias $\|\mathbb{E}[\boldsymbol{\Theta}] - \boldsymbol{\Theta}^*\|$	ESS $\theta_{1,1}$	ESS $\theta_{1,2}$
WHMC [34]	$1.28 \cdot 10^{-1}$	2 753	2 993
Proposed, $p = 0.1$	$7.02 \cdot 10^{-1}$	395	444
Proposed, $p = 0.9$	$\mathbf{4.61 \cdot 10^{-2}}$	**6 157**	**5 780**

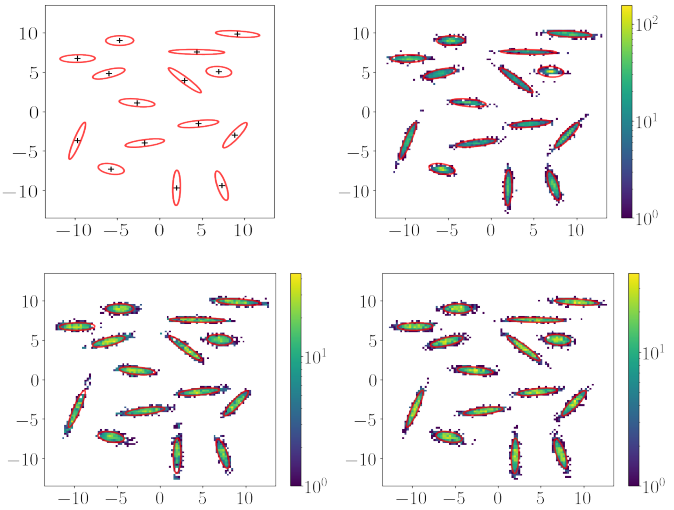

Fig. 1. Sampling results on GMM for WHMC (top right), the proposed kernel with $p = 0.1$ (bottom left) and $p = 0.9$ (bottom right). The red ellipses show the probability level at 2σ. All histograms are in logarithmic norm.

sampler achieves better performances than WHMC, despite the absence of information about the position of the modes $\boldsymbol{\mu}_i$. The high ESS values result from the 85% acceptance rate of the MTM kernel for $K = 50$. However, the MTM kernel with a fixed number of candidates K would not scale up to much higher dimensions. The probability to jump between modes is proportional to the volume of the high probability regions compared to the volume of $\mathcal{C}$, and thus decreases exponentially with the dimension of the problem. The proposed sampler would therefore fail to reach isolated modes in a high-dimensional GMM, whereas WHMC would succeed to do so by exploiting its additional information about the modes. However, the proposed approach focuses on scenarios where the parameter space can be partitioned into a collection of N subspaces of limited dimension D, typically $D \lesssim 10$. The MTM kernel thus remains out of reach from the curse of dimension thanks to the structure of the problem. As in this simple GMM example, the proposed sampler can then outperform WHMC, even without any prior information on the modes of a multimodal distribution.

B. Sensor localization

The sensor localization problem introduced in [17] is a common test case in multimodal sampling, e.g., in [32]–[34]. Three sensors have known locations and will serve as a reference to avoid ambiguities with respect to translation, rotation and negation. The goal is to estimate the unknown positions $\boldsymbol{\Theta} \in \mathbb{R}^{ND}$ of $N = 8$ sensors in dimension $D = 2$.

TABLE II
EFFECTIVE SAMPLE SIZE (ESS) ON THE SENSOR LOCALIZATION PROBLEM.

MCMC sampler	ESS min	ESS mean	ESS max
WHMC [34]	29	1 026	5 753
RDMC [33]	168	3 354	11 192
Proposed, $p = 0.1$	29	329	1 235
Proposed, $p = 0.9$	**299**	**3 561**	**16 789**

The observation matrix $\boldsymbol{Y} \in \mathbb{R}^{NL}$ collects noisy and partially censored pairwise distances between sensors located in a square $[0, 1]^2$, where $L = N+3$ is the total number of sensors. The distance to sensor ℓ feeds channel ℓ, so that the forward model is $f_\ell(\boldsymbol{\theta}_n) = \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_\ell\|$. Note that only $N+2$ distances will really be used since $f_\ell(\boldsymbol{\theta}_\ell) = 0$, and that we set $y_{n,\ell} = 0$ by convention. The probability of communication from sensor $\ell \in [\![1, L]\!]$ to sensor $n \in [\![1, N]\!]$ is set to $\exp\left\{-\frac{f_\ell(\boldsymbol{\theta}_n)^2}{2R^2}\right\}$ with $R = 0.3$. In absence of communication, the observation is censored, which is encoded by the binary latent variable $c_{n,\ell} = 1$. Otherwise, $c_{n,\ell} = 0$ when the observation occurs and is corrupted by a white Gaussian noise

$$y_{n,\ell} = f_\ell(\boldsymbol{\theta}_n) + \epsilon_{n,\ell}, \quad \text{with } \epsilon_{n,\ell} \sim \mathcal{N}(0, \sigma_\epsilon^2), \tag{27}$$

with $\sigma_\epsilon = 0.02$, leading to

$$-\log \pi\left(\boldsymbol{Y}|\boldsymbol{\Theta}\right) = \sum_{n=1}^{N}\sum_{\ell=1}^{L}(1 - c_{n,\ell})\left[\frac{(f_\ell(\boldsymbol{\theta}_n) - y_{n,\ell})^2}{2\sigma_\epsilon^2} + \frac{f_\ell(\boldsymbol{\theta}_n)^2}{2R^2}\right] + c_{n,\ell}\log\left[1 - \exp\left(-\frac{f_\ell(\boldsymbol{\theta}_n)^2}{2R^2}\right)\right], \tag{28}$$

The smoothed uniform prior on the square $\mathcal{C} = [-0.35, 1.2]^2$ is used as a prior on the location of each sensor. The corresponding penalty parameter δ introduced in (10) is set to 10^4. This prior is non-informative enough to match the results shown in [32]–[34]. The proposed sampler is compared to both Regeneration Darting Monte Carlo (RDMC) [33] and WHMC. A Markov chain of size 30 000 is generated by each algorithm, including 5 000 burn-in samples. The parameters of the PMALA kernel are set to $\alpha = 0.99$, $\eta = 10^{-5}$ and $\epsilon = 3 \times 10^{-3}$. The MTM kernel is selected with $p = 0.1$ or $p = 0.9$. Its proposal distribution q is set to the smooth uniform prior on $\mathcal{C}$. For each sensor, the high probability regions are small compared to $\mathcal{C}$. To obtain high acceptance rates for the MTM kernel, the number of candidates is set to $K = 1\,000$. Better proposal distributions can be obtained for this specific problem, which is beyond the scope of this experiment.

Fig. 2 shows the marginal distributions of each sensor position. The four samplers identified the same modes. Table II compares the samplers in terms of ESS. With $p = 0.9$, the proposed sampler yields better mixing capability than WHMC and RDMC. This is due to the partition of the $ND = 16$-dimensional problem into $N = 8$ simpler $D = 2$-dimensional problems. This divide-to-conquer strategy exploits the problem structure to fight the curse of dimension.

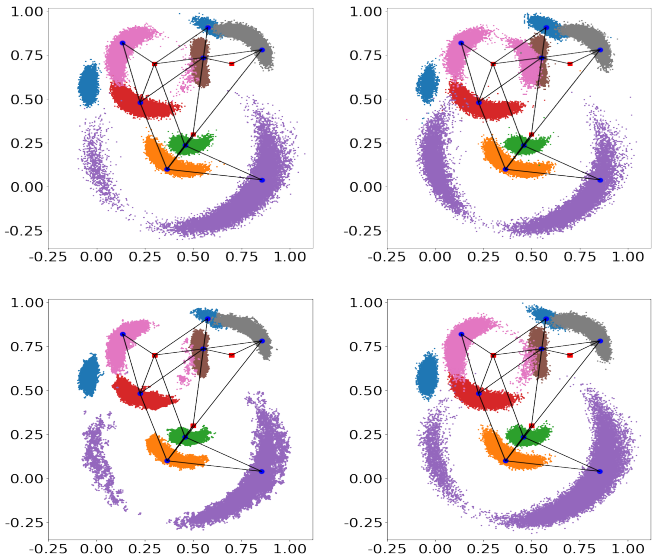

Fig. 2. Marginal distributions of the sensors positions for RDMC (top left), WHMC (top right), proposed with $p = 0.1$ (bottom left) and with $p = 0.9$ (bottom right). The graph shows the true position of all sensors. The sensors with a known position are in red and those whose position is inferred are in blue. The edges of the graph indicate which pairs of sensors are observed.

C. Realistic astrophysical synthetic data

The overall approach is now applied to a synthetic yet realistic inverse problem from astrophysics [4], [5]. The goal is to reconstruct maps of physical parameters of a molecular cloud from radio wave multispectral intensity maps. Each observation map contains $N = 4\,096$ pixels. Each pixel is associated to $D = 4$ physical parameters $\boldsymbol{\varphi} = (\kappa, P_{th}, G_0, A_V)$, so that the aim is to infer a set of parameters $\boldsymbol{\Phi} = (\boldsymbol{\varphi}_n)_{n=1}^N$ in dimension $N \times D = 16\,536$. The parameter κ is a nuisance parameter related to the conditions of observations. It's ground truth value is set to 1 over the whole map. The main parameters of interest are the thermal pressure P_{th}, the intensity of a UV radiative field G_0 and the visual extinction A_V, related to the cloud depth along the line of sight. The ground truth parameters $\boldsymbol{\Phi}^*$ are chosen according to a plausible astrophysical scenario [41]. The physics of the system is encoded within the Meudon PDR code [6], a large numerical simulator. This forward model features many properties that make inference difficult: it is a non-linear model that yields a multimodal posterior distribution, and the amplitude of observations as well as parameters $\boldsymbol{\varphi}$ span several decades. A discrete grid of values $\{(\boldsymbol{\varphi}[g], \boldsymbol{f}[g]), g \in \mathcal{G}\}$ is used to define a normalization process as well as the reduced model. To work with similar scales, the set of estimated parameters $\boldsymbol{\Theta}$ will correspond to normalized values $\boldsymbol{\theta}$ of $\log \boldsymbol{\varphi}$ with respect to empirical averages and variances of $\{\log \boldsymbol{\varphi}[g], g \in \mathcal{G}\}$. To avoid repeated expensive evaluations, the forward model $\boldsymbol{f}$ is reduced to an approximate model $\tilde{\boldsymbol{f}}$, as in (2). For each line ℓ, a polynomial approximation $\widetilde{P}_\ell$ of degree 6 is trained on collection $\{(\boldsymbol{\theta}[g], \log f_\ell[g]), g \in \mathcal{G}\}$. The approximation quality of the resulting $\tilde{\boldsymbol{f}}$ will be considered of sufficient quality to replace exact simulations everywhere. It is used to generate observation maps of $L = 10$ emission lines. For each

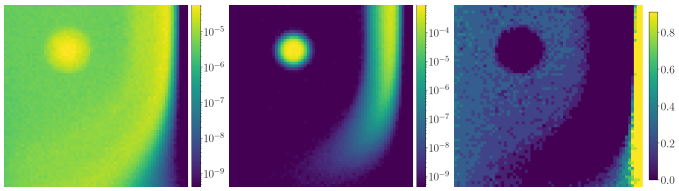

Fig. 3. Some observation maps of the astrophysical experiment. From left to right: line $\ell = 1$, line $\ell = 10$, proportion of censored lines per pixel.

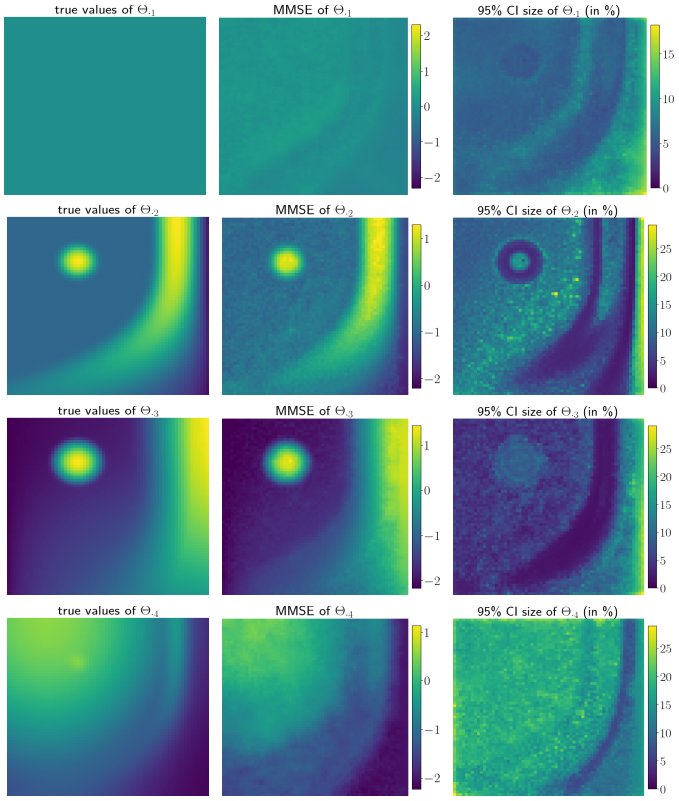

Fig. 4. Inference results: (left) ground truth $\boldsymbol{\Theta}^*$; (middle) MMSE estimate from the proposed transition kernel; (right) size of the 95% credibility intervals (CI) in % of the size of the validity intervals.

line ℓ, $\tilde{f}_\ell$ ranges from 10^{-18} to 10^{-2}. These maps are deteriorated by additive noise, multiplicative noise and censorship following the observation model (1). The standard deviation of the multiplicative noise is set to $\sigma_m = \log(1.1)$, which roughly represents a 10% alteration in average. For the additive noise, $\sigma_a = 1.38715 \cdot 10^{-10}$ so that the Signal-to-Noise Ratio (SNR) varies between -81 and 79 dB. Observations $y_{n,\ell}$ range from about 10^{-10} to 10^{-2}. The censorship level is set to $\omega = 3\sigma_a$. Fig. 3 shows the observation maps of two lines and the spatial distribution of censorship importance.

The likelihood approximation is obtained as indicated in Section II-C2, and its parameters $\boldsymbol{a}_\ell$ are adjusted as described in Appendix A. The validity set $\mathcal{C}$ of physical parameters is set as in [4], and the penalty parameter δ of the smooth uniform prior is set to 10^4. Given the smoothness of the true maps, for

each d the chosen spatial regularizer h is taken as

$$h(\boldsymbol{\Theta}_{\cdot d}) = \|\Delta\boldsymbol{\Theta}_{\cdot d}\|_2^2 = \sum_{n=1}^{N}\sum_{i\in V_n}(\theta_{n,d}-\theta_{i,d})^2, \tag{29}$$

where Δ is the discrete (5-point based) 2D Laplacian operator, and V_n is the set of neighbors of pixel n induced by Δ. The hyperparameter $\boldsymbol{\tau}$ from (10) is fixed to $\boldsymbol{\tau} = (10, 2, 3, 4)$.

Inference is carried out using 10 000 iterations of a Markov chain including 1 500 burn-in samples. The parameters of the proposed sampler are set to $\alpha = 0.99$, $\eta = 10^{-5}$ and $\epsilon = 10^{-6}$ for PMALA, and to $p = 0.5$ and $K = 50$ for MTM. Since the operator Δ only compares a pixel to its four neighbors and since the indicator prior and likelihood are pixel-wise, the set of pixels can be partitioned into two conditionally independent subsets of pixels. A two sites Chromatic Gibbs sampling [38] is therefore performed in the MTM kernel to speed up computations. Note that using the smooth uniform prior as a proposal distribution in MTM is inefficient due to the small size of high probability regions compared to the volume of $\mathcal{C}$. The proposal distribution q is based on the spatial prior (29) instead. For any pixel, one can show that the conditional spatial prior is a Gaussian distribution centered on the mean of the set of neighboring pixels V_n. Since maps are assumed to be smooth, the likelihood functions for a pixel n and its neighbors should correspond to similar modes in the parameters' domain. If the neighbors are not all in the same mode, the mean of the neighbors will in general not fall in a high probability region. Therefore, for a pixel n, the proposal distribution is defined as a Gaussian mixture whose modes are all the means of non-empty subsets $V \in \mathcal{P}(V_n)$ of V_n:

$$q(\boldsymbol{\theta}_n|\boldsymbol{\Theta}_{\setminus n}) \propto \prod_{d=1}^{D}\sum_{V\in\mathcal{P}(V_n)}\exp\left[-2\tau_d\sum_{i\in V}(\theta_{nd}-\theta_{id})^2\right] \tag{30}$$

$$\propto \prod_{d=1}^{D}\sum_{V\in\mathcal{P}(V_n)}\exp\left[-2\tau_d|V|\left(\theta_{nd}-\frac{1}{|V|}\sum_{i\in V}\theta_{id}\right)^2\right]. \tag{31}$$

Performance is assessed for the Minimum Mean Squared Error (MMSE) estimate $\widehat{\boldsymbol{\Theta}}$. Recall that the inferred parameters $\boldsymbol{\Theta}$ correspond to normalized logarithms of physical parameters $\boldsymbol{\Phi}$. Therefore, prediction errors on the D parameter maps $\boldsymbol{\Theta}_{\cdot d}$ are comparable. The quality of the reconstruction is quantified with the Mean Squared Error (MSE) $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_2^2$ and the Reconstruction Signal-to-Noise Ratio (R-SNR) $20\log_{10}\left(\frac{\|\boldsymbol{\Theta}^*\|}{\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^*\|}\right)$.

Fig. 4 shows the estimations results. The MMSE estimate $\widehat{\boldsymbol{\Theta}}$ (middle) is very close to the ground truth $\boldsymbol{\Theta}^*$ (left). The reconstructions are qualitatively very consistent with the underlying physics. The parameter $\boldsymbol{\Theta}_{\cdot 4}$, corresponding to $\varphi_4 = A_V$, is known by astrophysicists to be the most difficult to retrieve as high values lead to saturated line intensities. Such pixels appear in the top left corner of the ground truth map.

Table III shows the MSE and the R-SNR for each parameter $\boldsymbol{\Theta}_{\cdot d}$, and the relative size of the credibility intervals with respect to the associated (normalized) validity interval $\mathcal{C}$. As expected, the MSE is larger for $\boldsymbol{\Theta}_{\cdot 4}$ ($\leftrightarrow \varphi_4 = A_V$), and the relative size of its credibility intervals are overall the largest, about 16.2%. The problem is also very ill-posed for all parameters in pixels with very low SNR, where most of the lines are censored, see Fig. 3 (right). To interpret the results from an astrophysical viewpoint, performances are computed over two subsets of pixels with either less or more than 50% of censored lines. As expected, the credibility intervals of the latter are about twice as large as the former. Finally, all the parameters but A_V are well constrained for pixels with less than 50% of censored lines. The inference remains challenging since the posterior contains many local modes with high g values, but the proposal distribution q permits the Markov chain to successfully reach the mode of interest. The relative quadratic error results in an R-SNR between 15.5 dB and 23.4 dB. Credibility intervals at 95% level remain small, ranging from 5.7% to 9.3% of the admissible interval $\mathcal{C}$.

TABLE III
RECONSTRUCTION METRICS AND RELATIVE SIZE OF CREDIBLE INTERVALS FOR THE ASTROPHYSICS EXPERIMENT. THE R-SNR IS NOT DEFINED FOR $\boldsymbol{\Theta}_{\cdot 1}$, AS ITS GROUND TRUTH IS 0 EVERYWHERE.

	MMSE		Mean 95% credibility intervals size		
	MSE	R-SNR (dB)	censorship $\leq 50\%$	censorship $> 50\%$	overall
$\boldsymbol{\Theta}_{\cdot 1}$	0.017	–	6.1 %	11.9 %	6.8 %
$\boldsymbol{\Theta}_{\cdot 2}$	0.019	16.8	9.3 %	20.6 %	9.9 %
$\boldsymbol{\Theta}_{\cdot 3}$	0.009	23.4	5.7 %	19.8 %	6.5 %
$\boldsymbol{\Theta}_{\cdot 4}$	0.034	15.5	16.3 %	14.5 %	16.2 %

Combining all the difficulties addressed in this article, this astrophysical inverse problem illustrate the good performances of the proposed approach in a challenging scenario. The proposed likelihood approximation enabled handling the censorship and mixture of noises present in the observation model. Dealing with a multimodal posterior distribution, the MTM kernel allows the different modes to be visited, while the PMALA kernel permits to explore them efficiently. The proposed sampler provides high quality estimates and informative credibility intervals.

V. CONCLUSION

This work addresses a family of inverse problems that combine several difficulties: a non-linear black-box forward model, potentially non-injective, that covers multiple decades; observations damaged by both censorship and a mixture of additive and multiplicative noises. The likelihood is intractable and leads to a potentially multimodal posterior distribution. An approximation of the likelihood was proposed, based on a model reduction and an approximate parametric noise mixture model with controlled error. To efficiently sample from the resulting multimodal posterior, an original MCMC algorithm combining two kernels was proposed. The Gibbs-like MTM kernel permits jumps between modes, while the PMALA kernel efficiently explores the local geometry of each mode. The proposed sampler was shown to be competitive with state-of-the-art multimodal sampling methods on a Gaussian mixture model and a sensor localization problem. Motivated by astronomical observation, a more realistic application to a challenging inverse problem has shown the interest and the good performances of the proposed approach. Estimation

errors remain small and uncertainties are quantified. Future work includes applications to real astrophysical data such as the IRAM's Orion-B cloud observations [41] or the James Webb Spatial Telescope observations.

Acknowledgments

This work was partly supported by the CNRS 80Prime project OrionStat, by the ANR project "Chaire IA Sherlock" ANR-20-CHIA-0031-01 held by P. Chainais, by the Programme National "Physique et Chimie du Milieu Interstellaire" (PCMI) of CNRS/INSU with INC/INP, co-funded by CEA and CNES, and by the national support within the *programme d'investissements d'avenir* ANR-16-IDEX-0004 ULNE and Région HDF. We also thank all the ORION-B Consortium (https://www.iram.fr/~pety/ORION-B/team.html).

Appendix A
Choice of likelihood approximation parameters $\boldsymbol{a}_\ell$

For each channel ℓ, the parameter $\boldsymbol{a}_\ell = (a_{\ell,0}, a_{\ell,1})$ locates the frontiers between low, intermediate and high values regimes of $\widetilde{P}_\ell$ in the definition of λ (7). It has a critical influence on the approximation quality. It should be adjusted to $\widetilde{P}_\ell$, σ_a and σ_m. For simplicity, in this subsection, likelihood functions are conditioned with respect to $z = \widetilde{P}_\ell(\boldsymbol{\theta}) \in \mathbb{R}$ instead of $\boldsymbol{\theta} \in \mathbb{R}^D$. The true likelihood is not explicit, but the model (1) can be easily sampled from, and the approximation (8) is known.

The parameter $\boldsymbol{a}_\ell$ is set to obtain an approximation as close as possible to the true likelihood, with respect to some divergence criterion. The Kullback-Leibler (KL) divergence would be a natural choice. However, due to the number of decades spanned, the standard deviation of KL estimators is in practice larger than the quantity of interest [42], which prevents from performing optimization. The Kolmogorov-Smirnov (KS) distance is not affected by this property: for a given z, it only requires ordered samples $(y^{(i)})_{i=1}^M$. It reads

$$\widehat{D}_{\mathrm{KS}}(z, \boldsymbol{a}_\ell) = \sup_{y \in \mathbb{R}} \left| \widehat{F}_M(y|z) - \widetilde{F}(y|z, \boldsymbol{a}_\ell) \right|, \qquad (32)$$

where $\widehat{F}_M(\cdot|z)$ is the empirical cdf of the true likelihood $\pi(\cdot|z)$ estimated from M samples $y^{(i)}$, and $\widetilde{F}(\cdot|z, \boldsymbol{a}_\ell)$ is the cdf of the proposed approximation (8). Assuming that $\boldsymbol{\theta}$ follows a uniform distribution on $\mathcal{C}$ yields a distribution on z with pdf $\pi(z)$ which can be estimated by kernel density estimation (KDE). The function to minimize is

$$\varphi(\boldsymbol{a}_\ell) = \mathbb{E}_z \left[\widehat{D}_{\mathrm{KS}}(z, \boldsymbol{a}_\ell) \right] = \int \widehat{D}_{\mathrm{KS}}(z, \boldsymbol{a}_\ell) \pi(z) dz. \qquad (33)$$

An estimator $\widehat{\varphi}$ can be obtained using numerical integration on z over S bins. The higher M and S, the better the estimation accuracy. Minimizing $\widehat{\varphi}$ can be performed using a grid search, which is quite computationally intensive. A cheaper alternative is to use a Bayesian Optimization (BO) procedure [43]. This optimization was applied for each channel in the astrophysical application described in Section IV-C. Both grid search and BO approaches were used. The KDE of $\pi(z)$ was performed from 810 000 samples. The BO procedure

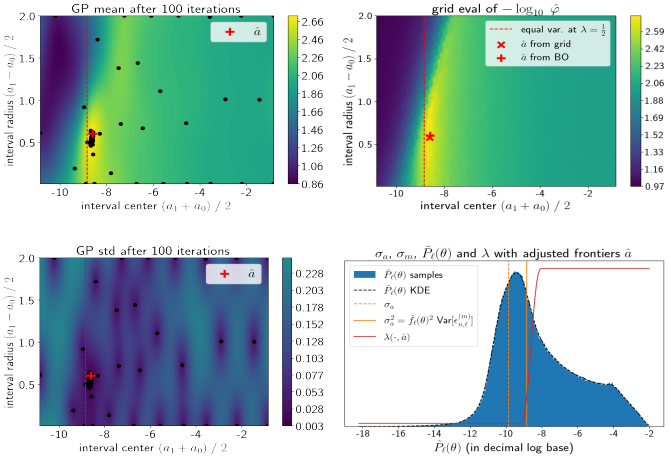

Fig. 5. Maximization of $-\log_{10}\widehat{\varphi}$ using both Bayesian Optimization (BO) and grid search for one channel of the astrophysical case detailed in IV-C. In BO, a Gaussian Process (GP) replaces the function to optimize (left column). The red dashed vertical bar represents the value of $\frac{a_0+a_1}{2}$ for which the additive and multiplicative noises have equal variances, i.e. $\sigma_a^2 = \tilde{f}_\ell(\boldsymbol{\theta})^2 \mathrm{Var}[\epsilon_{n,\ell}^{(m)}]$, at $\lambda = \frac{1}{2}$. For clarity, all scales are displayed in $\log_{10}$ scale, while computations are done in log scale.

was run with $S = 100$ and $M = 250\,000$ using [44] with default parameters. Fig. 5 shows the results for one channel. The proposed approximation with adjusted $\boldsymbol{a}_\ell$ is closer to the true likelihood than a purely additive Gaussian approximation, i.e., $a_{\ell,0} > \max_j z^{(j)}$, or a purely multiplicative lognormal approximations, i.e., $a_{\ell,1} < \min_j z^{(j)}$.

Appendix B
Sampling from smoothed indicator distribution

This section describes the algorithm to draw samples from the real-valued probability distribution with density $\pi(\theta) \propto \exp\left(-\delta\, \tilde{\imath}_{[l,u]}(\theta)\right)$, with $l < u$ and $\tilde{\imath}_{[l,u]}$ introduced in (9). To this aim, consider the generalized normal distribution $\mathcal{GN}(0, 1/\delta^4, 4)$ of probability density function [45]

$$p_{\mathcal{GN}}(\theta) = \frac{2\delta^{\frac{1}{4}}}{\Gamma(1/4)} \exp\left(-\delta\, \theta^4\right). \qquad (34)$$

Note that $\pi(\theta)$ is a continuous extension of a uniform distribution and of this generalized normal distribution at 0.

$$\pi(\theta) \propto \begin{cases} p_{\mathcal{GN}}(\theta - l) & \text{if } \theta < l, \\ p_{\mathcal{GN}}(0) & \text{if } \theta \in [l, u], \\ p_{\mathcal{GN}}(u - \theta) & \text{if } \theta > u. \end{cases} \qquad (35)$$

The normalizing constant of $\pi(\theta)$ is $1 + p_{\mathcal{GN}}(0)(u - l)$. The weight of the uniform section in the combination is therefore

$$w_{\mathrm{Unif}} = \frac{1}{1 + \frac{\Gamma(1/4)}{2} \frac{1}{\delta^{\frac{1}{4}}(u-l)}}. \qquad (36)$$

Algorithm 4 summarizes the procedure to sample from $\pi(\theta)$.

References

[1] C. Li, C. Chen, D. Carlson *et al.*, "Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks," *AAAI*, vol. 30, no. 1, Feb. 2016.

Algorithm 4: Sampling from smooth distribution (35)

Input: scale factor δ, bounds $l, u \in \mathbb{R}$ such that $u > l$
Output: sample θ

w_{Unif} // using (36)
$z \sim \mathcal{B}(w_{\text{Unif}})$
if $z = 1$ **then** $\theta \sim \text{Unif}([l, u])$
else
  $\theta \sim G\mathcal{N}(0, 1/\delta^4, 4)$ // using [46]
  if $\theta < 0$ **then** $\theta = \theta + l$ **else** $\theta = \theta + u$

[2] J. S. Liu, F. Liang, and W. H. Wong, "The Multiple-Try Method and Local Optimization in Metropolis Sampling," p. 15, 2021.

[3] D. M. Walker, D. Allingham, H. W. J. Lee *et al.*, "Parameter inference in small world network disease models with approximate Bayesian Computational methods," *Physica A: Statistical Mechanics and its Applications*, vol. 389, no. 3, pp. 540–548, Feb. 2010.

[4] R. Wu, E. Bron, T. Onaka *et al.*, "Constraining physical conditions for the PDR of Trumpler 14 in the Carina Nebula," *A&A*, vol. 618, p. A53, Oct. 2018.

[5] C. Joblin, E. Bron, C. Pinto *et al.*, "Structure of photodissociation fronts in star-forming regions revealed by *Herschel* observations of high-J CO emission lines," *A&A*, vol. 615, p. A129, Jul. 2018.

[6] F. Le Petit, C. Nehme, J. Le Bourlot *et al.*, "A Model for Atomic and Molecular Interstellar Gas: The Meudon PDR Code," *ASTROPHYS J SUPPL S*, vol. 164, no. 2, pp. 506–529, Jun. 2006.

[7] K. Krissian, C.-F. Westin, R. Kikinis *et al.*, "Oriented Speckle Reducing Anisotropic Diffusion," *IEEE Transactions on Image Processing*, vol. 16, no. 5, pp. 1412–1424, May 2007.

[8] S. Durand, J. Fadili, and M. Nikolova, "Multiplicative Noise Removal Using L1 Fidelity on Frame Coefficients," *J Math Imaging Vis*, vol. 36, no. 3, pp. 201–226, Mar. 2010.

[9] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, ser. Springer Texts in Statistics. New York, NY: Springer New York, 2004.

[10] M. Pereyra, P. Schniter, E. Chouzenoux *et al.*, "A Survey of Stochastic Simulation and Optimization Methods in Signal Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 2, pp. 224–241, Mar. 2016.

[11] D. Luengo, L. Martino, M. Bugallo *et al.*, "A survey of Monte Carlo methods for parameter estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2020, no. 1, p. 25, May 2020.

[12] G. O. Roberts and O. Stramer, "Langevin Diffusions and Metropolis-Hastings Algorithms," *Methodology and Computing in Applied Probability*, vol. 4, no. 4, pp. 337–357, Dec. 2002.

[13] R. Neal, "MCMC Using Hamiltonian Dynamics," in *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. Jones *et al.*, Eds. Chapman and Hall/CRC, May 2011, vol. 20116022.

[14] M. Girolami and B. Calderhead, "Riemann manifold Langevin and Hamiltonian Monte Carlo methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 2, pp. 123–214, 2011.

[15] T. Xifara, C. Sherlock, S. Livingstone *et al.*, "Langevin diffusions and the Metropolis-adjusted Langevin algorithm," *Statistics & Probability Letters*, vol. 91, pp. 14–19, Aug. 2014.

[16] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," pp. 26–31, 2012.

[17] A. Ihler, J. Fisher, R. Moses *et al.*, "Nonparametric belief propagation for self-localization of sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 4, pp. 809–819, Apr. 2005.

[18] P. Palud, P. Chainais, F. L. Petit *et al.*, "Mixture of noises and sampling of non-log-concave posterior distributions," in *2022 30th European Signal Processing Conference (EUSIPCO)*, Aug. 2022, pp. 2031–2035.

[19] M. A. Beaumont, W. Zhang, and D. J. Balding, "Approximate Bayesian Computation in Population Genetics," *Genetics*, vol. 162, no. 4, pp. 2025–2035, Dec. 2002.

[20] J. L. Peterson, K. D. Humbird, J. E. Field *et al.*, "Zonal flow generation in inertial confinement fusion implosions," *Phys. Plasmas*, vol. 24, no. 3, p. 032702, Mar. 2017.

[21] J. Kwan, K. Heitmann, S. Habib *et al.*, "Cosmic Emulation: Fast Predictions for the Galaxy Power Spectrum," *The Astrophysical Journal*, vol. 810, p. 35, Sep. 2015.

[22] M. F. Kasim, D. Watson-Parris, L. Deaconu *et al.*, "Building high accuracy emulators for scientific simulations with deep neural architecture search," *Mach. Learn.: Sci. Technol.*, vol. 3, no. 1, p. 015013, Dec. 2021.

[23] J. Bobin, R. C. Gertosio, C. Bobin *et al.*, "Non-linear interpolation learning for example-based inverse problem regularization," Jun. 2021.

[24] Y. Huang, M. Ng, and T. Zeng, "The Convex Relaxation Method on Deconvolution Model withMultiplicative Noise," *Communications in Computational Physics*, vol. 13, no. 4, pp. 1066–1092, Apr. 2013.

[25] R. Nicholson and J. P. Kaipio, "An Additive Approximation to Multiplicative Noise," *J Math Imaging Vis*, vol. 62, no. 9, pp. 1227–1237, Nov. 2020.

[26] J. Nocedal and S. J. Wright, *Numerical optimization*, 2nd ed., ser. Springer series in operations research. New York: Springer, 2006.

[27] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth *et al.*, "Equation of State Calculations by Fast Computing Machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, Jun. 1953.

[28] W. K. Hastings, "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[29] S. C. Kou, Q. Zhou, and W. H. Wong, "Equi-energy sampler with applications in statistical inference and statistical mechanics," *The Annals of Statistics*, vol. 34, no. 4, pp. 1581–1619, Aug. 2006.

[30] B. Miasojedow, E. Moulines, and M. Vihola, "An Adaptive Parallel Tempering Algorithm," *Journal of Computational and Graphical Statistics*, vol. 22, no. 3, pp. 649–664, 2013.

[31] I. Andricioaei, J. E. Straub, and A. F. Voter, "Smart Darting Monte Carlo," *J. Chem. Phys.*, vol. 114, no. 16, pp. 6994–7000, Apr. 2001.

[32] E. Pompe, C. Holmes, and K. Latuszynski, "A framework for adaptive MCMC targeting multimodal distributions," *Annals of Statistics*, vol. 48, pp. 2930–2952, Oct. 2020.

[33] S. Ahn, Y. Chen, and M. Welling, "Distributed and Adaptive Darting Monte Carlo through Regenerations," in *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*. PMLR, Apr. 2013, pp. 108–116.

[34] S. Lan, J. Streets, and B. Shahbaba, "Wormhole Hamiltonian Monte Carlo," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, ser. AAAI'14. Québec City, Québec, Canada: AAAI Press, Jul. 2014, pp. 1953–1959.

[35] W. R. Gilks, G. O. Roberts, and S. K. Sahu, "Adaptive Markov Chain Monte Carlo through Regeneration," *Journal of the American Statistical Association*, vol. 93, no. 443, pp. 1045–1054, 1998.

[36] L. Martino, "A review of multiple try MCMC algorithms for signal processing," *Digital Signal Processing*, vol. 75, pp. 134–152, Apr. 2018.

[37] L. Martino and J. Read, "On the flexibility of the design of multiple try Metropolis schemes," *Comput Stat*, vol. 28, no. 6, pp. 2797–2823, Dec. 2013.

[38] J. Gonzalez, Y. Low, A. Gretton *et al.*, "Parallel Gibbs Sampling: From Colored Fields to Thin Junction Trees," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, Jun. 2011, pp. 324–332.

[39] G. L. Jones, G. O. Roberts, and J. S. Rosenthal, "Convergence of Conditional Metropolis-Hastings Samplers," *Advances in Applied Probability*, vol. 46, no. 2, pp. 422–445, 2014.

[40] G. O. Roberts and R. L. Tweedie, "Geometric Convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis Algorithms," *Biometrika*, vol. 83, no. 1, pp. 95–110, 1996.

[41] J. Pety, V. Guzman, J. Orkisz *et al.*, "The anatomy of the Orion B Giant Molecular Cloud: A local template for studies of nearby galaxies," *Astronomy & Astrophysics*, vol. 599, Nov. 2016.

[42] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, no. 6, p. 066138, Jun. 2004.

[43] B. Shahriari, K. Swersky, Z. Wang *et al.*, "Taking the Human Out of the Loop: A Review of Bayesian Optimization," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148–175, Jan. 2016.

[44] F. Nogueira, "Bayesian Optimization: Open source constrained global optimization tool for Python," 2014–.

[45] S. Nadarajah, "A generalized normal distribution," *Journal of Applied Statistics*, vol. 32, no. 7, pp. 685–694, Sep. 2005.

[46] M. Nardon and P. Pianca, "Simulation techniques for generalized Gaussian densities," *Journal of Statistical Computation and Simulation*, vol. 79, no. 11, pp. 1317–1329, Nov. 2009.