Efficient Sampling through Variable Splitting-inspired Bayesian Hierarchical Models

Maxime Vono, Nicolas Dobigeon, Pierre Chainais

EFFICIENT SAMPLING THROUGH VARIABLE SPLITTING-INSPIRED BAYESIAN HIERARCHICAL MODELS

Maxime Vono, Nicolas Dobigeon

University of Toulouse, INP-ENSEEIHT
IRIT, CNRS, Toulouse, France

Pierre Chainais

University of Lille, CNRS, Centrale Lille
UMR 9189 - CRIStAL, Lille, France

ABSTRACT

Markov chain Monte Carlo (MCMC) methods are an important class of computation techniques to solve Bayesian inference problems. Much research has been dedicated to scale these algorithms in high-dimensional settings by relying on powerful optimization tools such as gradient information or proximity operators. In a similar vein, this paper proposes a new Bayesian hierarchical model to solve large scale inference problems by taking inspiration from variable splitting methods. Similarly to the latter, the derived Gibbs sampler permits to divide the initial sampling task into simpler ones. As a result, the proposed Bayesian framework can lead to a faster sampling scheme than state-of-the-art methods by embedding them. The strength of the proposed methodology is illustrated on two often-studied image processing problems.

***Index Terms*—** Bayesian inference, Gibbs sampler, high dimension, variable splitting.

1. INTRODUCTION

Solving signal/image processing and machine learning problems highly relies on efficient computational methods [1]. Among them, techniques based on variational optimization have received a lot of interest over the past decade leading to fast, efficient and distributed algorithms. For instance, stochastic optimization or Robbins-Monro algorithms [2] and distributed optimization algorithms such as the alternating direction method of multipliers (ADMM), dating back to [3, 4], were successfully resorted to deal with large datasets [5] and high-dimensional problems [6, 7]. As pointed out in [8], Bayesian approaches have not benefited from the same advances as in optimization. Nevertheless, by giving the opportunity to explore the posterior distribution of the variable of interest, those methods are of great interest in areas where models must be compared, e.g. for the analysis of gravitational waves [9], or where uncertainties have to be quantified [10]. This complete description of the variable to infer comes at a cost, that can be sometimes prohibitive compared to fast optimization-based techniques. This leads a lot of work to scale Bayesian methods, such as Markov chain Monte Carlo (MCMC) algorithms, enabling them to deal with more and more complex datasets and problems [11]. Among numerous improvements of MCMC methods, optimization-driven approaches attracted the interest of a lot of researchers. For instance, gradient information has been successfully used in MCMC methods based on diffusions, e.g. Hamiltonian Monte Carlo methods [12]. In addition, linear programming has been used to sample efficiently determinantal point processes [13] and convex optimization has been used to explore efficiently possibly non-smooth log-concave probability distributions [14, 15].

In this spirit, this paper proposes a set of methods to solve Bayesian inference problems by taking inspiration from variable splitting techniques. Variable splitting is classically resorted in methods (e.g. the ADMM) to solve large-scale optimization problems by dividing the difficulty over simpler sub-problems. To this purpose, Section 2 presents how variable splitting can be used within a Bayesian framework and derives an efficient Gibbs sampler where existing state-of-the-art approaches can be embedded. Section 3 shows the results of the proposed approach and of state-of-the-art optimization and MCMC methods applied to image processing problems. Finally, concluding remarks and considered further prospects are reported in Section 4.

2. VARIABLE SPLITTING-INSPIRED BAYESIAN INFERENCE

This section introduces the proposed Bayesian model built from an initial target distribution π and a variable splitting step. The main properties of the derived Bayesian hierarchical model are presented.

2.1. Problem statement

Let consider Bayesian inference problems where we are interested in estimating an unknown object $\mathbf{x} \in \mathbb{R}^d$ (e.g. a signal or parameters of a model) from observations $\mathbf{y}$ related to the variable to infer through a statistical model with likelihood $p(\mathbf{y}|\mathbf{x})$. Within this Bayesian setting, the uncertainty on $\mathbf{x}$ is modeled via a prior distribution $p(\mathbf{x})$ [16] leading to the posterior distribution

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}).$$

In the sequel, the latter is assumed to have the usual form

$$\pi(\mathbf{x}) \triangleq p(\mathbf{x}|\mathbf{y}) \propto \exp\left(-f_1(\mathbf{x}) - f_2(\mathbf{x})\right),$$

where f_1 and f_2 are two arbitrary functions such that π is well defined. Sampling directly from (1) can be difficult for different reasons. For instance, if f_1 and f_2 are not conjugate, one could rely on more sophisticated sampling schemes, such as Metropolis-Hastings (MH) algorithms [17]. The cost of the latter can be prohibitive in large-scale problems, especially when the likelihood function is a product of terms over a "big" dataset [18]. Additionally, in some cases, some efficient MCMC algorithms cannot be directly applied to sample from π [19]. In these challenging problems, variable splitting can provide an efficient surrogate method to simplify and/or improve the sampling from the target distribution (1). This technique is popular in optimization to tackle the initial difficulty by dividing the cost function $f_1 + f_2$ in a set of simpler ones. This is achieved by introducing auxiliary variables $\mathbf{z}$ to split the objective function. For

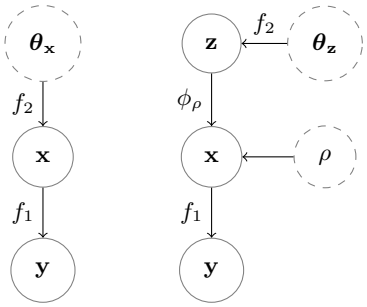

Fig. 1. DAG associated to the initial Bayesian model (left) and to the proposed model (right). Fixed parameters are represented with dashed circles. $\boldsymbol{\theta}_{\mathbf{x}}$ and $\boldsymbol{\theta}_{\mathbf{z}}$ stand for possible hyperparameters which are not discussed in this paper.

instance, maximum a posteriori (MAP) estimation under (1) leads to the optimization problem

$$\underset{\mathbf{x}\in\mathbb{R}^d}{\arg\min} \quad f_1(\mathbf{x}) + f_2(\mathbf{x}), \tag{1}$$

which can be rewritten using variable splitting as

$$\underset{\mathbf{x}\in\mathbb{R}^d,\mathbf{z}\in\mathbb{R}^d}{\arg\min} \quad f_1(\mathbf{x}) + f_2(\mathbf{z}) \quad \text{s.t.} \quad \mathbf{x} = \mathbf{z}. \tag{2}$$

This new formulation of MAP estimation yields several (two in this case) optimization sub-problems, one for each variable where f_1 and f_2 are dissociated [5].

2.2. Bayesian hierarchical model

Following this variable splitting trick, an auxiliary variable $\mathbf{z} \in \mathbb{R}^d$ is introduced to simplify the sampling from π. Thus sampling from the latter will be replaced by sampling from two simpler distributions (6) and (7). To this aim, let us consider a Bayesian hierarchical model through a joint distribution $\pi_\rho(\mathbf{x}, \mathbf{z})$ defined by

$$\pi_\rho(\mathbf{x}, \mathbf{z}) \propto \exp\left(-f_1(\mathbf{x}) - f_2(\mathbf{z}) - \phi_\rho(\mathbf{x}, \mathbf{z})\right), \tag{3}$$

where $\rho > 0$ and ϕ_ρ stands for a divergence where the discrepancy between $\mathbf{x}$ and $\mathbf{z}$ is controlled by ρ. Of course, ϕ_ρ has to be chosen such that π_ρ and its related probability distributions are Lebesgue-integrable. Possible choices for ϕ_ρ, that can be viewed as a coupling function, are discussed in Section 2.3. In cases where f_1 and f_2 stand for potential functions associated to the likelihood and to the prior respectively, fig.1 shows the directed acyclic graph (DAG) related to the proposed split model.

In general, the split distribution π_ρ in (3) does not correspond to an augmentation of the initial target distribution π. Thereby, the marginal distribution of $\mathbf{x}$ under π_ρ stands for an approximation of π. The corresponding approximation error, controlled by ρ, can be made arbitrarily small as depicted by Theorem 1 proven in [20].

Theorem 1 *[20, Theorem 1] Let $p_\rho = \int_{\mathbb{R}^d} \pi_\rho(\mathbf{x}, \mathbf{z})\mathrm{d}\mathbf{z}$ be the marginal of $\mathbf{x}$ under π_ρ. Assume that in the limiting case $\rho \to 0$, ϕ_ρ is such that*

$$\frac{\exp\left(-\phi_\rho(\mathbf{x}, \mathbf{z})\right)}{\int_{\mathbb{R}^d} \exp\left(-\phi_\rho(\mathbf{x}, \mathbf{z})\right) \mathrm{d}\mathbf{x}} \xrightarrow[\rho\to 0]{} \delta_{\mathbf{x}}(\mathbf{z}) \tag{4}$$

Then, under (4) *and by assuming that f_1 and f_2 are lower-bounded, p_ρ coincides with π when $\rho \to 0$, that is*

$$\|p_\rho - \pi\|_{\mathrm{TV}} \xrightarrow[\rho\to 0]{} 0. \tag{5}$$

Algorithm 1: Split Gibbs sampler (SGS)

Input: Functions f_1, f_2, hyperparam. ρ, total nb of iterations T_{MC}, nb 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 $\pi_\rho\left(\mathbf{x}|\mathbf{z}^{(t-1)}\right)$ (6) ;

  % Drawing the auxiliary variable

  Sample $\mathbf{z}^{(t)}$ according to $\pi_\rho\left(\mathbf{z}|\mathbf{x}^{(t)}\right)$ (7) ;

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 (3).

Generalizations of the proposed split scheme and more details can be found in [20, 21].

2.3. Gibbs sampler

The conditional distributions under the split distribution π_ρ write

$$\pi_\rho(\mathbf{x}|\mathbf{z}) \propto \exp\left(-f_1(\mathbf{x}) - \phi_\rho(\mathbf{x}, \mathbf{z})\right) \tag{6}$$

$$\pi_\rho(\mathbf{z}|\mathbf{x}) \propto \exp\left(-f_2(\mathbf{z}) - \phi_\rho(\mathbf{x}, \mathbf{z})\right). \tag{7}$$

Thus, the variable of interest $\mathbf{x}$ can be estimated through Gibbs sampling where each sampling step does not involve $f_1 + f_2$ but only a part of this initial potential function f_1 or f_2. Strong similarities with distributed optimization algorithms, e.g. the ADMM, are discussed in [20]. The Gibbs sampler derived to sample from the split distribution π_ρ is depicted in algo. 1 and called split Gibbs sampler (SGS). Our motivation to introduce variable splitting was the simplification of the inference and the derivation of a more efficient sampling scheme. To this purpose, the conditional distributions (6) and (7) must be simple to sample from compared to the sampling from π. Depending on the form and properties (e.g. convexity, differentiability) of the potential functions f_1 and f_2, the coupling function can be adaptively chosen. For instance, [20, 22] consider $\phi_\rho(\mathbf{x}) = (2\rho^2)^{-1} \|\mathbf{x} - \mathbf{z}\|_2^2$ for its gradient Lipschitz, differentiability and convexity properties while [21] invokes conjugacy arguments to choose ϕ_ρ. For a detailed discussion on the Gibbs sampling procedure, we refer the reader to [20, Section III.A].

3. EXPERIMENTS

This section illustrates the application of the proposed Bayesian framework on two linear Gaussian inverse problems encountered in image processing, namely image deblurring using either a total variation (TV) prior or a frame-based approach.

3.1. Linear Gaussian inverse problems

By considering either an analysis or a synthesis approach [23], linear Gaussian inverse problems involve the estimation of an unknown object $\mathbf{x}$ through the forward model

$$\mathbf{y} = \mathbf{P}\mathbf{x} + \mathbf{n}, \tag{8}$$

where $\mathbf{P}$ is a direct operator and $\mathbf{n} \sim \mathcal{N}\left(\mathbf{0}_d, \sigma^2\mathbf{I}_d\right)$ stands for noise. If f_1 denotes the potential associated to the likelihood, then for all $\mathbf{x} \in \mathbb{R}^d$,

$$f_1(\mathbf{x}) = \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{P}\mathbf{x}\|_2^2. \tag{9}$$

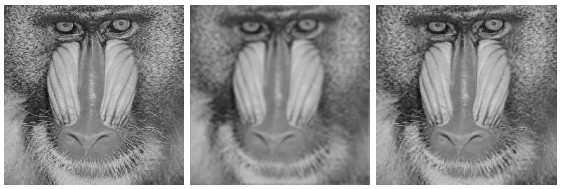

Fig. 2. Image deblurring with TV prior. (left) Original image, (middle) noisy and blurred image and (right) MMSE estimate computed with SGS.

By taking, as in [20], $\phi_\rho(\mathbf{x}) = \dfrac{1}{2\rho^2}\|\mathbf{x}-\mathbf{z}\|_2^2$ and by assuming that for all $\mathbf{x} \in \mathbb{R}^d$, the potential f_2 associated to the prior writes

$$f_2(\mathbf{x}) = \tau\psi(\mathbf{x}), \quad \tau > 0, \tag{10}$$

the conditional distributions (6) and (7) have the form

$$\pi_\rho(\mathbf{x}|\mathbf{z}) = \mathcal{N}\left(\boldsymbol{\mu}_\mathbf{x}, {\mathbf{Q}_\mathbf{x}}^{-1}\right) \tag{11}$$

$$\pi_\rho(\mathbf{z}|\mathbf{x}) \propto \exp\left(-\tau\psi(\mathbf{z}) - \frac{1}{2\rho^2}\|\mathbf{z}-\mathbf{x}\|_2^2\right), \text{ where} \tag{12}$$

$$\begin{cases} \mathbf{Q}_\mathbf{x} = \dfrac{1}{\sigma^2}\mathbf{P}^T\mathbf{P} + \dfrac{1}{\rho^2}\mathbf{I}_d & (13) \\ \boldsymbol{\mu}_\mathbf{x} = {\mathbf{Q}_\mathbf{x}}^{-1}\left(\dfrac{1}{\sigma^2}\mathbf{P}^T\mathbf{y} + \dfrac{1}{\rho^2}\mathbf{z}\right). & (14) \end{cases}$$

If the matrix $\mathbf{P}$ can be diagonalizable in a certain domain (e.g. Fourier domain), then sampling from (11) can be performed efficiently in this domain through the exact perturbation-optimization (E-PO) algorithm [24]. Additionally, if ψ corresponds to a convex and possibly non-smooth regularization function (e.g. TV), sampling from (12) can be conducted using the proximal MCMC algorithm P-MYULA which has well-understood theoretical properties.

3.2. Image deblurring with total variation prior

Problem formulation – In this first experiment, we consider an image deconvolution problem where an original image $\mathbf{x}$ of size 256×256 ($d = 65536$) is blurred via a 5×5 Gaussian blur kernel with standard deviation equal to 2, see fig. 2. Thereby, the corresponding operator $\mathbf{P} = \mathbf{B}$ is a circulant matrix, diagonalizable in the Fourier domain. The regularization potential ψ is the total variation (TV) function.

Experimental design – The proposed method is compared with the state-of-the-art proximal MCMC algorithm, namely proximal Moreau-Yoshida unadjusted Langevin algorithm (P-MYULA) [15]. Additionally, SGS and P-MYULA will be compared with their counterpart optimization algorithms namely the split augmented Lagrangian shrinkage algorithm (SALSA) [6] and the fast iterative shrinkage thresholding algorithm (FISTA) [25], respectively. The variance σ^2 of the Gaussian noise is set such that the signal-to-noise ratio (SNR) is equal to 40dB. The Lipschitz constant L_{f_1} associated to ∇f_1, needed to launch FISTA and P-MYULA, is defined by $L_{f_1} = \sigma^{-2}\lambda_{\max}(\mathbf{P}^T\mathbf{P})$ where $\lambda_{\max}(\mathbf{P}^T\mathbf{P})$ stands for the largest eigenvalue of $\mathbf{P}^T\mathbf{P}$.

Since SGS and P-MYULA do not target the same stationary distribution, comparing explicitly these two algorithms is not straightforward and highly depends on the reconstruction criteria used by the practitioner. Nethertheless, this experiment aims to demonstrate

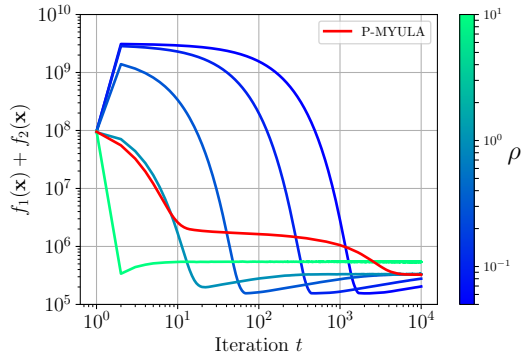

Fig. 3. Image deblurring with TV prior. Convergence of the Markov chains associated to SGS w.r.t ρ (from guppie green to blue) and P-MYULA (red) toward the typical set of π.

that SGS, by embedding P-MYULA, can lead to reliable approximate estimates with a lower computational cost.

The number of MCMC iterations has been set to $T_{\text{MC}} = 1.1 \times 10^4$ with $T_{\text{bi}} = 10^3$ burn-in iterations for SGS. For P-MYULA, due to slower mixing properties (see fig. 4), the number of MCMC iterations has been set to $T_{\text{MC}} = 10^5$ with $T_{\text{bi}} = 9 \times 10^4$ burn-in iterations. Thereby, 10^4 samples for each Markov chain are considered. The regularization parameter has been set to $\tau = 0.2$. Similarly to [6], the penalty parameter μ used in SALSA has been set to $\mu = 0.1\tau$. Sampling from (12) has been done with P-MYULA ($\lambda_{\text{MYULA}} = \rho^2$ and $\gamma_{\text{MYULA}} = \rho^2/4$ as prescribed in [15]) using Chambolle's algorithm [26] to compute the proximal operator associated to the total variation regularization function.

Choice of the hyperparameter ρ – Fig. 3 shows the convergence of the Markov chains associated to SGS and P-MYULA toward the typical set [27, Lemma 3.1.] of π. This convergence is directly related to the value of the hyperparameter ρ: small values lead to a slow convergence whereas high values lead to a fast convergence but with high asymptotic bias. Thus, the choice of ρ is a trade-off between good reconstruction properties and an efficient exploration of the parameters' space. According to our experiments, an intermediate value of $\rho \in [1, 3]$ seems to be a good trade-off. In the sequel, we set $\rho = 3$ for this experiment.

Results – Table 1 shows the performance results associated to both optimization and MCMC approaches averaged over 10 different runs. SGS and P-MYULA share roughly similar reconstruction results by comparing their minimum mean square error (MMSE) estimates although there are many other reconstruction metrics and summary statistics available. Interestingly, variable splitting-based approaches (SALSA and SGS) lead to a faster convergence compared to forward-backward splitting-based approaches (FISTA and P-MYULA). This behavior is related to two main consequences of the variable splitting step. First, as pointed out in [6], variable splitting leads to an algorithm that exploits second order information of the potential f_1 associated to the likelihood. In addition, the convergence rates of FISTA and P-MYULA are strongly related to the Lipschitz constant L_{f_1} of ∇f_1 which depends on $\mathbf{P}$. Thus, [15] shows that the dependence of the convergence rate of P-MYULA is of order $\mathcal{O}(L_{f_1}^2)$ where $L_{f_1} = 5.72$ in this experiment. By taking a variable splitting approach, SGS can embed P-MYULA which is now driven by a data-free Lipschitz constant $\rho^{-2} = 0.11$ that can be chosen carefully by the practitioner. Finally, fig. 4 shows the autocorrelation function (ACF) of the Markov chains associated to SGS

Table 1. Image deblurring with TV prior. Performance results for both optimization and simulation-based algorithms averaged over 10 runs. For MCMC algorithms, the SNR has been calculated with MMSE estimates.

	SALSA	FISTA	SGS	P-MYULA
time (s)	1	10	470	3600
time ($\times$ var. split.)	1	10	1	7.7
nb. iterations	22	214	$\sim 10^4$	10^5
SNR (dB)	17.87	17.86	18.36	17.97

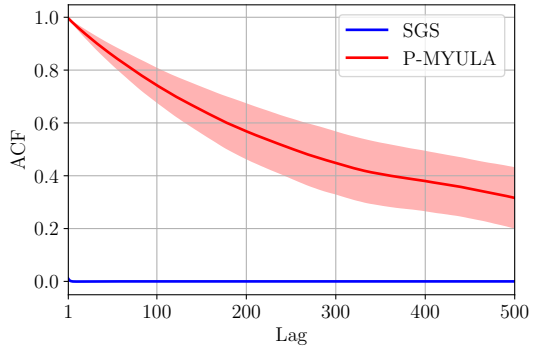

Fig. 4. Image deblurring with TV prior. Autocorrelation function of the Markov chains associated to SGS (blue) and P-MYULA (red) after the burn-in period. The shaded area corresponds to the standard deviation computed over 10 different runs.

and P-MYULA after the burn-in regime and by using $f_1 + f_2$ as a scalar summary. SGS appears to be more efficient than P-MYULA with again a well-chosen hyperparameter ρ.

3.3. Image deblurring in the wavelet domain

Problem formulation – In this second experiment, we consider the image deconvolution problem detailed in Section 3.2 and solve it by taking a synthesis approach. Thus, the direct operator $\mathbf{P}$ is now defined by $\mathbf{P} = \mathbf{BW}$, where the columns of $\mathbf{W}$ stand for the elements of a Haar wavelet frame with four levels. This *poor man's wavelet* has been chosen to illustrate the use of a frame-based approach (although more sophisticated frames can be considered). Since piecewise constant images (e.g. cartoon images) are sparse in the Haar wavelet domain; using this representation is similar to the TV regularization [28]. The objects of interest $\mathbf{x}$ are the coefficients of this image and the considered regularization potential is the ℓ_1 norm to promote sparsity inducing properties [29].
Experimental design – The proposed method is compared with P-MYULA. The total number of MCMC iterations and the number of burn-in iterations are set as in Section 3.2. The regularization parameter has been set to $\tau = 1$ and sampling from (12) has been performed by embedding P-MYULA and by using the soft-thresholding operator to compute the proximal operator of the ℓ_1 norm.
Choice of the hyperparameter ρ – The hyperparameter ρ has been set to $\rho = 1$ following the same type of arguments as in Section 3.2. Its influence on the convergence of the proposed sampler is depicted on fig. 5. Thus, similarly to the convergence behaviors shown in fig. 3, fig. 5 shows that the proposed sampler, by embedding P-MYULA, can accelerate its convergence toward the typical set of π.
Results – Fig. 6 shows the blurred and noisy observation, the MMSE

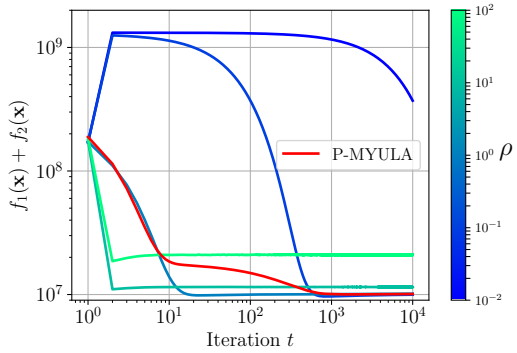

Fig. 5. Image deblurring with wavelets. Convergence of the Markov chains associated to SGS w.r.t. ρ (from guppie green to blue) and P-MYULA (red) toward the typical set of π.

Fig. 6. Image deblurring with wavelets. Noisy and blurred observation (left), MMSE estimate $\mathbf{W}\hat{\mathbf{x}}$ (middle, SNR = 21.52 dB) and 90% credibility intervals computed with SGS (right).

estimate of $\mathbf{Wx}$ and the 90% credibility intervals computed with SGS. Note that the latter cannot be obtained with an optimization-based method, e.g. SALSA or FISTA. Again, the proposed approach manages to deliver reliable results with a well-chosen hyperparameter ρ. Similarly to the performance results shown in Table 1 for the first experiment, SGS presents similar reconstruction performances as ADMM or P-MYULA and leads to an improvement in computational time. Some additional illustrations of the proposed approach can be found in [20, 22, 21] where it is shown that the derived Gibbs sampler is more efficient than state-of-the-art MCMC algorithms. Moreover, it can be distributed over multiple nodes with a well-chosen variable splitting strategy.

4. CONCLUSION

We have presented a new Bayesian framework inspired from variable splitting in optimization. Starting from an initial target distribution, a joint split distribution is defined leading to an approximate Bayesian hierarchical model in order to make the inference tractable and faster. In practice, the initial potential is split in a set of simpler ones that can be tackled in parallel and/or in distributed settings [21]. This comes at the cost of an approximation that is controlled and reliable. Thus, the proposed approach yields a very good compromise between performances and computational cost. Strong similarities between the associated Gibbs sampler and the ADMM can be pointed out, namely efficient and fast inference related to carefully chosen learning rates. The cost of the proposed approach compared to optimization-based algorithms is moderate and corresponds to the price to pay to get precious credibility intervals on the inferred parameters. A theoretical analysis of the proposed model is under consideration.

5. REFERENCES

[1] 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.

[2] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 09 1951.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in *Proc. Int. Conf. Machine Learning (ICML)*, 2011, pp. 681–688.

[9] 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.

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

[11] P. J. Green *et al.*, "Bayesian computation: a summary of the current state, and samples backwards and forwards," *Stat. Comput.*, vol. 25, no. 4, pp. 835–862, Jul 2015.

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

[13] G. Gautier, R. Bardenet, and M. Valko, "Zonotope hit-and-run for efficient sampling from projection DPPs," in *Proc. Int. Conf. Machine Learning (ICML)*, 2017.

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

[15] 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.

[16] C. P. Robert, *The Bayesian Choice: a decision-theoretic motivation*. Springer, 2001.

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

[18] R. Bardenet, A. Doucet, and C. Holmes, "On Markov chain Monte Carlo methods for tall data," *J. Mach. Learn. Res.*, May 2017.

[19] M. Vono, N. Dobigeon, and P. Chainais, "Bayesian image restoration under Poisson noise and log-concave prior," 2018. [Online]. Available: https://www.irit.fr/~Maxime.Vono/assets/pdf/appli_Poisson.pdf

[20] M. Vono, N. Dobigeon, and P. Chainais, "Split-and-augmented Gibbs sampler - Application to large-scale inference problems," *submitted*, 2018. [Online]. Available: https://arxiv.org/abs/1804.05809/

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

[22] 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.

[23] M. Elad, P. Milanfar, and R. Rubinstein, "Analysis versus synthesis in signal priors," vol. 23, pp. 947–968, 2007.

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

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

[27] M. Pereyra, "Maximum-a-posteriori estimation with Bayesian confidence regions," *SIAM J. Img. Sci.*, vol. 10, no. 1, pp. 285–302, March 2017.

[28] G. Steidl *et al.*, "On the equivalence of soft wavelet shrinkage, total variation diffusion, total variation regularization, and SIDEs," *SIAM J. Numer. Anal.*, vol. 42, no. 2, pp. 686–713, 2004.

[29] F. Bach *et al.*, "Optimization with sparsity-inducing penalties," *Found. Trends Mach. Learn.*, vol. 4, no. 1, pp. 1–106, Jan. 2012.