HAL
open science

Split-and-augmented Gibbs sampler - Application to large-scale inference problems

Maxime Vono, Nicolas Dobigeon, Pierre Chainais

To cite this version:

Maxime Vono, Nicolas Dobigeon, Pierre Chainais. Split-and-augmented Gibbs sampler - Application to large-scale inference problems. IEEE Transactions on Signal Processing, 2019, 67 (6), pp.1648-1661. 10.1109/TSP.2019.2894825 . hal-02438041

HAL Id: hal-02438041

https://hal.science/hal-02438041v1

Submitted on 14 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Split-and-augmented Gibbs sampler – Application to large-scale inference problems

Maxime Vono, Nicolas Dobigeon and Pierre Chainais

Abstract

This paper derives two new optimization-driven Monte Carlo algorithms inspired from variable splitting and data augmentation. In particular, the formulation of one of the proposed approaches is closely related to the alternating direction method of multipliers (ADMM) main steps. The proposed framework enables to derive faster and more efficient sampling schemes than the current state-of-the-art methods and can embed the latter. By sampling efficiently the parameter to infer as well as the hyperparameters of the problem, the generated samples can be used to approximate Bayesian estimators of the parameters to infer. Additionally, the proposed approach brings confidence intervals at a low cost contrary to optimization methods. Simulations on two often-studied signal processing problems illustrate the performance of the two proposed samplers. All results are compared to those obtained by recent state-of-the-art optimization and MCMC algorithms used to solve these problems.

Index Terms

Bayesian inference, data augmentation, high-dimensional problems, Markov chain Monte Carlo, variable splitting.

I. Introduction

Numerous machine learning, signal and image processing problems involve the estimation of a hidden object of interest $\mathbf{x} \in \mathbb{R}^N$ based on (noisy) observations $\mathbf{y} \in \mathbb{R}^M$. This unknown object of interest can stand for parameters of a given model in machine learning [1] or may represent a signal or image to be recovered within an inverse problem. With the increasing

Maxime Vono and Nicolas Dobigeon are with the University of Toulouse, IRIT/INP-ENSEEIHT, CNRS, 2 rue Charles Camichel, BP 7122, 31071 Toulouse cedex 7, France (e-mail: Maxime.Vono@irit.fr, Nicolas.Dobigeon@enseeiht.fr).

Pierre Chainais is with Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France (e-mail: Pierre.Chainais@centralelille.fr).

amount and variety of available data, solving such inference problems in high dimension becomes challenging and generally relies on sophisticated computational inference methods. Those methods are mainly based on stochastic simulation and variational optimization which are two powerful tools to perform inference in complex models [2]. An important class of stochastic simulation techniques is the family of the Markov chain Monte Carlo (MCMC) methods [3]. Within a Bayesian inference framework, MCMC algorithms have the great advantage of providing a comprehensive description of the posterior distribution of the parameter $\mathbf{x}$ to be inferred. Contrary to optimization techniques which generally provide a point estimate, this description permits the subsequent derivation of credibility intervals on the parameter $\mathbf{x}$. Nonetheless, note that optimization algorithms can also bring confidence information when the log-likelihood is supposed differentiable by relying on the theory of large samples [4]. These confidence measures are particulary important for inference problems where very few observations are available (e.g. in biology [5], physics [6] or astrophysics [7]) or when one is interested in extreme events (e.g. in hydrology [8] or cosmology [9]). For instance, MCMC methods have been recently used to conduct Bayesian inference on gravitational waves [10]. However, contrary to optimization techniques, MCMC methods may suffer from their high computational cost which can be prohibitive for high-dimensional problems. To overcome this limitation, a few attempts have been made to derive optimization-driven Monte Carlo methods. The Hamiltonian Monte Carlo method [11], also referred to as hybrid Monte Carlo, is an archetypal example of the successful use of variational analysis concepts (i.e., gradients) to facilitate the exploration of the target distribution. More recently, Pereyra [12] proposed an innovative combination of convex optimization and MCMC algorithms. Capitalizing on the advantages of proximal splitting recently popularized to solve large-scale inference problems [13]–[18], the proximal Monte Carlo method allows high-dimensional log-concave distributions to be sampled. For instance, this algorithm has been successfully used to conduct antisparse coding [19] and has been significantly improved in [20].

Concurrently, variable splitting methods, developed at least 70 years ago [21], have been recently and extensively used to solve large-scale inference problems of the form

$$\arg\min_{\mathbf{x}} \; f(\mathbf{x}) + g(\mathbf{x}), \tag{1}$$

where f commonly refers to a data fitting term and g stands for some regularization function which is often nonsmooth and/or even nonconvex. The main idea of those methods consists in

splitting the variable of interest $\mathbf{x}$ into a pair of variables $\mathbf{x}$ and $\mathbf{z}$ and then solving the counterpart minimization problem

$$\begin{aligned} &\arg\min_{\mathbf{x},\mathbf{z}} \; f(\mathbf{x}) + g(\mathbf{z}), \\ &\text{subject to } \mathbf{x} = \mathbf{z}. \end{aligned} \tag{2}$$

The equality constraint ensures that solving (2) is equivalent to solve the initial problem (1). Exploiting the variable splitting idea, the alternating direction method of multipliers (ADMM) [22], firstly introduced in [23], [24], has proven to be considerably faster than fast iterative thresholding-shrinkage algorithms (FISTA) [25] for solving high-dimensional inverse problems in signal/image processing [26], [27]. This increase in speed comes from the fact that ADMM uses a second-order information of the data fidelity term whereas ISTA or FISTA essentially only takes into account gradient information. The efficiency of ADMM makes it stand as a reference method in high-dimensional signal processing problems such as those encountered in hyperspectral imaging [28], [29]. This paper, in the same spirit as [12], attempts to reconcile optimization and Bayesian inference by proposing two new optimization-driven MCMC algorithms that do not sample directly from the usual target distribution

$$\pi(\mathbf{x}) \propto \exp\left[-f(\mathbf{x}) - g(\mathbf{x})\right], \tag{3}$$

which is assumed to be proper in the sequel. The first one is only based on the idea of variable splitting and considers a joint probability distribution $p(\mathbf{x}, \mathbf{z})$ which tends towards (3) in a limiting case. The main purpose is to work with two simpler distributions $\propto \exp\left[-f(\mathbf{x})\right]$ and $\exp\left[-g(\mathbf{z})\right]$ separately. A similar scheme was recently and independently proposed by [30] in order to distribute Monte Carlo methods on possibly multiple machines. The second proposed approach goes one step further by introducing an auxiliary variable $\mathbf{u} \in \mathbb{R}^N$ within a data augmentation scheme. The main rationales behind the proposed approaches are threefold. Firstly, fully Bayesian approaches allow other parameters (e.g. nuisance or regularization hyper-parameters) to be jointly estimated with the parameter of interest $\mathbf{x}$, avoiding their empirical and painful hand-tuning. Secondly, as emphasized above, samples generated by MCMC algorithms can be used to build confidence intervals on the estimated parameters contrary to optimization techniques that only provide a point estimate. Finally, variable splitting and data augmentation within the proposed approach pave the way towards faster and more efficient samplers.

To this purpose, Section II introduces the hierarchical Bayesian models associated to the proposed approaches. In particular, the main ingredients, namely variable splitting and data

augmentation, are presented. Section III derives the two resulting optimization-driven MCMC algorithms called SP (splitting) and SPA (splitting & augmentation). In particular, a parallel between ADMM and the proposed SPA algorithm is drawn. Section IV considers two often-studied inference problems encountered in signal processing that require to sample respectively from high-dimensional Gaussian and log-concave probability distributions. Section V illustrates the performance of the proposed algorithms on these inference problems. Finally, Section VI draws concluding remarks.

II. Model

This section introduces the proposed approach which aims at using variable splitting and data augmentation to accelerate and simplify the solving of large scale Bayesian inference problems. The main properties of the resulting joint distributions are introduced and its convergence properties towards the usual target distribution (3) are proven. Table I summarizes the main symbols used to define the proposed models.

TABLE I
LIST OF SYMBOLS.

Symbol	Description
$\mathbf{x}, \mathbf{z}, \mathbf{u}, N$	parameter of interest, auxiliary variables and their dimension
$\mathbf{y}, M$	observation vector and its dimension
f, g	data fitting term and regularization function
π	usual target distribution
ϕ_1, ϕ_2	functions associated to the split/augmented scheme
ρ, α	parameters of the proposed approaches
$\mathcal{N}$	normal distribution

A. Variable splitting

Within an optimization framework, variable splitting aims at individually using each term f and g of the objective function in an optimization sub-problem. This divide-to-conquer strategy generally yields simpler proximal operators and therefore an easier algorithm to implement [31].

Following the same intuition, in a Bayesian setting, variable splitting is expected to lead to simpler sampling steps and thereby to a more efficient sampler. Starting from the usual target distribution (3), the introduction of a splitting variable $\mathbf{z} \in \mathbb{R}^N$ leads to so-called *split distribution* defined by

$$\pi_\rho \triangleq p(\mathbf{x}, \mathbf{z}; \rho^2) \propto \exp\left[-f(\mathbf{x}) - g(\mathbf{z}) - \phi_1(\mathbf{x}, \mathbf{z}; \rho^2)\right] \tag{4}$$

where $\phi_1 : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^+$ is a divergence such that π_ρ defines a proper joint distribution and ρ is a positive parameter that controls the dissimilarity between $\mathbf{x}$ and $\mathbf{z}$. Interestingly, the associated conditional distributions that would be considered in a Gibbs algorithm scheme to sample according to (4) are

$$p(\mathbf{x}|\mathbf{z}; \rho^2) \propto \exp\left[-f(\mathbf{x}) - \phi_1(\mathbf{x}, \mathbf{z}; \rho^2)\right] \tag{5}$$

$$p(\mathbf{z}|\mathbf{x}; \rho^2) \propto \exp\left[-g(\mathbf{z}) - \phi_1(\mathbf{x}, \mathbf{z}; \rho^2)\right]. \tag{6}$$

Thus, this variable splitting allows f and g to be dissociated with the hope that these conditional distributions be easy to sample from. Indeed experiments in Section V will show that considering the split distribution π_ρ in (4) instead of π in (3) leads to a faster and more efficient algorithm.

It is worth noting that this variable splitting-based approach can be related to previous works [32], [33], revisited and extended in [34], which also introduced auxiliary variables to split the initial objective function. However, the aforementioned works considered an exact data augmentation scheme which is not the case here, see Theorem 1 below. In addition, this scheme was specifically designed for Bayesian models relying on a Gaussian likelihood function, which is much more restrictive than the target distribution (3) addressed here. Finally, the data augmentation scheme considered in [32], [33] may practically rise some computational difficulty since it requires closed-form expressions of the augmented prior, which could not be available in general. Nonetheless, note that both the latter and the proposed approaches can be interpreted as divide-to-conquer approaches ending up with simpler full conditional distributions.

Within a parallel setting, [30] proposed a similar variable-splitting Bayesian framework motivated by distributed computations when the likelihood function can be expressed as a sum of terms over a possibly big dataset. Their approach can be viewed as a particular instance of the proposed approach when $f(\mathbf{x}) = \sum_{i=1}^{b} f_i(\mathbf{x})$.

The directed acyclic graph (DAG) associated with the proposed splitting model is depicted in Fig. 1 in black and green. Note that sampling from (4) instead of (3) boils down to considering

another hierarchical Bayesian model. However, to ensure the relevance of this extended model and the associated distribution (4) with respect to the inference problem underlied by the target distribution (3), one can expect that ϕ_1 tend to zero when $\mathbf{z}$ is close to $\mathbf{x}$. Thus, if ϕ_1 is a divergence measure where the discrepancy between $\mathbf{x}$ and $\mathbf{z}$ is controlled by ρ^2, it has to satisfy the following assumption that is closely related to the equality constraint $\mathbf{x} = \mathbf{z}$ in variable splitting methods.

Assumption 1: Let $\mathbf{x}$ and $\mathbf{z}$ obeying the distribution (4). Then, ϕ_1 is assumed to be such that, for all $\mathbf{x}, \mathbf{z} \in \mathbb{R}^N$,

$$\lim_{\rho^2 \to 0} \frac{\exp\left[-\phi_1(\mathbf{x}, \mathbf{z}; \rho^2)\right]}{\int_{\mathbb{R}^N} \exp\left[-\phi_1(\mathbf{x}, \mathbf{z}; \rho^2)\right] \mathrm{d}\mathbf{z}} = \delta_{\mathbf{x}}(\mathbf{z}). \tag{7}$$

When this assumption is ensured, the usual target distribution (3) is expected to be recovered from the marginal distribution of $\mathbf{x}$ associated to (4) in the limiting case $\rho \to 0$. This expectation is met when a general form of the divergence ϕ_1 is chosen, as stated by the following theorem.

Theorem 1: Let $p_\rho(\mathbf{x}) = \int_{\mathbb{R}^N} \pi_\rho(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z}$. Then, under Assumption 1, the following result holds

$$\left\| \pi - p_\rho \right\|_{\mathrm{TV}} \xrightarrow[\rho^2 \to 0]{} 0. \tag{8}$$

Proof: See Appendix A. ■

Note that the convergence in total variation implies the convergence in distribution. Thereby, in the limiting case where ρ^2 tends to zero, the marginal distribution of $\mathbf{x}$ under π_ρ coincides with the usual target distribution π. In Section IV, the divergence ϕ_1 will be chosen quadratic. This choice is not a surprise since it is often used in optimization having the great advantage of being differentiable and convex.

B. Data augmentation

The first proposed approach introduces the idea of variable splitting only. It leads to a joint distribution (4) with an additional term ϕ_1 that controls the discrepancy between $\mathbf{x}$ and $\mathbf{z}$. Since ϕ_1 is governed by ρ^2, one might set ρ^2 to a small value to ensure that $\mathbf{x}$ and $\mathbf{z}$ will not be too far from each other (see Theorem 1). However, when sampling from (4) via its conditional distributions (5) and (6), the smaller ρ^2, the higher the correlation between samples, which may deteriorate mixing properties. One option to improve these mixing properties is to consider a data augmentation scheme. Such a strategy consists in introducing auxiliary variables within a target distribution: it is commonly used to build more efficient sampling algorithms [35] with

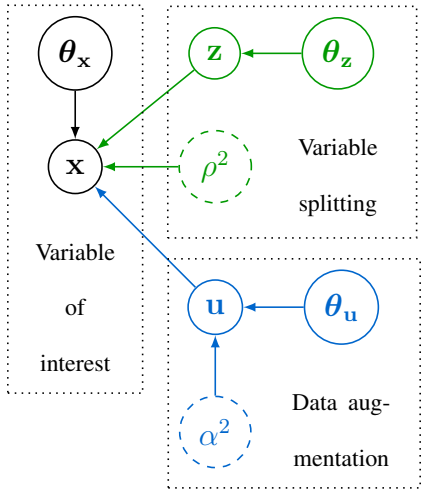

Fig. 1. DAGs associated with the usual and proposed hierarchical Bayesian models. In black: DAG associated to (3); in black and green: DAG associated to (4); in black, green and blue: DAG associated to (10). $\boldsymbol{\theta}_{\mathbf{x}}$, $\boldsymbol{\theta}_{\mathbf{z}}$ and $\boldsymbol{\theta}_{\mathbf{u}}$ stand for possible additional parameters that are not discussed in this paper. (User-defined parameters appear in dashed circles).

less interactions between MCMC draws. This issue was for instance discussed in [36], [37] for the Ising and Potts models. Along these lines, an additional variable $\mathbf{u} \in \mathbb{R}^N$ is introduced in the previous splitting model such that

$$\pi_{\rho,\alpha} \triangleq p(\mathbf{x}, \mathbf{z}, \mathbf{u}; \rho^2, \alpha^2) \tag{9}$$

$$\propto \exp\left[-f(\mathbf{x}) - g(\mathbf{z})\right] \tag{10}$$

$$\times \exp\left[-\phi_1(\mathbf{x}, \mathbf{z} - \mathbf{u}; \rho^2) - \phi_2(\mathbf{u}; \alpha^2)\right]$$

where ϕ_2 is a known function defined on $\mathbb{R}^N$ such that $\pi_{\rho,\alpha}$ defines a proper joint distribution and α is a positive parameter. The DAG associated with the so-called *split-augmented* distribution (10) is depicted in Fig. 1 with additional parameters drawn in blue compared to (4) in black & green only.

The conditional distributions associated with the joint split-augmented distribution (10) are

$$p(\mathbf{x}|\mathbf{z}, \mathbf{u}; \rho^2) \propto \exp\left[-f(\mathbf{x}) - \phi_1(\mathbf{x}, \mathbf{z} - \mathbf{u}; \rho^2)\right] \tag{11}$$

$$p(\mathbf{z}|\mathbf{x}, \mathbf{u}; \rho^2) \propto \exp\left[-g(\mathbf{z}) - \phi_1(\mathbf{x}, \mathbf{z} - \mathbf{u}; \rho^2)\right] \tag{12}$$

$$p(\mathbf{u}|\mathbf{x}, \mathbf{z}; \rho^2, \alpha^2) \propto \exp\left[-\phi_2(\mathbf{u}; \alpha^2)\right]$$

$$\times \exp\left[-\phi_1(\mathbf{x}, \mathbf{z} - \mathbf{u}; \rho^2)\right]. \tag{13}$$

The differences induced by data augmentation are clearly visible when comparing (5) and (6) with (11) and (12). Within a Gibbs sampler scheme, the auxiliary variable $\mathbf{u}$ could allow to decrease the correlation between $\mathbf{x}$ and $\mathbf{z}$ by giving an additional degree of freedom to each of the former variables. Indeed experiments in Section V will show that this data augmentation scheme leads to a sampler with better mixing properties compared to the sampler associated to π_ρ.

However, to assess the relevance of sampling from the split-augmented (SPA) distribution $\pi_{\rho,\alpha}$ in (10) instead of the split (SP) distribution π_ρ in (4), the introduction of $\mathbf{u}$ should not alter the joint distribution (4). Therefore ϕ_1 and ϕ_2 should obey the following assumption.

Assumption 2: Let $\mathbf{x}$, $\mathbf{z}$ and $\mathbf{u}$ obeying the distribution (10). Then, ϕ_2 and ϕ_1 are assumed to be such that for all $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{z} \in \mathbb{R}^N$,

$$\int_{\mathbb{R}^N} \exp\left[-\phi_1(\mathbf{x}, \mathbf{z}-\mathbf{u}; \rho^2) - \phi_2(\mathbf{u}; \alpha^2)\right] \mathrm{d}\mathbf{u} \propto \exp\left[-\phi_1(\mathbf{x}, \mathbf{z}; \eta^2(\rho, \alpha))\right]. \tag{14}$$

where $\eta(\rho, \alpha)$ plays the role of a parameter. In other words, this assumption ensures that a split distribution π_η of the form (4) can be obtained by marginalizing the split-augmented distribution $\pi_{\rho,\alpha}$ in (10) with respect to $\mathbf{u}$. For usual choices of ϕ_1 and ϕ_2, this assumption is satisfied, as stated in the following theorem.

Theorem 2: Let $\mathbf{x}$, $\mathbf{z}$ and $\mathbf{u}$ obeying the distribution (10). In the particular case where ϕ_1 is quadratic that is

$$\phi_1(\mathbf{x}, \mathbf{z}-\mathbf{u}; \rho^2) = \frac{1}{2\rho^2} \left\|\mathbf{x} - (\mathbf{z}-\mathbf{u})\right\|_2^2 \tag{15}$$

and ϕ_2 has the form

$$\phi_2(\mathbf{u}) = \frac{1}{2\alpha^2} \left\|\mathbf{u}\right\|_2^2, \tag{16}$$

Assumption 2 is verified with $\eta^2(\rho, \alpha) = \rho^2 + \alpha^2$ so that

$$\phi_1(\mathbf{x}, \mathbf{z}; \eta(\rho, \alpha)) = \frac{1}{2\left(\rho^2 + \alpha^2\right)} \left\|\mathbf{x} - \mathbf{z}\right\|_2^2. \tag{17}$$

Proof: The proof consists in a straightforward marginalization within a Gaussian model, which can be easily derived, e.g., from computations similar to those in [38, Chap. 10]. ■

In this particular case, it appears that a unique positive parameter $\eta^2(\rho, \alpha) = \rho^2 + \alpha^2$ drives the convergence of the marginal distribution of $\mathbf{x}$ w.r.t. the split distribution π_η, that is of the same form as (4), towards the target distribution π in (3). These quadratic forms of ϕ_1 and ϕ_2 play a

special role. They are closely related to the ADMM (see Section III-B) and will be considered in Section IV.

Eventually, we emphasize that the proposed splitting and data augmentation methods can be easily generalized to cases where there are more than two functions f and g, and when these functions involve distinct linear operators $\mathbf{K}_i$ (subsampling, blur, transform...). In this case, the target distribution can be written as $\pi(\mathbf{x}) \propto \exp\left[-\sum_i h_i(\mathbf{K}_i\mathbf{x})\right]$ where h_i can stand for data fitting, regularization or other types of functions and $\mathbf{K}_i \in \mathbb{R}^{k_i \times N}$ are arbitrary matrices, see Appendix B. For this general case, Theorem 1 holds and the proof can be easily derived with the same type of arguments as in Appendix A. Additionally, Assumption 2 is naturally extended by considering the marginalization of each auxiliary variable $\mathbf{u}_i$.

III. Inference

This section presents two MCMC algorithms to infer the parameter of interest $\mathbf{x}$ either from the split distribution π_ρ in (4) or from the split-augmented distribution $\pi_{\rho,\alpha}$ in (10). In particular, the proposed sampling strategies are discussed for two particular kinds of distributions frequently encountered in signal/image processing or machine learning problems. Additionally, a parallel between the proposed approach and the ADMM is drawn.

A. Gibbs samplers

Two MCMC algorithms, denoted SP (see Algo. 1) and SPA (see Algo. 2), respectively associated with the split and split-augmented distributions (4) and (10) are presented. These algorithms are special instances of Gibbs samplers where samples are alternatively drawn according to the conditional distributions of each variable. Precisely, SP consists in sampling according to (5) and (6), while SPA is defined by the conditional distributions (11)–(13).

As suggested in Section II, the splitting variable $\mathbf{z}$ has been introduced to build faster and simpler simulating schemes compared to the direct sampling from (3). If the conditional distributions of $\mathbf{x}$ and $\mathbf{z}$ are easy to sample from, one can apply Algo. 1 or Algo. 2 directly. If this is not the case despite the variable splitting strategy, one might use surrogates (e.g, Metropolis-Hastings [3] or data augmentation schemes) to sample efficiently from each conditional distribution.

To be more precise, the following paragraphs discuss the efficient sampling of two particular distributions of interest, namely Gaussian and log-concave distributions. These distributions are

Algorithm 1: SP

Input: Functions f, g, ϕ_1, ϕ_2, parameter ρ^2, total number of iterations T_{MC}, number of burn-in iterations T_{bi}, initialization $\mathbf{z}^{(0)}$

for $t \leftarrow 1$ **to** T_{MC} **do**

  % Drawing the variable of interest

  Sample $\mathbf{x}^{(t)}$ according to $p\left(\mathbf{x}|\mathbf{z}^{(t-1)};\rho^2\right)$ (5) ;

  % Drawing the splitting variable

  Sample $\mathbf{z}^{(t)}$ according to according to $p\left(\mathbf{z}|\mathbf{x}^{(t)};\rho^2\right)$ (6) ;

end

Output: Collection of samples $\left\{\mathbf{x}^{(t)},\mathbf{z}^{(t)}\right\}_{t=T_{\mathrm{bi}}+1}^{T_{\mathrm{MC}}}$ asymptotically distributed according to (4).

Algorithm 2: SPA

Input: Functions f, g, ϕ_1, ϕ_2, param. ρ^2, α^2, total nb of iterations T_{MC}, nb of burn-in iterations T_{bi}, initialization $\mathbf{z}^{(0)}$ & $\mathbf{u}^{(0)}$

for $t \leftarrow 1$ **to** T_{MC} **do**

  % Drawing the variable of interest

  Sample $\mathbf{x}^{(t)}$ according to $p\left(\mathbf{x}|\mathbf{z}^{(t-1)},\mathbf{u}^{(t-1)};\rho^2\right)$ (11) ;

  % Drawing the splitting variable

  Sample $\mathbf{z}^{(t)}$ according to $p\left(\mathbf{z}|\mathbf{x}^{(t)},\mathbf{u}^{(t-1)};\rho^2\right)$ (12) ;

  % Drawing the auxiliary variable

  Sample $\mathbf{u}^{(t)}$ according to $p\left(\mathbf{u}|\mathbf{x}^{(t)},\mathbf{z}^{(t)};\rho^2,\alpha^2\right)$ (13) ;

end

Output: Collection of samples $\left\{\mathbf{x}^{(t)},\mathbf{z}^{(t)},\mathbf{u}^{(t)}\right\}_{t=T_{\mathrm{bi}}+1}^{T_{\mathrm{MC}}}$ asymptotically distributed according to (10).

frequently encountered when addressing signal processing and machine learning problems, or may specifically result from the split and/or augment steps induced by the proposed schemes.

1) Gaussian distributions: When f stands for a data fitting term, it is often assumed to be quadratic since quadratic loss functions arise in a wide range of applicative contexts. Within a statistical framework, this choice leads to a likelihood function defined by a Gaussian probability distribution function. Following the same motivation, when g is associated with a penalization, it is often supposed to be quadratic, leading to a Tikhonov regularizer and a Gaussian prior distribution, e.g., used for ridge regression. More precisely, in a general formulation, f and g are assumed to have the form

$$f(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^T \mathbf{Q}_1 (\mathbf{x} - \boldsymbol{\mu}_1) \tag{18}$$

$$g(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^T \mathbf{Q}_2 (\mathbf{x} - \boldsymbol{\mu}_2). \tag{19}$$

where the $\mathbf{Q}_i$ are precision matrices. Then, the corresponding target posterior distribution π is also Gaussian

$$\pi(\mathbf{x}) = \mathcal{N}\left(\mathbf{m}, \mathbf{Q}^{-1}\right) \tag{20}$$

where

$$\begin{cases} \mathbf{Q} = \mathbf{Q}_1 + \mathbf{Q}_2 & (21) \\ \mathbf{m} = \mathbf{Q}^{-1}\left(\mathbf{Q}_1 \boldsymbol{\mu}_1 + \mathbf{Q}_2 \boldsymbol{\mu}_2\right). & (22) \end{cases}$$

If the two terms in (21) cannot be diagonalized in the same basis (e.g., the Fourier domain), then sampling directly from (20) can be computationally intensive since, e.g., it requires to invert the precision matrix $\mathbf{Q}$. In the very particular case where $\mathbf{Q}_1 = \mathbf{H}^T \boldsymbol{\Omega} \mathbf{H}$, if $\mathbf{Q}_2$ and $\mathbf{H}^T\mathbf{H}$ can be diagonalized in the same basis, then direct sampling from the posterior π can be achieved thanks to the specific auxiliary variable method proposed in [39], see also Section V-A. If these requirements are not met, this auxiliary method cannot be implemented. Conversely, the SP and SPA strategies proposed above can be applied to dissociate the precision matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ in the sampling procedure. Indeed, when the divergence ϕ_1 is also chosen quadratic, as in Theorem 1, the conditional distributions associated to $\mathbf{x}$ and $\mathbf{z}$ are Gaussian with precision matrices

$$\mathbf{Q}_{\mathbf{x}} = \mathbf{Q}_1 + \frac{1}{\rho^2}\mathbf{I}_N \tag{23}$$

$$\mathbf{Q}_{\mathbf{z}} = \mathbf{Q}_2 + \frac{1}{\rho^2}\mathbf{I}_N. \tag{24}$$

Again, this demonstrates the main interest of the splitting step which makes the two precision matrices appear in two separate distributions. Now, depending of the respective form of $\mathbf{Q}_1$ and $\mathbf{Q}_2$, one can directly sample from these conditional distributions or use surrogate methods [39]–[42], see Section IV-B and Appendix C-B for more details.

2) Non-smooth log-concave distributions: More generally, if the functions f and g are convex, then the conditional distributions of $\mathbf{x}$ and $\mathbf{z}$ involved in SP and SPA are log-concave. Additionally, when f (resp., g) is non-smooth, if the divergence ϕ_1 is convex, continuously differentiable and gradient Lipschitz, sampling from the conditional distribution associated with $\mathbf{x}$ (resp., $\mathbf{z}$) can be achieved thanks to the proximal Metropolis-adjusted Langevin algorithm (P-MALA) [12] or the proximal Moreau-Yoshida-unadjusted Langevin algorithm (P-MYULA) [20]. For instance, such cases can be encountered when f results from a loss function robust against outliers, e.g., for least absolute deviation regression, or when g stands for a sparsity-inducing regularization. P-MALA and P-MYULA are based on Langevin diffusion process and resort to proximal operators to build Markov chains with interesting convergence properties. The former uses an accept/reject step in order to correct the bias introduced by the considered approximations. On the other hand, the latter removes this Metropolis-Hasting correction step to accelerate the sampling and gives bounds on the convergence rate of the Markov chains.

To summarize, instead of sampling from (3) thanks to the direct use of the previously discussed state-of-the-art MCMC algorithms, the proposed approach aims at preparing and simplifying their implementations to sample according to the conditional distributions associated with the split and split-augmented distributions. In other words, adapted efficient methods are applied to conduct specific and simpler sampling steps where f and g are dissociated. Thereby, the proposed methodology does not aim at totally replacing efficient existing MCMC algorithms but can be interpreted as a "divide-and-conquer" approach that simplifies the task of each sampler to make the whole sampling algorithm faster.

B. When SPA meets ADMM

This "divide-and-conquer" idea is also at the heart of ADMM which allows simpler minimization sub-problems to be considered during the optimization process. This relation with the proposed approach is strengthened by another similarity between SPA and ADMM. More precisely, let consider the particular case where ϕ_1 and ϕ_2 have the forms (15) and (16) respectively

Algorithm 3: ADMM (scaled version)

Input: Functions f, g, penalty parameter ρ^2, initialization $t \leftarrow 0$ and $\mathbf{z}^{(0)}$,$\mathbf{u}^{(0)}$

while *stopping criterion not satisfied* **do**

  % Minimization w.r.t. $\mathbf{x}$

  $\mathbf{x}^{(t)} \in \arg\min_{\mathbf{x}} - \log p\left(\mathbf{x}|\mathbf{z}^{(t-1)}, \mathbf{u}^{(t-1)}; \rho^2\right)$;

  % Minimization w.r.t. $\mathbf{z}$

  $\mathbf{z}^{(t)} \in \arg\min_{\mathbf{z}} - \log p\left(\mathbf{z}|\mathbf{x}^{(t)}, \mathbf{u}^{(t-1)}; \rho^2\right)$;

  % Dual ascent

  $\mathbf{u}^{(t)} = \mathbf{u}^{(t-1)} + \mathbf{x}^{(t)} - \mathbf{z}^{(t)}$;

  % Updating iterations counter

  $t \leftarrow t + 1$;

end

Output: Approximate solution of the optimization problem $\hat{\mathbf{x}}$.

(in agreement with the assumptions required by Theorems 1 and 2), and assume that f and g are convex. Then, computing the MAP estimates instead of sampling in each step of Algo. 2 boils down to the ADMM [22], see Algo. 3. Within this optimization framework, $\mathbf{z}$ corresponds to the splitting variable, $\mathbf{u}$ stands for the scaled Lagrange multiplier and ρ^{-2} for the penalty parameter.

The ADMM is known to be an efficient optimization algorithm for high-dimensional problems. It simplifies the optimization problem by considering several simpler optimization sub-problems where advanced optimization tools and methods (e.g., proximal operators) can be embedded and applied efficiently. Additionally, it covers a large panel of optimization problems and can be generalized to the case where more than two functions f and g are considered. As noticed in the previous section, this generalization property also applies to the proposed SP and SPA methods, see Appendix B.

These advantages are retrieved using the proposed approach which draws a general framework to solve large-scale Bayesian inference problems. Finally, as it will be shown in Section V, the proposed SP and SPA algorithms need few fast iterations (akin to ADMM) to reach the same performance as state-of-the-art MCMC methods with good mixing properties.

IV. Application to linear Gaussian inverse problems

In this section, the proposed splitting-and-augmenting strategy is envisioned to address two particular instances of linear Gaussian inverse problems formulated within a Bayesian framework. It first defines the considered class of problems and then derives the proposed approaches on two often-studied particular cases. Note that only the derivation of the SPA algorithm is discussed since it naturally embeds SP. However, the conclusions made hereafter stand also for SP. In Section V, results of experiments associated to these two inverse problems will be reported and discussed.

A. Linear Gaussian inverse problems

Linear Gaussian inverse problems define an archetypal class of problems that could be efficiently tackled by the models and algorithms introduced in Sections II and III. Suppose that some noisy signals $\mathbf{y}$ are observed and one wants to infer an hidden parameter $\mathbf{x}$ under the linear model

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{e} \tag{25}$$

where $\mathbf{H}$ is a direct operator and $\mathbf{e}$ stands for noise or error modeling. Then, assuming that $\mathbf{e}$ is a Gaussian random vector with covariance matrix $\mathbf{\Omega}^{-1}$, the likelihood distribution associated with the observation vector $\mathbf{y}$ is

$$p\left(\mathbf{y}|\mathbf{x}\right) \propto \exp\left[-\frac{1}{2}\left(\mathbf{H}\mathbf{x}-\mathbf{y}\right)^T \mathbf{\Omega}\left(\mathbf{H}\mathbf{x}-\mathbf{y}\right)\right]. \tag{26}$$

In most applicative contexts, $\mathbf{H}$ is not invertible and inferring the unknown parameter vector $\mathbf{x}$ from the observation vector $\mathbf{y}$ under the linear model (25) is known to be an ill-posed inverse problem. To alleviate this issue, a convenient and widely admitted approach consists in adopting some sort of regularization. Within a Bayesian setting, this is done by assigning a prior distribution to the unknown parameter vector $\mathbf{x}$. Assuming that this prior distribution is given by the general form

$$p\left(\mathbf{x}\right) \propto \exp\left[-g(\mathbf{x})\right], \tag{27}$$

it follows by applying Bayes' rule that the posterior distribution of $\mathbf{x}$ has the same form as (3) where $f(\mathbf{x}) = \frac{1}{2}\left(\mathbf{H}\mathbf{x}-\mathbf{y}\right)^T \mathbf{\Omega}\left(\mathbf{H}\mathbf{x}-\mathbf{y}\right)$. As a consequence, the proposed methodology can be implemented to sample efficiently from a close approximation of this posterior distribution and use these samples to infer the hidden parameter $\mathbf{x}$. In the sequel, two standard problems

involving Gaussian and total variation (TV) prior distributions, respectively, are considered. One can easily verify that Assumptions 1 and 2 along with Theorem 1 hold for all these problems.

B. Deconvolution with a smooth prior

In the setup considered in this paragraph, the function g in (27) is chosen to be quadratic as in (19) with $\boldsymbol{\mu}_2 = \mathbf{0}_N$ and $\mathbf{Q}_2 = \gamma \mathbf{L}^T\mathbf{L}$, where $\mathbf{L}$ is a circulant matrix associated to a Laplacian filter. These choices lead to a frequently encountered smoothing conjugate Gaussian prior $\mathcal{N}\left(\mathbf{0}_N, \left(\gamma \mathbf{L}^T\mathbf{L}\right)^{-1}\right)$, for instance used in [43]–[45]. Note that this Gaussian prior distribution is degenerated since constant images are not penalized leading to the first eigenvalue of $\mathbf{Q}_2$ being equal to zero. Thus the posterior distribution (20) becomes

$$\pi(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{m}, \mathbf{Q}^{-1}\right) \tag{28}$$

where

$$\begin{cases} \mathbf{Q} = \mathbf{H}^T\boldsymbol{\Omega}\mathbf{H} + \gamma \mathbf{L}^T\mathbf{L} & (29) \\ \mathbf{m} = \mathbf{Q}^{-1}\mathbf{H}^T\boldsymbol{\Omega}\mathbf{y}. & (30) \end{cases}$$

Additionally, in the sequel, the operator $\mathbf{H}$ will be assumed to be an $N \times N$ circulant convolution matrix associated to a time/space-invariant blurring kernel. Finally, the noise covariance matrix is assumed to be diagonal, i.e., $\boldsymbol{\Omega}^{-1} = \mathrm{diag}[\sigma_1^2, \ldots, \sigma_N^2]$. Direct sampling according to the posterior distribution (28) is a challenging task, mainly due to the presence of the precision matrix $\boldsymbol{\Omega}$. Indeed, as emphasized in paragraph III-A1, the two terms in (29) cannot be diagonalized in the same basis (e.g. Fourier) which leads to computational problems in high dimension.

Conversely, assuming that ϕ_1 and ϕ_2 have the form (15) and (16) with parameters ρ and α, the proposed SPA Gibbs algorithm samples according to the conditional distributions

$$p(\mathbf{x}|\mathbf{z}, \mathbf{u}) = \mathcal{N}\left(\mathbf{m}_{\mathbf{x}}, {\mathbf{G}_{\mathbf{x}}}^{-1}\right) \tag{31}$$

$$p(\mathbf{z}|\mathbf{x}, \mathbf{u}) = \mathcal{N}\left(\mathbf{m}_{\mathbf{z}}, {\mathbf{G}_{\mathbf{z}}}^{-1}\right) \tag{32}$$

$$p(\mathbf{u}|\mathbf{x}, \mathbf{z}) = \mathcal{N}\left(\mathbf{m}_{\mathbf{u}}, {\mathbf{G}_{\mathbf{u}}}^{-1}\right) \tag{33}$$

where

$$\begin{cases} \mathbf{G}_{\mathbf{x}} = \mathbf{H}^T\boldsymbol{\Omega}\mathbf{H} + \dfrac{1}{\rho^2}\mathbf{I}_N & (34) \\ \mathbf{G}_{\mathbf{z}} = \gamma \mathbf{L}^T\mathbf{L} + \dfrac{1}{\rho^2}\mathbf{I}_N & (35) \\ \mathbf{G}_{\mathbf{u}} = \dfrac{\alpha^2 + \rho^2}{\alpha^2\rho^2}\mathbf{I}_N. & (36) \end{cases}$$

Thanks to the splitting-and-augmenting approach, these three sampling steps are much easier to handle than the direct sampling from the target posterior distribution (28). Indeed, sampling from (31) can be conducted by using the auxiliary method of [39] to deal separately with $\mathbf{H}^{\mathbf{T}}\mathbf{H}$ from the coupling induced by $\boldsymbol{\Omega}$ (see Appendix C-B). Additionally, sampling from (32) can be efficiently achieved in the Fourier domain (see Appendix C-A for details). Finally, sampling from (33) is straightforward since the covariance matrix is diagonal. Again, as previously noticed in Section III and more particularly in paragraph III-A1 dedicated to Gaussian distributions, the proposed splitting-and-augmenting allows specific and simpler sampling steps to be conducted where the difficulties inherent to f (here the Gaussian likelihood) and g (here the Gaussian prior) have been dissociated. The strategy developed in this paragraph will be experimentally assessed in paragraph V-A.

C. Image inpainting with total variation

TV has become an ubiquitous regularization to solve imaging problems [46]–[48]. Within the considered Bayesian framework, it consists in choosing the g function in (27) as $g(\mathbf{x}) = \beta \mathrm{TV}(\mathbf{x})$ where $\beta > 0$ and $\mathrm{TV}(\mathbf{x}) = \sum_{1 \leq i,j \leq N} \left\| (\nabla \mathbf{x})_{i,j} \right\|_2$ ($\nabla \mathbf{x}$ is the two-dimensional discrete gradient of $\mathbf{x}$). This type of prior is used for instance in image inpainting problems, which consist in recovering an original image $\mathbf{x} \in \mathbb{R}^N$ from the noisy and partial measurements $\mathbf{y} \in \mathbb{R}^M$ under the linear model (25). Note that, in general, $M \ll N$. Here, the noise is assumed to be white and Gaussian such that $\boldsymbol{\Omega}^{-1} = \sigma^2 \mathbf{I}_M$ and the operator $\mathbf{H}$ stands for the matrix associated with a damaging binary mask. Under this setting, the posterior distribution of $\mathbf{x}$ (3) becomes

$$p\left(\mathbf{x}|\mathbf{y}\right) \propto \exp\left[-\frac{1}{2\sigma^2} \left\|\mathbf{H}\mathbf{x} - \mathbf{y}\right\|_2^2 - \beta \mathrm{TV}(\mathbf{x})\right]. \tag{37}$$

Direct sampling from this posterior is a challenging task mainly due to *i)* the generally high dimension of the image to be recovered, *ii)* the non-conjugacy of the TV-based prior, leading to a non-standard posterior distribution and *iii)* the non-differentiability of g which precludes the use of some advanced simulation techniques, e.g., Hamiltonian Monte Carlo algorithms [11]. Conversely, instead of directly sampling from this posterior distribution, the proposed approach is applied. Again, assuming that ϕ_1 and ϕ_2 have the forms (15) and (16) with parameters ρ and

α, respectively, the conditional distributions associated to SPA are

$$p(\mathbf{x}|\mathbf{z},\mathbf{u}) \propto \exp\left[-\frac{1}{2\sigma^2}\left\|\mathbf{H}\mathbf{x}-\mathbf{y}\right\|_2^2\right] \times \exp\left[-\frac{1}{2\rho^2}\left\|\mathbf{x}-(\mathbf{z}-\mathbf{u})\right\|_2^2\right] \tag{38}$$

$$p(\mathbf{z}|\mathbf{x},\mathbf{u}) \propto \exp\left[-\beta\mathrm{TV}(\mathbf{z})-\frac{1}{2\rho^2}\left\|\mathbf{z}-(\mathbf{x}+\mathbf{u})\right\|_2^2\right] \tag{39}$$

$$p(\mathbf{u}|\mathbf{x},\mathbf{z}) \propto \exp\left[-\frac{1}{2\alpha^2}\left\|\mathbf{u}\right\|_2^2-\frac{1}{2\rho^2}\left\|\mathbf{u}-(\mathbf{z}-\mathbf{x})\right\|_2^2\right] \tag{40}$$

Here, assuming that ϕ_1 and ϕ_2 are quadratic allows to retrieve Gaussian distributions for (38) and (40). Sampling from (40) in high-dimension is not a problem since the covariance matrix is constant diagonal. However, the covariance matrix associated to (38) is $\left(\sigma^{-2}\mathbf{H}^T\mathbf{H}+\rho^{-2}\mathbf{I}_N\right)^{-1}$, which is more complex to handle. Hopefully, the direct operator $\mathbf{H}$ is a $M \times N$ binary matrix which can be obtained by taking a subset of rows of the identity matrix in dimension N. Due to this simple structure, $\mathbf{H}\mathbf{H}^T = \mathbf{I}_M$ and by using the Sherman-Morrison-Woodbury formula, it follows that

$$\left(\frac{1}{\sigma^2}\mathbf{H}^T\mathbf{H}+\frac{1}{\rho^2}\mathbf{I}_N\right)^{-1} = \rho^2\left(\mathbf{I}_N-\frac{\rho^2}{\sigma^2+\rho^2}\mathbf{H}^T\mathbf{H}\right). \tag{41}$$

The matrix $\mathbf{H}^T\mathbf{H}$ corresponds to an identity matrix with some zeros in the diagonal (corresponding to the missing pixels). Thereby, the covariance matrix (41) is diagonal and the sampling from (38) can be conducted efficiently with the exact perturbation-optimization (E-PO) algorithm [40].

As previously discussed in paragraph III-A2, the conditional distribution (39) being log-concave, one can sample efficiently from the latter in high-dimension with P-MALA or P-MYULA. In the sequel, P-MYULA will be preferred because its mixing properties are better than P-MALA and the estimation error is of the order of 1% using well-defined parameters [20]. As a conclusion, as advocated earlier, the proposed splitting-and-augmenting approach allows simpler sampling steps to be efficiently conducted thanks to dedicated algorithms.

V. Experiments

This section reports results of experiments aimed at comparing the proposed methodology with that of current state-of-the-art (optimization and Bayesian) methods for the inverse problems discussed in Section IV. All the results presented in this section have been obtained using MATLAB, on a computer equipped with an Intel Xeon 3.70 GHz processor, with 16.0 GB of

RAM, and running Windows 7. Other examples of the proposed approach on machine learning problems can be found in [30], [49].

A. Deconvolution with a smooth prior

1) Problem considered: The Gaussian sampling problem introduced in Section IV-B is considered. A blurred and noisy image $\mathbf{y} \in \mathbb{R}^M$ of size 512×512 ($M = 262144$) is observed. The purpose is then to recover the original image $\mathbf{x} \in \mathbb{R}^N$ of size 512×512 ($N = 262144$).

2) Experimental design: The diagonal elements σ_i^2 of the noise covariance matrix $\mathbf{\Omega}^{-1}$ have been randomly drawn according to the mixture $\sigma_i \sim (1-\beta)\delta_{\kappa_1} + \beta\delta_{\kappa_2}$ ($\kappa_1, \kappa_2 > 0$ and $0 < \beta < 1$) with $\beta = 0.35$, $\kappa_1 = 13$ and $\kappa_2 = 40$. This particular structure for $\mathbf{\Omega}^{-1}$ may be not physical but permits to show the interest of the proposed approach. The prior parameter γ has been set to $\gamma = 6 \times 10^{-3}$.

The proposed SP and SPA algorithms SP are compared to RJ-PO [41] and to the algorithms denoted AuxV1 and AuxV2 proposed in [39]. The parameters associated to SP and SPA have been set to $\rho = 20$ and $(\rho, \alpha) = (20, 1)$, respectively. RJ-PO has been run using conjugate gradient (CG) algorithm as the required linear solver whose tolerance has been adapted to reach an acceptance rate of 0.9. The number of burn-in iterations has been set to $T_{\text{bi}} = 200$ for AuxV1, RJ-PO, SP and SPA and to $T_{\text{bi}} = 2200$ for AuxV2 (due to its slower mixing properties, see below). For each MCMC algorithm, 800 samples obtained after the burn-in period have been used. The number of iterations T_{MC} and T_{bi} were empirically chosen by graphically inspecting the behavior of the Markov chains produced by the samplers.

The performances of the different approaches have been assessed by the signal-to-noise ratio (SNR) and the peak signal-to-noise ratio (PSNR)

$$\text{SNR} = 10 \log_{10} \frac{\|\mathbf{x}\|_2^2}{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2} \tag{42}$$

$$\text{PSNR} = 10 \log_{10} \frac{255^2}{N^{-1} \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2} \tag{43}$$

where $\hat{\mathbf{x}}$ refers to the MMSE estimate of $\mathbf{x}$ approximated by empirical averages of the samples generated by the MCMC algorithms. The performance results have been averaged over 25 Monte Carlo runs.

3) Results: Table II shows the average SNR and PSNR associated to the MMSE estimate for the different algorithms. The standard deviation associated to these results is the same for

TABLE II

GAUSSIAN SAMPLING: AVERAGE SNR AND PSNR (OVER 25 OBSERVATIONS) ASSOCIATED TO THE MMSE ESTIMATES.

	SNR (dB)	**PSNR (dB)**
RJ-PO	19.58	25.24
AuxV1	19.58	25.24
AuxV2	19.60	25.26
SP	19.58	25.23
SPA	19.58	25.23

the different methods and is equal to 0.02 and all the algorithms share similar performance results. However, we emphasize that the computational cost of each algorithm can differ widely as shown by Table III.

Table III presents the numerical complexity related to one iteration of each algorithm along with the average number of iterations performed and the average computational time for each algorithm (over the 25 Monte Carlo runs). The complexity of $\mathcal{N}$ refers to the sampling from an univariate normal distribution. The complexity of $O(N \log N)$ refers to the use of the Fourier transform as the matrices $\mathbf{H}$ and $\mathbf{L}$ are circulant and thereby diagonalizable in the Fourier domain. One can denote that SP, SPA, AuxV1 and AuxV2 share a roughly similar numerical complexity (for one iteration) whereas RJ-PO is slower because of the use of the CG method. The latter has a complexity of $O(N_{\text{CG}} N \log N)$ where N_{CG} is the number of iterations performed by the CG method. In this example, $N_{\text{CG}} = 155$ on average (after the burn-in period). On the other hand, the average computing times associated to each MCMC algorithm widely differ. RJ-PO is the slowest mainly due to the number of CG iterations performed at each iteration. AuxV1 appears to be the fastest. However, one has to recall that this algorithm was explicitly designed for this type of inference problems and cannot be used directly for more general Gaussian sampling tasks. SP and SPA appear to have reasonable computational costs compared to AuxV1. Finally, AuxV2 needs more iterations and thereby more time to reach the same level of performance as the other approaches. This algorithm can be used in more general cases than AuxV1 but appears to be roughly 3 times more costly than the proposed approach which covers a wider scope of sampling problems. This high computational cost is mainly related to the poor mixing properties of AuxV2 compared to the other methods as drawn by Fig. 2.

TABLE III

GAUSSIAN SAMPLING: COMPUTATIONAL COMPLEXITY RELATED TO ONE ITERATION, AVERAGE NUMBER OF ITERATIONS AND AVERAGE COMPUTATIONAL TIME FOR EACH ALGORITHM.

	computational complexity	**# iterations**	**time (s)**
RJ-PO	$O(N_{\mathrm{CG}} N \log N) + (M+N)\mathcal{N}$	10^3	4192
AuxV1	$O(N \log N) + 2N\mathcal{N}$	10^3	37
AuxV2	$O(N \log N) + 4N\mathcal{N}$	3×10^3	209
SP	$O(N \log N) + 3N\mathcal{N}$	10^3	62
SPA	$O(N \log N) + 4N\mathcal{N}$	10^3	86

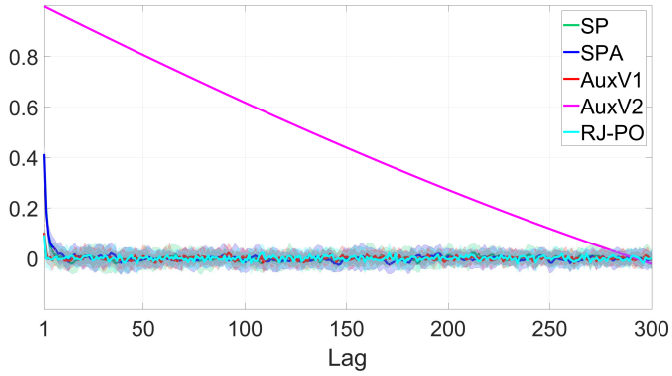

Fig. 2. Gaussian sampling: average chain autocorrelation functions of SP (green), SPA (blue), AuxV1 (red), AuxV2 (magenta) and RJ-PO (cyan). Shaded areas represent the intervals corresponding to the standard deviation computed over 25 trials.

Fig. 2 compares the autocorrelation functions (using $-\log \pi(\mathbf{x}|\mathbf{y})$ as a scalar summary) of AuxV1, AuxV2, RJ-PO, SP and SPA averaged over the 25 Monte Carlo runs, where only samples obtained after the burn-in period have been considered. The shaded regions depicted in Fig. 2 represent the standard deviation ranges associated to each MCMC algorithm. One can denote that all the algorithms share good mixing properties except AuxV2 which explores less efficiently the parameter space. This result is consistent with the findings highlighted in [39] which pointed out that the quality of the samples generated by RJ-PO and AuxV1 was better than those generated by AuxV2.

4) Discussion: For this specific experiment, the proposed general splitting-and-augmenting framework has shown that it can compete with efficient algorithms designed only for this type of sampling problems (e.g. AuxV1). Additionally, it proves to be more efficient than algorithms

Fig. 3. Set of 256×256 images used. From top left to bottom right: balloons, baboon, elaine, clock, donna, house, peppers, cameraman, boat.

designed for wider Gaussian sampling tasks (e.g. AuxV2 and RJ-PO). The performance of the proposed approach is strengthened by the fact that SP and SPA have also demonstrated to be more efficient than state-of-the-art MCMC algorithms designed to sample from other types of distributions, such as log-concave densities, as illustrated in the next paragraph V-B.

B. Image inpainting with total variation

1) Problem considered: The image inpainting problem introduced in Section IV-C and also addressed in [26] is considered here. Fig. 3 presents the nine 256×256 original gray-level images used for this experiment. The observation vector denoted $\mathbf{y}$ consists of 60% randomly selected of the original image pixels $\mathbf{x}$, corrupted by a white Gaussian noise with SNR of 40dB.

Fig. 4a and 4b present, as an example, the original Cameraman image and one of its associated observations where the missing pixels are depicted in white. The restoration results for this image are also presented in Fig. 4.

2) Experimental design: The two proposed algorithms SP and SPA, leading to sampling from (38)-(40), are compared with the split augmented Lagrangian shrinkage algorithm (SALSA) [26], which can be interpreted as a deterministic counterpart of SPA, as emphasized in paragraph III-B. SALSA solves the minimization problem resulting from the MAP inference associated with the posterior distribution (37) by using ADMM. These algorithms have been also compared with P-MYULA specifically designed to sample from possibly non-smooth log-concave distributions (see paragraph III-A2). The number of burn-in iterations has been set to $T_{\mathrm{bi}} = 200$ for SP and SPA and to $T_{\mathrm{bi}} = 95200$ for P-MYULA (due to slower mixing, see below). For each MCMC algorithm, 4800 samples obtained after the burn-in period have been used to approximate the MMSE estimator by empirical averaging.

Sampling from (39) has been done with P-MYULA ($\lambda = \rho^2$ and $\gamma = \rho^2/4$) using Chambolle's algorithm [50] to compute the proximal operator of g. The SP and SPA parameters have been

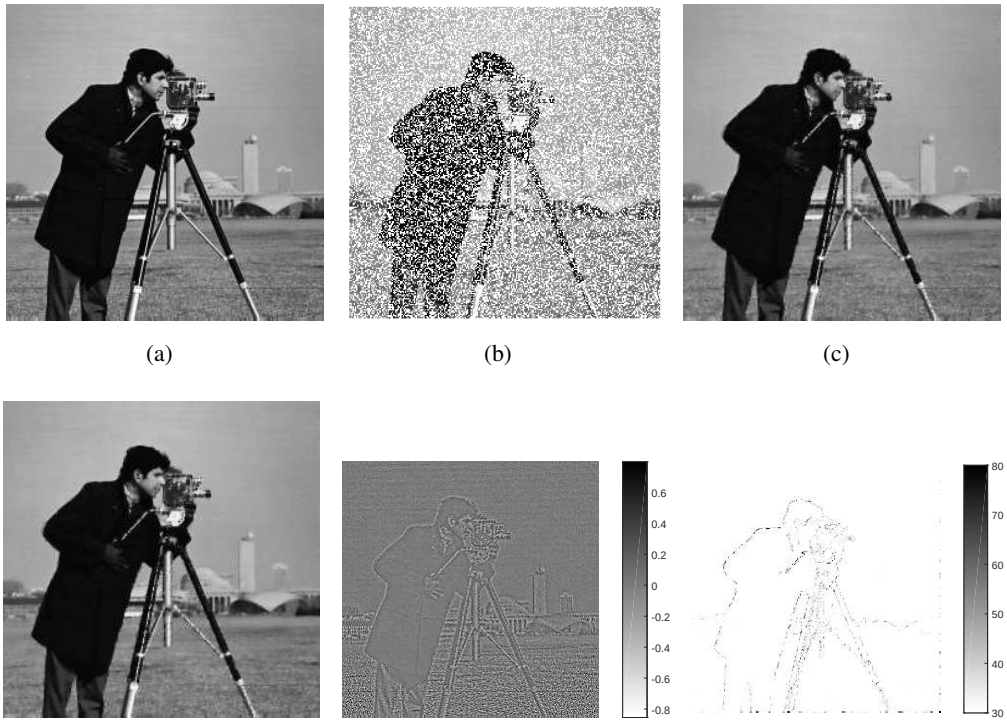

Fig. 4. Image inpainting with TV regularization using SPA: (a) original image; (b) noisy observation with missing pixels depicted in white; (c) MMSE estimate of $\mathbf{x}$; (d) MMSE estimate of $\mathbf{z}$; (e) MMSE estimate of $\mathbf{u}$; (f) Pixel-wise 90% credibility intervals.

set to $\rho = 2.8$, $\alpha = 1$ and $\beta = 0.2$ for Algo. 1 and to $\rho = 2$ and $\beta = 0.2$ for Algo. 2. In particular, the choice of ρ is discussed thereafter.

The performance of the estimators has been measured by computing the improvement in signal-to-noise ratio (ISNR) defined as

$$\mathrm{ISNR} = 10 \log_{10} \frac{\|\mathbf{x} - \mathbf{y}\|_2^2}{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2} \tag{44}$$

where $\hat{\mathbf{x}}$ refers to the MMSE (resp. MAP) estimate of $\mathbf{x}$ for SP, SPA and P-MYULA (resp., SALSA). This performance measure has been averaged over 25 Monte Carlo runs.

3) Influence of α: Fig. 5 highlights the potential benefit of the data augmentation step described in II-B. Thus, the autocorrelation functions associated to SP and SPA for different values of ρ and α are depicted. The latter were obtained by using 10^4 samples and by considering the Markov chains from their first iteration (no burn-in period has been considered here). The results are averaged over 10 independent runs. The standard deviations being very small, they are not depicted in Fig. 5. The effect of α for intermediate and large values of ρ ($\rho \geq 1$ in this

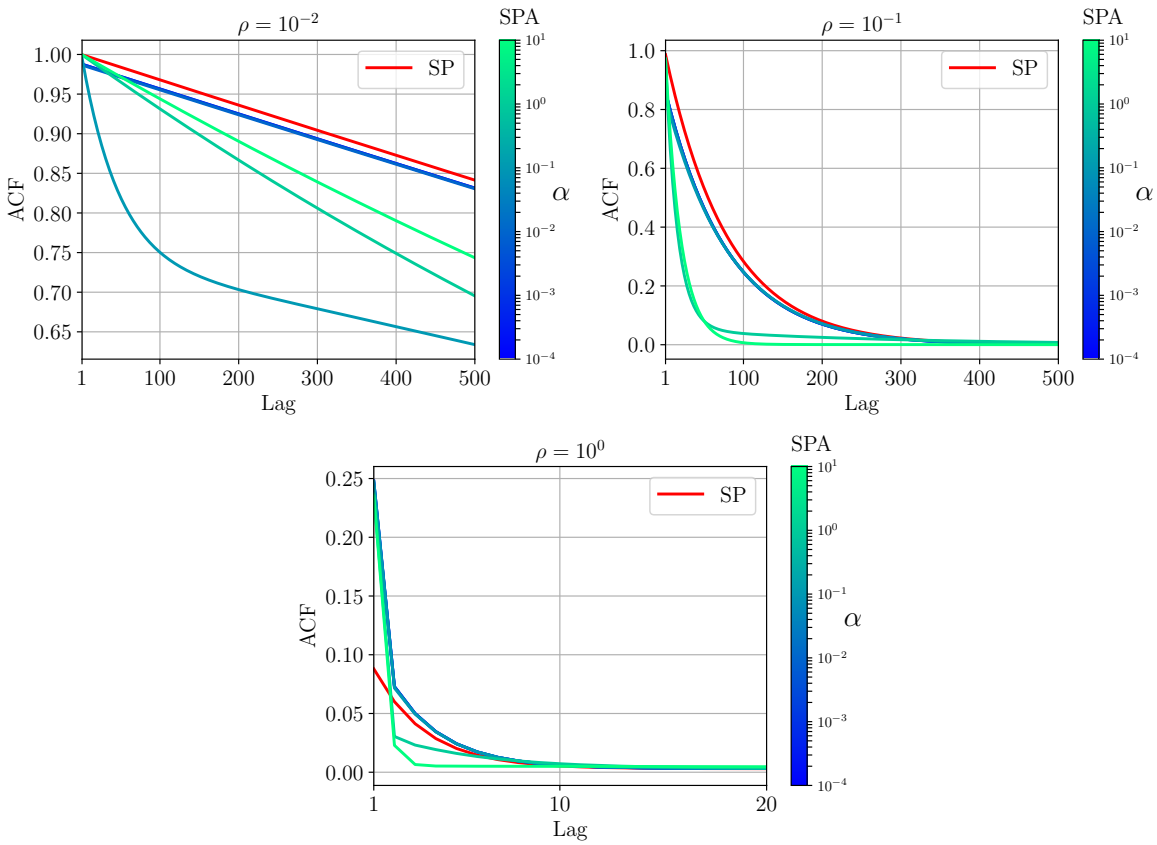

Fig. 5. Image inpainting: effect of the parameter α (associated to the data augmentation step) for different values of the parameter ρ on the autocorrelation functions of SPA (from guppie green to blue) and SP (red). The results were averaged over 10 independent runs.

case) is not significant. However, as ρ decreases, the impact of the data augmentation scheme governed by α on the autocorrelation function increases significantly. This behavior is expected since this data augmentation was introduced to bring an additional degree of freedom compared to the SP scheme when ρ is small. Although the limiting case $\rho \to 0$ is not considered in this experiment, it could be desired in some practical scenarios. In such cases, considering the data augmentation step proposed in the manuscript can bring a significant benefit concerning the exploration of the parameter space.

4) Influence of ρ: Fig. 6 shows the ISNR obtained with SPA on the Cameraman image w.r.t. the number of iterations and for different values of the parameter ρ ranging from $\rho = 1$ (blue) to $\rho = 8$ (yellow). High values of ρ (yellow to green) rapidly lead to a stable but not optimal ISNR with low variance. Conversely, small values of ρ (e.g. $\rho = 1$, dark blue) struggle to lead to

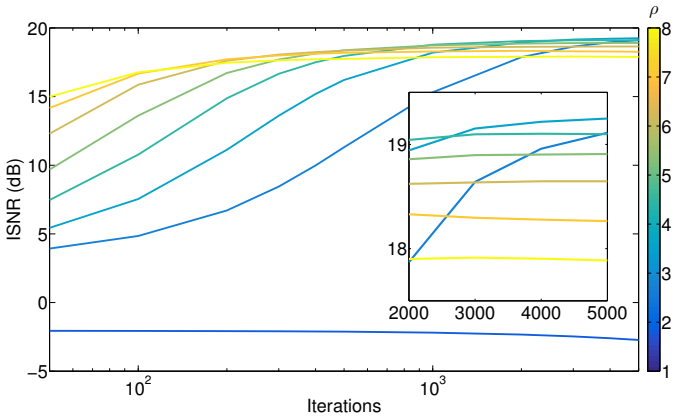

Fig. 6. Image inpainting: ISNR associated to SPA MMSE w.r.t. the number of iterations (in log-scale for the main figure and in normal-scale for the zoomed one) for different values of ρ.

an acceptable ISNR in a reasonable computational time. On the other hand, intermediate values of ρ (e.g. $\rho \in [2, 4]$, blue to green) appear to be a trade-off between speed and precision of the estimation. Thus, this range of values manages to lead, in a reasonable number of iterations, to an ISNR competing with the one obtained by SALSA (see Table IV).

5) Performance results: Table IV shows the average ISNR obtained with the different algorithms for each image depicted in Fig. 3. P-MYULA applied to the original target distribution (3) presents a lower ISNR on each image than the three other algorithms. However, when P-MYULA is used within the SP or SPA frameworks, it manages to reach average performance similar to SALSA. Note that the three MCMC approaches, contrary to the optimization algorithm SALSA, also carry credibility intervals for each pixel of the image to infer $\mathbf{x}$, see Fig. 4(f).

Table V presents the numerical complexity resulting from one iteration along with the average number of iterations performed and the average computational time for each algorithm. The complexity of $O(N^2)$ refers to matrix-vector multiplication, that of $O(N)$ to the use of a proximal operator and $\mathcal{N}$ stands for the sampling from an univariate normal distribution. Note that the number of iterations and thereby the computational time of SALSA has been adapted to each observation to reach a target reconstruction error. This has not been the case for the MCMC algorithms where the total number of iterations has been fixed beforehand. Note that the cost of one MCMC iteration is roughly equivalent to the cost of one iteration in an ADMM framework. The difference in computational time is mainly related to the number of iterations performed by each algorithm. P-MYULA took on average roughly 3400 longer time than SALSA. Much more

TABLE IV

IMAGE INPAINTING: AVERAGE RESULTS OVER 25 DIFFERENT OBSERVATION VECTORS IN TERMS OF ISNR FOR VARIOUS ALGORITHMS AND IMAGES. THE ISNR ASSOCIATED TO P-MYULA, SP AND SPA WAS COMPUTED WITH THE MMSE ESTIMATOR.

	SALSA	**P-MYULA**	**SP**	**SPA**
Balloons	26.18	23.00	26.19	26.18
Baboon	14.37	13.35	14.60	14.59
Elaine	23.61	21.21	23.86	23.84
Clock	25.72	24.50	25.45	25.42
Donna	24.71	21.69	23.87	23.82
House	20.21	19.59	20.43	20.43
Peppers	20.35	19.20	20.22	20.20
Cameraman	19.48	18.76	19.34	19.34
Boat	20.81	19.80	20.74	20.71

efficient, SP and SPA allowed to reduce the computing time w.r.t. P-MYULA by roughly 16 by embedding P-MYULA and by simplifying its task. This gain of computational time is mainly related to the Lipschitz constant of the gradient of the smooth potential used within P-MYULA. Indeed, the convergence of P-MYULA, similarly to forward-backward splitting algorithms in optimization, is driven by the Lipschitz constant of the gradient of the smooth term in the potential $f + g$. Namely, in this experiment, the Lipschitz constant L_f of ∇f is given by $L_f = \sigma^{-2}\lambda_{\max}(\mathbf{H}^T\mathbf{H})$, where $\lambda_{\max}(\mathbf{H}^T\mathbf{H})$ is the largest eigenvalue of $\mathbf{H}^T\mathbf{H}$. This constant is highly dependent on the problem, more precisely on the forward operator $\mathbf{H}$ and cannot be tuned. On the contrary, if the proposed variable splitting approach is used, P-MYULA is now embedded in the Gibbs sampling scheme and is used to sample from (39). In (39), the relevant Lipschitz constant is $L'_f = \rho^{-2}$: this constant now can be chosen carefully to improve the mixing and accelerate the convergence of P-MYULA within SPA, see Fig. (6).

Fig. 4 shows the results obtained by SPA on the Cameraman image. Those obtained by SP were similar and are omitted here for brevity. The MMSE estimators of $\mathbf{x}$ and $\mathbf{z}$ are very close, ensuring that the proposed variable splitting method behaves successfully. The variable splitting residuals contained in $\mathbf{u}$ appear to be close to 0 for most pixels but present a certain

TABLE V

IMAGE INPAINTING: COMPUTATIONAL COMPLEXITY RELATED TO ONE ITERATION, AVERAGE NUMBER OF ITERATIONS PERFORMED AND AVERAGE COMPUTATIONAL TIME FOR EACH ALGORITHM.

	computational complexity	**# iterations**	**time (s)**
SALSA	$O(N^2) + O(N)$	43	1
P-MYULA	$O(N^2) + O(N) + N\mathcal{N}$	10^5	3408
SP	$O(N^2) + O(N) + 3N\mathcal{N}$	5×10^3	207
SPA	$O(N^2) + O(N) + 4N\mathcal{N}$	5×10^3	215

structure. Thus, positive and negative residuals seem to share a complementary structure near the boundaries of objects in the image. This particular structure of the residuals is confirmed by the analysis of the credibility intervals: there is more uncertainty (of about 80 grey-levels) on the object contours of the image. The same conclusion was drawn in [12] when P-MALA was applied to an image deblurring problem with total variation.

Fig. 7 compares the average autocorrelation functions (using $-\log p(\mathbf{x}|\mathbf{y})$ as a scalar summary and obtained after the burn-in period) of SP, SPA and P-MYULA on the Cameraman image. The shaded regions depicted in Fig. 7 represent the standard deviation ranges associated to each MCMC algorithm. SP and SPA present better mixing properties than P-MYULA, showing that the proposed approaches successfully and more efficiently explore their respective parameter space. Additionally, although the average autocorrelation functions of SP and SPA are similar, the data augmentation scheme within SPA led to a Markov chain with more stable mixing properties over different observations (see the green and blue shaded areas). Note that the potential benefit of the data augmentation step detailed in Section II-B increases when ρ decreases.

6) Discussion: The expectations from MCMC algorithms like SP, SPA and P-MYULA are threefold. Firstly, to infer the hidden image $\mathbf{x}$, the MCMC methods are expected to efficiently explore the parameter space, in particular nearby the high potential regions. Secondly, the computational cost of these algorithms should remain reasonable compared to SALSA. Finally, they have to produce Markov chains with good mixing properties in order to explore the entire probability distribution and thus provide accurate credibility intervals.

Based on the previous results, SP and SPA appear as a very good trade-off between these three expectations: mixing properties, efficient exploration and reasonable computational cost.

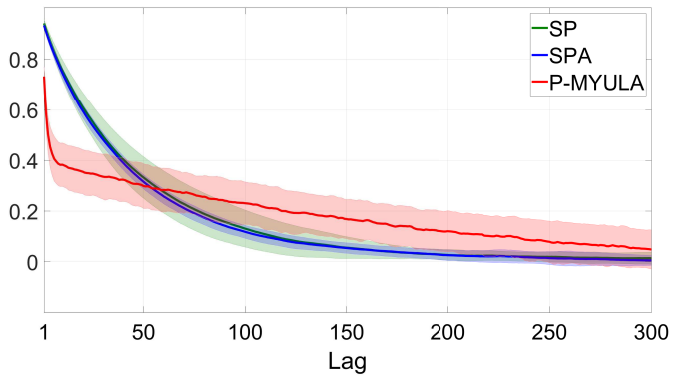

Fig. 7. Image inpainting: average chain autocorrelation functions of SP (green), SPA (blue) and P-MYULA (red). Shaded areas represent the intervals corresponding to the standard deviation computed over 25 trials.

The latter expectation is particulary satisfied. Yet, even though the computing times associated to the proposed approaches are reasonable, they are roughly 200 times higher than SALSA for a problem in high dimension ($N = 65536$). This overhead cost results from the exploration of the parameter space: this is the price to pay to derive confidence intervals on the inferred parameter, and it seems difficult to get cheaper methods.

VI. Conclusion

This paper introduced a new general Bayesian framework which aims at solving large-scale inference problems. To derive the proposed methodology, two new optimization-driven hierarchical Bayesian models and their associated MCMC algorithms, inspired from variable splitting and data augmentation, were introduced. Similarly to the ADMM in an optimization context, the proposed approach could be summarized as a "divide and conquer" method. Thus, the derived algorithms lead to simpler sampling steps so that efficient state-of-the-art MCMC algorithms can be embedded for each sampling task. Note that the proposed approach can also be used to distribute MCMC methods on multiples machines as detailed in [30].

The versatility and efficiency of the proposed algorithms have been assessed on two often-studied problems and compared to recent state-of-the-art optimization and sampling approaches. Based on these results, SP and SPA appear to be more efficient while sharing a large scope of applications. Additionally, their reasonably low computational cost compared to optimization algorithms helps to reduce the gap between optimization and simulation-based approaches while providing precious credibility intervals.

Future works will focus on other forms for the functions f, g, ϕ_1 and ϕ_2 to illustrate the broad scope of applications of the proposed approach. In particular, it will include inference problems involving non-convex target distributions. Finally, this paper presented SP and SPA as efficient algorithms designed to solve an inference problem. They could also be used to approximate complex target distributions. In this approximation context, future works will include a theoretical analysis of the proposed approach.

APPENDIX A
PROOF OF THEOREM 1

Proof: The usual target distribution (3) has the form

$$\pi(\mathbf{x}) = \frac{\exp\left[-f(\mathbf{x}) - g(\mathbf{x})\right]}{\int_{\mathbb{R}^N} \exp\left[-f(\mathbf{x}) - g(\mathbf{x})\right] \mathrm{d}\mathbf{x}}, \tag{45}$$

and has been assumed to define a proper probability distribution. By denoting

$$p_{\phi_1}(\mathbf{x}, \mathbf{z}; \rho^2) \triangleq \frac{\exp\left[-\phi_1(\mathbf{x}, \mathbf{z}; \rho^2)\right]}{\int_{\mathbb{R}^N} \exp\left[-\phi_1(\mathbf{x}, \mathbf{z}; \rho^2)\right] \mathrm{d}\mathbf{z}}, \tag{46}$$

the split-distribution (4) writes

$$\pi_\rho(\mathbf{x}, \mathbf{z}) = \frac{\exp\left[-f(\mathbf{x}) - g(\mathbf{z})\right] p_{\phi_1}(\mathbf{x}, \mathbf{z}; \rho^2)}{\int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \exp\left[-f(\mathbf{x}) - g(\mathbf{z})\right] p_{\phi_1}(\mathbf{x}, \mathbf{z}; \rho^2) \mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x}}. \tag{47}$$

Let define

$$p_\rho(\mathbf{x}) = \int_{\mathbb{R}^N} \pi_\rho(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z}. \tag{48}$$

Under the two distributions (45) and (48), we are interested in showing that

$$\left\|\pi - p_\rho\right\|_{\mathrm{TV}} = \int_{\mathbb{R}^N} \left|\pi(\mathbf{x}) - p_\rho(\mathbf{x})\right| \mathrm{d}\mathbf{x} \tag{49}$$

tends towards zero when $\rho^2 \to 0$.

Assumption 1 implies that

$$\begin{aligned} &\lim_{\rho^2 \to 0} \exp\left[-f(\mathbf{x}) - g(\mathbf{z})\right] p_{\phi_1}(\mathbf{x}, \mathbf{z}; \rho^2) \\ &= \exp\left[-f(\mathbf{x}) - g(\mathbf{z})\right] \delta_{\mathbf{x}}(\mathbf{z}). \end{aligned} \tag{50}$$

Since $\forall \rho > 0$, $\exp\left[-f(\mathbf{x}) - g(\mathbf{z})\right] p_{\phi_1}(\mathbf{x}, \mathbf{z}; \rho^2)$ has been supposed to be integrable, see Section II-A, it follows from the dominated convergence theorem that

$$\lim_{\rho^2 \to 0} \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \exp\left[-f(\mathbf{x}) - g(\mathbf{z})\right] p_{\phi_1}(\mathbf{x}, \mathbf{z}; \rho^2) \mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} \tag{51}$$

$$= \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \exp\left[-f(\mathbf{x}) - g(\mathbf{z})\right] \delta_{\mathbf{x}}(\mathbf{z}) \mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} \tag{52}$$

$$= \int_{\mathbb{R}^N} \exp\left[-f(\mathbf{x}) - g(\mathbf{x})\right] \mathrm{d}\mathbf{x}. \tag{53}$$

Combining (50) and (53), it follows

$$\lim_{\rho^2 \to 0} \pi_\rho(\mathbf{x}, \mathbf{z}) = \frac{\exp\left[-f(\mathbf{x}) - g(\mathbf{z})\right] \delta_{\mathbf{x}}(\mathbf{z})}{\int_{\mathbb{R}^N} \exp\left[-f(\mathbf{x}) - g(\mathbf{x})\right] \mathrm{d}\mathbf{x}}. \tag{54}$$

Using one more time the dominated convergence theorem, as in (52) and (54) leads for all $\mathbf{x} \in \mathbb{R}^N$ to

$$\lim_{\rho^2 \to 0} p_\rho(\mathbf{x}) = \frac{\exp\left[-f(\mathbf{x}) - g(\mathbf{x})\right]}{\int_{\mathbb{R}^N} \exp\left[-f(\mathbf{x}) - g(\mathbf{x})\right] \mathrm{d}\mathbf{x}} = \pi(\mathbf{x}). \tag{55}$$

Finally, Scheffé's lemma [51] ensures the convergence of p_ρ towards π in total variation, that is

$$\lim_{\rho^2 \to 0} \left\| \pi - p_\rho \right\|_{\mathrm{TV}} = \lim_{\rho^2 \to 0} \int_{\mathbb{R}^N} \left| \pi(\mathbf{x}) - p_\rho(\mathbf{x}) \right| \mathrm{d}\mathbf{x} = 0. \tag{56}$$

■

APPENDIX B
CASE OF MULTIPLE FUNCTIONS h_i

Assume that the problem considered involves the introduction of N_h functions h_i along with N_h observation operators $\mathbf{K}_i \in \mathbb{R}^{k_i \times N}$, $i \in \{1, \ldots, N_h\}$. Thereby, the usual target distribution takes the form

$$\pi(\mathbf{x}) \propto \exp\left[-\sum_{i=1}^{N_h} h_i(\mathbf{K}_i \mathbf{x})\right]. \tag{57}$$

Remark 1: In the case where $N_h = 2$ and $\mathbf{K}_1 = \mathbf{K}_2 = \mathbf{I}_N$, the usual target distribution defined in (3) is retrieved.

A. Derivation of SP

In order to simplify the sampling procedure, let introduce N_h splitting variables denoted $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N_h} \in \mathbb{R}^{k_i}$, a positive parameter ρ and N_h divergences ϕ_i defined on $\mathbb{R}^{k_i} \times \mathbb{R}^{k_i}$ such that the underlying joint probability distribution has the form

$$p(\mathbf{x}, \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N_h}; \rho^2) \propto \exp\left[-\sum_{i=1}^{N_h} h_i(\mathbf{z}_i) + \phi_i\left(\mathbf{K}_i\mathbf{x}, \mathbf{z}_i; \rho^2\right)\right]. \tag{58}$$

Thereby, the generalized SP implies the sampling from the conditional distributions

$$p(\mathbf{x}|\mathbf{z}_{i,i\in\{1,\ldots,N_h\}}; \rho^2) \propto \exp\left[-\sum_{i=1}^{N_h} \phi_i\left(\mathbf{K}_i\mathbf{x}, \mathbf{z}_i; \rho^2\right)\right], \tag{59}$$

$$p(\mathbf{z}_i|\mathbf{x}; \rho^2) \propto \exp\left[-h_i(\mathbf{z}_i) - \phi_i\left(\mathbf{K}_i\mathbf{x}, \mathbf{z}_i; \rho^2\right)\right], \tag{60}$$

for all $i \in \{1, \ldots, N_h\}$.

B. Derivation of SPA

In the same manner, let introduce N_h splitting and auxiliary variables denoted $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{N_h} \in \mathbb{R}^{k_i}$ and $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{N_h} \in \mathbb{R}^{k_i}$, respectively. Additionally, let introduce positive parameters ρ and α, N_h divergences ϕ_i defined on $\mathbb{R}^{k_i} \times \mathbb{R}^{k_i}$ and N_h functions ψ_i defined on $\mathbb{R}^{k_i}$ such that the underlying joint probability distribution has the form

$$p(\mathbf{x}, \mathbf{z}_{i,i\in\{1,\ldots,N_h\}}, \mathbf{u}_{i,i\in\{1,\ldots,N_h\}}; \rho^2, \alpha^2) \propto \exp\left[-\sum_{i=1}^{N_h} h_i(\mathbf{z}_i) + \phi_i\left(\mathbf{K}_i\mathbf{x}, \mathbf{z}_i - \mathbf{u}_i; \rho^2\right) + \psi_i(\mathbf{u}_i; \alpha^2)\right]. \tag{61}$$

The generalized SPA implies the sampling from the conditional distributions

$$p(\mathbf{x}|\mathbf{z}_i, \mathbf{u}_i; \rho^2) \propto \exp\left[-\sum_{i=1}^{N_h} \phi_i\left(\mathbf{K}_i\mathbf{x}, \mathbf{z}_i - \mathbf{u}_i; \rho^2\right)\right], \tag{62}$$

$$p(\mathbf{z}_i|\mathbf{x}, \mathbf{u}_i; \rho^2) \propto \exp\left[-h_i(\mathbf{z}_i) - \phi_i\left(\mathbf{K}_i\mathbf{x}, \mathbf{z}_i - \mathbf{u}_i; \rho^2\right)\right], \tag{63}$$

for all $i \in \{1, \ldots, N_h\}$, and

$$p(\mathbf{u}_i|\mathbf{x}, \mathbf{z}_i; \rho^2, \alpha^2) \propto \exp\left[-\psi_i(\mathbf{u}_i; \alpha^2) - \phi_i\left(\mathbf{K}_i\mathbf{x}, \mathbf{z}_i - \mathbf{u}_i; \rho^2\right)\right], \tag{64}$$

for all $i \in \{1, \ldots, N_h\}$.

APPENDIX C
EFFICIENT GAUSSIAN SAMPLING IN HIGH DIMENSION

In this Appendix, notations are those of Section IV-B. Suppose that one wants to sample efficiently from the high-dimensional Gaussian conditional distributions

$$p(\mathbf{z}|\mathbf{x},\mathbf{u}) = \mathcal{N}\left(\mathbf{m_z}, \mathbf{G_z}^{-1}\right) \tag{65}$$

$$p(\mathbf{x}|\mathbf{z},\mathbf{u}) = \mathcal{N}\left(\mathbf{m_x}, \mathbf{G_x}^{-1}\right) \tag{66}$$

where, in particular,

$$\begin{cases} \mathbf{G_z} = \gamma \mathbf{L}^T\mathbf{L} + \dfrac{1}{\rho^2}\mathbf{I}_N. & (67) \\ \mathbf{G_x} = \mathbf{H}^T\mathbf{\Omega}\mathbf{H} + \dfrac{1}{\rho^2}\mathbf{I}_N & (68) \end{cases}$$

A. *Efficient sampling from (65)*

The matrix $\mathbf{L}$ was assumed to be a circulant matrix. Thereby, the latter can be diagonalized in the Fourier domain such that

$$\mathbf{L} = \mathbf{F}^H\mathbf{\Lambda_L}\mathbf{F}, \tag{69}$$

where $\mathbf{F}$ and $\mathbf{F}^H$ are unitary matrices ($\mathbf{F}^H\mathbf{F} = \mathbf{F}\mathbf{F}^H = \mathbf{I}_N$) associated with the Fourier and inverse Fourier transforms. $\mathbf{\Lambda_L}$ is the diagonal counterpart of $\mathbf{L}$ in the Fourier domain. Using (69), the precision matrix defined in (67) has the form

$$\begin{aligned} \mathbf{G_z} &= \gamma \mathbf{F}^H{\mathbf{\Lambda_L}}^H\mathbf{F}\mathbf{F}^H\mathbf{\Lambda_L}\mathbf{F} + \frac{1}{\rho^2}\mathbf{I}_N \\ &= \gamma \mathbf{F}^H{\mathbf{\Lambda_L}}^H\mathbf{\Lambda_L}\mathbf{F} + \frac{1}{\rho^2}\mathbf{I}_N \end{aligned} \tag{70}$$

Then, the counterpart of $\mathbf{G_z}$ in the Fourier domain is diagonal and has the form

$$\mathbf{\Lambda_{G_z}} = \gamma {\mathbf{\Lambda_L}}^H\mathbf{\Lambda_L} + \frac{1}{\rho^2}\mathbf{I}_N. \tag{71}$$

Using (71), one can efficiently sample from (65) by drawing N independent Gaussian samples in the Fourier domain.

B. Efficient sampling from (66)

Unfortunately, although the matrix $\mathbf{H}$ was assumed circulant, the first term in (68) cannot be diagonalized in the Fourier domain. To cope with this problem, the auxiliary method of [39] is used. An additional variable $\mathbf{v} \in \mathbb{R}^N$ is introduced such that the conditional distributions of $\mathbf{x}$ and $\mathbf{v}$ are

$$p(\mathbf{x}|\mathbf{z}, \mathbf{u}, \mathbf{v}) = \mathcal{N}\left(\tilde{\mathbf{m}}_{\mathbf{x}}, \tilde{\mathbf{G}}_{\mathbf{x}}^{-1}\right) \tag{72}$$

$$p(\mathbf{v}|\mathbf{x}) = \mathcal{N}\left(\mathbf{m}_{\mathbf{v}}, {\mathbf{G}_{\mathbf{v}}}^{-1}\right) \tag{73}$$

where, in particular,

$$\begin{cases} \tilde{\mathbf{G}}_{\mathbf{x}} = \dfrac{1}{\mu_1}\mathbf{H}^T\mathbf{H} + \dfrac{1}{\rho^2}\mathbf{I}_N & (74) \\ {\mathbf{G}_{\mathbf{v}}}^{-1} = \dfrac{1}{\mu_1}\mathbf{I}_N - \mathbf{\Omega}. & (75) \end{cases}$$

Remark 2: The positive parameter μ_1 is such that $\mu_1 \|\mathbf{\Omega}\|_S < 1$ ($\|.\|_S$ stands for the spectral norm of a matrix) ensuring that (75) is positive definite.

As in Appendix C-B, the matrix $\mathbf{H}$ (assumed circulant) can be diagonalized in the Fourier domain. Under these two conditional distributions, $\mathbf{x}$ can be efficiently drawn in the Fourier domain and $\mathbf{v}$ can be efficiently sampled in $\mathbb{R}^N$ as $\mathbf{\Omega}$ was assumed diagonal.

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2001.

[2] M. Pereyra *et al.*, "A survey of stochastic simulation and optimization methods in signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 2, pp. 224–241, March 2016.

[3] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer, 2005.

[4] M. G. Kendall, *The Advanced Theory of Statistics*. Griffin, 1946.

[5] D. G. Hwang and P. Green, "Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution," *Proc. of the National Academy of Sciences*, vol. 101, no. 39, pp. 13 994–14 001, 2004.

[6] U. von Toussaint, "Bayesian inference in physics," *Rev. Mod. Phys.*, vol. 83, pp. 943–999, Sept. 2011.

[7] T. J. Loredo, *Promise of Bayesian Inference for Astrophysics*. Springer, 1992, pp. 275–297.

[8] D. S. Reis and J. R. Stedinger, "Bayesian MCMC flood frequency analysis with historical information," *Journal of Hydrology*, vol. 313, no. 1, pp. 97–116, 2005.

[9] R. Trotta, "Bayes in the sky: Bayesian inference and model selection in cosmology," *Contemporary Physics*, vol. 49, no. 2, pp. 71–104, 2008.

[10] J. Veitch *et al.*, "Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library," *Phys. Rev. D*, vol. 91, no. 4, Feb. 2015.

[11] S. Duane *et al.*, "Hybrid Monte Carlo," *Phys. Lett. B*, vol. 195, no. 2, pp. 216 – 222, 1987.

[12] M. Pereyra, "Proximal Markov chain Monte Carlo algorithms," *Stat. Comput.*, vol. 26, no. 4, pp. 745–760, July 2016.

[13] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *SIAM J. Multiscale Model. Simul.*, vol. 4, no. 4, pp. 1168–1200, 2005.

[14] M. A. T. Figueiredo and R. D. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, Aug. 2003.

[15] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.

[16] M. Elad, "Why simple shrinkage is still relevant for redundant representations?" *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5559–5569, Dec. 2006.

[17] E. T. Hale, W. Yin, and Y. Zhang, "Fixed-point continuation for ℓ_1-minimization: Methodology and convergence," *SIAM J. Optim.*, vol. 19, no. 3, pp. 1107–1130, 2008.

[18] J. M. Bioucas-Dias and M. A. T. Figueiredo, "A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2992–3004, Dec. 2007.

[19] C. Elvira, P. Chainais, and N. Dobigeon, "Bayesian antisparse coding," *IEEE Trans. Signal Process.*, vol. 65, no. 7, pp. 1660–1672, April 2017.

[20] A. Durmus, E. Moulines, and M. Pereyra, "Efficient Bayesian computation by proximal Markov chain Monte Carlo: When Langevin meets Moreau," *SIAM J. Imag. Sci.*, vol. 11, no. 1, pp. 473–506, 2018.

[21] R. Courant, "Variational methods for the solution of problems of equilibrium and vibrations," *Bull. Amer. Math. Soc.*, vol. 49, pp. 1–23, 1943.

[22] S. Boyd *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[23] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics with Applications*, vol. 2, no. 1, pp. 17 – 40, 1976.

[24] Glowinski, R. and Marroco, A., "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires," *R.A.I.R.O. Analyse Numrique*, vol. 9, pp. 41–76, 1975.

[25] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[26] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2345–2356, Sept. 2010.

[27] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 681–695, March 2011.

[28] P. A. Thouvenin, N. Dobigeon, and J. Y. Tourneret, "Hyperspectral unmixing with spectral variability using a perturbed linear mixing model," *IEEE Trans. Signal Process.*, vol. 64, no. 2, pp. 525–538, Jan. 2016.

[29] A. Halimi *et al.*, "Fast hyperspectral unmixing in presence of nonlinearity or mismodeling effects," *IEEE Trans. Comput. Imag.*, vol. 3, no. 2, pp. 146–159, June 2017.

[30] L. J. Rendell *et al.*, "Global consensus Monte Carlo," 2018. [Online]. Available: https://arxiv.org/abs/1807.09288/

[31] P. L. Combettes and J.-C. Pesquet, *Proximal Splitting Methods in Signal Processing*. Springer, 2011, pp. 185–212.

[32] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 14, no. 3, pp. 367–383, March 1992.

[33] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Trans. Image Process.*, vol. 4, no. 7, pp. 932–946, July 1995.

[34] J. Idier, "Convex half-quadratic criteria and interacting auxiliary variables for image restoration," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 1001–1009, July 2001.

[35] D. A. van Dyk and X.-L. Meng, "The art of data augmentation," *J. Comput. Graph. Stat.*, vol. 10, no. 1, pp. 1–50, 2001.

[36] J. Besag and P. J. Green, "Spatial statistics and Bayesian computation," *J. Roy. Stat. Soc. Ser. B*, vol. 55, no. 1, pp. 25–37, 1993.

[37] D. M. Higdon, "Auxiliary variable methods for Markov chain Monte Carlo with applications," *J. Amer. Stat. Assoc.*, vol. 93, no. 442, pp. 585–595, 1998.

[38] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation theory*. Englewood Cliffs NJ: Prentice Hall, 1993.

[39] Y. Marnissi *et al.*, "An auxiliary variable method for Markov chain Monte Carlo algorithms in high dimension," *Entropy*, vol. 20, no. 2, 2018.

[40] G. Papandreou and A. L. Yuille, "Gaussian sampling by local perturbations," in *Adv. in Neural Information Process. Systems*, 2010, pp. 1858–1866.

[41] C. Gilavert, S. Moussaoui, and J. Idier, "Efficient Gaussian sampling for solving large-scale inverse problems using MCMC," *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 70–80, Jan. 2015.

[42] O. Féron, F. Orieux, and J. F. Giovannelli, "Gradient scan Gibbs sampler: An efficient algorithm for high-dimensional Gaussian distributions," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 2, pp. 343–352, March 2016.

[43] R. Molina and B. D. Ripley, "Using spatial models as priors in astronomical image analysis," *J. Appl. Stat.*, vol. 16, no. 2, pp. 193–206, 1989.

[44] R. Molina, J. Mateos, and A. K. Katsaggelos, "Blind deconvolution using a variational approach to parameter, image, and blur estimation," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3715–3727, Dec. 2006.

[45] A. C. Likas and N. P. Galatsanos, "A variational approach for Bayesian blind image deconvolution," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2222–2233, Aug. 2004.

[46] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. Rev. D*, vol. 60, no. 1-4, pp. 259–268, Nov. 1992.

[47] D. Strong and T. Chan, "Edge-preserving and scale-dependent properties of total variation regularization," *Inverse Problems*, vol. 19, no. 6, pp. S165–S187, 2003.

[48] A. Chambolle *et al.*, "An introduction to total variation for image analysis," in *Theoretical Foundations and Numerical Methods for Sparse Recovery, De Gruyter*, 2010.

[49] M. Vono, N. Dobigeon, and P. Chainais, "Sparse Bayesian binary logistic regression using the split-and-augmented Gibbs sampler," in *Proc. IEEE Workshop Mach. Learning for Signal Process. (MLSP)*, 2018.

[50] A. Chambolle, "An algorithm for total variation minimization and applications," *J. Math. Imag. Vision*, vol. 20, no. 1, pp. 89–97, Jan. 2004.

[51] H. Scheffe, "A useful convergence theorem for probability distributions," *Ann. Math. Statist.*, vol. 18, no. 3, pp. 434–438, 09 1947.