A TV-Gaussian prior for infinite-dimensional Bayesian inverse problems and its numerical implementations

A TV-Gaussian prior for infinite-dimensional Bayesian inverse problems and its numerical implementations

Zhewei Yao[1]**, Zixi Hu**[1] **and Jinglai Li**[2,3]

[1] Department of Mathematics and Zhiyuan College, Shanghai Jiao Tong University, Shanghai 200240, People's Republic of China
[2] Institute of Natural Sciences, Department of Mathematics and MOE Key Laboratory of Scientific and Engineering Computing, Shanghai Jiao Tong University, Shanghai 200240, People's Republic of China

E-mail: zyaosjtu@gmail.com, hzx@sjtu.edu.cn and jinglaili@sjtu.edu.cn

CrossMark

Abstract

Many scientific and engineering problems require to perform Bayesian inferences in function spaces, where the unknowns are of infinite dimension. In such problems, choosing an appropriate prior distribution is an important task. In particular, when the function to infer is subject to sharp jumps, the commonly used Gaussian measures become unsuitable. On the other hand, the so-called total variation (TV) prior can only be defined in a finite-dimensional setting, and does not lead to a well-defined posterior measure in function spaces. In this work we present a TV-Gaussian (TG) prior to address such problems, where the TV term is used to detect sharp jumps of the function, and the Gaussian distribution is used as a reference measure so that it results in a well-defined posterior measure in the function space. We also present an efficient Markov Chain Monte Carlo (MCMC) algorithm to draw samples from the posterior distribution of the TG prior. With numerical examples we demonstrate the performance of the TG prior and the efficiency of the proposed MCMC algorithm.

---

[3] Author to whom any correspondence should be addressed.

1. Introduction

The Bayesian inference methods [12] for solving inverse problems have gained increasing popularity, largely due to their ability to quantify uncertainties in the estimation results. The Bayesian inverse problems have been extensively studied in the finite-dimensional setting [1, 15], and more recently, a rigorous Bayesian framework [24] is developed for the inverse problems in function spaces where the unknowns are of infinite dimension. In existing works that perform Bayesian inferences in function spaces, Gaussian measures are widely used as the prior distributions. Such a choice is in fact well justified as theoretical studies suggest that the Gaussian measures are well-behaved priors: the resulting posterior depends continuously on the data and it can be well approximated by finite-dimensional representations. However, in many practical problems such as image reconstructions, the true functions that one aims to infer are often subject to sharp jumps or even discontinues, which cannot be well modeled by Gaussian priors. In the deterministic inverse problem context, such functions are often estimated using the total variation (TV) regularization [23]. To this end, it is very intriguing to perform Bayesian inferences of the functions with sharp jumps, with a TV prior, the construction of which is rather straightforward in the finite-dimensional setting. Thus to use the prior in the infinite-dimensional setting, one first represents the unknown function with a finite-dimensional parametrization, for example, by discretizing the function on a pre-determined mesh grid, and then solve the resulting finite-dimensional inference problem with the TV prior. Such a method has been successfully applied to a variety of problems [2, 25]. A major issue of the TV prior is that, unlike the finite-dimensional Gaussian distribution, the posterior distribution of the TV prior may not converge to a well-defined infinite-dimensional measure as the discretization dimension increases, a property that is referred to as being discretization variant in [18]. In other word, the inference results depend on the discretization dimensions, which is certainly undesirable from both practical and theoretical points of view. A number of non-Gaussian priors therefore have been proposed to address the issue. For example, the work [17] proposes to use the Besov priors based on wavelet expansions in such problems and the theoretical properties of which are further investigated in [9]. In another work [13], a hierarchical Gaussian prior related to the Mumford–Shah regularization in the deterministic setting is developed.

We note that the priors mentioned above differ significantly from the Gaussian measures, in both theoretical properties and numerical implementations. As a result, many analysis techniques and numerical methods developed for the Gaussian priors cannot be easily extended to these non-Gaussian priors. In particular, as will be discussed later, some Markov Chain Monte Carlo (MCMC) algorithms developed for Gaussian priors may not be applied directly to, for example, the Besov ones. To this end, efforts have been made to developing particular MCMC algorithms for those non-Gaussian priors, e.g. [26]. As an alternative solution, in this work we propose a TV-Gaussian (TG) prior which is motivated by the elastic net regularization for linear regression problems [28], and the hybrid regularization method in image problems [4]. Namely, the prior includes a TV term to detect edges, and on the other hand, it uses Gaussian distributions as a reference measure, so that it leads to a well-defined posterior distribution in the function space.

MCMC simulations are widely used to draw samples from the posterior distribution in Bayesian inferences. It has been known that many standard MCMC algorithms, such as the random walk Metropolis–Hastings, can become arbitrarily slow as the discretization mesh of the unknown is refined [5, 22]. Namely the mixing time of an algorithm can increase to infinity as the dimension of the discretized parameter approaches to infinity, and in this case the algorithm is said to be *dimension-dependent*. A family of dimension-independent MCMC algorithms were presented in [5] by constructing a preconditioned Crank–Nicolson (pCN) discretization of a stochastic partial differential equation that preserves the reference measure. There are a number

of other algorithms for the infinite-dimensional problems that can further improve the sampling efficiency by incorporating the data information: the stochastic Newton MCMC [21], the dimension-independent likelihood-informed MCMC [6], and the adaptive independence sampler algorithm developed in [11], just to name a few. All the aforementioned algorithms are developed based on Gaussian priors, but thanks to the Gaussian reference measure, many of these algorithms, most notably the pCN method, can be directly applied to problems with our TG priors (some gradient based algorithms may require certain modifications). Moreover, by taking advantage of the special structure of the TG prior, we propose a simple splitting scheme to accelerate the pCN algorithm. Loosely speaking, the splitting scheme has two stages: in the first stage the sample is moved (accepted/rejected) several times only according to the TV term, and in the second stage it is finally accepted or rejected according to the likelihood function. We prove the detailed balance of the proposed splitting pCN (S-pCN) scheme and with numerical examples we demonstrate that the method can significantly improve the mixing rate by making very simple modifications to the standard pCN method.

To summarize, the main contributions of the work are two-fold: we propose a TV-Gaussian hybrid prior to handle unknowns that cannot be well modeled by standard Gaussian measures, and we also provide an efficient MCMC algorithm specifically designed for the proposed prior. The rest of the work is organized as the following. In section 2, we present the TG priors and provide some results regarding the theoretical properties of it. In section 3, we describe the S-pCN algorithm to efficiently sample from the proposed TG prior. Numerical examples are provided in section 4 to demonstrate the performance of the TG prior and the efficiency of the S-pCN algorithm. Finally section 5 offers some concluding remarks.

2. The TG priors

We describe the TG priors in this section, starting by a general introduction of Bayesian inverse problems in function spaces.

2.1. Problem setup

We consider a separable Hilbert space X with inner product $\langle \cdot, \cdot \rangle_X$. Our goal is to estimate the unknown $u \in X$ from data $y \in \mathbb{R}^m$. The data y is related to u via a forward model

$$y = G(u) + \zeta, \tag{1}$$

where $G : X \to \mathbb{R}^m$ and ζ is a m-dimensional zero mean Gaussian noise with covariance matrix Σ. In particular we assume that the data y is generated by applying the operator G to a truth $u^\dagger \in X$ and then adding noise to it. It is easy to see that, under this assumption, the likelihood function, i.e., the distribution of y conditional on u is

$$p(y|u) \propto \exp(-\Phi^y(u)),$$

where

$$\Phi^y(u) := \frac{1}{2}\|G(u) - y\|_\Sigma^2 = \frac{1}{2}\|\Sigma^{-1/2}(G(u) - y)\|_2^2, \tag{2}$$

is often referred to as the data fidelity term in deterministic inverse problems. In what follows, without causing any ambiguity, we shall drop the superscript y in Φ^y for simplicity. In the Bayesian inference we assume that the prior measure of u is μ_{pr}, and the posterior measure of u is provided by the Radon–Nikodym (R–N) derivative:

$$\frac{\mathrm{d}\mu^{y}}{\mathrm{d}\mu_{\mathrm{pr}}}(u) = \frac{1}{Z}\exp(-\Phi(u)), \tag{3}$$

where Z is a normalization constant. Equation (3) can be interpreted as the Bayes' rule in the infinite-dimensional setting.

As is mentioned in section 1, it is conventionally assumed that the prior $\mu_{\mathrm{pr}} = \mu_0$ where $\mu_0 = N(0, C_0)$, i.e., a Gaussian measure defined on X with (without loss of generality) zero mean and covariance operator C_0. Note that C_0 is symmetric positive and of trace class. The range of $C_0^{\frac{1}{2}}$

$$E = \left\{u = C_0^{\frac{1}{2}}x \mid x \in X\right\} \subset X,$$

which is a Hilbert space equipped with inner product [7]

$$\langle\cdot,\cdot\rangle_E = \langle C_0^{-\frac{1}{2}}\cdot, C_0^{-\frac{1}{2}}\cdot\rangle_X,$$

is called the Cameron–Martin space of measure μ_0. For the Gaussian prior, it has been proved that if G satisfies the following assumptions [24]:

Assumptions A.1.

(i) for every $\epsilon > 0$ there is $M = M(\epsilon) \in \mathbb{R}$ such that, for all $u \in X$,

$$\|G(u)\|_\Sigma \leqslant \exp(\epsilon\|u\|_X^2 + M),$$

(ii) for every $r > 0$ there is $K = K(r) > 0$ such that, for all $u_1, u_2 \in X$ with $\max\{\|u_1\|_X, \|u_2\|_X\} < r$

$$\|G(u_1) - G(u_2)\|_\Sigma \leqslant K\|u_1 - u_2\|_X,$$

the associated functional Φ satisfies assumptions 2.6 in [24]. As a result, the posterior μ^y is a well-defined probability measure on X (theorem 4.1 in [24]) and it is Lipschitz in the data y (theorem 4.2 in [24]). Moreover, under some additional assumptions, the posterior measure can be well approximated by a finite-dimensional representation (theorem 4.10 in [24]).

We also should note that the definition of the maximum *a posteriori* (MAP) estimator in finite-dimensional spaces does not apply here, as the measures μ^y and μ_0 are not absolutely continuous with respect to the Lebesgue measure; instead, the MAP estimator in X is defined as the minimizer of the Onsager–Machlup functional [8, 19]:

$$I(u) \coloneqq \Phi(u) + \frac{1}{2}\|u\|_E^2, \tag{4}$$

over the Cameron–Martin space E of μ_0. Here the Cameron–Martin norm $\|\cdot\|_E$ is given by

$$\|u\|_E = \|C_0^{-\frac{1}{2}}u\|_X.$$

2.2. A general class of the hybrid priors

We present a general class of hybrid priors in this section. The idea is rather straightforward: instead of simply letting $\mu_{\mathrm{pr}} = \mu_0$, we let

$$\frac{\mathrm{d}\mu_{\mathrm{pr}}}{\mathrm{d}\mu_0}(u) \propto \exp(-R(u)), \tag{5}$$

where $R(u)$ represents additional prior information (or regularization) on u. In what follows we shall refer to R as the additional regularization term. It follows immediately that the R–N derivative of μ^y with respect to μ_0 is

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(u) \propto \exp(-\Phi(u) - R(u)), \tag{6}$$

which returns to the conventional formulation with Gaussian priors. Next we shall show that equation (5) is a well-behaved prior under certain assumptions on R:

Assumptions A.2. The function $R : X \to \mathbb{R}$ has the following properties.

(i) For all $u \in X$, $R(u)$ is bounded from below, and without loss of generality we can simply assume $R(u) \geqslant 0$.
(ii) For every $r > 0$ there is a $K = K(r) > 0$ such that, for all $u \in X$ with $\|u\|_X < r$, $R(u) \leqslant K$.
(iii) For every $r > 0$, there is an $L(r) > 0$ such that, for all $u_1, u_2 \in X$ with $\max\{\|u_1\|_X, \|u_2\|_X\} < r$

$$|R(u_1) - R(u_2)| \leqslant L\|u_1 - u_2\|_X.$$

Note that the assumptions above are similar to the assumptions (2.6) in [24], with two major differences: first in assumption A.1 (i), we require R to be strictly bounded from below; secondly R is independent of the data y and so we do not have item (iv) in the assumptions in 2.6. The requirement that R is bounded from below is needed in the proof of our results regarding the finite-dimensional approximation of the posterior. It is easy to show that if Φ satisfy assumptions 2.6 in and R satisfy assumptions A.1, $\Phi + R$ satisfies the assumptions in 2.6 in [24]. As a result, μ^y is a well-defined measure on X and it is Lipschitz in the data y, which are stated in the following two theorems.

Theorem 2.1. *Let G satisfy assumptions A.1 and R satisfy assumptions A.2. Then μ^y given by equation (6) is a well-defined probability measure on X.*

Theorem 2.2. *Let G satisfy the assumption A.1 (i) and R satisfy the assumptions in A.2 (i) and (ii). Then μ^y given by equation (6) is Lipschitz in the data y, with respect to the Hellinger distance: if μ^y and $\mu^{y'}$ are two measures corresponding to data y and y' the there exists $C = C(r)$ such that, for all y, y' with $\max\{\|y\|_2, \|y'\|_2\} < r$*

$$d_{\mathrm{Hell}}(\mu^y, \mu^{y'}) \leqslant C\|y - y'\|_\Sigma.$$

Consequently the expectation of any polynomially bounded function $f : X \to E$ is continuous in y.

Theorems (2.1) and (2.2) are direct consequences of the fact that $\Phi + R$ satisfies the assumptions in 2.6 in [24] and so we omit the proofs here. Next we shall study the related issue of approximating the posterior μ^y with a measure defined in a finite-dimensional space, which is of essential importance for numerical implementations of the Bayesian inferences. In particular we consider the following approximation:

$$\frac{\mathrm{d}\mu^y_{N_1,N_2}}{\mathrm{d}\mu_0} \propto \exp\left(-\Phi_{N_1}(u) - R_{N_2}(u)\right), \tag{7}$$

where $\Phi_{N_1}(u)$ is a N_1 dimensional approximation of $\Phi(u)$ and $R_{N_2}(u)$ is a N_2 dimensional approximation of $R(u)$. The following theorem provides the convergence of $\mu^y_{N_1,N_2}$ to μ^y with respect to Hellinger distance under certain assumptions.

Theorem 2.3. *Assume that G and G_{N_1} satisfy assumption A.1 (i) with constants uniform in N_1, and R and R_{N_2} satisfy the assumptions in A.2 (i) and (ii) with constants uniform in N_2. Assume also that for $\forall \epsilon > 0$, there exist two positive sequences $\{a_{N_1}(\epsilon)\}$ and $\{b_{N_2}(\epsilon)\}$ converging to zero, such that $\mu_0(X_\epsilon) \geqslant 1 - \epsilon$ for $\forall N_1, N_2 \in \mathbb{N}$, where*

$$X_\epsilon = \{u \in X \mid |\Phi(u) - \Phi_{N_1}(u)| \leqslant a_{N_1}(\epsilon), |R(u) - R_{N_2}(u)| \leqslant b_{N_2}(\epsilon)\}.$$

Then we have

$$d_{\mathrm{Hell}}(\mu^y, \mu^y_{N_1,N_2}) \to 0 \qquad \text{as} \qquad N_1, N_2 \to +\infty.$$

Proof. For every $r > 0$ there is a $K_1 = K_1(r) > 0$ and a $K_2 = K_2(r) > 0$ such that, for all $u \in X$ with $\|u\|_X < r$, $\Phi(u) \leqslant K_1$ and $R(u) \leqslant K_2$. Let $r = \|y\|_\Sigma$ and $K(r) = K_1(r) + K_2(r)$, and we can show that the normalization constant Z for μ^y satisfies

$$Z \geqslant \int_{\{\|u\|_X < r\}} \exp(-K(r))\mu_0(\mathrm{d}u) = \exp(-K(r))\mu_0\{\|u\|_X < r\} = C.$$

Similarly, we can show that $Z_{N_1,N_2} \geqslant C$ where Z_{N1,N_2} is the normalization constant for $\mu^y_{N_1,N_2}$. Since for any $a > 0$ and $b > 0$, $|\exp(-a) - \exp(-b)| \leqslant \min\{1, |a - b|\}$, for any $\epsilon \in (0, 1)$

$$
\begin{aligned}
|Z - Z_{N_1,N_2}| &\leqslant \int_X |\exp(-\Phi(u) - R(u)) - \exp(-\Phi_{N_1}(u) - R_{N_2}(u))|\mathrm{d}\mu_0(u) \\
&\leqslant \int_{X \setminus X_\epsilon} \mu_0(\mathrm{d}u) + \int_{X_\epsilon} |\Phi(u) - \Phi_{N_1}(u)|\mu_0(\mathrm{d}u) \\
&\quad + \int_{X_\epsilon} |R(u) - R_{N_2}(u)|\mu_0(\mathrm{d}u) \\
&\leqslant \epsilon + a_{N_1}(\epsilon) + b_{N_2}(\epsilon).
\end{aligned}
$$

From the definition of Hellinger distance, we have

$$
\begin{aligned}
2d_{\mathrm{Hell}}(\mu^y, \mu^y_{N_1,N_2})^2 &= \int_X \left(\sqrt{\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}} - \sqrt{\frac{\mathrm{d}\mu^y_{N_1,N_2}}{\mathrm{d}\mu_0}}\right)^2 \mu_0(\mathrm{d}u) \\
&= \int_X \left(Z^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\Phi(u) - \frac{1}{2}R(u)\right)\right. \\
&\quad \left. - Z_{N_1,N_2}^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\Phi_{N_1}(u) - \frac{1}{2}R_{N_2}(u)\right)\right)^2 \mu_0(\mathrm{d}u) \leqslant I_1 + I_2 + I_3, \tag{8}
\end{aligned}
$$

where

$$I_1 = \int_{X \setminus X_\epsilon} \left(\frac{1}{\sqrt{Z}} \exp\left(-\frac{1}{2}\Phi(u) - \frac{1}{2}R(u) \right) \right.$$

$$\left. - \frac{1}{\sqrt{Z_{N_1,N_2}}} \exp\left(-\frac{1}{2}\Phi_{N_1}(u) - \frac{1}{2}R_{N_2}(u) \right) \right)^2 \mu_0(\mathrm{d}u),$$

$$I_2 = \frac{2}{Z} \int_{X_\epsilon} \left(\exp\left(-\frac{1}{2}\Phi(u) - \frac{1}{2}R(u) \right) - \exp\left(-\frac{1}{2}\Phi_{N_1}(u) - \frac{1}{2}R_{N_2}(u) \right) \right)^2 \mu_0(\mathrm{d}u),$$

$$I_3 = 2|Z^{-\frac{1}{2}} - (Z_{N_1,N_2})^{-\frac{1}{2}}|^2 \int_{X_\epsilon} \exp(-\Phi_{N_1}(u) - R_{N_2}(u)) \mu_0(\mathrm{d}u).$$

It is easy to show

$$I_1 \leqslant \int_{X \setminus X_\epsilon} (2C^{-\frac{1}{2}})^2 \mu_0(\mathrm{d}u) \leqslant C\epsilon,$$

$$I_2 \leqslant \frac{2}{C} \int_{X_\epsilon} (a_{N_1}(\epsilon) + b_{N_2}(\epsilon))^2 \mu_0(\mathrm{d}u) \leqslant C(a_{N_1}(\epsilon) + b_{N_2}(\epsilon))^2,$$

$$I_3 \leqslant C(Z^{-3} \wedge (Z_{N_1,N_2})^{-3})|Z - Z_{N_1,N_2}|^2 \int_{X_\epsilon} \mu_0(\mathrm{d}u) = C(\epsilon + a_{N_1}(\epsilon) + b_{N_2}(\epsilon))^2.$$

It follows immediately that

$$2\mathrm{d}_{\mathrm{Hell}}^2(\mu^y, \mu_{N_1,N_2}^y) \leqslant C(\epsilon + \epsilon^2 + (a_{N_1}(\epsilon) + b_{N_2}(\epsilon))^2 + \epsilon(a_{N_1}(\epsilon) + b_{N_2}(\epsilon))),$$

where C is a constant independent of N_1, N_2. Let N_1 and N_2 tend to $+\infty$, yielding

$$\lim_{N_1,N_2 \to +\infty} 2\mathrm{d}_{\mathrm{Hell}}(\mu, \mu_{N_1,N_2})^2 \leqslant C(\epsilon + \epsilon^2),$$

for any $\epsilon > 0$. Thus,

$$\lim_{N_1,N_2 \to +\infty} \mathrm{d}_{\mathrm{Hell}}(\mu^y, \mu_{N_1,N_2}^y) = 0,$$

which completes the proof.

We emphasize that the major difference between theorems 2.3 and 4.10 in [24] is that our assumption is weaker than that in theorem 4.10. The reason that we can use a weaker assumption is that we require R to be strictly bounded from below in assumption A.2 (i). This modification of assumptions is important for our work as that of the key assumption made in theorem 4.10 may not hold in our setting (see remark 2.5 for details).

In general, the approximation errors arise from two sources: representing u with a finite-dimensional basis, and solving the forward model G (or equivalently the functional Φ) approximately. Next we consider a special case where we assume that for a given finite-dimensional representation u_N of u, the functional $\Phi(u_N)$ can be computed exactly; this can be understood as the idealized formulation of the situation where for any given u_N one can choose a numerical scheme to compute the solution to a desired level of accuracy. In this setting, we can show that the finite-dimensional approximation μ_N^y converges to μ without assuming any additional conditions:

Corollary 2.4. *Let $\{e_k\}_{k=1}^{\infty}$ be a complete orthonormal basis of X*

$$u_N = \sum_{k=1}^{N} \langle u, e_k \rangle e_k \tag{9}$$

and

$$\frac{\mathrm{d}\mu_N^y}{\mu_0} = \exp\left(-\Phi(u_N) - R(u_N)\right).$$

Assume that G satisfies assumptions A.1 and R satisfies assumptions A.2. Then

$$d_{\mathrm{Hell}}(\mu^y, \mu_N^y) \to 0, \quad \text{as} \quad N \to \infty.$$

Proof.   Set

$$a_N = \mathbb{E}\|u - u_N\|_X^2 = \sum_{k=N+1}^{\infty} \mathbb{E}|\langle u, e_k \rangle|^2.$$

Since C_0 is in the trace class, $a_N \to 0$ as $N \to \infty$. By Markov's inequality, we have that, for any $\epsilon > 0$

$$\mu_0\left(\left\{\|u - u_N\|_X > \sqrt{\frac{2a_N}{\epsilon}}\right\}\right) \leqslant \frac{1}{2}\epsilon, \quad \text{for } \forall N \in \mathbb{N}. \tag{10}$$

For the given ϵ, there is a r_ϵ such that $\mu_0(\{u \in X \mid \|u\|_X > r_\epsilon\}) < \frac{1}{2}\epsilon$. It is easy to show that, for any $N \in \mathbb{N}$

$$\mu_0\left(\left\{u \in X \mid \|u\|_X \leqslant r_\epsilon, \|u - u_N\|_X \leqslant \sqrt{\frac{2a_N}{\epsilon}}\right\}\right) \geqslant 1 - \epsilon.$$

For conciseness, we define $\widetilde{X} = \{u \in X \mid \|u\|_X \leqslant r_\epsilon, \|u - u_N\|_X \leqslant \sqrt{\frac{2a_N}{\epsilon}}\}$.

Recall that Φ satisfies assumptions 2.6 in [7] and R satisfies assumptions A.2, and so we know that there are constants $L_\epsilon^\Phi, L_\epsilon^R > 0$, such that for any $u \in \widetilde{X}$

$$|\Phi(u) - \Phi(u_N)| \leqslant L_\epsilon^\Phi \|u - u_N\|_X \leqslant L_\epsilon^\Phi \sqrt{\frac{2a_N}{\epsilon}},$$

$$|R(u) - R(u_N)| \leqslant L_\epsilon^R \|u - u_N\|_X \leqslant L_\epsilon^R \sqrt{\frac{2a_N}{\epsilon}}.$$

Obviously, $L_\epsilon^\Phi \sqrt{\frac{2a_N}{\epsilon}}, L_\epsilon^R \sqrt{\frac{2a_N}{\epsilon}} \to 0$ as $N \to \infty$. As

$$\widetilde{X} \subset X_\epsilon = \left\{u \in X \mid |\Phi(u) - \Phi(u_N)| \leqslant L_\epsilon^\Phi \sqrt{\frac{2a_N}{\epsilon}}, |R(u) - R(u_N)| \leqslant L_\epsilon^R \sqrt{\frac{2a_N}{\epsilon}}\right\},$$

we have $\mu_0(X_\epsilon) \geqslant 1 - \epsilon$, and by theorem 2.3

$$d_{\mathrm{Hell}}(\mu^y, \mu_N^y) \to 0, \quad \text{as} \quad N \to \infty,$$

which completes the proof.

We reinstate that theorem 4.10 in [24] cannot be directly applied to this setting as its assumption may not necessarily be satisfied. Namely, to apply theorem 4.10, $\Phi + R$ must satisfy: for any $r > 0$, there is a $K = K(r) > 0$ such that for all u with $\|u\|_X < r$

$$|(\Phi(u) + R(u)) - (\Phi(u_N) + R(u_N))| < K\phi(N),$$

where $\phi(N) \to 0$ as $N \to \infty$. In the following remark, we construct a simple counterexample for such a condition.

Remark 2.5. Let u_N be given by equation (9) and let R satisfy $R(0) = 0$. Let Φ satisfy the assumption: for any $r > 0$, there is a $K = K(r) > 0$ such that for all u with $\|u\|_X < r$

$$|\Phi(u) - \Phi(u_N)| < K\phi(N),$$

where $\phi(N) \to 0$ as $N \to \infty$. For any given N, we consider $u = e_{N+1}$ and obviously $u_N = 0$. It follows immediately that

$$|(\Phi(u) + R(u)) - (\Phi(u_N) + R(u_N))| > R(e_{N+1}) - K\phi(N).$$

It should also be noted that, in this work we only consider finite-dimensional observation data for simplicity. A more general case is to consider infinite-dimensional data as well, and study the limiting behavior of the posterior distributions under measurement refinements. The matter has been discussed in [9, 17] for Besov priors.

2.3. The TG prior

Until this point, we have discussed the non-Gaussian priors in a rather general setting. In this section, we introduce the particular choice of the additional regularization term R for the edge-detection purposes. First we need to specify the space of functions X. Let Ω be a bounded open subset of $\mathbb{R}^q$ where $q \in \mathbb{N}$, and X be the Sobolev space $H^1(\Omega)$:

$$X = H^1(\Omega) = \{u \in L_2(\Omega) \mid \partial_x^\alpha u \in L_2(\Omega) \text{ for all } |\alpha| \leqslant 1 \},$$

where $\alpha = (\alpha_1, \ldots, \alpha_q)$ and $|\alpha| = \sum_{i=1}^q \alpha_i$, and the associated norm $\|\cdot\|_X = \|\cdot\|_{H^1}$ is

$$\|u\|_{H^1} = \sum_{|\alpha| \leqslant 1} \|\partial_x^\alpha u\|_{L^2(\Omega)}.$$

Naturally we choose the regularization term to be the TV seminorm

$$R(u) = \lambda \|u\|_{\text{TV}} = \lambda \int \|\nabla u\|_2 \, dx, \tag{11}$$

where λ is a prescribed positive constant. We show that the TV term (11) satisfies the required assumptions:

Lemma 2.6. *Equation (11) satisfies assumptions A.2.*

Proof.

(i): The assumption is trivially satisfied.
(ii): For all $u \in X$ with $\|u\|_X < r$, there exists a constant $C > 0$ such that

$$R(u) = \lambda \|u\|_{\text{TV}} \leqslant \lambda C \|u\|_X \leqslant \lambda C r = K(r).$$

(iii): For every $r > 0$ and all $u_1, u_2 \in X$, there is a constant $C > 0$ such that

$$|R(u_1) - R(u_2)| \leqslant \lambda \|(u_1 - u_2)\|_{\text{TV}} \leqslant \lambda C \|u_1 - u_2\|_X.$$

It follows directly from lemma 2.6 that theorems 2.1 and 2.2 and corollary 2.4 hold for the prior (11). Following the same procedure of [8], we can show that the MAP estimator for the TG prior is the minimizer of

$$I(u) := \Phi(u) + \lambda \|u\|_{\mathrm{TV}} + \frac{1}{2}\|u\|_E^2. \tag{12}$$

3. Splitting pCN

We now discuss the numerical implementation of the Bayesian inference with the TG priors. Typically the Bayesian inference is implemented with MCMC algorithms. The pCN algorithms and a family of dimension-independent MCMC schemes have been proposed to draw samples in the function spaces. As is mentioned in section 1, thanks to the special structure of the TG prior, the pCN algorithms can be applied directly to problems with this prior, requiring no modifications. Nevertheless, in this section, we propose to further improve the sampling efficiency by making very simple modifications to the pCN algorithms. The idea is based on the following two observations on the TG prior:

 (i) The TV term R is more sensitive to small local fluctuations of u than the data fidelity term Φ.
 (ii) The TV term R can be computed much more efficiently than the data fidelity term Φ. In fact, evaluating Φ requires to simulate forward model G which is often governed by some computationally intensive partial differential equations.

As a result of the first observation, to achieve a reasonable acceptance probability, for example, 20%, one has to use a very small stepsize in the pCN algorithm, resulting in poor mixing. On the other hand, the restriction of the stepsize is mainly due to the TV term. To address the issue, we introduce a splitting scheme to the standard pCN algorithm; namely, we perform the pCN in two stages: one for the TV term and one for the data fidelity term. The intuition behind the method is rather straightforward: the slowly varying and computationally intensive data fidelity term Φ is evaluated less frequently than the fast-varying and computationally efficient TV term R. A particle is moved and accepted or rejected k times according to the TV term. The sample resulting from the k short step moves is accepted or rejected using a Metropolis criterion based on Φ after the k short-range moves. Specifically, suppose the current state is u_{current}, the S-pCN proceeds as follows:

 (i) let $v_0 = u_{\mathrm{current}}$.
 (ii) for $i = 1$ to k perform the following iteration:
    (a) propose $v_{\mathrm{prop}} = \sqrt{1 - \beta^2}\, v_{i-1} + \beta w$, where $w \sim \mu_0$;
    (b) let $v_i = v_{\mathrm{prop}}$ with probability;

$$\mathrm{acc}_R(v_{\mathrm{prop}}, v_{i-1}) = \min\{1, \exp[-(R(v_{\mathrm{prop}}) - R(v_{i-1}))]\}; \tag{13}$$

    and let $v_i = v_{i-1}$ with probability $1 - \mathrm{acc}_R(v_{\mathrm{prop}}, v_{i-1})$.
    (c) return to step ((ii)(a));
 (iii) let $u_{\mathrm{next}} = v_k$ with probability

$$\mathrm{acc}_\Phi(v_k, u_{\mathrm{current}}) = \min\{1, \exp[-(\Phi(v_k) - \Phi(u_{\mathrm{current}}))]\} \tag{14}$$

and $u_{\mathrm{next}} = u_{\mathrm{current}}$ with probability $1 - \mathrm{acc}_\Phi(v, u_{\mathrm{current}})$.

We note the similarity of this splitting algorithm and the approximation-accelerated MCMC algorithms such as [3, 10], and especially the surrogate transition method in [20]. However, our algorithm simply splits the data fidelity term and the TV term, without requiring any approximations or surrogate models. In fact, if a surrogate is available for the forward model, it can be naturally incorporated in to the S-pCN algorithm to further improve the sampling efficiency. Nevertheless, constructing and implementing such a surrogate model is not in the scope of this work.

A possible interpretation of the S-pCN algorithm is that step 2 is responsible of generating a new proposal for the target distribution μ^y, which is rejected or accepted according to the normal MCMC rule in step 3. Under this interpretation, we need to show that the detailed balance condition is satisfied, which is stated by the following proposition.

Proposition 3.1. *Let $q(u, dv)$ is the proposal kernel corresponding to the v_k proposed in step 2 given u, and Q is its transition kernel given by*

$$Q(u, dv) = a(u, v_k)q(u, dv) + \delta_u(dv)\left(1 - \int acc_\Phi(z, u)q(u, dz)\right),$$

where $acc_\Phi(v_k, u)$ is given by equation (14) and δ_u is the point mass at u. We then have

$$\mu^y(du)Q(u, dv) = \mu^y(dv)Q(v_k, du).$$

The proof of the proposition is similar to the ergodicity proof of the preconditioned MCMC algorithm in [10], and thus is omitted here.

4. Numerical examples

4.1. A signal denoising problem

First we test the proposed prior on a signal denoising example. Namely, suppose we have a piecewise constant signal

$$u(t) = \begin{cases} 0, & 0 \leqslant t < 1/3; \\ 1, & 1/3 \leqslant t < 2/3; \\ 0, & 2/3 \leqslant t \leqslant 1. \end{cases}$$

We observe 23 data points equally distributed on [0, 1] each with independent Gaussian observation noise $N(0, 0.02^2)$. The true signal and the data points are shown in figure 1 (left). We first test the Gaussian priors, and specifically we choose the prior to be a zero-mean Gaussian measure with a squared exponential covariance:

$$K(t_1, t_2) = \gamma \exp\left[-\frac{1}{2}\left(\frac{t_1 - t_2}{d}\right)^2\right]. \tag{15}$$

Note that, with Gaussian prior, the posterior mean can be computed analytically, and here we computed it with $\gamma = 0.1$ and $d = 0.04, 0.08$ and 0.16. We plot the results in figure 1 (right), which clearly demonstrate that the Gaussian priors perform poorly for this function.

Next we compare the performance of our TG prior and that of the TV prior. As has been discussed, for the TV prior, we have to use a finite-dimensional formulation, and assume the density of the prior is
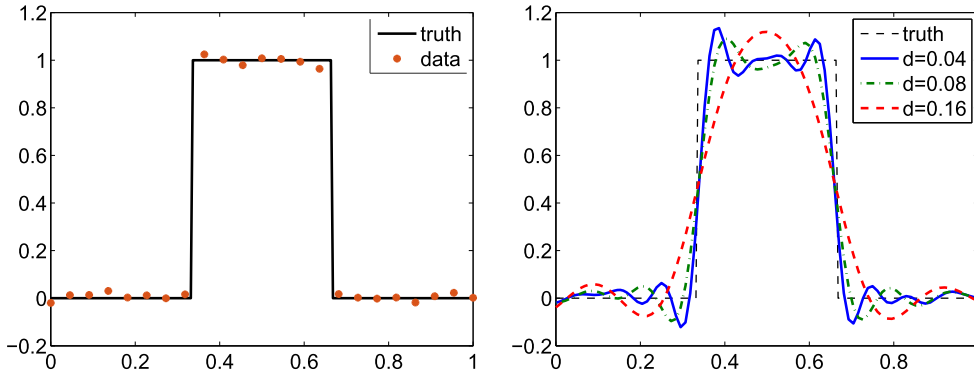
Figure 1. Left: the true signal (solid line) and the observed data points (dots). Right: the posterior mean of the Gaussian prior with $d = 0.04, 0.08, 0.16$.
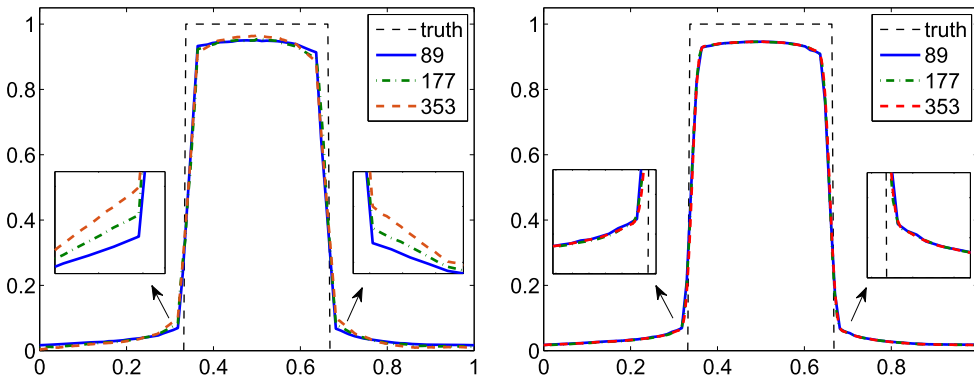

Figure 2. Left: the posterior mean results with the TV prior, computed with grid point numbers $N = 89, 177, 353$. Right: the posterior mean results with the TG prior, computed with grid point numbers $N = 89, 177, 353$. In both figures, the insets show the zoom-in view near the jumps.

$$p(u_N) \propto \exp(-\lambda \|u_N\|_{\text{TV}}),$$

where the regularization parameter is taken to be $\lambda = 500$. For the TG prior, we need to specify both the TV term and the Gaussian measure. For the TV term we set $\lambda = 500$ and for the Gaussian reference measure, we assume the covariance is again given by equation (15), with $d = 0.02$ and $\gamma = 0.1$. With either prior, we perform the inference using three different numbers of grid points $N = 89, 177, 353$, and so that we can see if the inference results depend on the dimensionality of the problem. As this example does not involve computationally demanding forward model, we choose to sample the posterior with the standard pCN algorithm. For the TV prior, we draw 10^8 samples for the cases $N = 89$ and $N = 177$ and 5×10^8 samples for $N = 353$. We use such large numbers of samples to ensure reliable estimates of the inference. We plot the posterior mean of the TV prior in figure 2 (left), and we can see (especially from the zoomed-in plots) that the results of the different numbers of grid points depart evidently from each other, which indicates that the inference results of the TV prior depends on the discretization dimensionality. These results are consistent with the findings reported in [17]. Next we draw 10^8 samples from the posterior
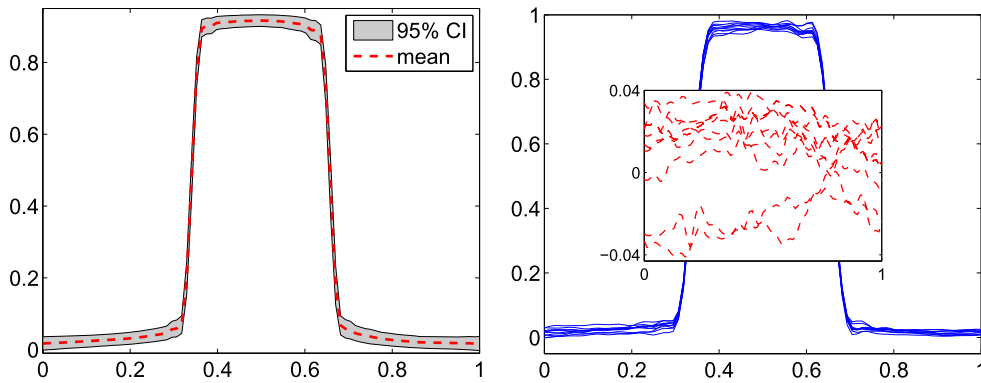
Figure 3. Left: the pointwise 95% credible interval (CI). Right: ten random samples drawn from the the posterior (solid lines) and from the prior (inset, dashed lines).
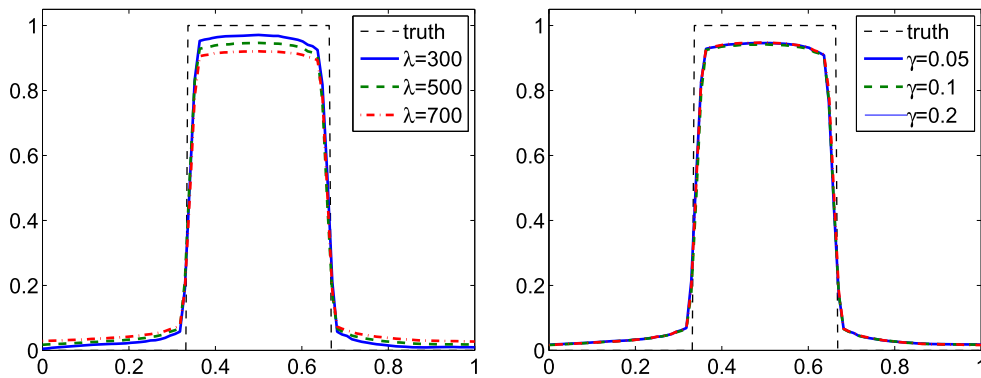

Figure 4. The posterior means of the TG prior, computed with various values of γ and λ. Left: the means computed with $\gamma = 0.1$ and $\lambda = 300, 500, 700$. Right: the means computed with $\lambda = 500$ and $\gamma = 0.05, 0.1, 0.2$.

with the TG prior, for all the three numbers of grid points, and plot the resulting posterior means in figure 2 (right). We can see that the results for the three different grid numbers look almost identical, suggesting that the results with the TG prior are independent of discretization dimensionality. We also can see that, compared to the standard Gaussian priors, the TG prior can much better detect the sharp jumps of the signal, thanks to the presence of the TV term. As is mentioned in the introduction, a major advantage of the Bayesian method is that it can quantify the uncertainty in the estimates. To show this, in figure 3 (left), we plot the 95% pointwise credible interval of the unknown, and in figure 3 we plot 10 random samples drawn from the prior and the posterior. These plot demonstrate the difference in the Bayesian and the deterministic methods for solving inverse problems.

Finally we want to exam how the inference results depend on the regularization parameters, i.e., λ and γ, for the TG prior. Specifically we perform the inferences with different values of λ and γ, each with 10^8 samples, and show the results in figure 4. In figure 4 (left), we show the posterior means computed with $\gamma = 0.1$ and $\lambda = 300, 500, 700$; in figure 4 (right), we show the posterior means computed with $\lambda = 500$ and $\gamma = 0.05, 0.1, 0.2$. Both
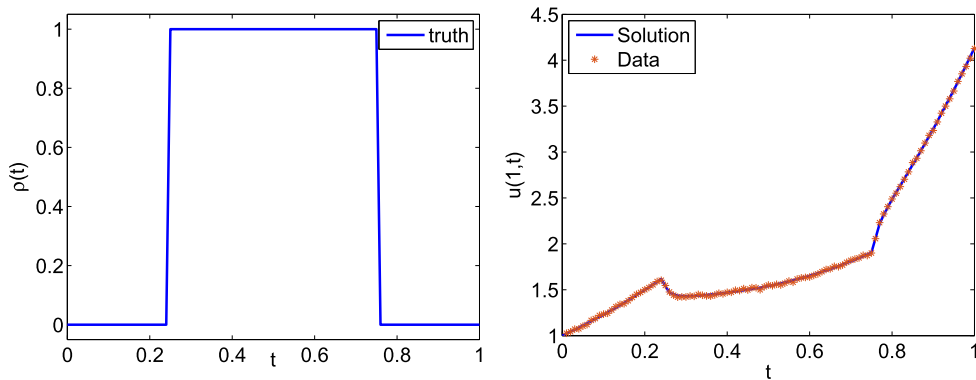
Figure 5. Left: the true Robin coefficient. Right: the true solution at $x = 1$ (solid line) and the observed data points (crosses).

figures suggest that the inference results are rather robust with respect to the values of these parameters.

Note, however, that this is only a simply toy problem, and in practice, the problems can be much more difficult: the observation noise can be much stronger and the forward operator can be much more ill-posed, etc. Thus, to further evaluate the performance of the TG prior, we consider a more complicated problem in the next example.

4.2. Estimating the Robin coefficients

Here we consider the one-dimensional heat conduction equation in the region $x \in [0, L]$

$$\frac{\partial u}{\partial t}(x, t) = \frac{\partial^2 u}{\partial x^2}(x, t), \tag{16a}$$

$$u(x, 0) = g(x), \tag{16b}$$

with the following Robin boundary conditions:

$$-\frac{\partial u}{\partial x}(0, t) + \rho(t)u(0, t) = h_0(t), \tag{16c}$$

$$-\frac{\partial u}{\partial x}(L, t) + \rho(t)u(L, t) = h_1(t). \tag{16d}$$

Suppose the functions $g(x)$, $h_0(t)$ and $h_1(t)$ are all known, and we want to estimate the unknown Robin coefficient $\rho(t)$ from certain measurements of the temperature $u(x, t)$. The Robin coefficient $\rho(t)$ characterizes thermal properties of the conductive medium on the interface which in turn provides information on certain physical processes near the boundary, and such problems have been extensively studied in the inverse problem literature, e. g, [14, 27].

To be specific, we assume that a temperature sensor is placed at one boundary $x = L$ of the medium, and we want to estimate $\rho(t)$ in the time interval $[0, T]$ from temperature records measured by the sensor during $[0, T]$. The resulting forward operator G satisfies the assumptions A.1, which can be derived from the theoretical results provided in [14]. In this example we choose $L = 1$, $T = 1$ and the functions to be
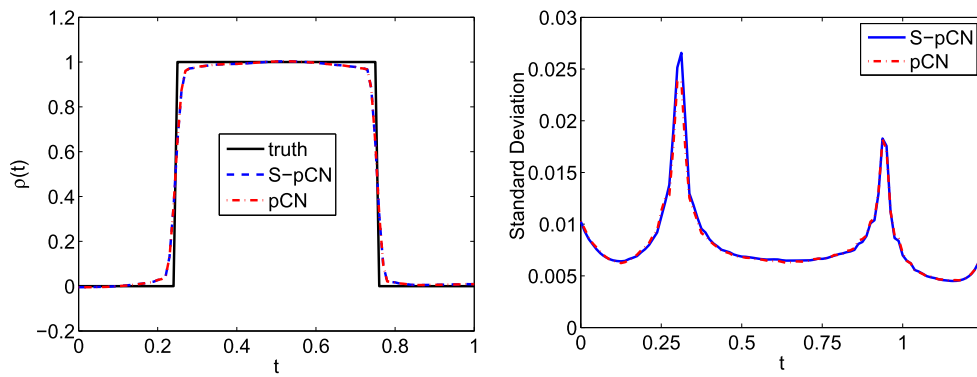
Figure 6. Left: the posterior mean computed by the S-pCN (dashed line) and by the standard pCN (dashed-dotted line), compared to the truth (solid line). Right: the posterior standard deviation computed with S-pCN (dashed) and with pCN (dashed–dotted).

$$g(x) = x^2 + 1, \quad h_0 = t(2t + 1), \quad h_1 = 2 + t(2t + 2).$$

Moreover, the temperature is measured 100 times (equally spaced) in the given time interval and the error in each measurement is assumed to be an independent zero-mean Gaussian random variable with variance 0.01^2. The 'true' Robin coefficient is a piece-wise constant function shown in figure 5 (left), and the data is generated by substituting the true Robin coefficient into equations (16), solving the equation and adding noise to the resulting solution, where both the clean solution and the noisy data are shown in figure 5 (right).

We now perform the Bayesian inference with the TG prior. In particular, we choose $\lambda = 300$ in the TV term, and for the Gaussian measure, the covariance is again given by equation (15) with $\gamma = 0.1$ and $d = 0.02$. Moreover, the number of grid points is taken to be 200 in this example, and the equation (16) is solved with the finite difference scheme used in [27]. We sample the posterior with the standard pCN and our S-pCN algorithms, and with either method, we draw 10^6 samples from the posterior with additional 0.5×10^6 samples used in the burn-in period. We set $\beta = 0.02$ in both algorithms, and in the S-pCN method we choose $k = 10$, i.e. 10 iterations being performed in the first stage. As a result the average acceptance rate in the pCN scheme is about 15% and that in the S-pCN scheme is about 40%.

In figure 6, we show the posterior mean (left) and standard deviation (right) computed by both methods, and we can see that the results of both methods agree rather well with each other, suggesting that both methods can correctly generate samples from the posterior distribution. We want to note that the posterior mean resulting from the TG prior can reasonably detect the jumps in the Robin coefficient; we have not optimized the hyperparameters in the prior and the results may be further improved if we do so. We now compare the efficiency performance of the two methods. First, we show in figure 7 the trace plots of the two methods for the unknown $\rho(t)$ at $t = 0.2$, 0.5 and 0.8. The trace plots provide a simple way to examine the convergence behavior of the MCMC algorithm. Long-term trends (i.e., low mixing rate) in the plot indicate that successive iterations are highly correlated and that the series of iterations have not converged. In this regard, the trace plots indicate that the chains produced by the S-pCN method mix much faster than those by the standard pCN. Next we examine the autocorrelation functions (ACF) of the chains generated by both methods. Once again we consider the points at $t = 0.2$, 0.5 and 0.8 and we plot the ACF for all the three points in figure 8. One can see from the figure that, for all three points, the ACF of the chain generated
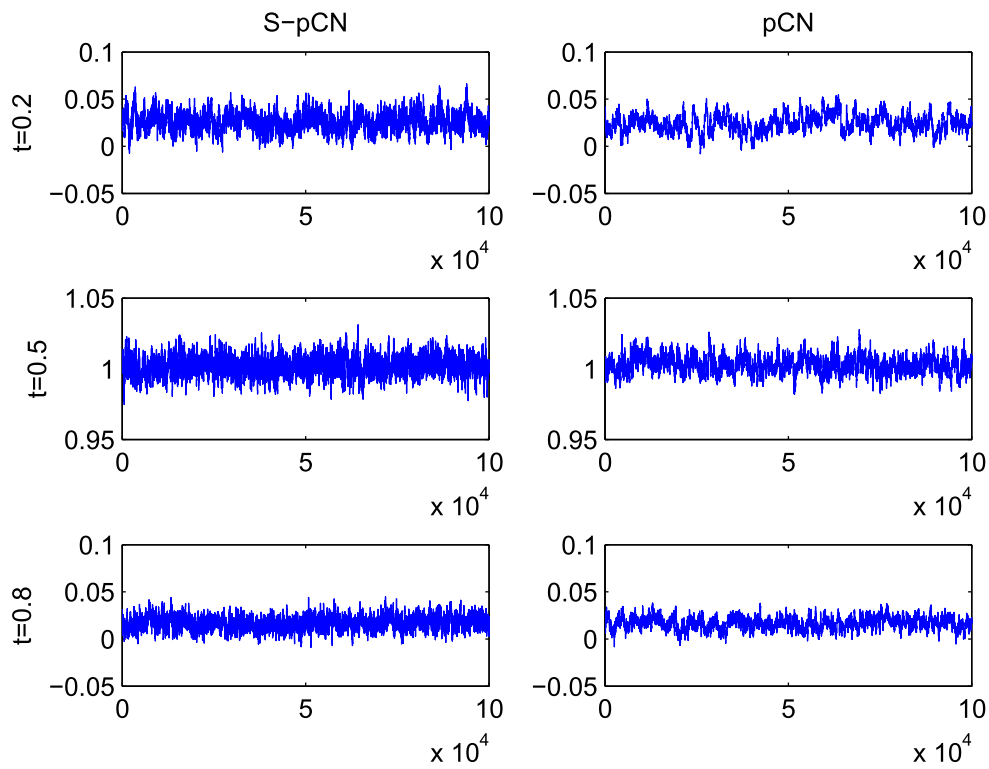
Figure 7. The trace plots for the points at $t = 0.2$, 0.5, and 0.8 with the S-pCN (left) and the pCN (right) methods.
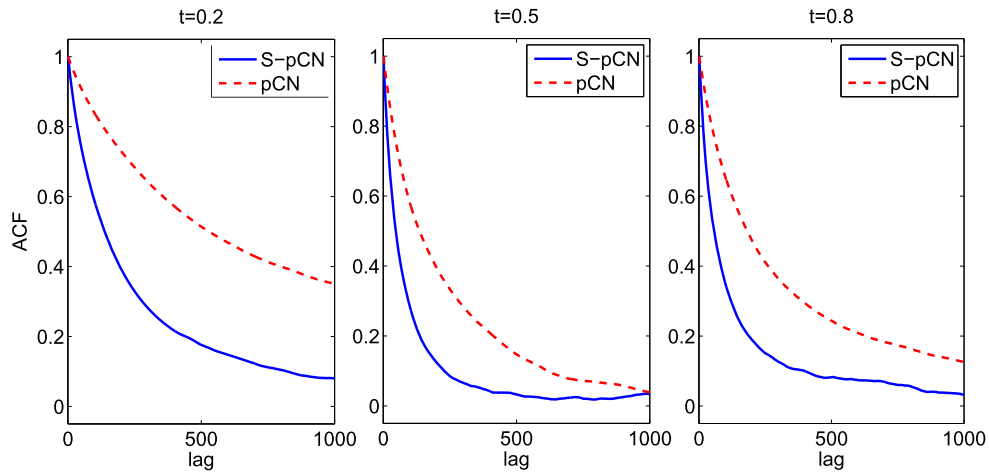

Figure 8. The ACF of ρ at $t = 0.2$, 0.5, 0.8.

by the S-pCN decreases much faster than that of the standard pCN, suggesting that the S-pCN method achieves a significantly better performance. Alternatively, we look at the ACF of lag 100 at all the grid points, which is plotted in figure 9 (left), and we can see that, the ACF of
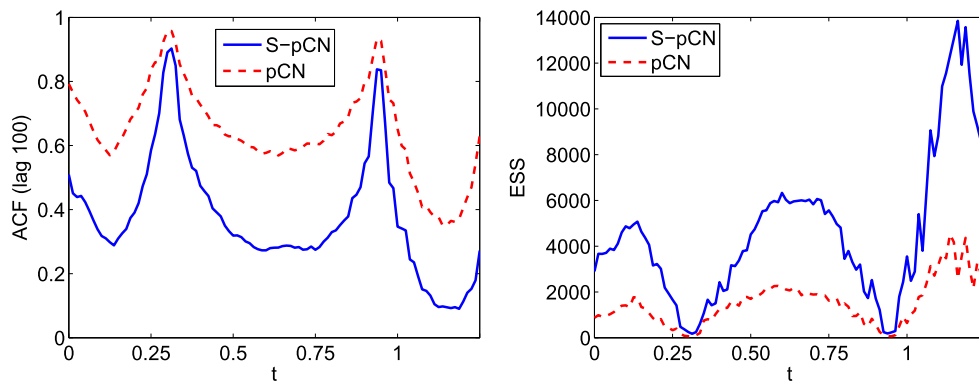
Figure 9. Left: the ACF of lag 100, at each grid point. Right: the ESS at each grid point.

the chain generated by the S-pCN is much lower than that of the standard pCN. The effective sample size (ESS) is another common measure of the sampling efficiency of MCMC [16]. ESS is computed by

$$\text{ESS} = \frac{N}{1 + 2\tau},$$

where τ is the integrated autocorrelation time and N is the total sample size, and it gives an estimate of the number of effectively independent draws in the chain. We computed the ESS of the unknown ρ at each grid point and show the results in figure 9 (right). The results show that the S-pCN algorithm on average produces significantly more effective independent samples than the standard pCN.

5. Conclusions

In summary, we have presented a TG prior for infinite-dimensional Bayesian inverse problems. We use the TV term to improve the ability to detect jumps and use the Gaussian reference measure to ensure that it results in a well defined posterior measure. Moreover, we show that the resulting posterior distributions depend continuously on data and more importantly can be well approximated by finite-dimensional representations. We also present the S-pCN algorithm which can significantly improve the sampling efficiency by simply splitting the standard pCN iterations into two stages. Finally we provide some numerical examples to demonstrate the performance of the TG prior and the efficiency of the S-pCN algorithm. We believe the proposed TG prior can be useful in many practical inverse problems involving functions with sharp jumps.

Several extension of the present work are possible. First, it is interesting to consider the connection between the proposed TG prior and the EN type regularization in deterministic inverse problems. As is shown, the MAP estimator of the TG prior yields the solution of the deterministic inverse problem with an EN type of regularization. We believe such a connection can provide some interesting theoretical results of the TG prior, the investigation of which is of our interest. Secondly, it should also be noted that, throughout the work we assume the regularization parameter λ is given, while in practice, determining λ can be a highly non-trivial task. A simply solution here is to determine λ with the techniques used in the deterministic setting, and then use the result directly in the Bayesian inference.

Nevertheless, developing rigorous and effective methods to determine the regularization parameter in the Bayesian setting is a problem of significance. A possible solution is to impose a prior on λ and formulate a hierarchical or empirical problem to estimate both λ and u. Finally, a very natural extension of the present work is to consider other choices of regularization term R, to reflect different prior information. We plan to investigate these issues in the future.

Acknowledgments

References

[1] Aster R C, Borchers B and Thurber C H 2013 *Parameter Estimation and Inverse Problems* (New York: Academic)
[2] Babacan S D, Molina R and Katsaggelos A K 2009 Variational Bayesian blind deconvolution using a total variation prior *IEEE Trans. Image Process.* **18** 12–26
[3] Christen J A and Fox C 2005 Markov chain Monte Carlo using an approximation *J. Comput. Graph. Stat.* **14** 795–810
[4] Compton R, Osher S and Bouchard L 2012 Hybrid regularization for MRI reconstruction with static field inhomogeneity correction *2012 9th IEEE Int. Symp. on Biomedical Imaging (ISBI)* (Piscataway, NJ: IEEE) pp 650–5
[5] Cotter S L *et al* 2013 MCMC methods for functions: modifying old algorithms to make them faster *Stat. Sci.* **28** 424–46
[6] Cui T, Law K J H and Marzouk Y M 2016 Dimension-independent likelihood-informed MCMC *J. Comput. Phys.* **304** 109–37
[7] Da Prato G 2006 *An Introduction to Infinite-Dimensional Analysis* (Berlin: Springer)
[8] Dashti M, Law K J H, Stuart A M and Voss J 2013 MAP estimators and their consistency in Bayesian nonparametric inverse problems *Inverse Problems* **29** 095017
[9] Dashti M, Harris S and Stuart A 2012 Besov prior for Bayesian inverse problems *Inverse Problems Imaging* **6** 183–200
[10] Efendiev Y, Hou T and Luo W 2006 Preconditioning Markov chain Monte Carlo simulations using coarse-scale models *SIAM J. Sci. Comput.* **28** 776–803
[11] Feng Z and Li J 2015 An adaptive independence sampler MCMC algorithm for infinite dimensional Bayesian inferences arXiv:1508.03283
[12] Gelman A, Carlin J, Stern H and Rubin D 2014 *Bayesian Data Analysis* vol 2 (London: Taylor and Francis)
[13] Helin T and Lassas M 2011 Hierarchical models in statistical inverse problems and the Mumford–Shah functional *Inverse Problems* **27** 015008
[14] Jin B and Lu X 2012 Numerical identification of a Robin coefficient in parabolic problems *Math. Comput.* **81** 1369–98
[15] Kaipio J and Somersalo E 2005 *Statistical and Computational Inverse Problems* vol 160 (Berlin: Springer)
[16] Kass R E, Carlin B P, Gelman A and Neal R M 1998 Markov chain Monte Carlo in practice: a roundtable discussion *Am. Stat.* **52** 93–100
[17] Lassas M, Saksman E and Siltanen S 2009 Discretization-invariant Bayesian inversion and Besov space priors *Inverse Problems Imaging* **3** 87–122
[18] Lassas M and Siltanen S 2004 Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Problems* **20** 1537
[19] Li J 2015 A note on the Karhunene–Loeve expansions for infinite-dimensional Bayesian inverse problems *Stat. Probab. Lett.* **106** 1–4
[20] Liu J S 2008 *Monte Carlo Strategies in Scientific Computing* (Berlin: Springer)

[21] Martin J, Wilcox L C, Burstedde C and Ghattas O 2012 A stochastic newton MCMC method for large-scale statistical inverse problems with application to seismic inversion *SIAM J. Sci. Comput.* **34** A1460–87

[22] Roberts G O *et al* 2001 Optimal scaling for various Metropolis–Hastings algorithms *Stat. Sci.* **16** 351–67

[23] Rudin L I, Osher S and Fatemi E 1992 Nonlinear total variation based noise removal algorithms *Physica* D **60** 259–68

[24] Stuart A M 2010 Inverse problems: a Bayesian perspective *Acta Numer.* **19** 451–559

[25] Vogel C R and Oman M E 1998 Fast robust total variation-based reconstruction of noisy, blurred images *IEEE Trans. Image Process.* **7** 813–24

[26] Vollmer S J 2015 Dimension-independent MCMC sampling for inverse problems with non-Gaussian priors *SIAM/ASA J. Uncertain. Quantification* **3** 535–61

[27] Yang F, Yan L and Wei T 2009 The identification of a Robin coefficient by a conjugate gradient method *Int. J. Numer. Methods Eng.* **78** 800–16

[28] Zou H and Hastie T 2005 Regularization and variable selection via the elastic net *J. R. Stat. Soc.* B **67** 301–20