Certified coordinate selection for high-dimensional Bayesian inversion with Laplace prior

Rafael Flock (raff@dtu.dk)
Technical University of Denmark
Yiqiu Dong
Technical University of Denmark
Felipe Uribe
Lappeenranta-Lahti University of Technology
Olivier Zahm
Grenoble Alpes University

Research Article

Keywords: Bayesian inverse problems, dimension reduction, MCMC, Laplace prior, sparsity.

Posted Date: October 25th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3471448/v1

Additional Declarations: No competing interests reported.

Certified coordinate selection for high-dimensional Bayesian inversion with Laplace prior

Rafael Flock[1*], Yiqiu Dong[1], Felipe Uribe[2], Olivier Zahm[3]

[1*]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, Building 324, Kongens Lyngby, 2800, Denmark.
[2]School of Engineering Science, Lappeenranta-Lahti University of Technology, Yliopistonkatu 34, Lappeenranta, 53850, Finland.
[3]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.

*Corresponding author(s). E-mail(s): raff@dtu.dk;
Contributing authors: yido@dtu.dk; felipe.uribe@lut.fi; olivier.zahm@inria.fr;

Abstract

We consider high-dimensional Bayesian inverse problems with arbitrary likelihood and product-form Laplace prior for which we provide a certified approximation of the posterior density in the Hellinger distance. The approximate posterior density differs from the prior density only in a small number of relevant coordinates that contribute the most to the update from the prior to the posterior. We propose and analyze a gradient-based diagnostic to identify these relevant coordinates. Although this diagnostic requires computing an expectation with respect to the posterior, we propose tractable methods for the classical case of a linear forward model with Gaussian likelihood. Our methods can be employed to estimate the diagnostic before solving the Bayesian inverse problem via, e.g., Markov chain Monte Carlo (MCMC) methods. After selecting the coordinates, the approximate posterior density can be efficiently inferred since most of its coordinates are only informed by the prior. Moreover, specialized MCMC methods, such as the pseudo-marginal MCMC algorithm, can be used to obtain less correlated samples when sampling the exact posterior density. We show the applicability of our method using a 1D signal deblurring problem and a high-dimensional 2D super-resolution problem.

Keywords: Bayesian inverse problems, dimension reduction, MCMC, Laplace prior, sparsity.

1 Introduction

Bayesian inverse problems arise in many applications in science and engineering. When performing Bayesian inversion, one tries to characterize a probability distribution on the unknown parameters of a model given some observed data. These data are typically subject to noise, and are often modeled as

$$y = A(x) + \varepsilon, \tag{1}$$

where $A : \mathbb{R}^d \to \mathbb{R}^m$ is the forward operator, $x \in \mathbb{R}^d$ is the unknown parameter and $\varepsilon \in \mathbb{R}^m$ is the noise. In this work, we are concerned with high-dimensional Bayesian inverse problems with the posterior density given by

$$\pi(x) \propto \mathcal{L}(x)\pi_0(x). \tag{2}$$

Here, $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ denotes the likelihood function of observing the data y given x, and $\pi_0(x)$ is

the prior density. The likelihood function models the relationship between forward operator, error model and data, e.g., (1). Note however, that our general framework applies to any likelihood function.

It is common that natural signals and images can be effectively represented in a sparse manner using adapted bases, like point source basis, wavelets basis, etc., see [1]. Then, one can use the so-called *synthesis formulation* $s = Wx$ to expand a signal s in a suitable basis W for which x is sparse [2]. In this case, the heavy-tailed Laplace distribution is a typical prior choice to enforce sparsity in x. Indeed, in [3] it was found that the marginals of wavelet coefficients of photographic images are well approximated by the Laplace distribution.

Other classical choices for heavy-tailed priors are α-stable distributions, such as the Cauchy distribution [4]. Heavy-tailed priors are especially popular in image reconstruction to preserve sharp edges. To this end, heavy-tailed priors can be imposed directly [5, 6] or via a hierarchical framework on the differences between pixels (see, e.g., [7]). In this paper, we consider a Laplace prior, i.e.,

$$\pi_0(x) \propto \exp\left(-\sum_{i=1}^{d} \delta_i |x_i|\right),$$

which has a product-form with rate parameters $\delta_i > 0$ for all $i = 1, \dots, d$. This prior leads to the posterior

$$\pi(x) \propto \mathcal{L}(x) \exp\left(-\sum_{i=1}^{d} \delta_i |x_i|\right). \tag{3}$$

We note that for the case where the forward operator is linear and the likelihood function is Gaussian, the posterior density (3) can be sampled via the Bayesian LASSO [8].

In real-world applications Bayesian inference is often performed in a high-dimensional parameter space. For instance, in imaging science, the number of pixels is very large, resulting in parameter spaces with dimensions of order $d = \mathcal{O}(10^4)$ or higher. When sampling from non-smooth densities as in (3), the proximal unadjusted Langevin (p-ULA) or proximal Metropolis-adjusted Langevin algorithm (p-MALA) [9, 10] can be used, but their performance deteriorates significantly with the dimensionality of the problem.

Inspired by the certified dimension reduction (CDR) methodology [11–13], we propose a new method, the *certified coordinate selection*, to select the components in x that contribute most to the update from the prior to the posterior density. Hence, the efficiency of the aforementioned sampling methods can be improved substantially by restricting them to perform inference on the selected components only.

In principle, the CDR method consists in replacing the likelihood function with a *ridge* approximation $x \mapsto \widetilde{\mathcal{L}}(U_r^\intercal x)$ for some matrix $U_r \in \mathbb{R}^{d\times r}$ with $r \ll d$ orthogonal columns, and some function $\widetilde{\mathcal{L}} : \mathbb{R}^r \to \mathbb{R}$. The matrix U_r is determined by minimizing an error bound on the Kullback-Leibler (KL) divergence obtained via logarithmic Sobolev inequalities. We term our method *certified coordinate selection* since in this work, $U_r^\intercal$ corresponds to a projection matrix on r selected coordinate axes.

The CDR method has been successfully applied within a number of Bayesian updating strategies, such as the cross-entropy method [14], Stein-variational gradient descent [15], transport maps [16], and in Bayesian inference applied to rare event estimation [17]. However, it has only been employed in cases where the prior is either Gaussian, or it is normalized by computing a map to push forward a standard Gaussian random variable to the original random variable [18].

Applying the CDR method on a posterior density as defined in (3) is not straightforward, because the Laplace prior does not meet the sufficient conditions for the required logarithmic Sobolev inequalities. To overcome this issue, we use the Hellinger distance instead of the KL divergence to bound the posterior approximation. This approach allows use to use Poincaré inequalities, which are satisfied by the Laplace prior [12, 13].

Our main contributions are as follows:

- We propose to select the relevant coordinates based on the following diagnostic

$$h_i = \frac{1}{\delta_i^2} \int_{\mathbb{R}^d} (\partial_i \log \mathcal{L}(x))^2 \pi(x) \mathrm{d}x,$$

and we show that the Hellinger distance between the exact and the approximated posterior density can be explicitly bounded using h_i.
- We proof that in the case of a Gaussian likelihood and a linear forward operator, we only need to estimate the posterior mean and the posterior covariance to compute the diagnostic.
- We show for the above case how a smoothing approximation to the Laplace prior can be used to compute a diagnostic and define an efficient proposal for the preconditioned Metropolis-adjusted Langevin algorithm (MALA).
- We test our methods on a 1D signal deblurring task, which is given in the synthesis formulation and a high-dimensional 2D super-resolution example.

The remainder of this paper is structured as follows. In section 2 we present the theoretical part of our method which comprises the posterior density approximation and its certification. In section 3 we outline the general approach to sample the approximate posterior and recall the pseudo-marginal Markov chain Monte Carlo (MCMC) algorithm, which can be used to sample the exact posterior. In section 4 we present detailed methodology for the case of Gaussian likelihood and linear forward operator. In section 5 we test our methods on two numerical examples: a 1D deblurring example and a 2D super-resolution example. We draw the conclusions in section 6.

2 Certified Coordinate Selection

In this section we first show how we approximate the posterior density and how this approximation can be controlled by an upper bound in the Hellinger distance. The result can then be used to compute a diagnostic $h \in \mathbb{R}^d$ which ranks the coordinates based on their contribution to the update from the prior to the posterior.

2.1 Posterior density approximation

We aim at identifying the set of components in the parameter vector $x \in \mathbb{R}^d$ that are most informed by the data relative to prior information. To this end, we define the coordinate splitting

$$x = (x_{\mathcal{I}}, x_{\mathcal{I}^c}), \tag{4}$$

where the set $\mathcal{I} \subset \{1, \dots, d\}$ contains the indices of the informed coordinates, and $\mathcal{I}^c = \{1, \dots, d\} \setminus \mathcal{I}$ includes the complementary indices. We refer to $x_{\mathcal{I}}$ as *selected coordinates*. Notice that if the likelihood is almost constant in $x_{\mathcal{I}^c}$, the update from prior to posterior happens mainly on $x_{\mathcal{I}}$.

To formalize this idea, let us introduce the posterior approximation

$$\widetilde{\pi}(x) = \pi(x_{\mathcal{I}})\pi_0(x_{\mathcal{I}^c}|x_{\mathcal{I}}), \tag{5}$$

where $\pi(x_{\mathcal{I}})$ is the posterior marginal and $\pi_0(x_{\mathcal{I}^c}|x_{\mathcal{I}})$ is the conditional prior. Compared to the exact posterior, which can be factorized as $\pi(x) = \pi(x_{\mathcal{I}})\pi(x_{\mathcal{I}^c}|x_{\mathcal{I}})$, the approximation $\widetilde{\pi}(x)$ essentially consists in replacing the conditional posterior $\pi(x_{\mathcal{I}^c}|x_{\mathcal{I}})$ with the conditional prior $\pi_0(x_{\mathcal{I}^c}|x_{\mathcal{I}})$. Combining (5) and (2), we can write

$$\widetilde{\pi}^\dagger(x) \propto \widetilde{\mathcal{L}}^\dagger(x_{\mathcal{I}})\pi_0(x), \tag{6}$$

where

$$\widetilde{\mathcal{L}}^\dagger(x_{\mathcal{I}}) = \int \mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c})\pi_0(x_{\mathcal{I}^c})\mathrm{d}x_{\mathcal{I}^c}. \tag{7}$$

Now we use the Hellinger distance $H(\cdot,\cdot)$ to certify the quality of the posterior approximation. It is defined as

$$H(\pi, \widetilde{\pi})^2 = \frac{1}{2}\int_{\mathbb{R}^d} \left(\sqrt{\pi(x)} - \sqrt{\widetilde{\pi}(x)}\right)^2 \mathrm{d}x. \tag{8}$$

The following proposition shows that for a given coordinate splitting the *reduced likelihood* $\widetilde{\mathcal{L}}^\dagger$ from (7) is quasi-optimal with respect to the Hellinger distance.

Proposition 1. *Let $\pi(x) \propto \mathcal{L}(x)\pi_0(x)$ be a probability density on $\mathbb{R}^d$ where $\pi_0(x)$ is a product-form density, and let $x = (x_{\mathcal{I}}, x_{\mathcal{I}^c})$ be any coordinate splitting. Then, the function $\widetilde{\mathcal{L}} : \mathbb{R}^{|\mathcal{I}|} \to \mathbb{R}_+$ which minimizes $H(\pi, \widetilde{\pi})$, where $\widetilde{\pi}(x) \propto \widetilde{\mathcal{L}}(x_{\mathcal{I}})\pi_0(x)$, is given by*

$$\widetilde{\mathcal{L}}^*(x_{\mathcal{I}}) = \left(\int_{\mathbb{R}^{\mathcal{I}^c}} \sqrt{\mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c})}\pi_0(x_{\mathcal{I}^c})\mathrm{d}x_{\mathcal{I}^c}\right)^2. \tag{9}$$

As a consequence, the function $\widetilde{\mathcal{L}}^\dagger$ defined in (7) is not optimal for the Hellinger distance, but still yields a quasi-optimal approximation in the following sense

$$H\big(\pi, \widetilde{\pi}^\dagger\big)^2 \leq 2H(\pi, \widetilde{\pi}^*)^2\,, \tag{10}$$

where $\widetilde{\pi}^\dagger(x) \propto \widetilde{\mathcal{L}}^\dagger(x_{\mathcal{I}})\pi_0(x)$ *and* $\widetilde{\pi}^*(x) \propto \widetilde{\mathcal{L}}^*(x_{\mathcal{I}})\pi_0(x)$.

Proof. See section A.1. □

2.2 Certifying the approximation

We now provide an upper bound on the Hellinger distance $H\big(\pi, \widetilde{\pi}^\dagger\big)$ for the posterior defined in (3).

Proposition 2. *Consider the probability density defined in (3). Given a coordinate splitting $x = (x_{\mathcal{I}}, x_{\mathcal{I}^c})$, the probability density $\widetilde{\pi}^\dagger(x)$ given in (6) satisfies*

$$H\big(\pi, \widetilde{\pi}^\dagger\big)^2 \leq 4\sum_{i\in\mathcal{I}^c} h_i, \tag{11}$$

where the diagnostic $h \in \mathbb{R}^d$ is given by

$$h_i = \frac{1}{\delta_i^2}\int_{\mathbb{R}^d} (\partial_i \log \mathcal{L}(x))^2 \pi(x)\mathrm{d}x. \tag{12}$$

Proof. See section A.2 □

Remark 1. *In case of non-negativity constraints on x, the analogue to the Laplace prior is the exponential prior. The upper bound on the Hellinger distance is the same for this case, since both distributions share the same Poincaré constant.*

With the diagnostic h, the coordinate splitting can be performed by finding $\mathcal{I}^c$ such that

$$4\sum_{i\in\mathcal{I}^c} h_i \leq \tau, \tag{13}$$

where τ is a given desired precision on the Hellinger distance. As a matter of fact, the set $\mathcal{I}$ contains the indices i associated with the $r(\tau)$ largest components in h.

Notice that the number of selected coordinates $r(\tau)$ can be abnormally large, especially if the bound (11) is loose. In this case, we set $r = \min(r(\tau), r_{\max})$ for a pre-given $r_{\max}$ and we let $\mathcal{I}$ contain the indices of the r largest components in h.

3 Sampling algorithms

In this section, we propose two algorithms for drawing samples from the approximate posterior density and the exact posterior density, respectively.

3.1 Sampling the approximate posterior density

The product-form of the prior allows us to write the optimal approximate posterior as

$$\widetilde{\pi}^*(x) \propto \widetilde{\mathcal{L}}^*(x_{\mathcal{I}})\pi_0(x_{\mathcal{I}})\pi_0(x_{\mathcal{I}^c}) \propto \widetilde{\pi}^*(x_{\mathcal{I}})\pi_0(x_{\mathcal{I}^c}),$$

and thus naturally suggests a simple sampling scheme where the main sampling effort is concentrated on the selected coordinates $x_{\mathcal{I}}$ [19]. The sampling method consists in firstly drawing samples $\{x_{\mathcal{I}}^i\}_{i=1}^N$ from the low-dimensional density $\widetilde{\pi}^*(x_{\mathcal{I}})$ using a MCMC method. Then, for each sample $x_{\mathcal{I}}^{(i)}$, we draw a sample $x_{\mathcal{I}^c}^{(i)}$ from the marginal prior $\pi_0(x_{\mathcal{I}^c})$. In the end, reassembling $x^{(i)} = (x_{\mathcal{I}}^{(i)}, x_{\mathcal{I}^c}^{(i)})$ yields samples from the approximate posterior $\widetilde{\pi}^*(x)$. We summarize this procedure in algorithm 1.

Algorithm 1 Sampling scheme for the approximate posterior density

Require: Number of samples N, index sets $\mathcal{I}$ and $\mathcal{I}^c$.

1: Sample $\{x_{\mathcal{I}}^{(i)}\}_{i=1}^N \sim \widetilde{\pi}^*(x_{\mathcal{I}}) \propto \widetilde{\mathcal{L}}^*(x_{\mathcal{I}})\pi_0(x_{\mathcal{I}})$, where $\widetilde{\mathcal{L}}^*(x_{\mathcal{I}})$ is given in (9).
2: Sample $\{x_{\mathcal{I}^c}^{(i)}\}_{i=1}^N$ directly from the prior.
3: Reassemble $\{x^{(i)}\}_{i=1}^N = \{(x_{\mathcal{I}}^{(i)}, x_{\mathcal{I}^c}^{(i)})\}_{i=1}^N$.
4: Return $\{x^{(i)}\}_{i=1}^N$.

In practice, the optimal reduced likelihood $\widetilde{\mathcal{L}}^*(x_{\mathcal{I}})$ (9) must be approximated to enable sampling from $\widetilde{\pi}^*(x_{\mathcal{I}})$. Since we expect the likelihood to be mostly flat in the directions of not selected coordinates, a natural approach is to fix $x_{\mathcal{I}^c}$ in (9) to the prior mean $\mu_0 = 0$. Then, we obtain the approximation $\widetilde{\mathcal{L}}^*(x_{\mathcal{I}}) \approx \mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c} = 0)$, which is computationally cheap while giving satisfactory results as has been shown in the numerical examples of [11].

3.2 Sampling the exact posterior via the pseudo-marginal MCMC algorithm

The approximate marginal posterior $\widetilde{\pi}^\dagger(x_\mathcal{I}) = \widetilde{\mathcal{L}}^\dagger(x_\mathcal{I})\pi_0(x_\mathcal{I})$ satisfies $\widetilde{\pi}^\dagger(x_\mathcal{I}) = \pi(x_\mathcal{I})$. Using this fact, a pseudo-marginal MCMC algorithm [19, 20] can be employed, which in combination with a so-called recycling step, samples the exact posterior. Note that the bound in (2) allows us to control the error of this quasi-optimal posterior approximation and therefore provides theoretical justification for using this sampling algorithm.

We outline the pseudo-marginal MCMC algorithm for the i-th iteration in the following. Given the state $x^{(i-1)} = (x_\mathcal{I}^{(i-1)}, x_{\mathcal{I}^c}^{(i-1)})$, a candidate $z_\mathcal{I}^{(i)}$ is drawn from a proposal distribution $q(\cdot|x_\mathcal{I}^{(i-1)})$, which targets $\widetilde{\pi}^\dagger(x_\mathcal{I})$. Then, the reduced likelihood $\widetilde{\mathcal{L}}^\dagger(z_\mathcal{I}^{(i)})$ is approximated with M freshly drawn samples $\{z_{\mathcal{I}^c}^{(i,j)}\}_{j=1}^M \sim \pi_0(z_{\mathcal{I}^c})$ as

$$\widetilde{\mathcal{L}}^\dagger(z_\mathcal{I}^{(i)}) \approx \frac{1}{M}\sum_{j=1}^{M} \mathcal{L}(z_\mathcal{I}^{(i)}, z_{\mathcal{I}^c}^{(i,j)}). \tag{14}$$

Thus, we obtain a set of candidate samples $\{z_\mathcal{I}^{(i)}, \{z_{\mathcal{I}^c}^{(i,j)}\}_{j=1}^M\}$, which is accepted with probability

$$\alpha = \min\left\{1, \frac{\pi_0(z_\mathcal{I}^{(i)})\widetilde{\mathcal{L}}^\dagger(z_\mathcal{I}^{(i)})q(x_\mathcal{I}^{(i-1)}|z_\mathcal{I}^{(i)})}{\pi_0(x_\mathcal{I}^{(i-1)})\widetilde{\mathcal{L}}^\dagger(x_\mathcal{I}^{(i-1)})q(z_\mathcal{I}^{(i)}|x_\mathcal{I}^{(i-1)})}\right\}. \tag{15}$$

At this point, the samples $x_\mathcal{I}^{(i)}$ follow $\widetilde{\pi}^\dagger(x_\mathcal{I}) = \pi(x_\mathcal{I})$. Now to obtain samples $x_{\mathcal{I}^c}^{(i)}$ from $\pi(x_{\mathcal{I}^c})$, we can use the following recycling step. We select $x_{\mathcal{I}^c}^{(i)}$ from $\{z_{\mathcal{I}^c}^{(i,j)}\}_{j=1}^M$ according to the discrete probability

$$\begin{aligned}\mathbb{P}\left(X_{\mathcal{I}^c}^{(i)} = x_{\mathcal{I}^c}^{(i,j)} | x_\mathcal{I}^{(i)}, \{x_{\mathcal{I}^c}^{(i,j)}\}_{j=1}^M\right) \\ = \frac{\mathcal{L}(x_\mathcal{I}^{(i)}, x_{\mathcal{I}^c}^{(i,j)})}{\sum_{j=1}^M \mathcal{L}\left(x_\mathcal{I}^{(i)}, x_{\mathcal{I}^c}^{(i,j)}\right)}.\end{aligned} \tag{16}$$

We summarize the pseudo-marginal MCMC algorithm in algorithm 2.

Algorithm 2 Pseudo-marginal MCMC

Require: Number of samples N, index sets $\mathcal{I}$ and $\mathcal{I}^c$, number of samples M for (14), initial state $\{x_\mathcal{I}^{(0)}, \{x_{\mathcal{I}^c}^{(0,j)}\}_{j=1}^M\}$, proposal density $q(\cdot|x_\mathcal{I})$.

for $i = 1, \cdots, N$ **do**
  Draw $z_\mathcal{I} \sim q(\cdot|x_\mathcal{I}^{(i-1)})$.
  Draw M i.i.d. samples $z_{\mathcal{I}^c}^{(j)} \sim \pi_0(\cdot)$.
  Compute $\widetilde{\mathcal{L}}^\dagger(z_\mathcal{I})$ via (14).
  Set $\{x_\mathcal{I}^{(i)}, \{x_{\mathcal{I}^c}^{(i,j)}\}_{j=1}^M\} = \{z_\mathcal{I}, \{z_{\mathcal{I}^c}^{(j)}\}_{j=1}^M\}$ with acceptance probability α (15).
end for
Return Markov chain $\{x_\mathcal{I}^{(i)}, \{x_{\mathcal{I}^c}^{(i,j)}\}_{j=1}^M\}_{i=1}^N$.
for $i = 1, \cdots, N$ **do** ▷ Recycling step
  Set $x_{\mathcal{I}^c}^{(i)} = x_{\mathcal{I}^c}^{(i,j)}$ with probability (16).
  Reassemble $x^{(i)} = (x_\mathcal{I}^{(i)}, x_{\mathcal{I}^c}^{(i)})$.
end for
Return Markov chain $\{x^{(i)}\}_{i=1}^N$.

4 Methodology for Gaussian likelihood and linear forward operator

In this section we describe a detailed application of the certified coordinate selection method for a posterior density in the form

$$\pi(x) \propto \exp\left(-\frac{1}{2}\|y - Ax\|_{\Sigma_{\text{obs}}^{-1}}^2 - \sum_{i=1}^{d}\delta_i|x_i|\right), \tag{17}$$

where the noise follows the Gaussian distribution $\mathcal{N}(0, \Sigma_{\text{obs}})$.

Note that to obtain h in (12) we need to compute an expectation over the posterior density. The next lemma shows that, for a linear forward operator with Gaussian likelihood, the diagnostic h admits a closed-form expression involving only the posterior mean and the posterior covariance.

Lemma 3. *Let $\mathcal{L}(x) \propto \exp(-\frac{1}{2}\|y - Ax\|_{\Sigma_{\text{obs}}^{-1}}^2)$ where $\Sigma_{\text{obs}} \in \mathbb{R}^{m\times m}$ is positive definite and assume the mean μ and the covariance Σ of the probability density $\pi(x) \propto \mathcal{L}(x)\pi_0(x)$ exist. Then we can compute the diagnostic h as*

$$\begin{aligned}h = \Lambda(\operatorname{diag}\left(A^\intercal\Sigma_{\text{obs}}^{-1}A\Sigma A^\intercal\Sigma_{\text{obs}}^{-1}A\right) \\ + (A^\intercal\Sigma_{\text{obs}}^{-1}(y - A\mu))^{\circ 2}),\end{aligned} \tag{18}$$

where $(\cdot)^{\circ 2}$ denotes entry-wise square and $\Lambda = \operatorname{diag}\left(1/\delta_1^2, \ldots, 1/\delta_d^2\right)$.

Proof. See section A.2.2. □

If the posterior mean and the posterior covariance are unknown, a first and intuitive choice is to replace them respectively by the prior mean $\mu_0 = 0$ and the prior covariance $\Sigma_0 = 2\Lambda$. This yields

$$\tilde{h}_{\text{prior}} = 2\,\text{diag}\left((\Lambda^{1/2}A^\intercal\Sigma_{\text{obs}}^{-1}A\Lambda^{1/2})^2\right) + \Lambda(A^\intercal\Sigma_{\text{obs}}^{-1}y)^{\circ 2}. \tag{19}$$

In the following, we show how a more precise estimate of the diagnostic compared to the prior-informed estimate of (19) can be obtained. To this end, we employ a Gaussian approximation at the maximum-a-posteriori (MAP) estimate. Note that the negative logarithm of (17) is strictly convex and that even for high-dimensional problems of this form its minimizer, i.e., the MAP-estimate, can be computed efficiently via convex optimization toolboxes.

4.1 MAP-approximated diagnostic

The density in (17) is unimodal and differs only from a Gaussian in that the norm in the prior is l^1 instead of l^2. This motivates us to employ a common strategy in Bayesian inversion where the posterior density is approximated by a Gaussian centered at the MAP-estimate x_{MAP} (e.g., [21]). That is, we estimate the mean as $\mu \approx x_{\text{MAP}}$, and the covariance matrix as

$$\Sigma^{-1} \approx H \coloneqq -\nabla^2 \log \pi(x_{\text{MAP}}). \tag{20}$$

Plugging in x_{MAP} for μ and H^{-1} for Σ in (18) we obtain

$$\tilde{h}_{\text{MAP}} = \Lambda(\text{diag}\left(A^\intercal\Sigma_{\text{obs}}^{-1}AH^{-1}A^\intercal\Sigma_{\text{obs}}^{-1}A\right) + (A^\intercal\Sigma_{\text{obs}}^{-1}(y - Ax_{\text{MAP}}))^{\circ 2}). \tag{21}$$

The non-differentiability of $|\cdot|$ poses an obstacle in computing (20). Inspired by [22], we use the approximation

$$|x| \approx \sqrt{x^2 + \varepsilon}, \tag{22}$$

where we can control the amount of smoothing around 0 with $0 < \varepsilon \ll 1$. With this, we obtain

$$H = A^\intercal\Sigma_{\text{obs}}^{-1}A + \text{diag}\left(\delta_i\varepsilon\left(\sqrt{x_{\text{MAP,i}}{}^2 + \varepsilon}\right)^{-3}\right), \tag{23}$$

where we now use $\text{diag}(\cdot)$ to describe a diagonal matrix with diagonal given by the vector argument.

It remains the question of how to choose ε. It appears natural to choose ε very small to obtain a good approximation to the absolute value. However, we have

$$\delta_i\varepsilon\left(\sqrt{x_{\text{MAP,i}}{}^2 + \varepsilon}\right)^{-3} \overset{x_{\text{MAP,i}}\to 0}{\to} \frac{\delta_i}{\sqrt{\varepsilon}} \overset{\varepsilon\to 0}{\to} \infty.$$

Since we expect $x_{\text{MAP,i}} \approx 0$ for many coordinates, ε must not be chosen too small to avoid fast, nearly non-smooth changes among the elements in H that lead to numerical instabilities when computing its inverse which is needed for (21). Hence, we set ε according to the following heuristic.

Observe that (23) resembles the inverse of the covariance matrix of a Gaussian posterior density constructed by a Gaussian likelihood with a linear forward operator and a Gaussian prior. In this light $\delta_i^{-1}\varepsilon^{-1}(\sqrt{x_{\text{MAP,i}}{}^2 + \varepsilon})^3$ represents the variance of the i-th component. Now our heuristic rule is that these variances should be at least as large as the smallest variance of the chosen Laplace prior. Therefore, we require

$$\min_{x_{\text{MAP,i}}} \|\delta\|_\infty^{-1}\varepsilon^{-1}\left(\sqrt{x_{\text{MAP,i}}{}^2 + \varepsilon}\right)^3 \geq \frac{2}{\|\delta\|_\infty^2}. \tag{24}$$

We can assume that $\min_{i=1,\dots,d} x_{\text{MAP,i}} = 0$ such that we obtain $\varepsilon \geq 4/\|\delta\|_\infty^2$.

4.2 Preconditioned MALA

The approximation in (22) enables employing MALA, since it allows for the computation of approximate gradients. MALA is derived by discretizating a Langevin diffusion equation and steers the sampling process by using gradient information of the log-target density [23]. While different algorithms have been developed to sample non-smooth log-densities (see, e.g., [10]), the proposed smoothing is simple to implement and computationally cheap. Moreover, combined with

the Metropolis step, we obtain convergence to the target density.

Here, we remark that if H (23) is invertible, the inverse H^{-1} can be used as preconditioner to make the MALA proposal more efficient in high dimensions [24, 25]. Furthermore, in the pseudo-marginal MCMC algorithm (algorithm 2), a preconditioner for a local MALA proposal for the update of $x_{\mathcal{I}}$ (line 2) can be obtained by projecting H^{-1} onto the selected coordinates.

5 Numerical experiments

In this section, we illustrate the performance of our methods in two different applications: a 1D deblurring problem and a 2D super-resolution problem. We use the Python package arviz [26] to compute the following sample statistics in our experiments: effective sample size (ESS), $\hat{R}$ (normalized R-hat), and credibility interval (CI) (see, e.g., [21] for definitions). After every MCMC simulation, we consider the maximal normalized R-hat diagnostic over all coordinates (i.e., $\max_i \hat{R}_i$), as an indicator for convergence of the Markov chains.

5.1 1D signal deblurring

The main purpose of this experiment is to demonstrate the applicability of our diagnostic when performing the coordinate splitting. Additionally, we show results when using the pseudo-marginal MCMC algorithm (algorithm 2), and when sampling from the approximate posterior (algorithm 1).

5.1.1 Problem description

The data y is obtained artificially via

$$y = Gs_{\text{true}} + e,$$

where $s_{\text{true}} \in \mathbb{R}^{1024}$ denotes the piece-wise constant ground truth, G is a Gaussian blur operator with the kernel width 27 and standard deviation 3, and $e \in \mathbb{R}^{1024}$ is a realization from $\mathcal{N}(0, \sigma_{\text{obs}}^2)$ with $\sigma_{\text{obs}} = 0.03$. The true signal and the data are shown in fig. 1.

We employ a 10-level Haar wavelet transform with periodic boundary condition and formulate the Bayesian inverse problem in the coefficients domain. Let W and $W^\dagger$ denote the discrete wavelet transform and the inverse discrete wavelet transform, respectively. The true coefficients $x_{\text{true}} = W^\dagger s_{\text{true}}$ are sparse with $\|x_{\text{true}}\|_0 = 60$, see fig. 2. The posterior density formulated with respect to the coefficients reads

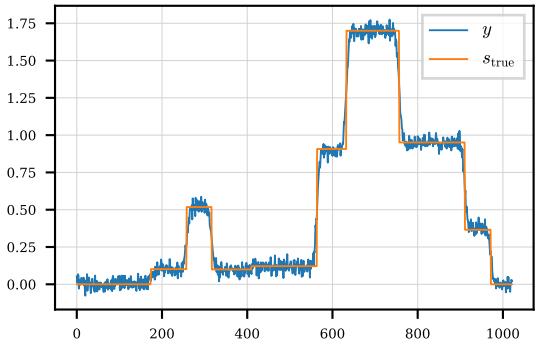

Fig. 1 True signal and data of the 1D example.

$$\pi(x) \propto \exp\left(-\frac{1}{2\sigma_{\text{obs}}^2}\|y - GWx\|_2^2 - \sum_{i=1}^{1024} \delta_i |x_i|\right). \tag{25}$$

Thus, following our previous notation for the forward operator, we have $A = GW$. We use the Python package pywt [27] to compute the discrete wavelet transforms.

To take the different scales of the wavelet coefficients into account, we chose different δ for each level of the wavelet basis. We define

$$\delta_i = c2^{\frac{\ell(i)}{2}}, \tag{26}$$

for some $c > 0$, where $\ell(i) \in \{1, \dots, 10\}$ denotes the level of the of the i-th wavelet coefficient. We plot δ_i for $c = 1$ in fig. 2. Note that for $c = 1$, the prior in (25) on x corresponds to a Besov-$\mathcal{B}^1_{11}$ prior [28, 29] on the signal s.

In the following, we illustrate the performance of the diagnostic and the pseudo-marginal MCMC (algorithm 2). We additionally study the influence of the global parameter c on the posterior by considering the cases $c \in \{1, 5, 25\}$.

5.1.2 Bound on the Hellinger distance

We compute x_{MAP} for δ defined in (26) with $c \in \{1, 5, 25\}$ by using the convex optimization Python package cvxpy [30, 31]. We show Wx_{MAP} in fig. 3 and see that all estimates are close to s_{true}.

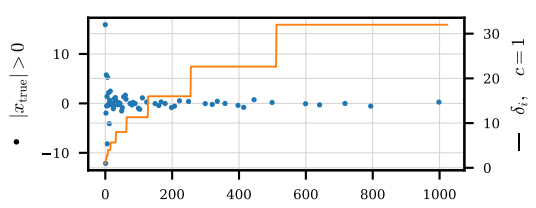

Fig. 2 Scatters (left y-axis): Haar wavelet coefficients of the true signal. Solid line (right y-axis): Rate parameters computed via (26) with $c = 1$.

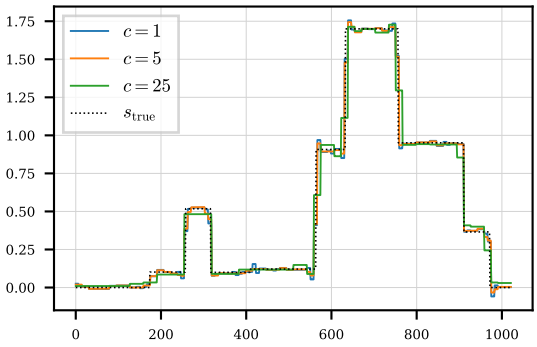

Fig. 3 Black dotted line: True signal. Colored lines: Wx_{MAP} for $c \in \{1, 5, 25\}$.

For the case $c = 5$, we compute a MAP-approximation $\tilde{h}_{\text{MAP}}$ via (21) and a prior-approximation $\tilde{h}_{\text{prior}}$ via (19). We compare the bound (11) obtained via $\tilde{h}_{\text{MAP}}$ and $\tilde{h}_{\text{prior}}$ to a reference solution that we compute via a reference diagnostic $\tilde{h}_{\text{ref}}$. We obtain $\tilde{h}_{\text{ref}}$ via the Monte Carlo approximation

$$\tilde{h}_{\text{ref}} = \frac{1}{N} \sum_{i=1}^{N} \nabla \log \mathcal{L}(x^{(i)})^{\circ 2},$$

where $\{x^{(i)}\}_{i=1}^{N}$ are posterior samples computed with MALA on the full dimension. The samples are obtained from 10 independent chains of 2×10^6 samples with an additional burn-in period of 10^5 samples, and we keep only every 100-th sample to reduce correlation. We obtain $\max_i \hat{R}_i = 1.02$ and a mean ESS of 500 over all chains and coordinates.

The computed bounds are shown in fig. 4. The curves are generated by first sorting the diagnostics in ascending order and then plotting their cumulative sums. The vertical lines indicate the indices of $\{i : x_{\text{true},i} \neq 0\}$.

Figure 4 shows that $\tilde{h}_{\text{MAP}}$ is a good approximation to $\tilde{h}_{\text{ref}}$, and the slight differences can mainly be seen in the vertical lines. The concentration of vertical lines on the right suggests that most of the indices $\{i : x_{\text{true},i} \neq 0\}$ tend to be included in $\mathcal{I}$ in all cases. Recall that the diagnostic reveals the coordinates where the update from prior to posterior information is most evident. Consequently, the remaining vertical lines scattered across the graph correspond to coordinates where the data cannot be distinguished easily from prior information.

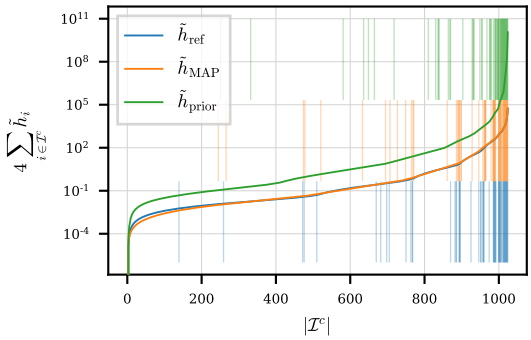

Fig. 4 Upper bounds computed via $\tilde{h}_{\text{MAP}}$, $\tilde{h}_{\text{prior}}$ and $\tilde{h}_{\text{ref}}$ for $c = 5$. The bounds are computed by sorting the diagnostics in ascending order followed by a cumulative sum. The vertical lines indicate the indices $\{i : x_{\text{true},i} \neq 0\}$.

In fig. 5 we show the bound obtained via the MAP- approximated diagnostic for $c \in \{1, 5, 25\}$. As the prior becomes tighter with larger c, the same bound can be retained while more coordinates are included in $\mathcal{I}^c$, and consequently, a more efficient dimension reduction becomes possible.

5.1.3 Sampling the exact posterior

We sample the exact posterior density (25) for $c \in \{1, 5, 25\}$ with MALA on the full dimensional space, as well as with the pseudo-marginal MCMC algorithm (algorithm 2) with MALA-proposals on the selected coordinates. In the following, we refer to these methods as 'MALA' and 'PM-MALA'.

To compute $\nabla \log \pi$ for the MALA proposals, we use an approximated gradient, which is derived from the smoothing in (22). Furthermore, we require the adjoints of W and G for the gradients. We compute W^* with the technique from [32],

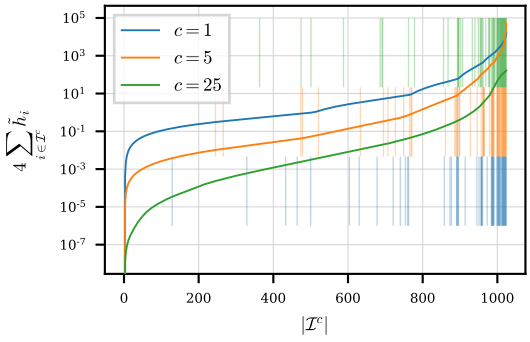

Fig. 5 Upper bounds computed via $\tilde{h}_{\text{MAP}}$ for $c \in \{1, 5, 25\}$. The bounds are computed by sorting the diagnostics in ascending order followed by a cumulative sum. The vertical lines indicate the indices $\{i : x_{\text{true},i} \neq 0\}$.

which involves handling the padding of the boundary conditions manually. Regarding the blurring operator, we have $G^* = G$.

We obtain H^{-1} via (23) and use only its diagonal as preconditioner for the MALA proposals to save computational cost. For PM-MALA, where the MALA-proposals are employed to update $x_{\mathcal{I}}$, we project H^{-1} on the selected coordinates and also use only its diagonal. We sample $M = 5$ vectors of $x_{\mathcal{I}^c}$ in each iteration inspired by the numerical experiments in [19].

We sample 10 independent chains for each configuration with the following settings. We use a burn-in period of 10^5 samples during which we adapt the step size to achieve a fixed acceptance rate. Following [33], we target an acceptance rate of 0.574 for x in the MALA runs and for $x_{\mathcal{I}}$ in the PM-MALA runs. Note that we need to select enough coordinates in order to achieve a stable acceptance rate during the PM-MALA iterations. Based on some pilot runs, we select $n_{\mathcal{I}} \in \{800, 200, 150\}$ for $c \in \{1, 5, 25\}$, respectively. We compute 10^6 samples in each run and save every 50-th sample to decrease correlation.

In fig. 6 we show the 99% CI for $c \in \{1, 5, 25\}$ in the signal space and observe that the CI becomes narrower with increasing c. However, reduced uncertainty due to large c enables more efficient pseudo-marginal MCMC sampling since $\mathcal{I}$ can be chosen smaller with increasing c whilst still obtaining good mixing. This can be seen in table 1, where mixing in terms of ESS can be orders of magnitude larger for PM-MALA than for MALA on the full dimension. In particular, ESS for coordinate indices in $\mathcal{I}^c$ is close to the total amount of samples since the proposals for $x_{\mathcal{I}^c}$ are drawn independently.

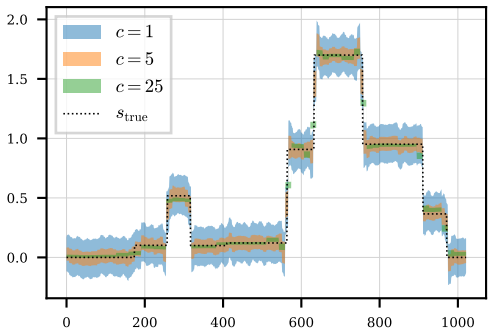

Fig. 6 Sample means and 99% CI for $c \in \{1, 5, 25\}$ in signal space.

Table 1 also shows that running MALA on the reduced dimensional space allows for larger step sizes, which results in improved mixing in $x_{\mathcal{I}}$. We note that PM-MALA takes on average about twice as much time as MALA which, however, is justified by the greatly improved ESS. Moreover, the increasing $\max_i \hat{R}_i$ when using MALA on the full dimension mean that the chains converge slower. On the contrary, all chains of PM-MALA have a $\max_i \hat{R}_i$ close to 1 and thus converge within the sampling time.

5.1.4 Sampling the approximate posterior

In this section we use algorithm 1 to sample the approximated posterior (6) and employ a MALA-proposal to sample $x_{\mathcal{I}}$ in line 1. As outlined in section 3.1, we approximate the optimal reduced likelihood (9) by fixing not selected coordinates to the prior mean such that the approximated posterior reads

$$\widetilde{\pi}(x) \propto \mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c} = 0)\pi_0(x). \tag{27}$$

We choose $c = 5$ and select $|\mathcal{I}| = 400$ coordinates. Hence, we expect $H(\pi, \widetilde{\pi})^2 \leq 2.49 \times 10^{-2}$ according to $\tilde{h}_{\text{ref}}$ in fig. 4. An estimation of the Hellinger distance based on samples from $\widetilde{\pi}$ would allow for assessing the quality of the approximate posterior and for checking the tightness of our

Table 1 Comparison between MALA on the full dimensional space and MALA within the pseudo-marginal MCMC algorithm for the 1D example. $\text{ESS}_{\mathcal{I}}$ and $\text{ESS}_{\mathcal{I}^c}$ are the mean effective sample sizes for all coordinates in $\mathcal{I}$ and in $\mathcal{I}^c$, respectively. $\text{ESS}_{\mathcal{I}}$, $\text{ESS}_{\mathcal{I}^c}$ and wall-clock time are shown as mean $\pm$ standard deviation across the chains.

c	1		5		25	
Method	PM-MALA	MALA	PM-MALA	MALA	PM-MALA	MALA
$\lvert\mathcal{I}\rvert$	800	-	200	-	150	-
$\max_i \hat{R}_i$	1.00	1.00	1.00	1.08	1.01	1.38
$\text{ESS}_{\mathcal{I}}$	1523 ± 166	1139 ± 100	3233 ± 414	132 ± 30	1353 ± 136	64 ± 21
$\text{ESS}_{\mathcal{I}^c}$	19749 ± 28	1521 ± 136	19732 ± 18	287 ± 68	19767 ± 14	111 ± 37
time [min]	58.3 ± 3.1	34.0 ± 1.1	69.4 ± 0.6	34.4 ± 1.0	73.2 ± 2.6	34.1 ± 0.2
step size	1.7×10^{-2}	1.5×10^{-2}	8.9×10^{-3}	4.5×10^{-4}	3.7×10^{-4}	1.7×10^{-5}

bound (11), which we estimate through the MAP-approximated diagnostic. However, computing a numerical estimate of the Hellinger distance based on samples is hard since it tends to be unstable due to the unknown normalizing constants of π and $\widetilde{\pi}$.

Instead, we can obtain a numerical estimate of another bound on the Hellinger distance based on samples from $\widetilde{\pi}$, which is independent of the normalizing constants as

$$H(\pi, \widetilde{\pi})^2 \leq 2 \int \left(\sqrt{\frac{\rho(x)}{\tilde{\rho}(x)}} - 1 \right)^2 \widetilde{\pi}(x) \mathrm{d}x \tag{28}$$

$$\approx \frac{2}{N} \sum_{i=1}^{N} \left(\sqrt{\frac{\rho(x^{(i)})}{\tilde{\rho}(x^{(i)})}} - 1 \right)^2 \quad x^{(i)} \sim \widetilde{\pi}(x), \tag{29}$$

where ρ and $\tilde{\rho}$ are the unnormalized exact and approximated posterior density, respectively. See section A.3 for the derivation of (28). While we can use this bound to assess the quality of the approximate posterior, (29) does not allow for any conclusions on the tightness of our bound (11), which we estimate through the MAP-approximated diagnostic.

As in the previous section, we compute a preconditioner for the MALA-proposals by projecting the diagonal of H^{-1} onto the selected coordinates. We sample again 10 independent chains of 2×10^6 samples and an additional burn-in period of 10^5 samples with adapting step size targeting an acceptance rate of 0.574. We thin the chains to decrease auto-correlation by keeping only every 100-th sample.

We obtain $\max_i \hat{R}_i = 1.00$ and the following mean and standard deviation across the chains for the mean ESS of each chain: 2995 ± 49. These diagnostics suggest converged chains and little correlation. Employing (28) we get the following mean upper bound and standard deviation across the chains: $H(\pi, \widetilde{\pi})^2 \leq 4.73 \times 10^{-2} \pm 5.10 \times 10^{-4}$. These results suggest that the approximate posterior is close to the exact posterior and that the estimate (28) is indeed stable. Further, we note that the bound estimated via $\tilde{h}_{\text{ref}}$, 2.49×10^{-2}, is tighter than the sample-approximated bound.

5.2 2D super-resolution microscopy

The purpose of this experiment is to show that our coordinate selection method works well in high dimensions and that the approximate posterior can be used to perform efficient inference. The test problem is inspired by the application of stochastic optical reconstruction microscopy (STORM) from [34]. A similar example was considered in the Bayesian context in [35]. STORM is a super-resolution microscopy method based on single-molecule stochastic switching, where the goal is to detect molecule positions in live cell imaging. The images are obtained by a microscope detecting the photon count of the (fluorescence) photoactivated molecules.

5.2.1 Problem description

We consider a microscopic image $y \in \mathbb{R}^m$, which is obtained from a 2D pixel-array by concatenation in the usual column-wise fashion. Here, we set $m = 32^2 = 1024$. In STORM, we want to estimate precise molecule positions by computing a super-resolution image $x \in \mathbb{R}^d$. In this example, we set the oversampling ratio $k = 4$, which leads to $d = mk^2 = 16384$. Based on the kernel from the optical measurement instrument given in [34],

we generate the forward operator $A \in \mathbb{R}^{m\times d}$. The data y is obtained via

$$y = Ax_{\text{true}} + e, \tag{30}$$

where $e \in \mathbb{R}^m$ is simulated from $\mathcal{N}(0, \sigma^2_{\text{obs}})$.

Similar as in [34], we generate the ground-truth image x_{true} for the high photon count case with 50 uniformly distributed molecules on a field of size $4\mu\text{m} \times 4\mu\text{m}$. The intensity of each molecule is simulated from a lognormal distribution with mode 3000 and standard deviation 1700. In fig. 7 we show the ground-truth image and the data, which is obtained according to (30) with $\sigma_{\text{obs}} = 30$.

We use a Laplace prior due to the sparse behavior of x_{true}, which leads to the posterior density

$$\pi(x) \propto \exp\left(-\frac{1}{2\sigma^2_{\text{obs}}}\|y - Ax\|_2^2 - \delta\|x\|_1\right). \tag{31}$$

In fig. 7 we also show the MAP-estimate with $\delta = 1.275$, where δ is chosen based on the visual quality after some pilot runs.

5.2.2 Bound on the Hellinger distance

We use the MAP-approximation (21) to estimate the diagnostic and to compute the bound on the Hellinger distance, which we show in the left panel in fig. 8. It is obvious that by using the MAP-approximation we can detect the coordinates of interest very accurately, which may be due to the good quality of the MAP-estimate. Although we obtain very large bounds on the Hellinger distance for this example, we can still employ the diagnostic to detect the most relevant coordinates to perform uncertainty quantification on the molecule positions.

5.2.3 Sampling the approximate posterior and uncertainty quantification

We use the MAP-approximated diagnostic to select 1000 coordinates, which we show in the center of fig. 8. The posterior density is again approximated as in (27). We sample the posterior density by using the No-U-Turn sampler (NUTS) [36] implemented in the Python package pyro [37]. After sampling 10 independent chains with 20000 burn-in samples and 80000 posterior samples for each chain, we obtain converged chains and samples with low correlation with $\max_i \hat{R}_i = 1.01$, and an averaged ESS (over all chains and components) equal to 451.

In this example, we cannot estimate the bound on the Hellinger distance via (28), since the ratio $\frac{\rho(x^{(i)})}{\tilde{\rho}(x^{(i)})}$ computed with our samples $\{x^{(i)}\}_{i=1}^N \sim \widetilde{\pi}(x)$ is unstable. However, as it can be observed from the center figure in fig. 8, our diagnostic is able to select the correct molecule coordinates and the relevant neighbourhoods around them. Therefore, we can still use the samples from $\widetilde{\pi}(x)$ to perform uncertainty quantification on the intensity of the photons and on the true molecule positions as follows.

To illustrate the uncertainty in the intensity, we plot the 99% CI for the selected molecules in the right figure of fig. 8. The large ranges in CI can be contributed to the large ranges in photon intensity. Further, we observe that the approximated posterior tends to have larger CI at the true molecule positions, which are marked in red.

Now we estimate the uncertainty in the true molecule positions in the super-resolution grid by applying the following procedure. We select the coordinates of the 50 largest posterior means as the detected molecule positions. Then, we displace each identified molecule vertically and horizontally until its posterior mean leaves the 99% CI. On average, we obtain an uncertainty of the molecule positions of ± 3.82 pixels, corresponding to 118.4nm, in both horizontal and vertical direction. We note that this result is in agreement with the results in [34].

6 Conclusions

We outlined a coordinate selection method for high-dimensional Bayesian inverse problems with product-form Laplace prior. Inspired by the CDR methodology, we defined an approximate posterior density by replacing the likelihood with a ridge approximation. The ridge approximation is constructed such that it varies mainly on the coordinates which contribute mostly to the update from the prior to the posterior. Based on a bound in the Hellinger distance between the exact and approximate posterior density, we then derived a

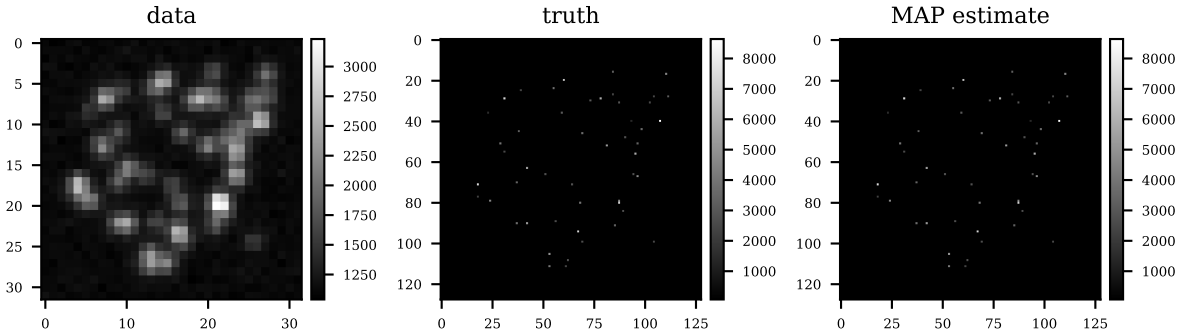

Fig. 7 Left: Data computed via (30). Center: Truth in super-resolution. Right: MAP-estimate.

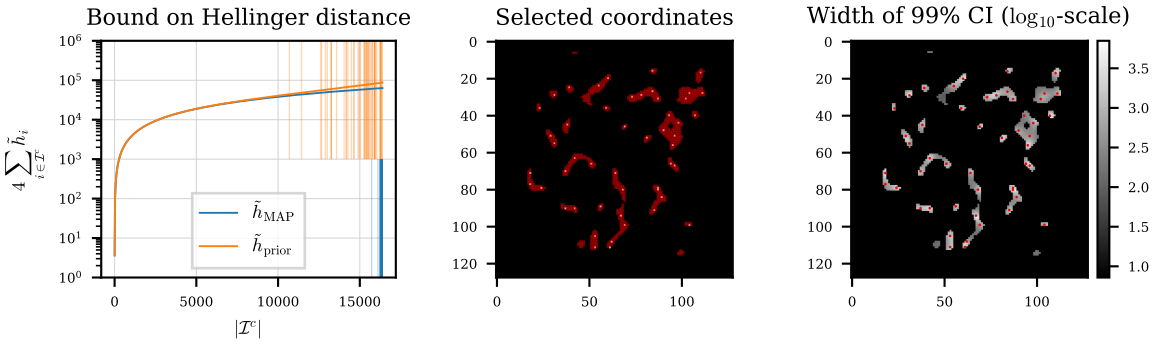

Fig. 8 Left: Upper bounds computed via $\tilde{h}_{\text{MAP}}$ and $\tilde{h}_{\text{prior}}$. The bounds are computed by sorting the diagnostics in ascending order followed by a cumulative sum. The vertical lines indicate the indices of the molecule positions. Center: True molecule positions (white scatters) and 1000 selected indices (red). Right: 99% CI in $\log_{10}$-scale for better visibility and true molecule positions (red scatters).

diagnostic vector $h \in \mathbb{R}^d$, which can be used to select the important coordinates.

After performing the coordinate selection, it is relatively easy to sample the approximate posterior density. An additional advantage of our coordinate splitting is that advanced MCMC algorithms, such as delayed acceptance MCMC or pseudo-marginal MCMC, can be employed to sample the exact posterior.

The computation of h involves, however, integrating over the posterior density. For the case of a linear forward operator with additive Gaussian error, we presented a tractable methodology for estimating the diagnostic h before performing Bayesian inference via, e.g., MCMC methods.

The numerical results indicate that our methodology, which estimates the diagnostic based on a MAP-estimate, succeeds in revealing the most important coordinates. This enabled us to sample the approximate posterior very efficiently. Furthermore, the coordinate splitting allowed us to employ the pseudo-marginal MCMC algorithm to sample the exact posterior density. Here our results show that the pseudo-marginal MCMC algorithm with MALA-proposals on the selected coordinates performs significantly better in terms of convergence and correlation of the sample chains when compared to MALA on the full-dimensional space.

Our methodology for estimating the diagnostic based on a MAP-estimate hinges on a smoothing approximation of the prior. This introduces an additional parameter ε which we fix following a heuristic rule. However, other ways for estimating the diagnostic not only in the linear-Gaussian case, but also for more general problems with non-linear forward operator and/or non-Gaussian likelihood should be investigated.

Moreover, we approximate the optimal reduced likelihood by setting the non-selected coordinates to zero (the prior mean). While this approximation yields good results in the first example, the approximation deteriorates in the second high-dimensional example. Therefore, better approximations to the optimal reduced likelihood should be explored as well.

7 Acknowledgments

R.F. and Y.D. were supported by a Villum Investigator grant (no. 25893) from The Villum Foundation, and F.U. was supported by the Research Council of Finland (project number 353095).

Appendix A Proofs

A.1 Proposition 1

We begin by introducing the normalizing constants for π, $\widetilde{\pi}$ and $\widetilde{\pi}^*$:

$$\begin{aligned}
Z &= \int_{\mathbb{R}_{\mathcal{I}^c}} \int_{\mathbb{R}_{\mathcal{I}}} \mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c}) \pi_0(x_{\mathcal{I}}, x_{\mathcal{I}^c}) \mathrm{d}x_{\mathcal{I}} \mathrm{d}x_{\mathcal{I}^c} \\
\widetilde{Z} &= \int_{\mathbb{R}_{\mathcal{I}^c}} \int_{\mathbb{R}_{\mathcal{I}}} \widetilde{\mathcal{L}}(x_{\mathcal{I}}) \pi_0(x_{\mathcal{I}}, x_{\mathcal{I}^c}) \mathrm{d}x_{\mathcal{I}} \mathrm{d}x_{\mathcal{I}^c} \\
&= \int_{\mathbb{R}_{\mathcal{I}}} \widetilde{\mathcal{L}}(x_{\mathcal{I}}) \pi_0(x_{\mathcal{I}}) \mathrm{d}x_{\mathcal{I}} \\
\widetilde{Z}^* &= \int_{\mathbb{R}_{\mathcal{I}}} \widetilde{\mathcal{L}}^*(x_{\mathcal{I}}) \pi_0(x_{\mathcal{I}}) \mathrm{d}x_{\mathcal{I}}
\end{aligned}$$

where we have used that the prior has product-form. Recall that for a coordinate splitting $x = (x_{\mathcal{I}}, x_{\mathcal{I}^c})$ we want to control the approximation $\widetilde{\pi}(x) \propto \widetilde{\mathcal{L}}(x_{\mathcal{I}}) \pi_0(x)$ for $\pi(x) \propto \mathcal{L}(x) \pi_0(x)$ in the Hellinger distance.

We split the remaining proof into two parts. In the first part we will show that the choice

$$\widetilde{\mathcal{L}}^*(x_{\mathcal{I}}) = \left(\int_{\mathbb{R}_{\mathcal{I}^c}} \sqrt{\mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c})} \pi_0(x_{\mathcal{I}^c}) \mathrm{d}x_{\mathcal{I}^c} \right)^2$$

for $\widetilde{\mathcal{L}}$ minimizes the Hellinger distance and is thus the optimal reduced likelihood function. In the second part we will show that $\widetilde{\mathcal{L}}^\dagger$ is quasi-optimal with respect to the Hellinger distance.

A.1.1 Optimal reduced likelihood function

For an approximated posterior defined by any reduced likelihood function, we can write

$$\begin{aligned}
&H(\pi, \widetilde{\pi})^2 \\
&= \frac{1}{2} \int_{\mathbb{R}^d} \left(\sqrt{\pi(x_{\mathcal{I}}, x_{\mathcal{I}^c})} - \sqrt{\widetilde{\pi}(x_{\mathcal{I}}, x_{\mathcal{I}^c})} \right)^2 \mathrm{d}x \\
&= 1 - \int_{\mathbb{R}^d} \sqrt{\pi(x_{\mathcal{I}}, x_{\mathcal{I}^c}) \widetilde{\pi}(x_{\mathcal{I}}, x_{\mathcal{I}^c})} \mathrm{d}x \\
&= 1 - \frac{1}{\sqrt{Z\widetilde{Z}}} \int_{\mathbb{R}^d} \sqrt{\mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c}) \widetilde{\mathcal{L}}(x_{\mathcal{I}})} \pi_0(x_{\mathcal{I}}, x_{\mathcal{I}^c}) \mathrm{d}x \\
&= 1 - \frac{1}{\sqrt{Z\widetilde{Z}}} \int_{\mathbb{R}_{\mathcal{I}}} \int_{\mathbb{R}_{\mathcal{I}^c}} \sqrt{\mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c})} \pi_0(x_{\mathcal{I}^c}) \mathrm{d}x_{\mathcal{I}^c} \\
&\times \sqrt{\widetilde{\mathcal{L}}(x_{\mathcal{I}})} \pi_0(x_{\mathcal{I}}) \mathrm{d}x_{\mathcal{I}} \\
&= 1 - \frac{1}{\sqrt{Z\widetilde{Z}}} \int_{\mathbb{R}_{\mathcal{I}}} \sqrt{\widetilde{\mathcal{L}}^*(x_{\mathcal{I}}) \widetilde{\mathcal{L}}(x_{\mathcal{I}})} \pi_0(x_{\mathcal{I}}) \mathrm{d}x_{\mathcal{I}} \\
&= 1 - \frac{\sqrt{\widetilde{Z}^*}}{\sqrt{Z}} \int_{\mathbb{R}_{\mathcal{I}}} \sqrt{\frac{\widetilde{\mathcal{L}}^*(x_{\mathcal{I}}) \pi_0(x_{\mathcal{I}})}{\widetilde{Z}^*} \frac{\widetilde{\mathcal{L}}(x_{\mathcal{I}}) \pi_0(x_{\mathcal{I}})}{\widetilde{Z}}} \mathrm{d}x_{\mathcal{I}} \\
&= 1 - \frac{\sqrt{\widetilde{Z}^*}}{\sqrt{Z}} (1 - H(\widetilde{\pi}^*, \widetilde{\pi})^2).
\end{aligned}$$

On the contrary, if we use the optimal reduced likelihood, the Hellinger distance reads

$$\begin{aligned}
&H(\pi, \widetilde{\pi}^*)^2 \\
&= 1 - \frac{1}{\sqrt{Z\widetilde{Z}^*}} \int_{\mathbb{R}^d} \sqrt{\widetilde{\mathcal{L}}^*(x_{\mathcal{I}}) \mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c}) \pi_0(x_{\mathcal{I}}, x_{\mathcal{I}^c})} \mathrm{d}x \\
&= 1 - \frac{1}{\sqrt{Z\widetilde{Z}^*}} \int_{\mathbb{R}_{\mathcal{I}}} \sqrt{\widetilde{\mathcal{L}}^*(x_{\mathcal{I}})} \\
&\times \int_{\mathbb{R}_{\mathcal{I}^c}} \sqrt{\mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c})} \pi_0(x_{\mathcal{I}^c}) \mathrm{d}x_{\mathcal{I}^c} \pi_0(x_{\mathcal{I}}) \mathrm{d}x_{\mathcal{I}}
\end{aligned}$$

$$= 1 - \frac{1}{\sqrt{Z\widetilde{Z}^*}} \int_{\mathbb{R}_{\mathcal{I}}} \widetilde{\mathcal{L}}^*(x_{\mathcal{I}})\pi_0(x_{\mathcal{I}})\mathrm{d}x_{\mathcal{I}}$$
$$= 1 - \frac{\sqrt{\widetilde{Z}^*}}{\sqrt{Z}}.$$

Combining the two results, we obtain

$$H(\pi, \widetilde{\pi})^2 = H(\pi, \widetilde{\pi}^*)^2 + \frac{\sqrt{\widetilde{Z}^*}}{\sqrt{Z}} H(\widetilde{\pi}^*, \widetilde{\pi})^2,$$

which concludes the first part of the proof.

A.1.2 Quasi-optimal reduced likelihood function

For the reduced likelihood $\widetilde{\mathcal{L}}^\dagger(x_{\mathcal{I}})$ we have $\widetilde{Z}^\dagger = Z$. Further, we can write

$$\begin{aligned}
&\mathrm{Var}_{\pi_0(x_{\mathcal{I}^c})}\left(\sqrt{\mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c})}\right) = \\
&= \int_{\mathbb{R}_{\mathcal{I}^c}} \mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c})\pi_0(x_{\mathcal{I}^c})\mathrm{d}x_{\mathcal{I}^c} \\
&- \left(\int_{\mathbb{R}_{\mathcal{I}^c}} \sqrt{\mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c})}\pi_0(x_{\mathcal{I}^c})\right)^2 \\
&= \widetilde{\mathcal{L}}^\dagger(x_{\mathcal{I}}) - \widetilde{\mathcal{L}}^*(x_{\mathcal{I}}),
\end{aligned}$$

which implies $\widetilde{\mathcal{L}}^*(x_{\mathcal{I}}) \leq \widetilde{\mathcal{L}}^\dagger(x_{\mathcal{I}})$ and $\widetilde{Z}^* \leq \widetilde{Z}^\dagger$ (which we already know because $\widetilde{Z}^\dagger = Z$ and from section A.1 we have $0 \leq H(\pi, \widetilde{\pi}^*)^2 = 1 - \frac{\sqrt{\widetilde{Z}^*}}{\sqrt{Z}}$).

With this we can write

$$\begin{aligned}
&H\left(\pi, \widetilde{\pi}^\dagger\right)^2 \\
&= 1 - \int_{\mathbb{R}_{\mathcal{I}}} \int_{\mathbb{R}_{\mathcal{I}^c}} \sqrt{\pi(x_{\mathcal{I}}, x_{\mathcal{I}^c})\widetilde{\pi}^\dagger(x_{\mathcal{I}}, x_{\mathcal{I}^c})}\mathrm{d}x_{\mathcal{I}^c}\mathrm{d}x_{\mathcal{I}} \\
&= 1 - \frac{1}{Z}\int_{\mathbb{R}_{\mathcal{I}}} \int_{\mathbb{R}_{\mathcal{I}^c}} \sqrt{\mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c})}\pi_0(x_{\mathcal{I}^c})\mathrm{d}x_{\mathcal{I}^c} \\
&\times \sqrt{\widetilde{\mathcal{L}}^\dagger(x_{\mathcal{I}})}\pi_0(x_{\mathcal{I}})\mathrm{d}x_{\mathcal{I}} \\
&= 1 - \frac{1}{Z}\int_{\mathbb{R}_{\mathcal{I}}} \sqrt{\widetilde{\mathcal{L}}^*(x_{\mathcal{I}})\widetilde{\mathcal{L}}^\dagger(x_{\mathcal{I}})}\pi_0(x_{\mathcal{I}})\mathrm{d}x_{\mathcal{I}} \\
&\leq 1 - \frac{1}{Z}\int_{\mathbb{R}_{\mathcal{I}}} \sqrt{\widetilde{\mathcal{L}}^*(x_{\mathcal{I}})}\pi_0(x_{\mathcal{I}})\mathrm{d}x_{\mathcal{I}} \\
&= 1 - \frac{\widetilde{Z}^*}{Z} \qquad \text{(A1)} \\
&= \left(1 - \frac{\sqrt{\widetilde{Z}^*}}{\sqrt{Z}}\right)\left(1 + \frac{\sqrt{\widetilde{Z}^*}}{\sqrt{Z}}\right) \\
&\leq 2H(\pi, \widetilde{\pi}^*)^2,
\end{aligned}$$

which concludes the proof.

A.2 Proposition 2

Resuming from (A1) we have

$$\begin{aligned}
&H\left(\pi, \widetilde{\pi}^\dagger\right)^2 \overset{\text{(A1)}}{\leq} \frac{Z - \widetilde{Z}^*}{Z} \\
&= \frac{1}{Z}\int_{\mathbb{R}_{\mathcal{I}}} \left(\widetilde{\mathcal{L}}^\dagger(x_{\mathcal{I}}) - \widetilde{\mathcal{L}}^*(x_{\mathcal{I}})\right)\pi_0(x_{\mathcal{I}})\mathrm{d}x_{\mathcal{I}} \\
&= \frac{1}{Z}\int_{\mathbb{R}_{\mathcal{I}}} \mathrm{Var}_{\pi_0(x_{\mathcal{I}^c})}\left(\sqrt{\mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c})}\right)\pi_0(x_{\mathcal{I}})\mathrm{d}x_{\mathcal{I}}
\end{aligned}$$

According to (6), $\pi_0(x_{\mathcal{I}^c})$ satisfies the Poincaré inequality so that

$$\begin{aligned}
&\mathrm{Var}_{\pi_0(x_{\mathcal{I}^c})}\left(h(x_{\mathcal{I}}, x_{\mathcal{I}^c})\right) \leq \\
&4\int_{\mathbb{R}_{\mathcal{I}^c}} \|\nabla_{x_{\mathcal{I}^c}} h(x_{\mathcal{I}}, x_{\mathcal{I}^c})\|_\Lambda^2 \pi_0(x_{\mathcal{I}^c})\mathrm{d}x_{\mathcal{I}^c},
\end{aligned}$$

where $\Lambda = \mathrm{diag}\left(1/\delta_1^2, 1/\delta_2^2, \ldots, 1/\delta_d^2\right)$. We let $h(x_{\mathcal{I}}, x_{\mathcal{I}^c}) = \sqrt{\mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c})}$, so that

$$\begin{aligned}
&\nabla_{x_{\mathcal{I}^c}} h(x_{\mathcal{I}}, x_{\mathcal{I}^c}) \\
&= \frac{1}{2}\mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c})^{-1/2}\nabla_{x_{\mathcal{I}^c}} \mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c}) \\
&= \frac{1}{2}\mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c})^{1/2}\nabla_{x_{\mathcal{I}^c}} \log \mathcal{L}(x_{\mathcal{I}}, x_{\mathcal{I}^c}).
\end{aligned}$$

Hence, we obtain

$$\begin{aligned}
&H\left(\pi, \widetilde{\pi}^\dagger\right)^2 \\
&\leq \frac{4}{Z}\int_{\mathbb{R}^d} \|\nabla_{x_{\mathcal{I}^c}} \log \mathcal{L}(x)\|_\Lambda^2 \mathcal{L}(x)\pi_0(x)\mathrm{d}x \\
&= 4\int_{\mathbb{R}^d} \|\nabla_{x_{\mathcal{I}^c}} \log \mathcal{L}(x)\|_\Lambda^2 \pi(x)\mathrm{d}x \\
&= 4\sum_{i \in \mathcal{I}^c} h_i,
\end{aligned}$$

$$\text{where} \quad h_i = \frac{1}{\delta_i^2}\int_{\mathbb{R}^d} (\partial_i \log \mathcal{L}(x))^2 \pi(x)\mathrm{d}x. \qquad \text{(A2)}$$

A.2.1 Poincaré inequality

The following proposition is a restatement of proposition 4.4.1 in [38].

Proposition 4. *For the exponential probability density function $\nu(x) = \delta \exp(-\delta x)$ on $\mathbb{R}_+$ with*

rate parameter $\delta > 0$ and any differentiable function $f : \mathbb{R} \to \mathbb{R}$ with $f(0) = 0$, the Poincaré inequality reads

$$\mathrm{Var}_\nu(f) \leq \frac{4}{\delta^2} \mathrm{E}_\nu\left[f'^2\right].$$

Proof. Using $\mathrm{Var}_\nu(f) = \mathrm{E}_\nu\left[f^2\right] - \mathrm{E}_\nu[f]^2$ we can write

$$\begin{aligned}\mathrm{Var}_\nu(f) &\leq \int_{\mathbb{R}_+} f(x)^2 \delta \exp(-\delta x)\mathrm{d}x \\ &= 2\delta \int_{\mathbb{R}_+}\int_0^x f(t)f'(t)\mathrm{d}t \exp(-\delta x)\mathrm{d}x \\ &= 2\int_{\mathbb{R}_+} f(x)f'(x)\delta\exp(-\delta x)\mathrm{d}x.\end{aligned}$$

To go from line 2 to 3, observe that for $h(t) = f(t)f'(t)$ we have

$$\begin{aligned}&\left(\int_0^x h(t)\mathrm{d}t \exp(-\delta x)\right)' \\ &= h(x)\exp(-\delta x) - \delta\int_0^x h(t)\mathrm{d}t\exp(-\delta x) = 0.\end{aligned}$$

After applying a Cauchy-Schwarz inequality, we obtain the desired result. □

Corollary 5. *For the Laplace probability density function $\mu(x) = \frac{\delta}{2}\exp(-\delta|x|)$ with rate parameter $\delta > 0$ on $\mathbb{R}$ and any differentiable function $f : \mathbb{R} \to \mathbb{R}$, the Poincaré inequality reads*

$$\mathrm{Var}_\mu(f) \leq \frac{4}{\delta^2}\mathrm{E}_\mu\left[f'^2\right].$$

Proof. In the following we assume without loss of generality $f(0) = 0$. For any such f we have

$$\begin{aligned}&\mathrm{Var}_\mu(f) \\ &= \int_{\mathbb{R}_-} f^2\frac{\delta}{2}\exp(\delta x)\mathrm{d}x + \int_{\mathbb{R}_+} f^2\frac{\delta}{2}\exp(-\delta x)\mathrm{d}x \\ &= \frac{1}{2}\mathrm{Var}_\nu(f(-x)) + \frac{1}{2}\mathrm{Var}_\nu(f(x)),\end{aligned}$$

where $\nu = \mathrm{Exponential}(\delta)$. Applying the Poincaré inequality for ν yields

$$\begin{aligned}\mathrm{Var}_\mu(f) &\leq \frac{2}{\delta^2}\mathrm{E}_\nu[f'(-x)] + \frac{2}{\delta^2}\mathrm{E}_\nu[f'(x)] \\ &= \frac{4}{\delta^2}\mathrm{E}_\mu[f'(x)].\end{aligned}$$

□

Corollary 6. *For the product-form Laplace probability density function $\mu(x) = \prod_{i=1}^d \delta_i \exp(-\delta_i|x_i|)$ on $\mathbb{R}^d$ with rate parameters $\delta_1, \delta_2, \ldots \delta_d > 0$, the Poincaré inequality*

$$\mathrm{Var}_\mu(f) \leq \sum_{i=1}^d \frac{4}{\delta_i^2}\mathrm{E}_\mu\left[(\partial_i f)^2\right],$$

holds for any differentiable function $f : \mathbb{R}^d \to \mathbb{R}$.

Proof. The result follows directly from the stability under products of Poincaré inequalities, see [38]. For the sake of completeness, we give the proof for $d = 2$ (a simple recursion permits to extend to any $d > 2$).

Let $f : x \mapsto f(x_1, x_2)$ be a differentiable function. Without loss of generality, we assume f is centered $\int f\mathrm{d}\mu = 0$. For $F(x_1) = \int f(x_1, x_2)\mathrm{d}\mu_2(x_2)$ the Poincaré inequality permits to write

$$\begin{aligned}&\mathrm{Var}_\mu(f) \\ &= \int (f(x_1,x_2) - F(x_1))^2\mathrm{d}\mu(x) + \int F(x_1)^2\mathrm{d}\mu_1(x_1) \\ &= \int \mathrm{Var}_{\mu_2}(f(x_1,\cdot))\,\mathrm{d}\mu_1(x_1) + \mathrm{Var}_{\mu_1}(F) \\ &\leq \frac{4}{\delta_2^2}\int(\partial_2 f(x))^2\mathrm{d}\mu(x) + \frac{4}{\delta_1^2}\int(\partial_1 F(x_1))^2\mathrm{d}\mu_1(x_1) \\ &\leq \frac{4}{\delta_2^2}\int(\partial_2 f(x))^2\mathrm{d}\mu(x) + \frac{4}{\delta_1^2}\int(\partial_1 f(x))^2\mathrm{d}\mu(x)\end{aligned}$$

where we used a Jensen's inequality for the last step. This gives the result for $d = 2$. □

A.2.2 Lemma 3

We consider the case of a Gaussian likelihood (for fixed data y) of the form

$$\mathcal{L}(x) \propto \exp\left(-\frac{1}{2}\|y - Ax\|^2_{\Sigma^{-1}_{\mathrm{obs}}}\right).$$

Then, continuing from (A2) we can write

$$h = \mathrm{diag}\left(\int_{\mathbb{R}^d} \Lambda^{1/2}\nabla\log\mathcal{L}(x)\nabla\log\mathcal{L}(x)^\intercal\Lambda^{1/2}\pi(x)\mathrm{d}x\right)$$

$$= \Lambda \operatorname{diag}\left(A^\intercal \Sigma_{\text{obs}}^{-1} \int (y - Ax)(y - Ax)^\intercal \pi(x) \mathrm{d}x \Sigma_{\text{obs}}^{-1} A\right).$$

Given the covariance Σ_{XX} and mean μ_X of the probability density $\pi(x)$ we substitute $z = (y - Ax)$. Then, the mean and covariance of Z read $\mu_Z = y - A\mu_X$ and $\Sigma_{ZZ} = A\Sigma_{XX}A^\intercal$, respectively. Therefore,

$$\begin{aligned} h &= \Lambda \operatorname{diag}\left(A^\intercal \Sigma_{\text{obs}}^{-1} \int_{\mathbb{R}^d} zz^\intercal \pi(z) \mathrm{d}z \Sigma_{\text{obs}}^{-1} A\right) \\ &= \Lambda \operatorname{diag}\left(A^\intercal \Sigma_{\text{obs}}^{-1} (\Sigma_{ZZ} + \mu_Z \mu_Z^\intercal) \Sigma_{\text{obs}}^{-1} A\right) \\ &= \Lambda \operatorname{diag}\left(A^\intercal \Sigma_{\text{obs}}^{-1} (A\Sigma_{XX}A^\intercal \right. \\ &\quad \left. + (y - A\mu_X)(y - A\mu_X)^\intercal) \Sigma_{\text{obs}}^{-1} A\right). \end{aligned}$$

A.3 Numerical estimation of a bound on the Hellinger distance

For two probability densities $\pi = \frac{\rho}{Z}$ and $\widetilde{\pi} = \frac{\tilde{\rho}}{\widetilde{Z}}$ where Z and $\widetilde{Z}$ are the normalization constants, we can write

$$\begin{aligned} &\sqrt{2H(\pi, \widetilde{\pi})} \\ &= \|\sqrt{\pi} - \sqrt{\widetilde{\pi}}\|_{L^2} \\ &= \|\sqrt{\rho/Z} - \sqrt{\tilde{\rho}/\widetilde{Z}}\|_{L^2} \\ &= \frac{1}{\sqrt{Z}} \|\sqrt{\rho} - \sqrt{\tilde{\rho}}\sqrt{Z/\widetilde{Z}} \pm \sqrt{\tilde{\rho}}\|_{L^2} \\ &\leq \frac{1}{\sqrt{Z}} \left(\|\sqrt{\rho} - \sqrt{\tilde{\rho}}\|_{L^2} + |1 - \sqrt{Z/\widetilde{Z}}| \|\sqrt{\tilde{\rho}}\|_{L^2}\right). \end{aligned}$$

Furthermore, we have

$$\|\sqrt{\tilde{\rho}}\|_{L^2}^2 = \int \tilde{\rho} \mathrm{d}x = \tilde{Z},$$

so that we obtain

$$\sqrt{2H(\pi, \widetilde{\pi})} \leq \frac{1}{\sqrt{Z}} \left(\|\sqrt{\rho} - \sqrt{\tilde{\rho}}\|_{L^2} + |\sqrt{\widetilde{Z}} - \sqrt{Z}|\right).$$

In addition,

$$\begin{aligned} \left|\tilde{Z} - Z\right| &= |\int \rho - \tilde{\rho} \mathrm{d}x| \\ &\leq \sqrt{\int (\sqrt{\rho} - \sqrt{\tilde{\rho}})^2 \mathrm{d}x} \sqrt{\int (\sqrt{\rho} + \sqrt{\tilde{\rho}})^2 \mathrm{d}x} \\ &= \|\sqrt{\rho} - \sqrt{\tilde{\rho}}\|_{L^2} \|\sqrt{\rho} + \sqrt{\tilde{\rho}}\|_{L^2} \\ &\leq \|\sqrt{\rho} - \sqrt{\tilde{\rho}}\|_{L^2} (\|\sqrt{\rho}\|_{L^2} + \|\sqrt{\tilde{\rho}}\|_{L^2}) \\ &\leq \|\sqrt{\rho} - \sqrt{\tilde{\rho}}\|_{L^2} (\sqrt{Z} + \sqrt{\widetilde{Z}}). \end{aligned}$$

Because $|\tilde{Z} - Z| = |\sqrt{\widetilde{Z}} - \sqrt{Z}||\sqrt{\widetilde{Z}} + \sqrt{Z}|$ we get

$$|\sqrt{\widetilde{Z}} - \sqrt{Z}| \leq \|\sqrt{\rho} - \sqrt{\tilde{\rho}}\|_{L^2}.$$

So, in the end we have

$$H(\pi, \widetilde{\pi})^2 \leq \frac{2}{Z} \|\sqrt{\rho} - \sqrt{\tilde{\rho}}\|_{L^2}^2.$$

Therefore,

$$\begin{aligned} H(\pi, \widetilde{\pi})^2 &\leq \frac{2}{\widetilde{Z}} \int \left(\sqrt{\rho} - \sqrt{\tilde{\rho}}\right)^2 \mathrm{d}x \\ &= 2 \int \left(\sqrt{\frac{\rho}{\tilde{\rho}}} - 1\right)^2 \widetilde{\pi} \mathrm{d}x \\ &\approx \frac{2}{N} \sum_{i=1}^{N} \left(\sqrt{\frac{\rho(x^{(i)})}{\tilde{\rho}(x^{(i)})}} - 1\right)^2. \end{aligned}$$

References

[1] Cai, X., Pereyra, M., McEwen, J.D.: Uncertainty quantification for radio interferometric imaging – I. Proximal MCMC methods. Monthly Notices of the Royal Astronomical Society **480**(3), 4154–4169 (2018) https://doi.org/10.1093/mnras/sty2004

[2] Elad, M., Milanfar, P., Rubinstein, R.: Analysis versus synthesis in signal priors. Inverse Problems **23**(3), 947–968 (2007) https://doi.org/10.1088/0266-5611/23/3/007

[3] Simoncelli, E.P.: Modeling the joint statistics of images in the wavelet domain. In: Unser, M.A., Aldroubi, A., Laine, A.F. (eds.) SPIE's International Symposium on Optical Science, Engineering, And Instrumentation, Denver, CO, pp. 188–195 (1999). https://doi.org/10.1117/12.366779

[4] Suuronen, J., Soto, T., Chada, N.K., Roininen, L.: Bayesian inversion with α-stable priors. Inverse Problems **39**(10), 105007 (2023)

[5] Hosseini, B.: Well-posed Bayesian inverse problems with infinitely divisible and heavy-tailed prior measures. SIAM/ASA Journal on Uncertainty Quantification **5**(1), 1024–1060 (2017)

[6] Markkanen, M., Roininen, L., Huttunen, J.M., Lasanen, S.: Cauchy difference priors for edge-preserving Bayesian inversion. Journal of Inverse and Ill-posed Problems **27**(2), 225–240 (2019)

[7] Uribe, F., Dong, Y., Hansen, P.C.: Horseshoe Priors for Edge-Preserving Linear Bayesian Inversion. arXiv (2022)

[8] Park, T., Casella, G.: The Bayesian Lasso. Journal of the American Statistical Association **103**(482), 681–686 (2008) https://doi.org/10.1198/016214508000000337

[9] Pereyra, M.: Proximal Markov chain Monte Carlo algorithms. Statistics and Computing **26**(4), 745–760 (2016) https://doi.org/10.1007/s11222-015-9567-4

[10] Lau, T.T.-K., Liu, H., Pock, T.: Non-Log-Concave and Nonsmooth Sampling via Langevin Monte Carlo Algorithms. arXiv (2023)

[11] Zahm, O., Cui, T., Law, K., Spantini, A., Marzouk, Y.: Certified dimension reduction in nonlinear Bayesian inverse problems. Mathematics of Computation **91**(336), 1789–1835 (2022) https://doi.org/10.1090/mcom/3737

[12] Cui, T., Tong, X.T.: A Unified Performance Analysis of Likelihood-Informed Subspace Methods. arXiv (2021)

[13] Li, M.T., Marzouk, Y., Zahm, O.: Principal feature detection via ϕ-sobolev inequalities. arXiv preprint arXiv:2305.06172 (2023)

[14] Ehre, M., Flock, R., Fußeder, M., Papaioannou, I., Straub, D.: Certified dimension reduction for Bayesian updating with the cross-entropy method. SIAM/ASA Journal on Uncertainty Quantification **11**(1), 358–388 (2023)

[15] Chen, P., Ghattas, O.: Projected Stein Variational Gradient Descent

[16] Brennan, M.C., Bigoni, D., Zahm, O., Spantini, A., Marzouk, Y.: Greedy inference with structure-exploiting lazy maps

[17] Uribe, F., Papaioannou, I., Marzouk, Y.M., Straub, D.: Cross-Entropy-Based Importance Sampling with Failure-Informed Dimension Reduction for Rare Event Simulation. arXiv (2020)

[18] Cui, T., Tong, X.T., Zahm, O.: Prior normalization for certified likelihood-informed subspace detection of Bayesian inverse problems. Inverse Problems **38**(12), 124002 (2022) https://doi.org/10.1088/1361-6420/ac9582

[19] Cui, T., Zahm, O.: Data-free likelihood-informed dimension reduction of Bayesian inverse problems. Inverse Problems **37**(4), 045009 (2021) https://doi.org/10.1088/1361-6420/abeafb

[20] Andrieu, C., Roberts, G.O.: The pseudo-marginal approach for efficient Monte Carlo computations (2009)

[21] Murphy, K.P.: Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machine Learning Series. MIT Press, Cambridge, MA (2012)

[22] Vogel, C.R.: Computational Methods for Inverse Problems. Society for Industrial and Applied Mathematics, ??? (2002). https://doi.org/10.1137/1.9780898717570

[23] Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Springer Texts in Statistics. Springer New York, New York, NY (2004). https://doi.org/10.1007/978-1-4757-4145-2

[24] Martin, J., Wilcox, L.C., Burstedde, C., Ghattas, O.: A Stochastic Newton MCMC Method for Large-Scale Statistical Inverse Problems with Application to Seismic Inversion. SIAM Journal on Scientific Computing **34**(3), 1460–1487 (2012) https://doi.org/10.1137/110845598

[25] Petra, N., Martin, J., Stadler, G., Ghattas, O.: A Computational Framework for Infinite-Dimensional Bayesian Inverse Problems, Part II: Stochastic Newton MCMC with Application to Ice Sheet Flow Inverse Problems. SIAM Journal on Scientific Computing **36**(4), 1525–1555 (2014) https://doi.org/10.1137/130934805

[26] Kumar, R., Carroll, C., Hartikainen, A., Martin, O.: ArviZ a unified library for exploratory analysis of Bayesian models in Python. Journal of Open Source Software **4**(33), 1143 (2019) https://doi.org/10.21105/joss.01143

[27] Lee, G.R., Gommers, R., Waselewski, F., Wohlfahrt, K., O’Leary, A.: PyWavelets: A Python package for wavelet analysis. Journal of Open Source Software **4**(36), 1237 (2019) https://doi.org/10.21105/joss.01237

[28] Kolehmainen, V., Lassas, M., Niinimäki, K., Siltanen, S.: Sparsity-promoting bayesian inversion. Inverse Problems **28**, 025005 (2012) https://doi.org/10.1088/0266-5611/28/2/025005

[29] Lassas, M., Siltanen, S.: Discretization-invariant bayesian inversion and besov space priors. Inverse Problems and Imaging **3** (2009) https://doi.org/10.3934/ipi.2009.3.87

[30] Diamond, S., Boyd, S.: CVXPY: A Python-embedded modeling language for convex optimization. Journal of Machine Learning Research **17**(83), 1–5 (2016)

[31] Agrawal, A., Verschueren, R., Diamond, S., Boyd, S.: A rewriting system for convex optimization problems. Journal of Control and Decision **5**(1), 42–60 (2018)

[32] Folberth, J., Becker, S.: Efficient Adjoint Computation for Wavelet and Convolution Operators [Lecture Notes]. IEEE Signal Processing Magazine **33**(6), 135–147 (2016) https://doi.org/10.1109/MSP.2016.2594277

[33] Roberts, G.O., Rosenthal, J.S.: Optimal scaling for various Metropolis-Hastings algorithms. Statistical Science **16**(4) (2001) https://doi.org/10.1214/ss/1015346320

[34] Zhu, L., Zhang, W., Elnatan, D., Huang, B.: Faster STORM using compressed sensing. Nature Methods **9**(7), 721–723 (2012) https://doi.org/10.1038/nmeth.1978

[35] Durmus, A., Moulines, É., Pereyra, M.: Efficient Bayesian Computation by Proximal Markov Chain Monte Carlo: When Langevin Meets Moreau. SIAM Journal on Imaging Sciences **11**(1), 473–506 (2018) https://doi.org/10.1137/16M1108340

[36] Hoffman, M.D., Gelman, A.: The No-U-turn sampler

[37] Bingham, E., Chen, J.P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., Goodman, N.D.: Pyro: Deep Universal Probabilistic Programming. arXiv (2018)

[38] Bakry, D., Gentil, I., Ledoux, M.: Analysis and Geometry of Markov Diffusion Operators. Grundlehren Der Mathematischen Wissenschaften, vol. 348. Springer International Publishing, Cham (2014). https://doi.org/10.1007/978-3-319-00227-9