

The split Gibbs sampler revisited: improvements to its algorithmic structure and augmented target distribution

Marcelo Pereyra^{2,3}
m.pereyra@hw.ac.uk

Luis A. Vargas-Mieles^{1,2,3}
lv375@cam.ac.uk

Konstantinos C. Zygalakis^{1,3}
k.zygalakis@ed.ac.uk

May 4, 2023

Abstract

Developing efficient Bayesian computation algorithms for imaging inverse problems is challenging due to the dimensionality involved and because Bayesian imaging models are often not smooth. Current state-of-the-art methods often address these difficulties by replacing the posterior density with a smooth approximation that is amenable to efficient exploration by using Langevin Markov chain Monte Carlo (MCMC) methods. Such methods rely on gradient or proximal operators to exploit geometric information about the target posterior density and scale efficiently to large problems. An alternative approach is based on data augmentation and relaxation, where auxiliary variables are introduced in order to construct an approximate augmented posterior distribution that is amenable to efficient exploration by Gibbs sampling. This paper proposes a new accelerated proximal MCMC method called latent space SK-ROCK (ls SK-ROCK), which tightly combines the benefits of the two aforementioned strategies. Additionally, instead of viewing the augmented posterior distribution as an approximation of the original model, we propose to consider it as a generalisation of this model. Following on from this, we empirically show that there is a range of values for the relaxation parameter for which the accuracy of the model improves, and propose a stochastic optimisation algorithm to automatically identify the optimal amount of relaxation for a given problem. In this regime, ls SK-ROCK converges faster than competing approaches from the state of the art, and also achieves better accuracy since the underlying augmented Bayesian model has a higher Bayesian evidence. The proposed methodology is demonstrated with a range of numerical experiments related to image deblurring and inpainting, as well as with comparisons with alternative approaches from the state of the art. An open-source implementation of the proposed MCMC methods is available from <https://github.com/luisvargasmieles/ls-MCMC>.

Keywords: Bayesian inference, inverse problems, image processing, Markov chain Monte Carlo methods, mathematical imaging, proximal algorithms, uncertainty quantification.

AMS subject classifications: 62F15, 65C40, 65C60, 65J22, 68U10, 68W25

1 Introduction

The problem of estimating an unknown image from noisy and/or incomplete data is central to imaging sciences [16, 7]. Canonical examples include, for example, noise removal [31], image inpainting [55], image deblurring [20], medical imaging [29, 3, 37], astronomical imaging [13, 14]. Estimation by direct inversion of the forward model relating the unknown image to the data is not usually possible, inasmuch the inverse problem is often severely ill-conditioned or ill-posed. The literature describes a range of mathematical frameworks to incorporate regularisation and formulate well-posed solutions (see, e.g., [28, 48, 7]).

We consider Bayesian statistical solutions to imaging inverse problems [28]. In this case, the solution is a probability distribution characterising our knowledge about the value of the unknown image of interest given the observed data. This distribution can then be used to derive image estimators, calibrate unknown model parameters, perform uncertainty quantification analyses, or Bayesian model selection (see, e.g., [24, 50, 42, 12]).

There are three main strategies to perform Bayesian computation in imaging inverse problems. One strategy relies on optimisation methods, which typically scale efficiently to large models and offer detailed convergence guarantees, but can only support

¹School of Mathematics, University of Edinburgh, James Clerk Maxwell Building, Edinburgh, EH9 3FD, Scotland, UK.

²School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, Scotland, UK.

³Maxwell Institute for Mathematical Sciences, The Bayes Centre, 47 Potterrow, EH8 9BT, Edinburgh, Scotland, UK.

maximum-a-posteriori (MAP) estimation and a very limited range of other inferences [16, 14, 44]. Another strategy is to iteratively construct an approximation of the distribution of interest by fitting a tractable surrogate model. Prominent examples of this strategy include variational Bayesian inference [9, 41, 8] and expectation propagation [36, 56]. This approach can be very powerful for some models, but it often requires a careful model-specific implementation, it has weaker guarantees, and in some cases it can exhibit local convergence issues resulting in poor inferences. The third approach, which we adopt in this paper, is to directly approximate the distribution of interest by using a Markov chain Monte Carlo (MCMC) sampling method [45]. This allows computing the expectations and probabilities of interest in a highly reliable way. However, the application of MCMC methods comes at the expense of a higher computational cost. Reducing the cost of MCMC Bayesian computation in imaging has been a focus of significant efforts recently (see, e.g., [24, 42, 51]). Leaving computation strategies aside, it is possible to gain valuable insights about a Bayesian imaging model through mathematical analysis. A prominent example is Gaussian denoising under a total variation prior (see [33, 34]).

There are two main challenges in designing efficient MCMC algorithms for Bayesian imaging: the dimensionality of the problem and the fact that imaging models are often not smooth (this makes it difficult to directly apply gradient-based Markov Chain Monte Carlo methods). A first attempt to address these problems was the proposal of proximal MCMC methods [39], such as the so-called Moreau Yosida unadjusted Langevin algorithm (MYULA) [24] that combines ideas from Langevin gradient-based sampling with ideas from the field of non-smooth convex optimization. MYULA and its variants represent a significant improvement in computational efficiency and have good theoretical convergence guarantees. However, they are computationally inefficient for problems that are severely ill-conditioned because of step-size restrictions that leads to a slow exploration of the solution space. Recently, two different approaches were proposed in order to accelerate the convergence of proximal MCMC algorithms: the proximal stochastic Runge-Kutta-Chebyshev method (SK-ROCK) [2, 42], which carefully combines s gradient evaluations to achieve an s^2 -increase in the step-size, and the Split Gibbs Sampler (SGS) [51, 53], which is based on an augmentation and relaxation scheme that can significantly improve convergence speed at the expense of some estimation bias.

This paper explores two natural questions. First, how do SGS and SK-ROCK compare methodologically and empirically. Second, if the two methods can be combined in order to yield even more efficient MCMC methods. We address these questions in the following way:

1. Rather than viewing the model augmentation and relaxation strategy of [43, 51, 53] as an approximation, we propose to regard the augmented model as a generalisation of the original model. We show empirically that there is a range of relaxation values for which the accuracy of the model improves. In this regime, relaxation leads to better convergence properties and better accuracy. Beyond this regime, the accuracy of the relaxed model deteriorates rapidly.
2. Given the critical role of the amount of relaxation, we build on [50] to propose an empirical Bayesian method to automatically estimate the value of the relaxation parameter by maximum marginal likelihood estimation.
3. We formally identify a relationship between SGS and MYULA by re-expressing SGS as a discrete-time approximation of a Langevin stochastic differential equation (SDE) closely related to MYULA.
4. Having connected SGS and MYULA at the level of the SDE, we propose two novel MCMC methods for Bayesian imaging: 1) an integration of SGS and MYULA that improves on both SGS and MYULA; and 2) an integration of SGS and SK-ROCK that outperform SK-ROCK, the previously fastest method in the literature.

The remainder of the paper is organised as follows: In Section 2 we introduce the models considered in this work, and recall the state-of-the-art MCMC methods to sample from them. In Section 3 we revisit the augmented model and show empirically that there is a subset of hyperparameter values that enhances this model, while we also present a method to automatically compute an optimal choice of these hyperparameters. In Section 4 we establish a formal connection between MYULA and SGS which allows us to propose two novel and more efficient sampling algorithms for the augmented model. Section 5 illustrates the proposed methodologies with two experiments related to image deblurring and image inpainting, where we report detailed comparisons with state-of-the-art algorithms. Conclusions and perspectives for future work are reported in Section 6.

2 Problem statement

2.1 Bayesian inference and imaging inverse problems

Let $x \in \mathbb{R}^d$ be the image we are interested in estimating and y the available observation related to x by a statistical model with likelihood function

$$p(y|x) \propto e^{-f_y(x)}.$$

In this work, we pay special attention to problems where the estimation of x given y is ill-posed or ill-conditioned¹. We address this difficulty by considering the Bayesian framework, where we regularize the estimation problem by specifying a prior distribution, given for any x and $\theta \in (0, +\infty)^{d'}$ by

$$p(x|\theta) \propto e^{-\theta^\top g(x)},$$

for some vector of statistics $g : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$, where θ parametrises this prior distribution and controls the level of imposed regularity. Using Bayes' theorem, we derive the posterior distribution

$$p(x|y, \theta) = \frac{p(y|x)p(x|\theta)}{p(y|\theta)} = \frac{\exp[-f_y(x) - \theta^\top g(x)]}{\int_{\mathbb{R}^d} \exp[-f_y(x') - \theta^\top g(x')] dx'}, \quad (2.1)$$

which models the knowledge we have about x given the observed data y .

We focus on Bayesian computational methodology for log-concave models of the form (2.1), where f_y and g satisfy the following conditions:

1. $f_y : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex and Lipschitz continuously differentiable with constant L_f
2. $(g_i)_{i \in \{1, \dots, d'\}} : \mathbb{R}^d \rightarrow \mathbb{R}$ is proper, convex, and lower semi continuous, but potentially non-smooth.

Models with these characteristics are widely adopted in the imaging community due to the variety of currently available Bayesian optimization tools that exploit the convexity properties of $p(x|y, \theta)$ [44, 38], such as the computation of the MAP estimate, which can be formulated as a convex optimization problem [40] given by

$$\hat{x}_{\text{MAP}} = \underset{x}{\operatorname{argmax}} p(x|y, \theta) = \underset{x}{\operatorname{argmin}} f_y(x) + \theta^\top g(x), \quad (2.2)$$

and can be solved by using state-of-the-art convex optimization algorithms [16]. However, there are other more complicated Bayesian analyses beyond point estimation, such as model calibration, Bayesian model selection and hypothesis testing [46], which cannot be addressed by using optimization algorithms and typically require the calculation of probabilities and expectations w.r.t $p(x|y, \theta)$. This is challenging in imaging problems since it requires calculating intractable integrals on \mathbb{R}^d . In this case, the application of MCMC methods [45, 6] and, in particular, proximal MCMC methods [39, 24, 42, 51, 53], specialised for non-smooth log-concave distributions, is the preferred approach.

2.2 Sampling via the Langevin diffusion

We consider proximal MCMC methods derived from the discretization of the overdamped Langevin diffusion process, which we discuss below. Assume that we are interested in calculating probabilities w.r.t. a smooth distribution with density $\pi(x)$, and consider the stochastic differential equation (SDE)

$$dX_t = \nabla \log \pi(X_t) dt + \sqrt{2} dW_t, \quad (2.3)$$

where W_t is a d -dimensional Brownian motion. Under mild assumptions on $\pi(x)$, this SDE has a unique strong solution and admits $\pi(x)$ as its unique invariant density [47, Theorem 2.1]. However, in imaging applications it is usually not possible to solve (2.3) exactly, so a numerical approximation needs to be employed. Most works consider the Euler-Maruyama (EM) scheme given by

$$X_{n+1} = X_n + \delta \nabla \log \pi(X_n) + \sqrt{2\delta} Z_{n+1}, \quad (2.4)$$

where $\delta > 0$ is a given step-size and $(Z_{n+1})_{n \geq 0}$ is an i.i.d. sequence of d -dimensional standard Gaussian random vectors. This recursion is known as the unadjusted Langevin algorithm (ULA) and has been shown to be a highly efficient method for high-dimensional Bayesian inference when $\pi(x)$ is log-concave and smooth with $\nabla \log \pi(x)$ L -Lipschitz continuous and $\delta < 1/L$ [23, 22]. However, in many imaging models, $\pi(x)$ is not smooth and hence appropriate adjustments need to be made to ULA.

2.2.1 Moreau-Yosida unadjusted Langevin algorithm (MYULA)

Proximal MCMC methods [39] deal with the non-differentiability of $\pi(x)$ by replacing $\pi(x)$ with a smooth approximation $\pi^\lambda(x)$ which, by construction, satisfies the required conditions of ULA. In this case $p(x|y, \theta)$ in (2.1) is replaced by $p^\lambda(x|y, \theta)$ defined as

$$p^\lambda(x|y, \theta) = \frac{p(y|x)p^\lambda(x|\theta)}{p^\lambda(y|\theta)} = \frac{\exp[-f_y(x) - \theta^\top g^\lambda(x)]}{\int_{\mathbb{R}^d} \exp[-f_y(x') - \theta^\top g^\lambda(x')] dx'}, \quad (2.5)$$

¹ Either the problem does not admit a unique solution that changes continuously with y , or there exists a unique solution but it is not stable w.r.t. small perturbations in y .

where

$$g^\lambda(x) = [g_1^\lambda(x), \dots, g_{d'}^\lambda(x)],$$

that is, each non-smooth term $g_i(x)$, $i \in \{1, \dots, d'\}$ is replaced by its Moreau-Yosida envelope², defined as

$$g_i^\lambda(x) = \min_{u \in \mathbb{R}^d} \left\{ g_i(u) + \frac{1}{2\lambda} \|x - u\|^2 \right\}. \quad (2.6)$$

This leads to a smooth posterior (2.5) which has the following properties:

- $p^\lambda(x|y, \theta)$ is log-concave and Lipschitz continuously differentiable with gradient

$$\begin{aligned} \nabla \log p^\lambda(x|y, \theta) &= -\nabla f_y(x) - \nabla(\theta^\top g^\lambda(x)), \\ &= -\nabla f_y(x) - \frac{1}{\lambda} \sum_{i=1}^p \left(x - \text{prox}_{\theta_i g_i}^\lambda(x) \right), \end{aligned}$$

with Lipschitz constant $L = L_f + p/\lambda$, and for every $x \in \mathbb{R}^d$

$$\text{prox}_{g_i}^\lambda(x) = \underset{u \in \mathbb{R}^d}{\text{argmin}} \left\{ g_i(u) + \frac{1}{2\lambda} \|x - u\|^2 \right\}.$$

- $p^\lambda(x|y, \theta)$ converges in total variation to $p(x|y, \theta)$ [24, Proposition 3.1], i.e.,

$$\lim_{\lambda \rightarrow 0} \|p^\lambda(x|y, \theta) - p(x|y, \theta)\|_{\text{TV}} = 0.$$

Applying the ULA scheme to the smooth posterior approximation $p^\lambda(x|y, \theta)$ leads to the recursion

$$X_{n+1} = X_n - \delta \nabla f_y(X_n) - \frac{\delta}{\lambda} \sum_{i=1}^{d'} \left(X_n - \text{prox}_{\theta_i g_i}^\lambda(X_n) \right) + \sqrt{2\delta} Z_{n+1},$$

which is known as the Moreau-Yosida unadjusted Langevin algorithm (MYULA) [24]. The main benefit of the MYULA is that now since $p^\lambda(x|y, \theta)$ is smooth and preserves log-concavity, the results from [23, 22] apply hence providing an efficient method for imaging applications.

One of the main computational bottlenecks of ULA and MYULA is the fact that, in order to converge, one needs to choose $\delta \leq 1/L$ where L is the Lipschitz constant of $\nabla \log p^\lambda(x|y, \theta)$, given by $L = L_f + d'/\lambda$ (see [24, Theorem 3.2]). This step-size restriction is not problematic for moderate values of L , for example in denoising problems. However, in problems of the form $y = Ax + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_m)$ and $\sigma > 0$ where the forward operator A is very poorly conditioned, the MYULA sampler will have poor mixing properties, particularly on the subspace of \mathbb{R}^d where x has high uncertainty (this is related to the *slow* components of the Markov chain). A similar situation occurs when the problem requires a high level of accuracy in the approximated prior term $g_i^\lambda(x)$ (i.e., λ is very small), for example in problems where one has to enforce domain constraints on the solution space.

2.2.2 SK-ROCK

A natural way of overcoming the step-size limitation of the MYULA is to adopt a discretization scheme for the Langevin Diffusion (2.3) with better numerical stability properties. In particular, an explicit stochastic Runge-Kutta-Chebyshev discretization of the Langevin SDE was proposed in [2] called SK-ROCK, and a proximal variant suitable for computational imaging problems was recently proposed in [42]. This highly advanced Runge-Kutta stochastic integration scheme extends the deterministic Chebyshev method [1] to sampling processes and has been shown to require fewer gradient evaluations than ULA or MYULA to reach a given level of accuracy. In particular, it has been shown in [42, Proposition 3.1] that, in order to be ε close to a multivariate Gaussian target distribution (in the 2-Wasserstein distance), SK-ROCK requires $\mathcal{O}(\sqrt{\kappa})$ gradient evaluations (where κ is the condition number of $\nabla \log \pi(x)$) instead of $\mathcal{O}(\kappa)$ required by the ULA, similar to accelerated optimization methods.

This is possible due to the fact that SK-ROCK uses s gradient evaluations per iteration at carefully chosen extrapolated points, which allows using a much larger step-size than the MYULA and thus having a faster decorrelation of the Markov chain in a stable manner. This is one of the great advantages of SK-ROCK in very ill-posed and ill-conditioned models, compared to MYULA, accelerating its convergence while maintaining its stability, thus improving the quality of the generated samples.

²If the calculation of the proximal operator of the sum of some elements of g is possible, it is not necessary to replace each of these elements of the vector g with its corresponding Moreau-Yosida envelope. In addition, if there is some $g_k(x)$, $k \in \{1, \dots, d'\}$ that is Lipschitz differentiable, its gradient can be computed directly.

The SK-ROCK scheme is shown in Algorithm 1, where T_s denotes the Chebyshev polynomial of order s of the first kind, defined recursively by $T_{k+1} = 2xT_k(x) - T_{k-1}(x)$ with $T_0(x) = 1$ and $T_1(x) = x$. The two main parameters of the algorithm are the number of stages $s \in \mathbb{N}^*$ and the step-size $\delta \in (0, \delta_s^{\max}]$, where the range of admissible values for δ is controlled by s : for any $s \in \mathbb{N}^*$, the maximum allowed step-size is given by $\delta_s^{\max} = l_s/(L_f + 1/\lambda)$ with $l_s = [(s - 0.5)^2(2 - 4/3\eta) - 1.5]$ and $\eta = 0.05$ [2]. Please see Section 4.4 for guidelines for setting s and δ .

Algorithm 1 SK-ROCK

- 1: **Input:** $X_0 \in \mathbb{R}^d$, $\lambda > 0$, $n, s \in \mathbb{N}$, $\eta = 0.05$.
 - 2: **Compute** $l_s = (s - 0.5)^2(2 - 4/3\eta) - 1.5$,
 - 3: **Compute** $\omega_0 = 1 + \eta/s^2$, $\omega_1 = T_s(\omega_0)/T'_s(\omega_0)$,
 - 4: **Compute** $\mu_1 = \omega_1/\omega_0$, $\nu_1 = s\omega_1/2$, $k_1 = s\omega_1/\omega_0$,
 - 5: **Choose** $\delta \in (0, \delta_s^{\max}]$, where $\delta_s^{\max} = l_s/(L_f + 1/\lambda)$,
 - 6: **for** $i = 0 : n - 1$ **do**
 - 7: **Set** $\tilde{X}_0 = X_i$,
 - 8: **Sample** $\xi_{i+1} \sim \mathcal{N}(0, 2\delta\mathbb{I}_d)$,
 - 9: **Compute** $\tilde{X}_1 = \tilde{X}_0 + \mu_1\delta\nabla \log p^\lambda(\tilde{X}_0 + \nu_1\xi_{i+1}|y, \theta) + k_1\xi_{i+1}$,
 - 10: **for** $j = 2 : s$ **do**
 - 11: **Compute** $\mu_j = 2\omega_1T_{j-1}(\omega_0)/T_j(\omega_0)$, $\nu_j = 2\omega_0T_{j-1}(\omega_0)/T_j(\omega_0)$, $k_j = 1 - \nu_j$,
 - 12: **Compute** $\tilde{X}_j = \mu_j\delta\nabla \log p^\lambda(\tilde{X}_{j-1}|y, \theta) + \nu_j\tilde{X}_{j-1} + k_j\tilde{X}_{j-2}$,
 - 13: **end for**
 - 14: **Set** $X_{i+1} = \tilde{X}_s$,
 - 15: **end for**
 - 16: **Output:** Samples X_1, \dots, X_n .
-

2.3 Sampling via augmentation: Split Gibbs sampler (SGS)

A separate line of research seeks to address the limitation of ULA (and MYULA) by introducing an auxiliary variable $z \in \mathbb{R}^d$ and operating on the augmented state-space (x, z) . This allows to relax the original model (2.1) and instead uses the following augmented posterior

$$\begin{aligned}
 p(x, z|y, \theta, \rho^2) &= \frac{p(y|x)p(x, z|\theta, \rho^2)}{p(y|\theta, \rho^2)} = \frac{p(y|x)p(x|z, \rho^2)p(z|\theta)}{\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(y|x)p(x|z, \rho^2)p(z|\theta)dx dz} \\
 &= \frac{\exp[-f_y(x) - \theta^\top g(z) - \frac{1}{2\rho^2}\|x - z\|^2]}{\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \exp[-f_y(x) - \theta^\top g(z) - \frac{1}{2\rho^2}\|x - z\|^2] dx dz}, \quad \rho^2 > 0,
 \end{aligned} \tag{2.7}$$

where

$$p(y|x) = \frac{\exp(-f_y(x))}{\int_{\mathbb{R}^d} \exp(-f_y(x))dx}, \quad p(z|\theta) = \frac{\exp(-\theta^\top g(z))}{\int_{\mathbb{R}^d} \exp(-\theta^\top g(z'))dz'} \tag{2.8}$$

$$p(x|z, \rho^2) = \frac{\exp(-\|x - z\|^2/2\rho^2)}{\int_{\mathbb{R}^d} \exp(-\|x - z\|^2/2\rho^2)dx} = \frac{\exp(-\|x - z\|^2/2\rho^2)}{(2\pi\rho^2)^{d/2}}, \tag{2.9}$$

where ρ^2 controls the correlation between the variable of interest x and the auxiliary variable z , and f_y, g are the same as in Section 2.1. If we now consider the marginal posterior distribution

$$p(x|y, \theta, \rho^2) = \int_{\mathbb{R}^d} p(x, z|y, \theta, \rho^2) dz, \tag{2.10}$$

it is possible to show that it converges in total variation to the original posterior $p(x|y, \theta)$ as $\rho^2 \rightarrow 0$.

This approach was first introduced several decades ago as a way of calculating maximum likelihood estimates from incomplete data [19], and as an efficient method for sampling from posterior distributions [49] (see [25] for a review of these techniques). In

the current literature, this model was revisited by [43] in the context of consensus Monte Carlo in distributed settings and applied to imaging inverse problems in [51], where its similarities to the algorithmic structure of the Alternating Direction Method of Multipliers (ADMM) optimization algorithm [10] were also discussed.

From a computational point of view, as in the case of MYULA, because g is not differentiable one needs to approximate $p(x, z|y, \theta, \rho^2)$ by

$$p^\lambda(x, z|y, \theta, \rho^2) \propto \exp \left[-f_y(x) - \theta^\top g^\lambda(z) - \frac{1}{2\rho^2} \|x - z\|^2 \right], \quad \rho^2 > 0. \quad (2.11)$$

To sample (2.11), [51, 53] proposed a Gibbs-like splitting strategy scheme, applied on the following conditional distributions

$$p(x|y, z, \rho^2) \propto \exp \left[-f_y(x) - \frac{1}{2\rho^2} \|x - z\|^2 \right], \quad (2.12)$$

$$p^\lambda(z|x, \theta, \rho^2) \propto \exp \left[-\theta^\top g^\lambda(z) - \frac{1}{2\rho^2} \|x - z\|^2 \right]. \quad (2.13)$$

This method is known as the split Gibbs sampler (SGS). See Algorithm 2. In the case where the likelihood is Gaussian one can

Algorithm 2 SGS

- 1: Input: $X_0, Z_0 \in \mathbb{R}^d$, $\lambda, \rho^2 > 0$, $n \in \mathbb{N}$.
 - 2: **for** $i = 0 : n - 1$ **do**
 - 3: Sample $X_{i+1} \sim p(x|y, Z_i, \rho^2)$ according to (2.12),
 - 4: Compute $Z_{i+1} = Z_i - \delta \sum_{k=1}^{d'} [Z_i - \text{prox}_{\theta_k g_k}^\lambda(Z_i)] / \lambda - \delta(Z_i - X_{i+1}) / \rho^2 + \sqrt{2\delta} \zeta_{i+1}$; where $\zeta_{i+1} \sim \mathcal{N}(0, \mathbb{I}_d)$,
 - 5: **end for**
 - 6: Output: Samples X_1, \dots, X_n .
-

exactly sample from (2.12) [27] (for a review and comparison of existing Gaussian sampling approaches, see [52]). Additionally, when the iterates Z_i are also sampled exactly from $p(z|y, X_i, \rho^2)$ (i.e., by replacing Step 4 by an exact sampler), the resulting scheme is provably ergodic and can be used for approximate inference w.r.t. $p(x|y)$ [53]. Leaving exact sampling aside, a main benefit of this splitting approach is that the step-size one needs to set for the proximal MCMC method used for sampling (2.13) will be independent of the Lipschitz constant associated with the likelihood distribution, and will only depend on the parameters λ and ρ^2 (i.e., the step-size now depends on the Lipschitz constant indirectly via λ and ρ^2 , there is still some dependence w.r.t. the Lipschitz constant through the bias incurred). This can lead to faster sampling algorithms compared to MYULA for suitably chosen values of the parameter ρ^2 [51], albeit for a biased posterior distribution.

3 Enhancing Bayesian imaging models by smoothing

As discussed previously, the augmented model (2.7) was originally proposed as a relaxation of (2.1) that allows for a faster exploration of the target distribution, at the expense of some additional bias when compared to the original model. One then might think that $\rho^2 = 0$ represents the best model for inference (at the expense of higher computing cost). However, we have found empirically that this is not the case.

As an illustration, Figure 1(a) shows the estimation mean-squared error (MSE) for a Bayesian image deblurring problem (the details of this experiment will be explained in Section 5.1). The error is computed w.r.t. the posterior mean, as estimated by an adaptation of the SK-ROCK method to target (2.11) (see Section 4.3 for details), using a value of $\theta = 4.4 \times 10^{-2}$ estimated by [50, Algorithm 1], and by using different values for ρ^2 . Recalling that increasing ρ^2 improves convergence speed, one can clearly identify a regime of small values of ρ^2 for which convergence speed improves without a deterioration in estimation accuracy (in fact, there is a mild improvement). Beyond this range, the estimation MSE deteriorates dramatically. This suggests the need for a method to automatically set the value of ρ^2 .

We propose an empirical Bayesian method to estimate optimal values for θ and ρ^2 directly from y by maximum marginal likelihood estimation (MMLE)

$$(\theta_*, \rho_*^2) = \underset{\theta \in \Theta, \rho^2 \in \Omega}{\operatorname{argmax}} p(y|\theta, \rho^2), \quad (3.1)$$

where $\Theta \subset (0, +\infty)^{d'}$, $\Omega \subset (0, +\infty)$ are compact convex sets, and $p(y|\theta, \rho^2)$ is defined in (2.7). To solve (3.1), we modify the stochastic approximation proximal gradient (SAPG) algorithm of [50]. By maximising the model evidence, (3.1) seeks to select the best model to perform inference within the class of posterior distributions parametrised by $\theta \in \Theta, \rho^2 \in \Omega$ [54].

3.1 Computing the optimal values for θ and ρ^2

We adopt the approach of [50] to solve (3.1) and estimate optimal values for θ and ρ^2 in (2.7). The method [50] was proposed for models of the form (2.1), so we will now adapt it to the augmented model (2.7).

We are interested in estimating the parameters $\theta \in \Theta$, $\rho^2 \in \Omega$ by MMLE (3.1). If we had access to the gradients $\nabla_{\rho^2} \log p(y|\theta, \rho^2)$ and $\nabla_{\theta} \log p(y|\theta, \rho^2)$, then we could construct an iterative algorithm that converges to the solution of (3.1) by using the projected gradient algorithm [32]

$$\begin{aligned}\rho_{n+1}^2 &= \Pi_{\Omega} [\rho_n^2 + \gamma'_n \nabla_{\rho^2} \log p(y|\theta_n, \rho_n^2)] \\ \theta_{n+1} &= \Pi_{\Theta} [\theta_n + \gamma_n \nabla_{\theta} \log p(y|\theta_n, \rho_n^2)],\end{aligned}$$

where Π_{Ω} and Π_{Θ} are the projection onto Ω and Θ respectively, and $(\gamma'_n, \gamma_n)_{n \in \mathbb{N}}$ are sequences of non-increasing step-sizes such that $\sum_{n \in \mathbb{N}} \gamma'_n \rightarrow +\infty$ and $\sum_{n \in \mathbb{N}} \gamma_n^2 < \infty$, and similarly $\sum_{n \in \mathbb{N}} \gamma_n \rightarrow +\infty$ and $\sum_{n \in \mathbb{N}} \gamma_n^2 < \infty$, (see Section 4.4 for details). However, due to the complexity of the model, $\nabla_{\rho^2} \log p(y|\theta, \rho^2)$ and $\nabla_{\theta} \log p(y|\theta, \rho^2)$ are intractable.

As shown in [50], one can construct carefully designed stochastic estimates of these gradients that satisfy the conditions for the solution to converge to (3.1). To build these stochastic estimators, we are going to express the gradients as expectations by applying the Fisher's identity [21, Proposition D.4]. More precisely, we have that

$$\nabla_{\rho^2} \log p(y|\theta, \rho^2) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(x, z|y, \theta, \rho^2) \nabla_{\rho^2} \log p(x, z, y|\theta, \rho^2) dx dz,$$

and

$$\nabla_{\theta} \log p(y|\theta, \rho^2) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(x, z|y, \theta, \rho^2) \nabla_{\theta} \log p(x, z, y|\theta, \rho^2) dx dz.$$

We can approximate these expectations by using MCMC. In fact, we will see that one MCMC sample will suffice to obtain an estimate of the gradient accurate enough to converge asymptotically to (3.1). As $p(x, z, y|\theta, \rho^2) = p(y|x)p(x, z|\theta, \rho^2) = p(y|x)p(x|z, \rho^2)p(z|\theta)$, we have

$$\nabla_{\rho^2} \log p(y|\theta, \rho^2) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(x, z|y, \theta, \rho^2) \nabla_{\rho^2} \log p(x|z, \rho^2) dx dz,$$

and

$$\nabla_{\theta} \log p(y|\theta, \rho^2) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(x, z|y, \theta, \rho^2) \nabla_{\theta} \log p(z|\theta) dx dz.$$

Replacing (2.9) in $p(x|z, \rho^2)$ we obtain

$$\nabla_{\rho^2} \log p(y|\theta, \rho^2) = A_{\theta, \rho^2}(y) - \frac{d}{2\rho^2},$$

where

$$A_{\theta, \rho^2}(y) = \mathbb{E}_{x, z|y, \theta, \rho^2} \left[\frac{\|x - z\|^2}{2(\rho^2)^2} \right] = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(x, z|y, \theta, \rho^2) \frac{\|x - z\|^2}{2(\rho^2)^2} dx dz,$$

and similarly, replacing (2.8) in $p(z|\theta)$ gives

$$\nabla_{\theta} \log p(y|\theta, \rho^2) = -B_{\theta, \rho^2}(y) - C_{\theta, \rho^2}(y),$$

where

$$\begin{aligned}B_{\theta, \rho^2}(y) &= \mathbb{E}_{x, z|y, \theta, \rho^2} [g(z)] = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(x, z|y, \theta, \rho^2) g(z) dx dz, \\ C_{\theta, \rho^2}(y) &= \mathbb{E}_{x, z|y, \theta, \rho^2} \left[\nabla_{\theta} \log \left(\int_{\mathbb{R}^d} \exp(-\theta^T g(z)) dz \right) \right] \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} p(x, z|y, \theta, \rho^2) \nabla_{\theta} \log \left[\int_{\mathbb{R}^d} \exp(-\theta^T g(z)) dz \right] dx dz.\end{aligned}$$

Because of the complexity of the model, $A_{\theta, \rho^2}(y)$ and $B_{\theta, \rho^2}(y)$ are not available analytically and need to be approximated by MCMC computation (e.g., by using the methods we develop in Section 4). With respect to $C_{\theta, \rho^2}(y)$, and more precisely, the integral between brackets, we can follow a similar procedure as in [50, Section 3.2.1]. In particular, if we consider the case where each $g_i(z)$ is α_i positively homogeneous³, which is the case for many regularizers such as ℓ_1 , ℓ_2 or TV, we have that

$$\frac{\partial \log p(y|\theta, \rho^2)}{\partial \theta^{(i)}} = \frac{d}{\alpha_i \theta^{(i)}} - \mathbb{E}_{x, z|y, \theta^{(i)}, \rho^2} [g_i(z)].$$

³ $g(x)$ is α positively homogeneous if, for any $x \in \mathbb{R}^d$ and $t > 0$, $g(tx) = t^\alpha g(x)$.

(See [50] for more details and [50, Section 3.2] for the case of inhomogeneous regularizers).

Following on from this, and by using Monte Carlo approximations of $A_{\theta, \rho^2}(y)$ and $B_{\theta, \rho^2}(y)$, we construct an SAPG algorithm [26, 50] to solve (3.1) and produce optimal estimates of θ and ρ^2 . This method is presented in Algorithm 3. We refer the reader to Section 4.4 for guidelines on setting the step-size $(\gamma'_n, \gamma_n)_{n \in \mathbb{N}}$ and the weights $(w_n)_{n \in \mathbb{N}}$ for the averages computed in the final step of Algorithm 3. See also [18, 17] for details about the convergence properties of this kind of SAPG algorithm.

Algorithm 3 SAPG algorithm for the augmented model (2.7)

```

1: Input:  $X_0^0, Z_0^0 \in \mathbb{R}^d, \theta_0, \rho_0^2, \gamma_0, \gamma'_0 \in \mathbb{R}, \lambda > 0, m, n \in \mathbb{N}$ .
2: for  $i = 0 : m - 1$  do
3:   if  $i > 0$  then
4:     Set  $X_i^{(0)} = X_{i-1}^{(n)}$ ,
5:   end if
6:   for  $j = 0 : n - 1$  do
7:     Sample  $X_{i+1}^{(j+1)}, Z_{i+1}^{(j+1)}$  according to Algorithm 4,
8:   end for
9:   for  $j = 1 : d'$  do
10:    Set  $\theta_{i+1}^{(j)} = \Pi_{\Theta} \left[ \theta_i^{(j)} + \frac{\gamma_{i+1}^{(j)}}{n} \sum_{k=1}^n \left\{ \frac{d}{\alpha_j \theta_i^{(j)}} - g_j(Z_{i+1}^{(k)}) \right\} \right]$ ,
11:   end for
12:   Set  $\rho_{i+1}^2 = \Pi_{\Omega} \left[ \rho_i^2 + \frac{\gamma'_{i+1}}{n} \sum_{k=1}^n \left\{ \|X_{i+1}^{(k)} - Z_{i+1}^{(k)}\|^2 / 2(\rho_i^2)^2 - d / (2\rho_i^2) \right\} \right]$ ,
13: end for
14: Output:  $\bar{\theta}_m^{(j)} = \sum_{k=0}^m w_k \theta_k^{(j)} / \sum_{k=0}^m w_k$  for  $j \in \{1, \dots, d'\}$ ,  $\bar{\rho}_m^2 = \sum_{k=0}^m w_k \rho_k^2 / \sum_{k=0}^m w_k$ .

```

To illustrate Algorithm 3 in action, Figure 1 shows the value of ρ^2 estimated by the algorithm for the image deblurring problem. Observe that the MMLE estimate is close to the value that produces the best estimation MSE in this case. This is in agreement with the results reported in [50] for other problems. Lastly, notice that Step 10 of Algorithm 3 involves the original prior, with terms $\{g_j\}_{j=1}^{d'}$, and not the smooth approximations $\{g_j^\lambda\}_{j=1}^{d'}$, as the SAPG scheme to estimate θ and ρ^2 is not affected by the non-smoothness $p(z|\theta)$ w.r.t z (instead, it requires some smoothness of the Markov kernels w.r.t. θ , see [17, 18] for technical details). The approximation $\{g_j^\lambda\}_{j=1}^{d'}$ is used within the MYULA step of Step 7 Algorithm 3, which does require a smooth prior. This mismatch introduces some bias, which is controlled by using a small value of λ [17]. One could use $\{g_j^\lambda\}_{j=1}^{d'}$ instead of $\{g_j\}_{j=1}^{d'}$ within Step 10. However, doing so would prevent the SAPG scheme from exploiting the homogeneity of $\{g_j\}_{j=1}^{d'}$, resulting in a significantly more expensive SAPG scheme (see [50, Algorithm 3]).

4 Reinterpretation of SGS as noisy MYULA & new MCMC methods

4.1 Noisy MYULA

We now proceed to show that the SGS algorithm 2 can be viewed as a noisy version of MYULA. This link will be crucial in allowing us to write it as a noisy discretisation of an SDE, which will help us to propose more efficient MCMC methods for sampling (2.7).

First, note that the marginal of z computed from (2.11) can be written as follows

$$p^\lambda(z|y, \theta, \rho^2) = \int_{\mathbb{R}^d} p^\lambda(x, z|y, \theta, \rho^2) dx \propto p(y|z, \rho^2) p^\lambda(z|\theta),$$

where

$$p(y|z, \rho^2) \propto \int_{\mathbb{R}^d} \exp \left[-f_y(x) - \frac{1}{2\rho^2} \|x - z\|^2 \right] dx, \quad p^\lambda(z|\theta) \propto \exp[-\theta^\top g^\lambda(z)].$$

Notice that in the case where $f_y(x)$ is quadratic, $p(y|z, \rho^2)$ is Gaussian with eigenvalues in its covariance matrix shifted by ρ^2 , when compared with the covariance of $f_y(x)$. Now applying the MYULA to $p^\lambda(z|y, \theta, \rho^2)$, we have that

$$Z_{n+1} = Z_n + \delta \nabla_z \log p^\lambda(Z_n|\theta) + \delta \nabla_z \log p(y|Z_n, \rho^2) + \sqrt{2\delta} \zeta_{n+1}, \quad (4.1)$$

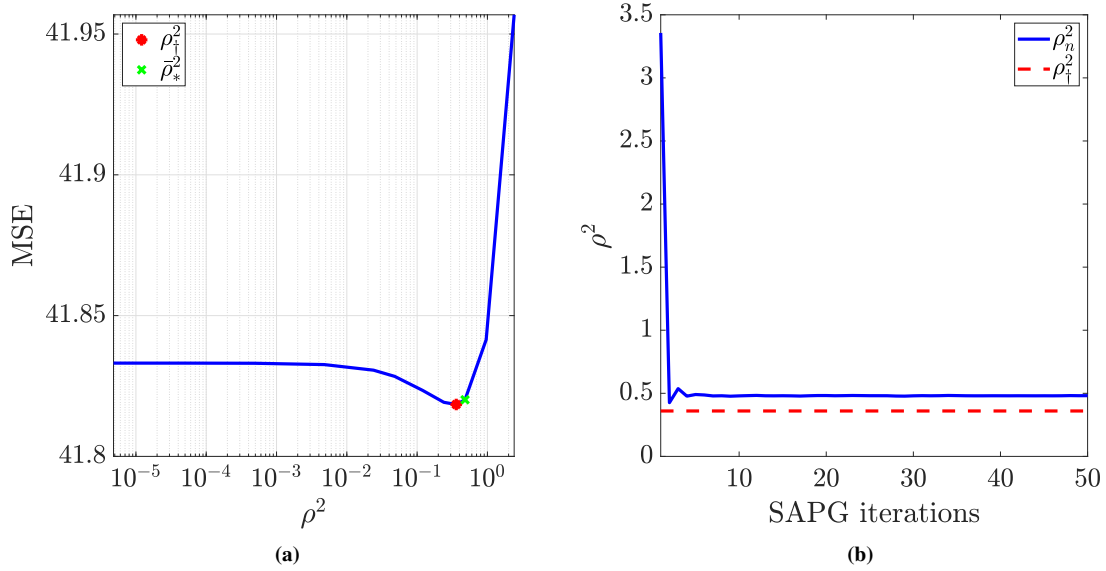


Figure 1: Image deblurring experiment: (a) MSE between the true image and the posterior mean estimated using Algorithm 5, for some values of ρ^2 . In red, the optimal value of ρ^2 that minimises the MSE, and in green, the value of ρ^2 found by Algorithm 3. (b) Iterations of SAPG algorithm to estimate ρ^2 .

where $(\zeta_{n+1})_{n \geq 0}$ is an i.i.d. sequence of d -dimensional standard Gaussian random vectors. Due to the complexity of the model, it is difficult to compute $\nabla_z \log p(y|z, \rho^2)$, however, we can express it as an expectation by using Fisher's identity [21, Proposition D.4] as follows

$$\begin{aligned} \nabla_z \log p(y|z, \rho^2) &= \int_{\mathbb{R}^d} p(x|y, z, \rho^2) \nabla_z \log p(x, y|z, \rho^2) dx \\ &= \mathbb{E}_{x|y, z, \rho^2} [\nabla_z \log p(x, y|z, \rho^2)]. \end{aligned}$$

As $p(x, y|z, \rho^2) = p(y|x)p(x|z, \rho^2)$, we have

$$\begin{aligned} \nabla_z \log p(x, y|z, \rho^2) &= \mathbb{E}_{x|y, z, \rho^2} [\nabla_z \log p(x|z, \rho^2)] \\ &= \frac{1}{\rho^2} \mathbb{E}_{x|y, z, \rho^2} (x - z). \end{aligned}$$

Using this expression in (4.1) we obtain

$$Z_{n+1} = Z_n - \delta \nabla_z p^\lambda(Z_n|\theta) - \frac{\delta}{\rho^2} \mathbb{E}_{x|y, z, \rho^2} (Z_n - x) + \sqrt{2\delta} \zeta_{n+1}, \quad (4.2)$$

We are now ready to explicitly establish the connection to SGS. SGS stems from dealing with the presence of the expectation in this algorithm by replace it by a Monte Carlo empirical average, i.e.,

$$\mathbb{E}_{x|y, z, \rho^2} (Z_n - x) \approx Z_n - \frac{1}{N} \sum_{i=1}^N X^{(i)}, \text{ where } X^{(i)} \sim p(x|y, Z_n; \rho). \quad (4.3)$$

More precisely, to recover SGS we take $N = 1$ and substitute in (4.2) to obtain

$$Z_{n+1} = Z_n - \delta \nabla_z p^\lambda(Z_n|\theta) - \frac{\delta}{\rho^2} (Z_n - X^{(1)}) + \sqrt{2\delta} \zeta_{n+1}. \quad (4.4)$$

Since $X^{(1)}$ is an exact sample from $p(x|y, Z_n, \rho^2)$, (4.4) corresponds to the fourth line of Algorithm 2.

This establishes that SGS is equivalent to a noisy version of MYULA that relies on one sample from $p(x|y, z, \rho^2)$ to compute a stochastic estimate of the gradient $\nabla_z \log p(y|z, \rho^2)$ via (4.3). Using multiple samples from $p(x|y, Z_n; \rho)$ would improve the estimation of the expectation (4.3) and hence the behaviour of the algorithm. Alternatively, in the experiments considered in this paper $p(x|y, Z_n; \rho)$ is Gaussian, and hence this expectation can be calculated exactly. This is exploited in the MCMC methods proposed below.

4.2 Latent space MYULA

We established that SGS is equivalent to MYULA targeting the marginal of z with an inexact (i.e., stochastic) estimate of the gradient. Replacing this stochastic estimate with its exact value in Algorithm 2 produces the following recursion

$$\begin{aligned} X_{\text{grad}}^{i+1} &= \mathbb{E}_{x|y, Z_i, \rho^2}[x], \\ Z_{i+1} &= Z_i - \frac{\delta}{\lambda} \sum_{k=1}^{d'} [Z_i - \text{prox}_{\theta_k g_k}^\lambda(Z_i)] - \delta(Z_i - X_{\text{grad}}^{i+1})/\rho^2 + \sqrt{2\delta}\zeta_{i+1}, \end{aligned} \quad (4.5)$$

where $\zeta_{i+1} \sim \mathcal{N}(0, \mathbb{I}_d)$.

We now discuss how to use samples $\{Z_i\}_{i \geq 1}^m$ to compute expectations w.r.t. the marginal of interest $x|y, \theta, \rho^2$. More precisely, consider the computation of an expectation $\mathbb{E}_{x|y, \theta, \rho^2}[h(x)]$ for some function h w.r.t. the posterior distribution $p^\lambda(x|y, \theta, \rho^2)$ defined in (2.10) by using (4.5). Formally,

$$\mathbb{E}_{x|y, \theta, \rho^2}[h(x)] = \int_{\mathbb{R}^d} h(x) \int_{\mathbb{R}^d} p^\lambda(x, z|y, \theta, \rho^2) dz dx.$$

Using the fact that $p^\lambda(x, z|y, \theta, \rho^2) = p(x|y, z, \rho^2)p^\lambda(z|\theta)$ we have that

$$\begin{aligned} \mathbb{E}_{x|y, \theta, \rho^2}[h(x)] &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} h(x) p(x|y, z, \rho^2) p^\lambda(z|\theta) dz dx \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} h(x) p(x|y, z, \rho^2) dx p^\lambda(z|\theta) dz \\ &= \mathbb{E}_{z|\theta} [\mathbb{E}_{x|y, z, \rho^2}(h(x))]. \end{aligned} \quad (4.6)$$

In cases where $\mathbb{E}_{x|y, \theta, \rho^2}[h(x)]$ is available analytically, we suggest using a Rao-Blackwellised estimator of the form [45]

$$\mathbb{E}_{x|y, \theta, \rho^2}[h(x)] \approx \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{x|y, Z_i, \rho^2}[h(x)].$$

The computation of $\mathbb{E}_{x|y, Z_i, \rho^2}[h(x)]$ can be done as a postprocessing step, or alternatively within the iterations of the sampler. If $\mathbb{E}_{x|y, \theta, \rho^2}[h(x)]$ is not available analytically, we would draw samples from the conditional $x|y, Z_i, \theta, \rho^2$ and apply a standard Monte Carlo estimator.

We are now ready to present our first new MCMC method, summarised in Algorithm 4 below. We henceforth refer to this method as *latent space MYULA* (ls-MYULA), since it corresponds to MYULA applied to the marginal of the latent variable z .

Algorithm 4 ls-MYULA

- 1: Input: $X_0, Z_0 \in \mathbb{R}^d$, $\lambda, \rho > 0$, $m \in \mathbb{N}$.
 - 2: **for** $i = 0 : m - 1$ **do**
 - 3: Compute $X_{\text{grad}}^{i+1} = \mathbb{E}_{x|y, Z_i, \rho^2}[x]$,
 - 4: Compute $Z_{i+1} = Z_i - \frac{\delta}{\lambda} \sum_{k=1}^{d'} [Z_i - \text{prox}_{\theta_k g_k}^\lambda(Z_i)]/\lambda - \delta(Z_i - X_{\text{grad}}^{i+1})/\rho^2 + \sqrt{2\delta}\zeta_{i+1}$; where $\zeta_{i+1} \sim \mathcal{N}(0, \mathbb{I}_d)$,
 - 5: Compute $\hat{h}_{i+1} = \mathbb{E}_{x|y, Z_{i+1}, \rho^2}[h(x)]$,
 - 6: **end for**
 - 7: Output: an estimator of $\mathbb{E}_{x|y, \theta, \rho^2}[h(x)]$ given by $\{\sum_{k=1}^m \hat{h}_k\}/m$.
-

Remark 4.1. The underlying assumption in Algorithm 4 is that one can explicitly calculate $\mathbb{E}_{x|y, Z_i, \rho^2}[x]$ which, for example, is the case when the expectation represents the first moment of a Gaussian distribution, which corresponds to the likelihood models we consider in our experiments. In cases where $\mathbb{E}_{x|y, Z_i, \rho^2}[x]$ is intractable, we recommend to replace the expectation by its corresponding MCMC estimation, i.e.,

$$\mathbb{E}_{x|y, Z_i, \rho^2}[x] \approx \frac{1}{M} \sum_{i=1}^M X_i \text{ where } X_i \sim p(x|y, Z_i, \rho^2).$$

Remark 4.2. Notice that Algorithm 4 relies implicitly on two forms of smoothing, which operate differently and provide complementary benefits. On the one hand, through the MYULA construction, $\{g_k\}_{k=1}^d$ is replaced by the (smooth) Moreau-Yosida envelopes $\{g_k^\lambda\}_{k=1}^d$. On the other, the use of the auxiliary variable z introduces smoothing on the marginal prior $p(x|\rho^2) = \int p(x|z, \rho^2)p(z)dz$, where $p(x|z, \rho^2)$ acts as a Gaussian smoothing kernel. It can thus appear that there is some redundancy and that a single smoothing mechanism would suffice. However, Algorithm 4 operates on the latent space, and from that perspective, the smoothing related to $p(x|z, \rho^2)$ acts on the likelihood function of z given y , and not to the prior of z . This can lead to significant benefits in terms of convergence speed, as illustrated in Section 5. This effect can be analysed in detail in the case of the Gaussian likelihood function, where the smoothing introduced by $p(x|z, \rho^2)$ shifts the eigenvalues of the likelihood covariance matrix by ρ^2 . As a result, the likelihood of y w.r.t. z is by construction strongly log-concave. The same remark holds for the latent space SK-ROCK algorithm described below. Also note that the bias introduced by this additional smoothing is undone exactly when the samples are mapped from the latent space of z to the canonical space of x .

4.3 Latent space SK-ROCK

In the same way that an exact MYULA discretization is more beneficial than the stochastic MYULA discretization used in SGS, we can further improve results by using an exact SK-ROCK discretization which, as we described in Section 2.2.2, has many important advantages compared to MYULA. In particular, we present this method in Algorithm 5, and we will refer to it as *latent space SK-ROCK* (ls-SK-ROCK). The main difference between this algorithm and Algorithm 1 is that the conditional expectation $\mathbb{E}_{x|y, \tilde{Z}_j, \rho^2}[x]$ is computed on each internal stage s .

Algorithm 5 ls-SK-ROCK

- 1: Input: $X_0, Z_0 \in \mathbb{R}^d$, $\lambda, \rho > 0$, $m, s \in \mathbb{N}$, $\eta = 0.05$.
 - 2: Compute $l_s = (s - 0.5)^2(2 - 4/3\eta) - 1.5$,
 - 3: Compute $\omega_0 = 1 + \eta/s^2$, $\omega_1 = T_s(\omega_0)/T'_s(\omega_0)$,
 - 4: Compute $\mu_1 = \omega_1/\omega_0$, $\nu_1 = s\omega_1/2$, $k_1 = s\omega_1/\omega_0$,
 - 5: Choose $\delta \in (0, \delta_s^{\max})$, where $\delta_s^{\max} = l_s/(1/(\rho^2 + L_f^{-1}) + 1/\lambda)$,
 - 6: **for** $i = 0 : m - 1$ **do**
 - 7: Set $\tilde{X}_{\text{grad}}^0 = X_{\text{grad}}^i$, $\tilde{Z}_0 = Z_i$,
 - 8: Sample $\xi_{i+1} \sim \mathcal{N}(0, 2\delta\mathbb{I}_d)$,
 - 9: Compute $\tilde{X}_{\text{grad}}^1 = \mathbb{E}_{x|y, \tilde{Z}_0 + \nu_1\xi_{i+1}, \rho^2}[x]$,
 - 10: Compute $\Lambda(\tilde{Z}_0) = \sum_{k=1}^d [\tilde{Z}_0 + \nu_1\xi_{i+1} - \text{prox}_{\theta_k g_k}^\lambda(\tilde{Z}_0 + \nu_1\xi_{i+1})]/\lambda + (\tilde{Z}_0 + \nu_1\xi_{i+1} - \tilde{X}_{\text{grad}}^1)/\rho^2$,
 - 11: Compute $\tilde{Z}_1 = \tilde{Z}_0 - \mu_1\delta\Lambda(\tilde{Z}_0) + k_1^2\xi_{i+1}$,
 - 12: **for** $j = 2 : s$ **do**
 - 13: Compute $\mu_j = 2\omega_1 T_{j-1}(\omega_0)/T_j(\omega_0)$, $\nu_j = 2\omega_0 T_{j-1}(\omega_0)/T_j(\omega_0)$, $k_j = 1 - \nu_j$,
 - 14: Compute $\tilde{X}_{\text{grad}}^j = \mathbb{E}_{x|y, \tilde{Z}_{j-1}, \rho^2}[x]$,
 - 15: Compute $\Lambda(\tilde{Z}_{j-1}) = \sum_{k=1}^d [\tilde{Z}_{j-1} - \text{prox}_{\theta_k g_k}^\lambda(\tilde{Z}_{j-1})]/\lambda + (\tilde{Z}_{j-1} - \tilde{X}_{\text{grad}}^j)/\rho^2$,
 - 16: Compute $\tilde{Z}_j = -\mu_j\delta\Lambda(\tilde{Z}_{j-1}) + \nu_j\tilde{Z}_{j-1} + k_j\tilde{Z}_{j-2}$,
 - 17: **end for**
 - 18: Set $X_{\text{grad}}^{i+1} = \tilde{X}_{\text{grad}}^s$, $Z_{i+1} = \tilde{Z}_s$, $\hat{h}_{i+1} = \mathbb{E}_{x|y, Z_{i+1}, \rho^2}[h(x)]$,
 - 19: **end for**
 - 20: Output: an estimator of $\mathbb{E}_{x|y, \theta, \rho^2}[h(x)]$ given by $\{\sum_{k=1}^m \hat{h}_k\}/m$.
-

4.4 Implementation guidelines

Setting λ

As the priors of the experiments performed in this work are non-differentiable, we will use the Moreau-Yosida envelope defined in (2.6) with $\lambda \in [L_f^{-1}, 10L_f^{-1}]$. We chose $\lambda = L_f^{-1}$ in our numerical experiments, however, we have found numerically that values of $\lambda = 5L_f^{-1}$ or $\lambda = 10L_f^{-1}$ lead to faster convergence at the cost of additional bias.

Setting $\gamma_i^{(j)}$, γ'_i and n

With respect to Algorithm 3, it is suggested in [50] to set $\gamma_i^{(j)} = C_0^{(j)} i^{-p}$ and $\gamma'_i = C'_0 i^{-p}$ where $p \in [0.6, 0.9]$ (in the experiments performed in this paper, we have set $p = 0.8$), $C_0^{(j)}$ and C'_0 starting with $(\theta_0^{(j)} d)^{-1}$ and $(\rho_0^2 d)^{-1}$ respectively, and then readjusting it if necessary. With respect to n , we have followed the recommendation in [50] and used a single sample (i.e., $n = 1$) on each iteration (we did not observe significant difference for larger values of n). For more details about these parameters and their effect on the convergence properties of the algorithms please see [18, Section 4.2] and [17, Theorem 1].

Setting w_n

Following [50], we recommend setting w_n as follows

$$w_n = \begin{cases} 0 & \text{if } n < N_0, \\ 1 & \text{if } N_0 \leq n \leq N_1, \\ \gamma_n & \text{otherwise,} \end{cases}$$

where N_0 is the number of initial iterations to be discarded (when $n < N_0$, the values of ρ_n^2 and θ_n are still bouncing and not stabilized), $n \in [N_0, N_1]$ corresponds to the averaging estimation phase, in which the values of ρ_n^2 and θ_n have stabilized and start converging, and $n > N_1$ is known as the refinement phase where we use decreasing weights to enhance the accuracy of the estimator (see [50, Section 3.3.1] for details).

Setting an stopping criteria

It is recommended to supervise the evolution of $|\bar{\theta}_{m+1} - \bar{\theta}_m|/\bar{\theta}_m$ and $|\bar{\rho}_{m+1}^2 - \bar{\rho}_m^2|/\bar{\rho}_m^2$ in the execution of Algorithm 3 until they reach a tolerance level β to stop the algorithm execution. In our imaging experiments, we set $\beta = 10^{-4}$ but we have observed that $\beta = 10^{-3}$ is often enough to reach an acceptable estimate of the hyper-parameters in small computational times.

Other implementation considerations

In the implementation of the SAPG method, it is important to update the step-size of the MCMC method to sample X_i and Z_i within each iteration of the SAPG scheme, as the maximum step-size depends on the value of ρ_i^2 .

Regarding the Lipschitz constant L one needs to compute the step-size for MYULA and SK-ROCK algorithms, the model (2.5) has $L = \lambda^{-1} + L_f$. With respect to the augmented model (2.10), the Lipschitz constant is $L_a = \lambda^{-1} + (\rho^2 + L_f^{-1})^{-1}$ which we use to implement SGS, ls-MYULA and ls-SK-ROCK. Therefore, we set the step-size of the MCMC methods to $1/L$ for MYULA, to $1/L_a$ for SGS and ls-MYULA, and to δ_s^{\max} for SK-ROCK and ls-SK-ROCK, where δ_s^{\max} can be found in Algorithms 1 and 5, respectively. Regarding the SK-ROCK and ls-SK-ROCK algorithms, their most important parameters are the number of internal stages $s \in \mathbb{N}$ and the maximum step-size δ_s^{\max} . Note the increase in the step-size of Algorithm 5 compared to Algorithm 1, since $\delta_s^{\max} = l_s/(1/(\rho^2 + L_f^{-1}) + 1/\lambda) = l_s/L_a$ in ls-SK-ROCK, compared to $\delta_s^{\max} = l_s/(L_f + 1/\lambda) = l_s/L$ in SK-ROCK (with $l_s = [(s - 0.5)^2(2 - 4/3\eta) - 1.5]$ and $\eta = 0.05$ in both algorithms). This leads to a shift of ρ^2 in δ_s^{\max} that allows ls-SK-ROCK to converge faster. We have empirically observed that a good bias-variance trade-off is achieved by taking $\delta \in [\delta_{max}/2, \delta_{max})$ and $s \in \{5, \dots, 15\}$.

It is worth highlighting at this point that the improvement in Lipschitz constant experiences by the latent-space model allows ls-MYULA and ls-SK-ROCK to take larger step-sizes than MYULA and SK-ROCK, respectively. This leads to an improvement in convergence speed and therefore to higher computational efficiency, without noticeable additional bias.

5 Numerical experiments

We now illustrate the improvement that can be obtained by sampling the augmented model (2.7) using Algorithms 4 and 5 together with an optimal estimate of ρ^2 using Algorithm 3. To evaluate the performance of the methods in a variety of situations, we perform two imaging experiments related to *image deblurring* (whose model is strongly log-concave) and *image inpainting* (whose model is weakly log-concave). We implement these algorithms as described in the implementation guidelines (see Section 4.4).



Figure 2: Image deblurring experiments: Test images x and their corresponding noisy and blurred observations y .

For a fair comparison the results we show have been plotted as a function of the number of gradient evaluations, i.e., the number of times $\nabla \log p^\lambda(x|y, \theta)$ and $\nabla_z \log p^\lambda(z|y, \theta, \rho^2)$ are computed in our algorithms. The plots we show include the evolution of the MCMC samples in the burn-in stage using the scalar statistic $\log p(X_n|y, \theta)$ for MYULA and SK-ROCK, and $\log p(X_n^{\text{grad}}|y, \theta)$ for SGS, ls-MYULA and ls-SKROCK. We have also plotted the progression of the mean-squared error (MSE) between the posterior mean and the true image, when all the algorithms have reached stationarity, including the MAP estimate defined in (2.2) and computed using a highly efficient optimization algorithm called SALSA [5, 4] for the *image deblurring* and *image inpainting* experiments.

We also provide pixel-wise standard deviation plots as a way of quantifying the uncertainty in the delivered solution. We have also computed standard deviation plots performing downsampling by averaging the samples by a factor of $2 \times j$ where $j = \{1, 2, 4\}$, which allows us to observe the uncertainty in image structures at different scales. Finally, we also show autocorrelation plots of the slowest component of the samples produced by each of the methods, applying a 1-in- s thinning to the MYULA, SGS and ls-MYULA chains to equal the number of gradient evaluations between the mentioned methods (one gradient evaluation per iteration) and SK-ROCK/ls-SKROCK methods (s gradient evaluations per iteration). The chain's slowest component was identified by computing the approximated eigenvalue decomposition of the posterior covariance matrix and projecting the samples onto the leading eigenvector. Now using the slowest component, we have also computed effective sample sizes (ESS) of the five algorithms discussed in this paper, where the sum is truncated at lag k when the lag- k autocorrelation reaches a value less than 0.05.

For completeness, in Table 7 we have also provided computing times of all the experiments. These results have been obtained on an Intel core i5-8350U@1.70GHz workstation running MATLAB R2018a.

5.1 Image deblurring

To examine the performance of the MCMC methods in different scenarios, we consider a deblurring problem with two test images: *cameraman*, and the training image #61060 from the Berkeley Segmentation Dataset and Benchmark [35], we henceforth refer to this image as *skier*. Both images have a size of $d = 256 \times 256$ pixels. A uniform blur operator H of size 5×5 is applied to the true image $x \in \mathbb{R}^d$ and then additive Gaussian noise η is added with a sigma-to-noise level of 40dB, to produce an observation $y \in \mathbb{R}^d$ related to the true image by $y = Hx + \eta$. As H is nearly singular, the problem becomes ill-conditioned. So, to promote

regularity, we have used the isotropic TV pseudonorm as a prior, given by $\text{TV}(x) = \sum_i \sqrt{(\Delta_i^h x)^2 + (\Delta_i^v x)^2}$, where Δ_i^h, Δ_i^v denote horizontal and vertical first-order local difference operators. This leads to the following posterior distributions

$$p(x|y, \theta) \propto \exp[-\|y - Hx\|^2/2\sigma^2 - \theta\text{TV}(x)] \quad (5.1)$$

$$p(x, z|y, \theta, \rho^2) \propto \exp[-\|y - Hx\|^2/2\sigma^2 - \theta\text{TV}(z) - \|x - z\|^2/2\rho^2], \quad (5.2)$$

where $f_y(x) = \|y - Hx\|^2/2\sigma^2$ and $g(x) = \text{TV}(x)$. Figures 2(a), (b) show the test images for this experiment and Figures 2(c), (d) show the corresponding observations y for each image.

Table 1: Values for θ and ρ^2 estimated using Algorithm 3 for (5.1) and (5.2) in the image deblurring experiments, together with the corresponding Lipschitz constants L and L_a .

Experiment	θ	ρ^2	σ^2	$L = 1/\lambda + 1/\sigma^2$	$L_a = 1/\lambda + (\sigma^2 + \rho^2)^{-1}$
cameraman	0.044	0.480	0.335	5.959	4.205
skier	0.044	0.250	0.175	11.440	8.078

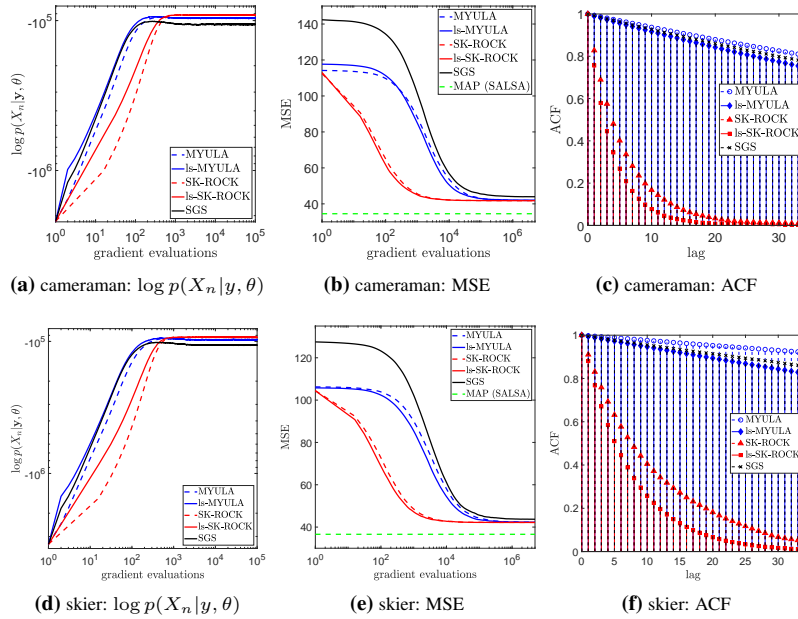


Figure 3: Image deblurring experiments: (a),(d) Convergence to the typical set of the posterior distribution (5.1) and (5.2) for the first 10^5 MYULA, SGS and ls-MYULA samples, and the first $10^5/s$ SK-ROCK and ls-SK-ROCK samples ($s = 15$). (b),(e) MSE between the mean of the algorithms and the true image, measured using 5×10^6 MYULA, SGS and ls-MYULA samples, and $5 \times 10^6/s$ SK-ROCK and ls-SK-ROCK samples ($s = 15$), in stationary regime. (c),(f) Autocorrelation function for the values of the slowest component of the samples.

We begin estimating optimal values for θ and ρ^2 for the given models implementing Algorithm 3 setting $\gamma_i = \gamma'_i = 10 \times i^{-0.8}/d$, $\theta_0 = 0.04$, $\rho_0^2 = L_f^{-1} = \sigma^2$ and $X_0 = Z_0 = H^T y$. The corresponding results for the parameters estimation are given in Table 1, together with the Lipschitz constants L and L_a required to sample (5.1) and (5.2) respectively. We then generate 5×10^6 samples using MYULA and $5 \times 10^6/s$ samples using SK-ROCK (with $s = 15$) from (5.1), and 5×10^6 samples using SGS and ls-MYULA and $5 \times 10^6/s$ samples using ls-SK-ROCK (with $s = 15$) from (5.2). The results of these experiments are plotted in 3. We note from the evolution of the MSE (when the chains have reached the typical set of the target distributions) that ls-MYULA and ls-SK-ROCK outperform SGS in terms of the convergence speed of the posterior mean. The improvement of ls-MYULA w.r.t. SGS illustrates the benefits of using an exact MYULA implementation rather than a noisy one, as shown in Section 4. The minor improvements between ls-MYULA and MYULA, and between SK-ROCK and ls-SK-ROCK, are due to the effect of $\rho^2 > 0$, which does not have a significant impact on the estimation of the posterior mean. The step-sizes used by each method are reported in Table 2.

Table 2: Image deblurring experiment: Summary of the values for the step-size δ for each of the MCMC methods applied to the two imaging experiments: cameraman and skier.

MCMC method	Cameraman	Skier
MYULA	0.167	0.087
SK-ROCK ($s = 15$)	67.959	35.402
SGS	0.237	0.124
ls-MYULA	0.237	0.124
ls-SK-ROCK ($s = 15$)	96.294	50.161

Table 3: Image deblurring experiments: Effective sample sizes of the slowest component, after generating 15×10^3 samples using the five algorithms discussed in this work, and the speed increase (i.e., speed-up) achieved by the algorithms w.r.t. MYULA.

	MYULA		SK-ROCK		SGS		ls-MYULA		ls-SK-ROCK	
	ESS	Speed-up	ESS	Speed-up	ESS	Speed-up	ESS	Speed-up	ESS	Speed-up
Cam.	46	-	1175	25.54	49	1.07	54	1.17	1604	34.87
Skier	18	-	636	35.33	35	1.94	37	2.06	924	51.33

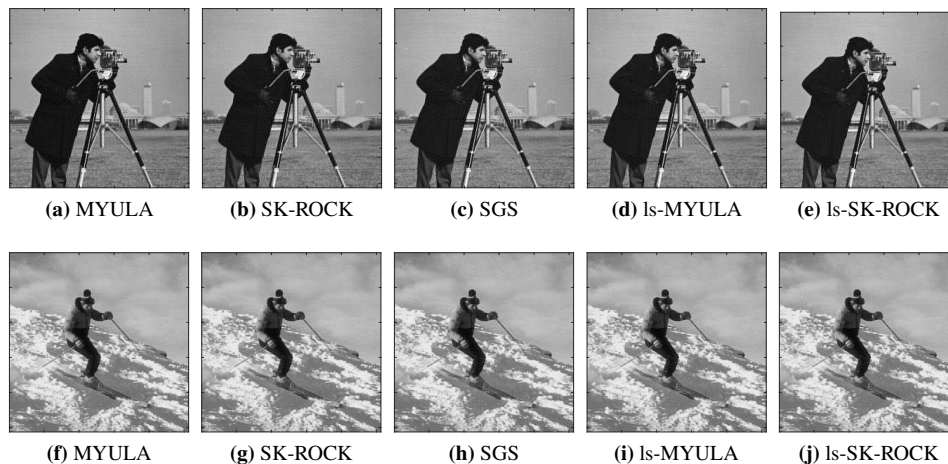


Figure 4: Image deblurring experiments: MMSE computed using 5×10^6 MYULA, SGS and ls-MYULA samples, and $5 \times 10^6/s$ SK-ROCK and ls-SK-ROCK samples, in stationarity.

We also plot the autocorrelation function of the slowest component from the chains of the MCMC algorithms, this is shown in Figure 3(c),(f) and, as can be seen, ls-SK-ROCK presents the fastest decay. In addition, Table 3 shows the effective sample sizes (ESS) associated with these autocorrelation plots, and one can notice that ls-SK-ROCK reaches the largest ESS. The comparison of the autocorrelation function and ESS for the slowest mixing component for MYULA and ls-MYULA illustrates the benefits of operating on the latent space (and the effect of $\rho^2 > 0$). A similar comparison between ls-MYULA and SGS illustrates the efficiency cost that SGS incurs due to the use of an inexact gradient. For completeness, we also illustrated in Figure 4 the minimum mean-square estimator (MMSE) of all the MCMC methods for all two deblurring experiments.

Finally, Figure 5 shows the marginal posterior standard deviation of the cameraman deblurring experiment at different scales. One can notice that edges show higher uncertainty, which is expected due to the nature of the forward operator. As can be seen, ls-MYULA and, in particular, ls-SK-ROCK outperform SGS in terms of delivering comparable estimates in less computational time, showing the benefit of using these algorithms to sample in a more efficient way the augmented posterior distribution.

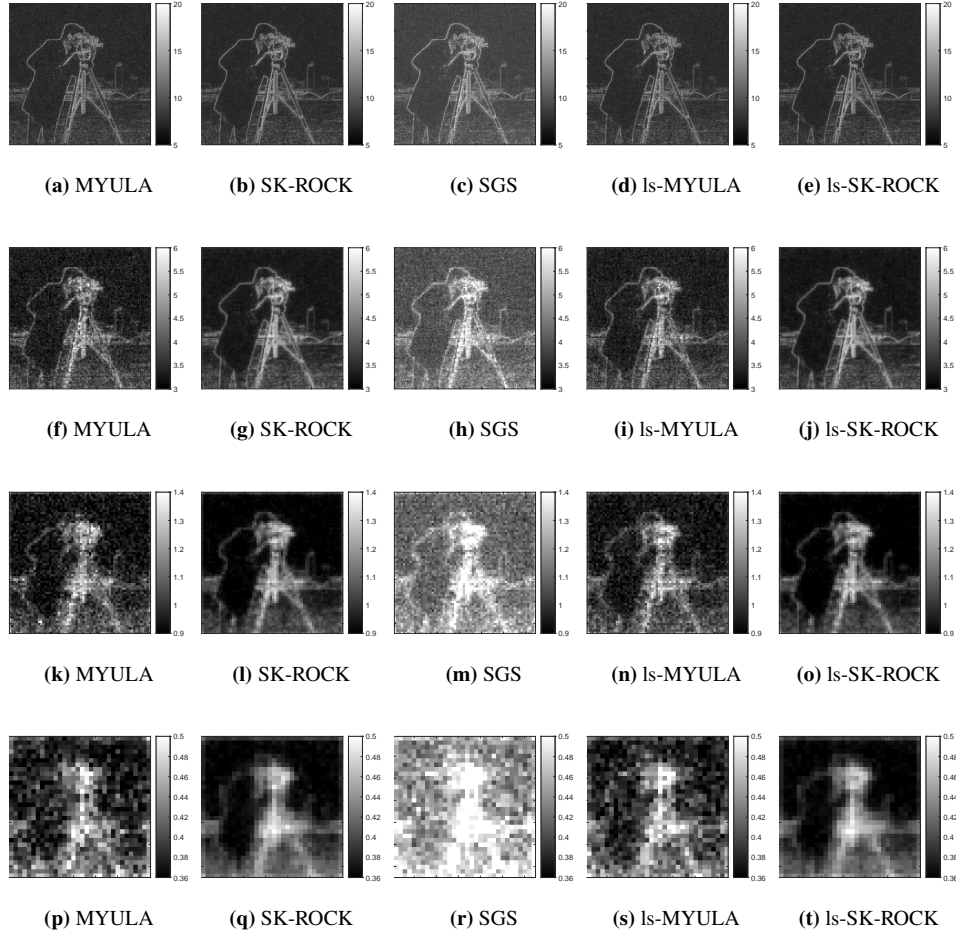


Figure 5: Image deblurring experiments - cameraman: pixel-wise standard deviation computed using 10^4 MYULA, SGS and ls-MYULA samples with a 1-in-15 thinning, and 10^4 SK-ROCK and ls-SK-ROCK samples with $s = 15$, using (a)-(e) the original sample size (256×256) and with downsampling by a factor of (f)-(j) 2, (k)-(o) 4 and (p)-(t) 8.

5.2 Image inpainting

We now perform an image inpainting experiment, which consists of randomly selecting 60% of the image pixels $x \in \mathbb{R}^d$ to form the observation vector $y \in \mathbb{R}^m$ ($m < d$) and then adding Gaussian noise with a SNR level of 40dB (the observation images y are illustrated in Figure 6). To test the MCMC methods in different regimes, we will use the same two test images given in Section 5.1 (cameraman and skier) and illustrated in Figure 2(a)-(b). For this experiment, we consider the following models

$$p(x|y, \theta) \propto \exp[-\|y - Ax\|^2/2\sigma^2 - \theta \text{TV}(x)] \quad (5.3)$$

$$p(x, z|y, \theta, \rho^2) \propto \exp[-\|y - Ax\|^2/2\sigma^2 - \theta \text{TV}(z) - \|x - z\|^2/2\rho^2], \quad (5.4)$$

where $f_y(x) = \|y - Ax\|^2/2\sigma^2$, $A \in \mathbb{R}^{m \times d}$ is a rectangular matrix obtained by taking a random subset of rows from the identity matrix in dimension d , and $g(x) = \text{TV}(x)$, previously defined in Section 5.1.

We first proceed to estimate optimal hyperparameters θ and ρ^2 for (5.3) and (5.4) using Algorithm 3 setting $\gamma_i = \gamma'_i = 10 \times i^{-0.8}/d$, $\theta_0 = 0.5$, $\rho_0^2 = L_f^{-1}/2 = \sigma^2/2$ and $X_0 = Z_0 = A^\top y$. The estimated parameter values can be seen in Table 4, together with the Lipschitz constants L and L_a required to sample (5.3) and (5.4) respectively.

Having obtained our estimates from SAGP algorithm for the values of the hyperparameters, we proceed to generate 5×10^6 MYULA samples and $5 \times 10^6/s$ SK-ROCK samples (with $s = 15$) from (5.3) and 5×10^6 SGS and ls-MYULA samples, and $5 \times 10^6/s$ ls-SK-ROCK samples (with $s = 15$) from (5.4). The step-sizes for each method are reported in Table 5.

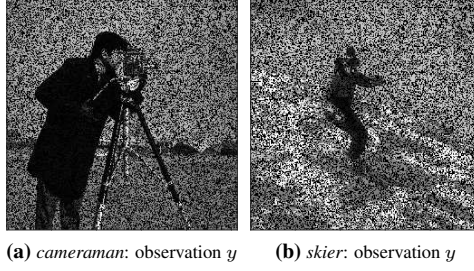


Figure 6: Image inpainting experiments: noisy and incomplete observations y (pixels in black represent unobserved components).

Table 4: Values for θ and ρ^2 estimated using Algorithm 3 for (5.3) and (5.4) in the image inpainting experiments

Experiment	θ	ρ^2	σ^2	$L = 1/\lambda + 1/\sigma^2$	$L_a = 1/\lambda + (\sigma^2 + \rho^2)^{-1}$
cameraman	0.058	0.65	0.388	5.146	3.530
skier	0.052	0.37	0.175	9.071	6.220

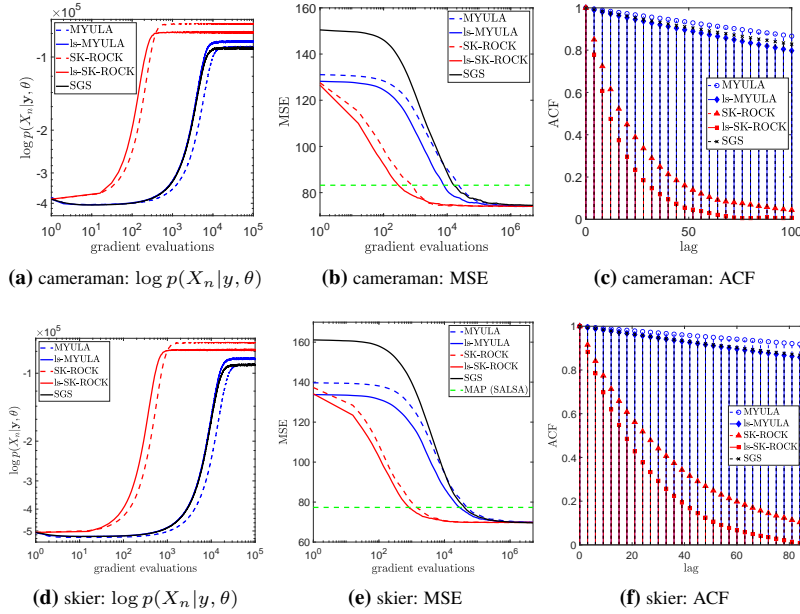


Figure 7: Image inpainting experiments: (a),(d) Convergence to the typical set of the posterior distribution (5.3) and (5.4) for the first 10^5 MYULA, SGS and ls-MYULA samples, and the first $10^5/s$ SK-ROCK and ls-SK-ROCK samples ($s = 15$). (b),(e) MSE between the mean of the algorithms and the true image x in stationarity and, as can be seen, ls-SK-ROCK is computationally efficient in being the fastest method to reach the MMSE in all two experiments, even outperforming the MAP estimate in terms of accuracy in all two experiments; moreover, the improvement of ls-MYULA over SGS, in terms of accuracy is evident similar to our previous results. (c),(f) Autocorrelation function for the slowest component of the samples.

With the generated samples, we proceed to plot the results of these experiments in Figure 7. We first notice the acceleration one can get from ls-SK-ROCK from the convergence to equilibrium of the MCMC samples in the burn-in stage represented by the evolution of the scalar estimate $\log p(X_n|y, \theta)$. Then, we illustrate the evolution of the mean-squared error (MSE) between the mean of the samples and the true image x in stationarity and, as can be seen, ls-SK-ROCK is computationally efficient in being the fastest method to reach the MMSE in all two experiments, even outperforming the MAP estimate in terms of accuracy in all two experiments; moreover, the improvement of ls-MYULA over SGS, in terms of accuracy is evident similar to our previous results.

We also plot the autocorrelation function of the pixel values for the slowest component in Figure 7(c),(f) and, as can be seen, the ACF of the ls-SK-ROCK samples decays faster than all the other MCMC methods. Again, a comparison of the autocorrelation

function for the slowest mixing component shows the benefits of operating on the latent space (and the effect of $\rho^2 > 0$), as well as the efficiency cost that SGS incurs because of the use of an inexact gradient in this case. For completeness, we also illustrate in Figure 8 the MMSE of all the MCMC methods for all two inpainting experiments and, as in previous numerical results, we can see in Figure 7(b),(e) that ls-SK-ROCK is the fastest method in compute this estimate.

Table 5: Image inpainting experiment: Summary of the values for the step-size δ for each of the MCMC methods applied to the two imaging experiments: cameraman and skier.

MCMC method	Cameraman	Skier
MYULA	0.194	0.110
SK-ROCK ($s = 15$)	78.698	44.646
SGS	0.283	0.161
ls-MYULA	0.283	0.161
ls-SK-ROCK ($s = 15$)	114.717	65.105

Finally, Figure 9 presents uncertainty quantification plots by showing pixel-wise standard deviation estimates for the cameraman inpainting experiment. In this case, the uncertainty is concentrated on the unobserved pixels, which is expected given the nature of the inpainting problem. One can notice that ls-MYULA and ls-SK-ROCK deliver comparable estimates in less computational time than SGS, showing the good performance of these algorithms in sampling the augmented posterior distribution.

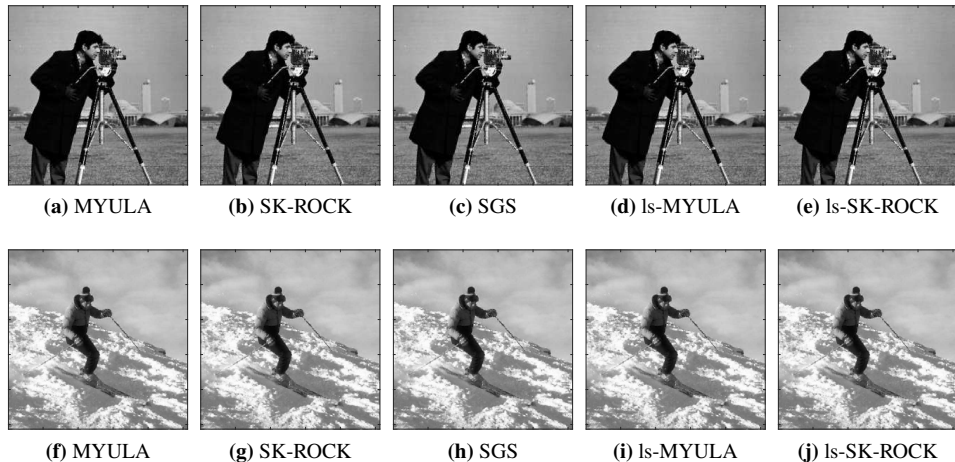


Figure 8: MMSE for the image inpainting experiment.

Table 6: Image inpainting experiments: Effective sample sizes of the slowest component, after generating 15×10^3 samples using the five algorithms discussed in this work, and the speed increase (i.e., speed-up) achieved by the algorithms w.r.t. MYULA.

	MYULA		SK-ROCK		SGS		ls-MYULA		ls-SK-ROCK	
	ESS	Speed-up	ESS	Speed-up	ESS	Speed-up	ESS	Speed-up	ESS	Speed-up
Cam.	14	-	281	20.07	14	1	21	1.5	448	32
Skier	8	-	212	26.5	11	1.38	15	1.88	320	40

5.3 Image deblurring with a total generalized variation prior

We conclude this section with an experiment related to image deblurring with a total generalised variation prior. The experiment setup is akin to Section 5.1, except that the prior is now given by

$$p(x|\theta^{(1)}, \theta^{(2)}) \propto \exp \left[-\text{TGV}_{\theta^{(1)}, \theta^{(2)}}^2(x) - \varepsilon \|x\|_2^2 \right], \quad (5.5)$$

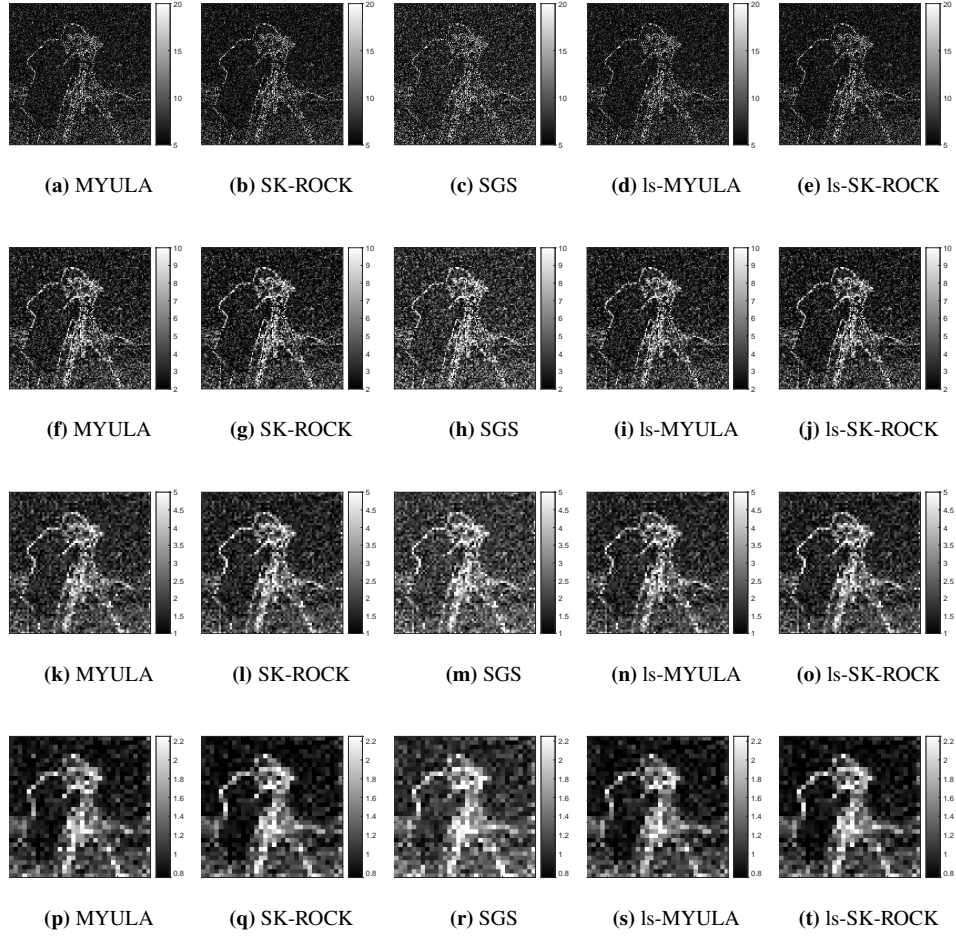


Figure 9: Image inpainting experiments - cameraman: pixel-wise standard deviation computed using 10^4 MYULA, SGS and ls-MYULA samples with a 1-in-15 thinning, and 10^4 SK-ROCK and ls-SK-ROCK samples with $s = 15$, using (a)-(e) the original sample size (256×256) and with downsampling by a factor of (f)-(j) 2, (k)-(o) 4 and (p)-(t) 8.

where $\varepsilon > 0$ and where $\text{TGV}_{\theta^{(1)}, \theta^{(2)}}^2(x)$ is the so-called total generalized variation (TGV) regulariser [11, 15], defined for every $\theta^{(1)}, \theta^{(2)} \in [0, +\infty)^2$, and $x \in \mathbb{R}^d$ by

$$\text{TGV}_{\theta^{(1)}, \theta^{(2)}}^2(x) = \min_{u \in \mathbb{R}^{2d}} \{ \theta^{(1)} \|u\|_{1,2} + \theta^{(2)} \|J(\Delta x - u)\|_{1, \text{Frob.}} \},$$

where $\Delta = (\Delta^v, \Delta^h)$ computes the first-order vertical and horizontal pixel differences, while the second-order information of the image-gradient vector field is computed by the Jacobian matrix J . The incorporation of second-order information removes the characteristic stair-casing artefacts commonly associated with (non-generalised) TV regularisation [11].

Image deblurring with a TGV prior is challenging because the results are highly sensitive to the choice of $\theta^{(1)}$ and $\theta^{(2)}$. However, these parameters are difficult to set a priori. Their direct estimation from y by maximum marginal likelihood estimation is also difficult because the TGV prior does not belong to the exponential family. [50, Section 4.4] proposes to address this difficulty by using a SAPG scheme to compute a pseudo-maximum marginal likelihood estimator for $\theta^{(1)}$ and $\theta^{(2)}$, which we also adopt in this paper (we refer the reader to [50, Section 4.4] for more details). Also note that the term $\varepsilon \|x\|_2^2$ is required so that $p(x|\theta^{(1)}, \theta^{(2)})$ defines a proper prior, in practice ε is very small (we use $\varepsilon = 10^{-10}$).

In a manner akin to Section 5.1, we consider a deblurring problem with the `skier` test images of size $d = 128 \times 128$ pixels, which is a patch of Image 2(b) (a reduced size image is considered in this experiment because the evaluation of the TGV proximal operator is highly computationally expensive). The observation y is generated by applying a uniform blur operator H of size 5×5

Table 7: Summary of the execution times (in seconds) to produce one sample (i.e., after one iteration) on each of the MCMC algorithms implemented for each experiment.

Imaging Experiment	MYULA	SK-ROCK (s=15)	ls-MYULA	ls-SK-ROCK (s=15)	SGS
deblurring	3.8×10^{-2}	6.1×10^{-1}	4.3×10^{-2}	6.1×10^{-1}	4.7×10^{-2}
inpainting	4.1×10^{-2}	5.8×10^{-1}	3.8×10^{-2}	5.6×10^{-1}	4.7×10^{-2}

to the true image $x \in \mathbb{R}^d$, followed by additive Gaussian noise with a sigma-to-noise level of 30dB. This experiment considers the following posterior distributions

$$p(x|y, \theta^{(1)}, \theta^{(2)}) \propto \exp \left[-\|y - Hx\|^2 / 2\sigma^2 - p(x|\theta^{(1)}, \theta^{(2)}) \right] \quad (5.6)$$

$$p(x, z|y, \theta^{(1)}, \theta^{(2)}, \rho^2) \propto \exp \left[-\|y - Hx\|^2 / 2\sigma^2 - p(z|\theta^{(1)}, \theta^{(2)}) - \|x - z\|^2 / 2\rho^2 \right], \quad (5.7)$$

where $p(x|\theta^{(1)}, \theta^{(2)})$ is defined in (5.5), and $p(z|\theta^{(1)}, \theta^{(2)})$ is the same TGV prior applied to the latent variable z instead of x . Figure 10(a) shows the test images for this experiment and Figure 10(b) shows the corresponding observations y .

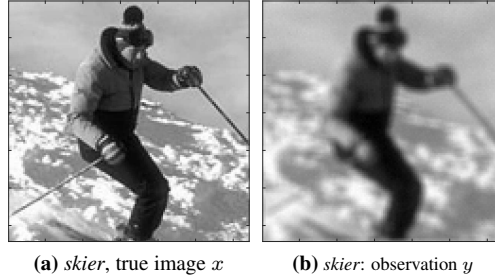


Figure 10: Image deblurring experiment with TGV prior: Test image x (patch from Image 2(b) of size 128×128) and its corresponding noisy and blurred observation y .

We begin estimating optimal values for $\theta^{(1)}, \theta^{(2)}$ and ρ^2 for the given models implementing a variant of the Algorithm 3 for inhomogeneous regularizers (See [50, Section 3.2.3] for details), setting $\gamma_i^{(1)} = 100 \times i^{-0.8}/d$, and $\gamma_i^{(2)} = \gamma'_i = i^{-0.8}/d$, $\theta_0^{(1)} = \theta_0^{(2)} = 5$, $\rho_0^2 = L_f^{-1} = \sigma^2$ and $X_0 = Z_0 = H^\top y$. The corresponding results for the estimated parameters are given in Table 8, together with the Lipschitz constants L and L_a required to sample (5.6) and (5.7) respectively. We then generate 1.5×10^5 samples using MYULA and $1.5 \times 10^5/s$ samples using SK-ROCK (with $s = 15$ and $\delta = \delta_s^{\max}/2$) from (5.6), and 1.5×10^5 samples using SGS and ls-MYULA and $1.5 \times 10^5/s$ samples using ls-SK-ROCK (with $s = 15$ and $\delta = \delta_s^{\max}/2$) from (5.7). The results of these experiments are plotted in Figure 11. In particular, we note from the evolution of the MSE (when the chains have reached the typical set of the target distributions) that ls-MYULA and ls-SK-ROCK outperform SGS, as we are using an exact MYULA implementation rather than a noisy one, as shown in Section 4. We have also reported the step-size used by each method in Table 9.

Table 8: Values for $\theta^{(1)}, \theta^{(2)}$ and ρ^2 estimated using Algorithm 3 for (5.6) and (5.7) in the image deblurring experiment with TGV, together with the corresponding Lipschitz constants L and L_a .

$\theta^{(1)}$	$\theta^{(2)}$	ρ^2	σ^2	$L = 1/\lambda + 1/\sigma^2$	$L_a = 1/\lambda + (\sigma^2 + \rho^2)^{-1}$
4.46	7.38	7.15×10^{-5}	5.04×10^{-5}	3.97×10^4	2.81×10^4

We also plot the autocorrelation function of the slowest component from the chains of the MCMC algorithms, this is shown in Figure 11(c) and, as can be seen, ls-SK-ROCK presents the fastest decay. In addition, we also illustrated in Figure 12 the minimum mean-square estimator (MMSE) of all the MCMC methods for this experiment.

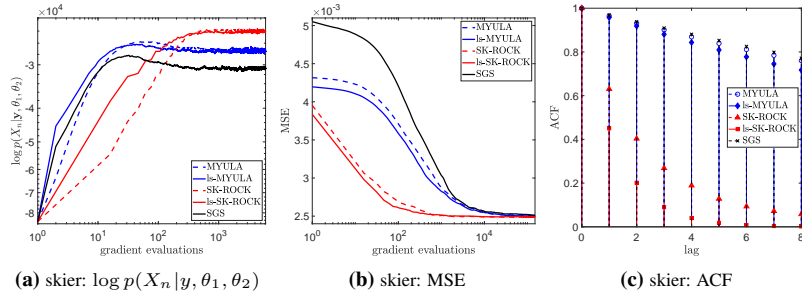


Figure 11: Image deblurring experiment with TGV prior: (a) Convergence to the typical set of the posterior distribution (5.6) and (5.7) for the first 5×10^3 MYULA, SGS and ls-MYULA samples, and the first $5 \times 10^3/s$ SK-ROCK and ls-SK-ROCK samples ($s = 15$). (b) MSE between the mean of the algorithms and the true image, measured using 1.5×10^5 MYULA, SGS and ls-MYULA samples, and $1.5 \times 10^5/s$ SK-ROCK and ls-SK-ROCK samples ($s = 15$), in stationary regime. (c) Autocorrelation function for the values of the slowest component of the samples.

Table 9: Image deblurring experiment with TGV: Summary of the values for the step-size δ for each of the MCMC methods applied to this experiment.

MCMC method	δ
MYULA	2.51×10^{-5}
SK-ROCK ($s = 15$)	5.10×10^{-3}
SGS	3.56×10^{-5}
ls-MYULA	3.56×10^{-5}
ls-SK-ROCK ($s = 15$)	7.21×10^{-3}

As can be seen, ls-MYULA and, in particular, ls-SK-ROCK outperform SGS in terms of delivering comparable estimates in less computational time, showing the benefit of using these algorithms to sample in a more efficient way the augmented posterior distribution.

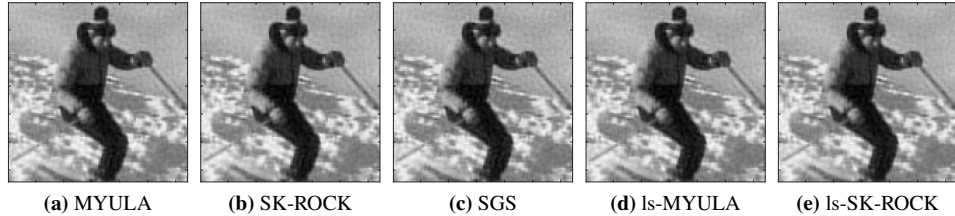


Figure 12: MMSE for the image deblurring experiment with TGV prior.

6 Conclusions

We presented a strategy to combine MYULA or proximal SK-ROCK with augmentation and relaxation in the manner of SGS. This was achieved by first establishing that SGS is equivalent to a noisy ULA scheme applied to the marginal distribution of the latent variable z in an augmented Bayesian model $x, z | y, \theta, \rho^2$. This then naturally led to two new samplers that apply MYULA and SK-ROCK to the latent marginal distribution $z | y, \theta, \rho^2$. Probabilities and expectations w.r.t. the primal marginal $x | y, \theta, \rho^2$ are then straightforwardly computed by using a Rao-Blackwellised Monte Carlo estimator. Moreover, we also observed empirically that there is a range of values for ρ^2 for which convergence speed and model quality improve (in the sense of the model evidence). Increasing ρ^2 beyond this range leads to improvements in convergence speed at the expense of significant estimation bias. We therefore proposed to adopt an empirical Bayesian approach and set ρ^2 , together with the regularisation parameter θ , by maximum marginal likelihood estimation from y . This was achieved by using an SAPG scheme that converges in very few iterations. We

illustrated the benefits from adopting the proposed methodology with two experiments, image deblurring and image inpainting. The results showed that the new proximal SK-ROCK algorithm that benefits from augmentation and relaxation outperforms the other methods from the state of the art in terms of computational efficiency. Future work will focus on extending the proposed approach to plug-and-play priors encoded by neural network denoisers [30], and on establishing non-asymptotic convergence results for the methods. Another perspective for future work is to compare the proposed MCMC methods with deterministic Bayesian computation strategies based on approximate message passing and expectation propagation algorithms [41], and to explore ways in which the methods presented in this paper could be used to compute message passing or expectation propagation update steps within these schemes.

References

- [1] A. Abdulle. “Explicit Stabilized Runge–Kutta Methods”. In: *Encyclopedia of Applied and Computational Mathematics*. Berlin, Heidelberg: Springer, 2015, pp. 460–468. DOI: 10.1007/978-3-540-70529-1_100.
- [2] A. Abdulle, I. Almuslimani, and G. Vilmart. “Optimal Explicit Stabilized Integrator of Weak Order 1 for Stiff and Ergodic Stochastic Differential Equations”. In: *SIAM/ASA Journal on Uncertainty Quantification* 6.2 (2018), pp. 937–964. DOI: 10.1137/17M1145859.
- [3] J. Adler and O. Öktem. “Learned Primal-Dual Reconstruction”. In: *IEEE Transactions on Medical Imaging* 37.6 (2018), pp. 1322–1332. DOI: 10.1109/TMI.2018.2799231.
- [4] 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”. In: *IEEE Transactions on Image Processing* 20.3 (2011), pp. 681–695. DOI: 10.1109/TIP.2010.2076294.
- [5] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo. “Fast Image Recovery Using Variable Splitting and Constrained Optimization”. In: *IEEE Transactions on Image Processing* 19.9 (2010), pp. 2345–2356. DOI: 10.1109/TIP.2010.2047910.
- [6] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. “An Introduction to MCMC for Machine Learning”. In: *Machine Learning* 50 (2003), pp. 5–43. DOI: 10.1023/A:1020281327116.
- [7] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb. “Solving inverse problems using data-driven models”. In: *Acta Numerica* 28 (May 2019), pp. 1–174. DOI: 10.1017/S0962492919000059.
- [8] S. D. Babacan, R. Molina, and A. K. Katsaggelos. “Bayesian Compressive Sensing Using Laplace Priors”. In: *IEEE Transactions on Image Processing* 19.1 (2010), pp. 53–63. DOI: 10.1109/TIP.2009.2032894.
- [9] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. “Variational Inference: A Review for Statisticians”. In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877. DOI: 10.1080/01621459.2017.1285773.
- [10] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers”. In: *Foundations and Trends in Machine Learning* 3.1 (Jan. 2011), pp. 1–122. DOI: 10.1561/2200000016.
- [11] K. Bredies, K. Kunisch, and T. Pock. “Total Generalized Variation”. In: *SIAM Journal on Imaging Sciences* 3.3 (Sept. 2010), pp. 492–526. DOI: 10.1137/090769521.
- [12] X. Cai, J. D. McEwen, and M. Pereyra. *Proximal nested sampling for high-dimensional Bayesian model selection*. 2021. DOI: 10.48550/ARXIV.2106.03646.
- [13] X. Cai, M. Pereyra, and J. D. McEwen. “Uncertainty quantification for radio interferometric imaging – I. Proximal MCMC methods”. In: *Monthly Notices of the Royal Astronomical Society* 480.3 (2018), pp. 4154–4169. DOI: 10.1093/mnras/sty2004.
- [14] X. Cai, M. Pereyra, and J. D. McEwen. “Uncertainty quantification for radio interferometric imaging: II. MAP estimation”. In: *Monthly Notices of the Royal Astronomical Society* 480.3 (2018), pp. 4170–4182. DOI: 10.1093/mnras/sty2015.
- [15] A. Chambolle and P.-L. Lions. “Image recovery via total variation minimization and related problems”. In: *Numerische Mathematik* 76.2 (1997), pp. 167–188. DOI: 10.1007/s002110050258.

- [16] A. Chambolle and T. Pock. “An introduction to continuous optimization for imaging”. In: *Acta Numerica* 25 (May 2016), pp. 161–319. DOI: 10.1017/S096249291600009X.
- [17] V. De Bortoli, A. Durmus, M. Pereyra, and A. F. Vidal. “Efficient stochastic optimisation by unadjusted Langevin Monte Carlo”. In: *Statistics and Computing* 31.3 (2021). DOI: 10.1007/s11222-020-09986-y.
- [18] V. De Bortoli, A. Durmus, M. Pereyra, and A. F. Vidal. “Maximum Likelihood Estimation of Regularization Parameters in High-Dimensional Inverse Problems: An Empirical Bayesian Approach. Part II: Theoretical Analysis”. In: *SIAM Journal on Imaging Sciences* 13.4 (2020), pp. 1990–2028. DOI: 10.1137/20M1339842.
- [19] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum Likelihood from Incomplete Data Via the EM Algorithm”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22. DOI: 10.1111/j.2517-6161.1977.tb01600.x.
- [20] W. Dong, L. Zhang, G. Shi, and X. Wu. “Image Deblurring and Super-Resolution by Adaptive Sparse Domain Selection and Adaptive Regularization”. In: *IEEE Transactions on Image Processing* 20.7 (2011), pp. 1838–1857. DOI: 10.1109/TIP.2011.2108306.
- [21] R. Douc, É. Moulines, and D. Stoffer. *Nonlinear Time Series: Theory, Methods and Applications with R Examples*. Texts in statistical science. Boca Raton, FL: Chapman & Hall/CRC, 2014.
- [22] A. Durmus and É. Moulines. “High-dimensional Bayesian inference via the unadjusted Langevin algorithm”. In: *Bernoulli* 25 (Sept. 2019), pp. 2854–2882. DOI: 10.3150/18-BEJ1073.
- [23] A. Durmus and É. Moulines. “Nonasymptotic convergence analysis for the unadjusted Langevin algorithm”. In: *Ann. Appl. Probab.* 27.3 (July 2017), pp. 1551–1587. DOI: 10.1214/16-AAP1238.
- [24] A. Durmus, É. Moulines, and M. Pereyra. “Efficient Bayesian Computation by Proximal Markov Chain Monte Carlo: When Langevin Meets Moreau”. In: *SIAM Journal on Imaging Sciences* 11.1 (2018), pp. 473–506. DOI: 10.1137/16M1108340.
- [25] D. A. van Dyk and X.-L. Meng. “The Art of Data Augmentation”. In: *Journal of Computational and Graphical Statistics* 10.1 (2001), pp. 1–50. DOI: 10.1198/10618600152418584.
- [26] G. Fort, E. Ollier, and A. Samson. “Stochastic proximal-gradient algorithms for penalized mixed models”. In: *Statistics and Computing* 29 (Mar. 2019), pp. 231–253. DOI: 10.1007/s11222-018-9805-7.
- [27] C. Gilavert, S. Moussaoui, and J. Idier. “Efficient Gaussian Sampling for Solving Large-Scale Inverse Problems Using MCMC”. In: *IEEE Transactions on Signal Processing* 63.1 (2015), pp. 70–80. DOI: 10.1109/TSP.2014.2367457.
- [28] J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*. Vol. 160. Applied Mathematical Sciences. Springer-Verlag, 2006.
- [29] L. F. Lang, S. Neumayer, O. Öktem, and C.-B. Schönlieb. “Template-Based Image Reconstruction from Sparse Tomographic Data”. In: *Applied Mathematics & Optimization* 82.3 (2020), pp. 1081–1109. DOI: 10.1007/s00245-019-095
- [30] R. Laumont, V. D. Bortoli, A. Almansa, J. Delon, A. Durmus, and M. Pereyra. “Bayesian Imaging Using Plug & Play Priors: When Langevin Meets Tweedie”. In: *SIAM Journal on Imaging Sciences* 15.2 (2022), pp. 701–737. DOI: 10.1137/21M1406349.
- [31] M. Lebrun, A. Buades, and J. M. Morel. “A Nonlocal Bayesian Image Denoising Algorithm”. In: *SIAM Journal on Imaging Sciences* 6.3 (2013), pp. 1665–1688. DOI: 10.1137/120874989.
- [32] E. Levitin and B. Polyak. “Constrained minimization methods”. In: *USSR Computational Mathematics and Mathematical Physics* 6.5 (1966), pp. 1–50. DOI: 10.1016/0041-5553(66)90114-5.
- [33] C. Louchet and L. Moisan. “Posterior Expectation of the Total Variation Model: Properties and Experiments”. In: *SIAM Journal on Imaging Sciences* 6.4 (2013), pp. 2640–2684. DOI: 10.1137/120902276.
- [34] C. Louchet and L. Moisan. “Total Variation denoising using iterated conditional expectation”. In: *2014 22nd European Signal Processing Conference (EUSIPCO)*. 2014, pp. 1592–1596.
- [35] D. Martin, C. Fowlkes, D. Tal, and J. Malik. “A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics”. In: *Proc. 8th Int’l Conf. Computer Vision*. Vol. 2. July 2001, pp. 416–423.

- [36] T. P. Minka. “Expectation Propagation for Approximate Bayesian Inference”. In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. UAI’01. Seattle, Washington: Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.
- [37] M. Pereyra, N. Dobigeon, H. Batatia, and J. Tourneret. “Segmentation of Skin Lesions in 2-D and 3-D Ultrasound Images Using a Spatially Coherent Generalized Rayleigh Mixture Model”. In: *IEEE Transactions on Medical Imaging* 31.8 (2012), pp. 1509–1520. DOI: 10.1109/TMI.2012.2190617.
- [38] M. Pereyra. “Maximum-a-Posteriori Estimation with Bayesian Confidence Regions”. In: *SIAM Journal on Imaging Sciences* 10.1 (2017), pp. 285–302. DOI: 10.1137/16M1071249.
- [39] M. Pereyra. “Proximal Markov chain Monte Carlo algorithms”. In: *Statistics and Computing* 26.4 (2016), pp. 745–760. DOI: 10.1007/s11222-015-9567-4.
- [40] M. Pereyra. “Revisiting Maximum-A-Posteriori Estimation in Log-Concave Models”. In: *SIAM Journal on Imaging Sciences* 12.1 (Mar. 2019), pp. 650–670. DOI: 10.1137/18M1174076.
- [41] M. Pereyra, P. Schniter, É. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A. O. Hero, and S. McLaughlin. “A Survey of Stochastic Simulation and Optimization Methods in Signal Processing”. In: *IEEE Journal of Selected Topics in Signal Processing* 10.2 (2016), pp. 224–241. DOI: 10.1109/JSTSP.2015.2496908.
- [42] M. Pereyra, L. Vargas, and K. C. Zygalakis. “Accelerating Proximal Markov Chain Monte Carlo by Using an Explicit Stabilized Method”. In: *SIAM Journal on Imaging Sciences* 13.2 (2020), pp. 905–935. DOI: 10.1137/19M1283719.
- [43] L. J. Rendell, A. M. Johansen, A. Lee, and N. Whiteley. “Global Consensus Monte Carlo”. In: *Journal of Computational and Graphical Statistics* (2020), pp. 1–11. DOI: 10.1080/10618600.2020.1811105.
- [44] A. Repetti, M. Pereyra, and Y. Wiaux. “Scalable Bayesian Uncertainty Quantification in Imaging Inverse Problems via Convex Optimization”. In: *SIAM Journal on Imaging Sciences* 12.1 (Jan. 2019), pp. 87–118. DOI: 10.1137/18M1173629.
- [45] C. Robert and G. Casella. *Monte Carlo statistical methods*. 2nd ed. Springer-Verlag, 2004.
- [46] C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York, NY: Springer, 2007. DOI: 10.1007/0-387-71599-1_5.
- [47] G. O. Roberts and R. L. Tweedie. “Exponential convergence of Langevin distributions and their discrete approximations”. In: *Bernoulli* 2.4 (Dec. 1996), pp. 341–363. DOI: 10.2307/3318418.
- [48] O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, and F. Lenzen. *Variational Methods in Imaging*. New York, NY: Springer, 2009. DOI: doi:10.1007/978-0-387-69277-7.
- [49] M. A. Tanner and W. H. Wong. “The Calculation of Posterior Distributions by Data Augmentation”. In: *Journal of the American Statistical Association* 82.398 (1987), pp. 528–540. DOI: 10.1080/01621459.1987.10478458.
- [50] A. F. Vidal, V. D. Bortoli, M. Pereyra, and A. Durmus. “Maximum Likelihood Estimation of Regularization Parameters in High-Dimensional Inverse Problems: An Empirical Bayesian Approach Part I: Methodology and Experiments”. In: *SIAM Journal on Imaging Sciences* 13.4 (2020), pp. 1945–1989. DOI: 10.1137/20M1339829.
- [51] M. Vono, N. Dobigeon, and P. Chainais. “Split-and-Augmented Gibbs Sampler—Application to Large-Scale Inference Problems”. In: *IEEE Transactions on Signal Processing* 67.6 (2019), pp. 1648–1661.
- [52] M. Vono, N. Dobigeon, and P. Chainais. “High-Dimensional Gaussian Sampling: A Review and a Unifying Approach Based on a Stochastic Proximal Point Algorithm”. In: *SIAM Review* 64.1 (2022), pp. 3–56. DOI: 10.1137/20M1371026.
- [53] M. Vono, D. Paulin, and A. Doucet. “Efficient MCMC Sampling with Dimension-Free Convergence Rate using ADMM-type Splitting”. In: *Journal of Machine Learning Research* 23.25 (2022), pp. 1–69. URL: <http://jmlr.org/papers>
- [54] H. White. “Maximum Likelihood Estimation of Misspecified Models”. In: *Econometrica* 50.1 (1982), pp. 1–25. DOI: 10.2307/1912526.

- [55] J. Xie, L. Xu, and E. Chen. “Image Denoising and Inpainting with Deep Neural Networks”. In: *Advances in Neural Information Processing Systems* 25. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Vol. 25. Curran Associates, Inc., 2012, pp. 341–349.
- [56] D. Yao, S. McLaughlin, and Y. Altmann. “Patch-Based Image Restoration Using Expectation Propagation”. In: *SIAM Journal on Imaging Sciences* 15.1 (2022), pp. 192–227. DOI: 10.1137/21M1427541.