Data-Free Likelihood-Informed Dimension Reduction of Bayesian Inverse Problems

Tiangang Cui, Olivier Zahm

Data-Free Likelihood-Informed Dimension Reduction of Bayesian Inverse Problems

Tiangang Cui[1], **Olivier Zahm**[2]

[1]School of Mathematics, Monash University, VIC 3800, Australia
[2]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

E-mail: tiangang.cui@monash.edu, olivier.zahm@inria.fr

Abstract. Identifying a low-dimensional informed parameter subspace offers a viable path to alleviating the dimensionality challenge in the sampled-based solution to large-scale Bayesian inverse problems. This paper introduces a novel gradient-based dimension reduction method in which the informed subspace does not depend on the data. This permits online-offline computational strategy where the expensive low-dimensional structure of the problem is detected in an offline phase, meaning before observing the data. This strategy is particularly relevant for multiple inversion problems as the same informed subspace can be reused. The proposed approach allows to control the approximation error (in expectation over the data) of the posterior distribution. We also present sampling strategies which exploit the informed subspace to draw efficiently samples from the exact posterior distribution. The method is successfully illustrated on two numerical examples: a PDE-based inverse problem with a Gaussian process prior and a tomography problem with Poisson data and a Besov-$\mathcal{B}^2_{11}$ prior.

1. Introduction

The Bayesian approach to inverse problems builds a probabilistic representation of the parameter of interest conditioned on observed data. Denoting the parameter and data by $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^m$, respectively, the solution to the inverse problem is encapsulated in the posterior distribution, which has the probability density function (pdf)

$$\pi^y_{\text{pos}}(x) = \frac{1}{\pi_{\text{data}}(y)} \mathcal{L}^y(x) \pi_{\text{pr}}(x), \tag{1}$$

where $\pi_{\text{pr}}(x)$ denotes the prior density, $\mathcal{L}^y(x)$ is the likelihood function, and $\pi_{\text{data}}(y)$ is the marginal density of the data y that can be expressed as

$$\pi_{\text{data}}(y) = \int_{\mathbb{R}^d} \mathcal{L}^y(x) \pi_{\text{pr}}(x) \, \mathrm{d}x. \tag{2}$$

This way, one can encode the posterior into summary statistics, for example, moments, quantiles, or probabilities of some events of interest [31, 51, 52], to provide parameter inference and associated uncertainty quantification.

In practice, computing these summary statistics requires dedicated methods to efficiently characterize the posterior distribution. Markov chain Monte Carlo (MCMC) methods [9, 38],

originating with the Metropolis-Hastings algorithm [29, 41], and sequential Monte Carlo methods [22] have been developed as workhorses in this context. However, many inverse problems have high-dimensional or infinite-dimensional parameter space, which present a significant hurdle to the applicability of MCMC, SMC, and other related sampling methods in general.

The efficiency of these sampling methods, measured by the required number of posterior density evaluations, may deteriorate with the dimension of the parameter space, see [47, 48] and references therein. Even with the rather strong log-concave assumption, start-of-the-art MCMC methods can still be sensitive to the dimension of the problem, see for instance [20, 24, 25].

One promising way to alleviating the challenge of dimensionality is to exploit the effectively low-dimensional structures of the posterior distribution. Such low-dimensional structures can be used to construct certified low-dimensional approximations of the posterior distribution [50, 54] and efficient MCMC proposals that are robust in the parameter dimension [5, 6, 17, 16, 33]. There exists several ways to detect low-dimensional structures. A widely accepted method is to utilize the regularity of the prior, in which the dominant eigenvectors of the prior covariance operator [40] can be used to define such a low-dimensional subspace. This prior-based dimension reduction also plays a key role in the analysis of high-dimensional integration methods such as [26, 27]. In addition to the prior regularity, the limited accuracy of the observations and the ill-posed nature of the forward model often allow one to express the posterior as a low-dimensional update from the prior. Methods such as the active subspace (AS) [13, 14] and the likelihood-informed subspace (LIS) [18, 19, 54] utilize gradients of the forward model and/or of the likelihood function in order to better identify the low-dimensional structure of the problem. We refer to [19, 54] for an overview and a comparison of the existing methods.

The success of AS and LIS relies on the computation of the gradient or the Hessian of the log-likelihood function. Since the likelihood function depends on the observed data, the resulting subspaces need to be reconstructed each time a new data set is observed. This can add a significant computational burden to the solution of inverse problems. In this paper, we present a new *data-free* strategy for constructing the informed subspace in which the computationally costly subspace construction can be performed in an *offline phase*, meaning before observing any data sets. In the subsequent *online phase*, the data set is observed and the precomputed informed subspace is utilized to accelerate the inversion process. This computational strategy is particularly relevant for real-time systems such as medical imaging where multiple inversions are needed.

The rest of the paper is organized as follows. To begin, we introduce the problem setting in Section 2. In Section 3, we present a new data-free likelihood-informed approach to construct the subspace. Denoting the Fisher information matrix of the likelihood function by $\mathcal{I}(x) = \int (\nabla \log \mathcal{L}^y(x))(\nabla \log \mathcal{L}^y(x))^\top \mathcal{L}^y(x) \mathrm{d}y$, this approach defines the informed subspace as the rank-r dominant eigenspace of the matrix

$$H = \int \mathcal{I}(x)\, \pi_{\mathrm{pr}}(x) \mathrm{d}x, \tag{3}$$

with $r \ll d$. This definition makes no particular assumption on the likelihood function, so it can be applied to a wide range of measurement processes, e.g., Gaussian likelihood and Poisson likelihood. It also does not involve any particular data set y, and hence can be constructed offline. Given the informed subspace, we approximate the posterior density $\pi^y_{\mathrm{pos}}(x)$ by

$$\widetilde{\pi}^y_{\mathrm{pos}}(x) = \widetilde{\pi}^y_{\mathrm{pos}}(x_r) \pi_{\mathrm{pr}}(x_\perp | x_r), \tag{4}$$

where x_r and $x_\perp$ denote respectively the informed and the non-informed components of x. We prove that the expected Kullback-Leibler (KL) divergence of the full posterior from its approximation is

bounded as

$$\mathbb{E}[D_{\mathrm{KL}}(\pi^Y_{\mathrm{pos}}||\widetilde{\pi}^Y_{\mathrm{pos}})] \leq \kappa \sum_{i>r} \lambda_i(H), \tag{5}$$

where the expectation is taken over the data $Y \sim \pi_{\mathrm{data}}(y)$, κ being the subspace Poincaré constant of the prior [53, 54] and $\lambda_i(H)$ the i-th largest eigenvalue of H. This way, a problem with a fast decay in the spectrum of H yields an accurate low-dimensional posterior approximation in expectation over the data.

In Section 4, we restrict the analysis to Gaussian likelihood. In this case, we show that the vector-valued extension [53] of the AS method [12], which reduces parameter dimensions via approximating forward models, also leads to the same data-free informed subspace as that obtained using (3). We can further show that, although the likelihood-informed approach and AS employ different approximations to the posterior density, the resulting approximations share the same structure as shown in (4) and follow the same error bound as in (5).

As suggested by (4), the factorized form of the approximate posterior densities allows for dimension-robust sampling. One can explore the low-dimensional intractable parameter reduced posterior $\widetilde{\pi}^y_{\mathrm{pos}}(x_r)$ using methods such as MCMC, followed by direct sampling of the high-dimensional but tractable conditional prior $\pi_{\mathrm{pr}}(x_\perp|x_r)$. This strategy has been previously investigated, see [18, 54] and references therein. We provide a brief summary to this existing sampling strategy in Section 5. Despite the accelerated sampling offered by the informed subspace, the resulting inference results are subject to the dimension truncation error that is bounded in (5). In Section 6, by integrating the pseudo-marginal approach [1, 2] and the surrogate transition approach [11, 38, 39] into the abovementioned sampling strategy, we present new *exact inference* algorithms that can enjoy the same subspace acceleration while target on the full posterior. Our exact inference algorithms only require minor modifications to the sampling strategy of [18, 54].

While our dimension reduction method readily apply for Gaussian priors, its application to non-Gaussian priors might not be straightforward. In Section 7, we show how to use the propose method for problems with Besov priors [21, 32, 34] which are commonly used in image reconstruction problems.

We demonstrate the accuracy of the proposed data-free LIS and the efficiency of new sampling strategies on a range of problems. These include the identification of the diffusion coefficient of a two-dimensional elliptic partial differential equation (PDE) with a Gaussian prior in Section 8 and Positron emission tomography (PET) with Poisson data and a Besov prior in Section 9.

2. Problem setting

For high-dimensional ill-posed inverse problems, the data are often informative only along a few directions in the parameter space. To detect and exploit this low-dimensional structure, we introduce a projector $P_r \in \mathbb{R}^{d\times d}$ of rank $r \ll d$ such that $\mathrm{Im}(P_r)$ is the informed subspace and $\mathrm{Ker}(P_r)$ the non-informed one. This splits the parameter space as

$$\mathbb{R}^d = \mathrm{Im}(P_r) \oplus \mathrm{Ker}(P_r),$$

where the subspaces $\mathrm{Im}(P_r)$ and $\mathrm{Ker}(P_r)$ are not necessarily orthogonal unless P_r is orthogonal. The fact that the data are only informative in $\mathrm{Im}(P_r)$ means there exists an approximation to the posterior density $\pi^y_{\mathrm{pos}}(x) \propto \mathcal{L}^y(x)\pi_{\mathrm{pr}}(x)$ under the form

$$\widetilde{\pi}^y_{\mathrm{pos}}(x) \propto \widetilde{\mathcal{L}}^y(P_r x)\pi_{\mathrm{pr}}(x), \tag{6}$$

in which the likelihood function $x \mapsto \mathcal{L}^y(x)$ is replaced by a ridge function $x \mapsto \widetilde{\mathcal{L}}^y(P_r x)$. A ridge function [46] is a function which is constant on a subspace, here $\mathrm{Ker}(P_r)$. Let $x_r = P_r x$ and $x_\perp = (I_d - P_r)x$ be the components of x in $\mathrm{Im}(P_r)$ and $\mathrm{Ker}(P_r)$, respectively. We have the parameter decomposition

$$x = x_r + x_\perp.$$

Using a slight abuse of notation, we factorize the prior density as $\pi_{\mathrm{pr}}(x) = \pi_{\mathrm{pr}}(x_r)\pi_{\mathrm{pr}}(x_\perp|x_r)$, where

$$\pi_{\mathrm{pr}}(x_r) = \int_{\mathrm{Ker}(P_r)} \pi_{\mathrm{pr}}(x_r + x'_\perp)\mathrm{d}x'_\perp \quad \text{and} \quad \pi_{\mathrm{pr}}(x_\perp|x_r) = \pi_{\mathrm{pr}}(x_r + x_\perp)/\pi_{\mathrm{pr}}(x_r)$$

denote the marginal prior and the conditional prior. The approximate posterior (6) writes

$$\widetilde{\pi}^y_{\mathrm{pos}}(x_r + x_\perp) \propto \underbrace{\left(\widetilde{\mathcal{L}}^y(x_r)\pi_{\mathrm{pr}}(x_r)\right)}_{\widetilde{\pi}^y_{\mathrm{pos}}(x_r)} \pi_{\mathrm{pr}}(x_\perp|x_r).$$

This factorization shows that, under the approximate posterior density, the Bayesian update is effective on the informed subspace $\mathrm{Im}(P_r)$ (first term $\widetilde{\pi}^y_{\mathrm{pos}}(x_r)$), while the non-informed subspace $\mathrm{Ker}(P_r)$ is characterized by the prior (second term $\pi_{\mathrm{pr}}(x_\perp|x_r)$). This property will be exploited later on to design efficient sampling strategies for exploring both the approximate posterior and the full posterior.

The challenge of dimension reduction is to construct both the low-rank projector P_r and the ridge approximation $\widetilde{\mathcal{L}}^y$ such that the KL divergence of the full posterior from its approximation

$$D_{\mathrm{KL}}(\pi^y_{\mathrm{pos}}||\widetilde{\pi}^y_{\mathrm{pos}}) = \int \log\left(\frac{\pi^y_{\mathrm{pos}}(x)}{\widetilde{\pi}^y_{\mathrm{pos}}(x)}\right)\pi^y_{\mathrm{pos}}(x)\mathrm{d}x,$$

can be controlled. In this work, we specifically focus on constructing a projector P_r which is independent on the data y and which allows to bound $D_{\mathrm{KL}}(\pi^y_{\mathrm{pos}}||\widetilde{\pi}^y_{\mathrm{pos}})$.

3. Dimension reduction via optimal parameter-reduced likelihood

In this section, we first briefly review the optimal parameter-reduced likelihood and the data-dependent LIS proposed in [54], and then we will introduce the data-free LIS.

3.1. Optimal parameter-reduced likelihood using a given projector

As shown in Section 2.1 of [54], for a given data set y and a given projector P_r, the parameter-reduced likelihood function

$$\mathcal{L}^{*,y}(x_r) = \int_{\mathrm{Ker}(P_r)} \mathcal{L}^y(x_r + x_\perp)\pi_{\mathrm{pr}}(x_\perp|x_r)\,\mathrm{d}x_\perp, \tag{7}$$

is an optimal approximation in the sense that it minimizes $\widetilde{\mathcal{L}}^y \mapsto D_{\mathrm{KL}}(\pi^y_{\mathrm{pos}}||\widetilde{\pi}^y_{\mathrm{pos}})$. We denote by

$$\pi^{*,y}_{\mathrm{pos}}(x) \propto \mathcal{L}^{*,y}(P_r x)\pi_{\mathrm{pr}}(x),$$

the resulting approximate posterior density. The marginal density $\pi^{*,y}_{\text{pos}}(x_r) = \int_{\text{Ker}(P_r)} \pi^{*,y}_{\text{pos}}(x_r + x_\perp)\mathrm{d}x_\perp$ can be expressed as

$$\pi^{*,y}_{\text{pos}}(x_r) \propto \mathcal{L}^{*,y}(x_r)\pi_{\text{pr}}(x_r) \overset{(7)}{\propto} \int_{\text{Ker}(P_r)} \mathcal{L}^{y}(x_r + x_\perp)\pi_{\text{pr}}(x_r + x_\perp)\mathrm{d}x_\perp \overset{(1)}{\propto} \pi^{y}_{\text{pos}}(x_r), \tag{8}$$

for all $x_r \in \text{Im}(P_r)$, where $\pi^{y}_{\text{pos}}(x_r) = \int_{\text{Ker}(P_r)} \pi^{y}_{\text{pos}}(x_r + x'_\perp)\mathrm{d}x'_\perp$ is the marginal density of the full posterior. Thus, for any projector P_r and any data y, the approximate posterior $\pi^{*,y}_{\text{pos}}$ and the full posterior π^{y}_{pos} have the same marginal density on $\text{Im}(P_r)$. In summary we have

$$\begin{aligned}\pi^{y}_{\text{pos}}(x) &= \pi^{y}_{\text{pos}}(x_r)\pi^{y}_{\text{pos}}(x_\perp|x_r),\\ \pi^{*,y}_{\text{pos}}(x) &\overset{(8)}{=} \pi^{y}_{\text{pos}}(x_r)\pi_{\text{pr}}(x_\perp|x_r),\end{aligned}$$

which shows that the optimal approximation $\pi^{*,y}_{\text{pos}}(x)$ to $\pi^{y}_{\text{pos}}(x)$ replaces the conditional posterior $\pi^{y}_{\text{pos}}(x_\perp|x_r)$ with the conditional prior $\pi_{\text{pr}}(x_\perp|x_r)$.

3.2. Data-dependent dimension reduction

We denote by $P_r = P^y_r$ a projector built by a data-dependent approach. Ideally, we would like to build P^y_r that minimizes $D_{\text{KL}}(\pi^{y}_{\text{pos}}||\pi^{*,y}_{\text{pos}})$ over the manifold of rank-r projectors. However, this non-convex minimization problem can be challenge to solve. Instead, the strategy proposed in [54] minimizes an upper bound of the KL divergence obtained by logarithmic Sobolev inequalities, in which the following assumption on the prior density is adopted.

Assumption 3.1 (Subspace logarithmic Sobolev inequality)**.** There exists a symmetric positive definite matrix $\Gamma \in \mathbb{R}^{d\times d}$ and a scalar $\kappa > 0$ such that for any projector $P_r \in \mathbb{R}^{d\times d}$ and for any continuously differentiable function $h : \mathbb{R}^d \to \mathbb{R}$ the inequality

$$\int_{\mathbb{R}^d} h(x)^2 \log\left(\frac{h(x)^2}{h_{P_r}(x)^2}\right)\pi_{\text{pr}}(x)\mathrm{d}x \le 2\kappa \int_{\mathbb{R}^d} \|(I_d - P_r)^\top \nabla h(x)\|^2_{\Gamma^{-1}}\pi_{\text{pr}}(x)\mathrm{d}x,$$

holds, where $h^2_{P_r}$ is the conditional expectation of h^2 given by $h_{P_r}(x)^2 = \int_{\text{Ker}(P_r)} h(P_r x + x_\perp)^2\pi_{\text{pr}}(x_\perp|P_r x)\,\mathrm{d}x_\perp$. Here the norm $\|\cdot\|_{\Gamma^{-1}}$ is defined by $\|v\|^2_{\Gamma^{-1}} = v^\top\Gamma^{-1}v$ for any $v \in \mathbb{R}^d$.

Theorem 1 in [54] gives sufficient conditions on the prior density such that Assumption 3.1 holds. In particular, any Gaussian prior $\pi_{\text{pr}} = \mathcal{N}(m_{\text{pr}}, \Sigma_{\text{pr}})$ with mean $m_{\text{pr}} \in \mathbb{R}^d$ and non-singular covariance matrix $\Sigma_{\text{pr}} \in \mathbb{R}^{d\times d}$ satisfies Assumption 3.1 with $\kappa = 1$ and $\Gamma = \Sigma^{-1}_{\text{pr}}$. As shown in [54, example 2], any Gaussian mixture also satisfies this assumption, but with a constant κ which might not be accessible in practice. We refer to [28, 36] for nicely written introductions to logarithmic Sobolev inequalities and examples of distributions which satisfy it.

Proposition 3.2. Suppose π_{pr} satisfies Assumption 3.1 and the likelihood function $\mathcal{L}^y$ is continuously differentiable. Then, for any projector $P_r \in \mathbb{R}^{d\times d}$, the posterior approximation $\pi^{*,y}_{\text{pos}}(x) \propto \big(\mathcal{L}^{*,y}(x_r)\pi_{\text{pr}}(x_r)\big)\pi_{\text{pr}}(x_\perp|x_r)$ induced by the optimal parameter-reduced likelihood as in (7) satisfies

$$D_{\text{KL}}(\pi^{y}_{\text{pos}}||\pi^{*,y}_{\text{pos}}) \le \frac{\kappa}{2}\,\text{trace}\left(\Gamma^{-1}(I_d - P_r^\top)H(y)(I_d - P_r)\right), \tag{9}$$

where the matrix $H(y) \in \mathbb{R}^{d\times d}$ is defined by

$$H(y) = \int_{\mathbb{R}^d} \left(\nabla \log \mathcal{L}^y(x)\right)\left(\nabla \log \mathcal{L}^y(x)\right)^\top \pi^y_{\text{pos}}(x)\mathrm{d}x. \tag{10}$$

Proof. See the proof of Corollary 1 in [54]. □

Proposition 3.2 gives an upper bound on $D_{\text{KL}}(\pi^y_{\text{pos}}||\pi^{*,y}_{\text{pos}})$. The minimizer of this bound

$$P^y_r \in \underset{\substack{P_r\in\mathbb{R}^{d\times d}\\ \text{rank-}r \text{ projector}}}{\arg\min} \quad \text{trace}\left(\Gamma^{-1}(I_d - P_r^\top)H(y)(I_d - P_r)\right),$$

can be obtained from the leading generalized eigenvectors of the matrix pair $(H(y), \Gamma)$, see [53, Proposition 2.6]. Let $(\lambda^y_i, v^y_i) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^d$ denotes the i-th eigenpair of $(H(y), \Gamma)$ such that $H(y)v^y_i = \lambda^y_i \Gamma v^y_i$, with $(v^y_i)^\top \Gamma v^y_j = \delta_{i,j}$ and $\lambda^y_i \geq \lambda^y_j$ for all $i \leq j$. The image and the kernel of P^y_r are respectively defined as

$$\begin{aligned} \text{Im}(P^y_r) &= \text{span}\{v^y_1, \ldots, v^y_r\}, \\ \text{Ker}(P^y_r) &= \text{span}\{v^y_{r+1}, \ldots, v^y_d\}. \end{aligned} \tag{11}$$

The resulting projector P^y_r yields an approximate posterior density $\pi^{*,y}_{\text{pos}}$ that satisfies

$$D_{\text{KL}}(\pi^y_{\text{pos}}||\pi^{*,y}_{\text{pos}}) \leq \frac{\kappa}{2}\sum_{i=r+1}^{d} \lambda^y_i.$$

The above relation can be used to choose the rank $r = \text{rank}(P^y_r)$ to guarantee that the $D_{\text{KL}}(\pi^y_{\text{pos}}||\pi^{*,y}_{\text{pos}})$ is bounded below some user-defined tolerance. A rapid decay in the spectrum $(\lambda^y_1, \lambda^y_2, \ldots)$ ensures that one can choose a rank r that is much lower than the original dimension d. Note that the projector P^y_r may not be unique, unless there exists a spectral gap $\lambda^y_r > \lambda^y_{r+1}$ which ensures the r-dimensional dominant eigenspace of $(H(y), \Gamma)$ is unique.

Remark 3.3 (Coordinate selection). The projector defined in (11) is, in general, not aligned with the canonical coordinates. However, in some parametrizations—for example, different components of x represent physical quantities of different nature—we may prefer coordinate selection than subspace identification to make the dimension reduction more interpretable. Denoting the i-th canonical basis vector of $\mathbb{R}^d$ by e_i, we let $P^y_r = \sum_{i\in\mathcal{I}} e_i e_i^\top$, be the projector of rank $r = \#\mathcal{I}$, which extracts the components of x indexed by the index set $\mathcal{I} \subset \{1, \ldots, d\}$ such that

$$\begin{aligned} \text{Im}(P^y_r) &= \text{span}\{x_i : i \in \mathcal{I}\}, \\ \text{Ker}(P^y_r) &= \text{span}\{x_i : i \notin \mathcal{I}\}. \end{aligned}$$

Using such a projector, the bound (9) becomes

$$D_{\text{KL}}(\pi^y_{\text{pos}}||\pi^{*,y}_{\text{pos}}) \leq \frac{\kappa}{2}\sum_{i\notin\mathcal{I}} (\Gamma^{-1})_{ii} H(y)_{ii},$$

which suggests to define the index set $\mathcal{I}$ that selects the r largest values of $(\Gamma^{-1})_{ii}H(y)_{ii}$.

Because of the dependency on the data set y, the projector P_r^y must be built after a data set has been observed, see Algorithm 1. For scenarios where one wants to solve multiple inverse problems with multiple data sets, the matrix $H(y)$ and the resulting projector have to be reconstructed for each data set. This can be a computationally challenging task. In addition, $H(y)$ is defined as an expectation over the high-dimensional posterior distribution, which further raises the computational burden.

Algorithm 1: Data-dependent dimension reduction.

Requires: π_{pr} satisfying Assumption 3.1, tolerance $\varepsilon > 0$ and maximal rank $r_{\max}$
Online phase: given the data y **do:**
- Compute the matrix $H(y)$ using (10).
- Compute the generalized eigendecomposition $H(y)v_i^y = \lambda_i^y \Gamma v_i^y$.
- Find the smallest r such that $\frac{\kappa}{2}\sum_{i=r+1}^{d} \lambda_i^y \leq \varepsilon$. If $r \geq r_{\max}$, set $r = r_{\max}$.
- Assemble the projector P_r^y using (11).
- Define the conditional expectation $\mathcal{L}^{*,y}(x_r)$ defined in (7).

Return: Approximate posterior $\pi_{\text{pos}}^{*,y}(x) \propto \mathcal{L}^{*,y}(P_r^y x)\pi_{\text{pr}}(x)$.

3.3. Data-free dimension reduction

To overcome the abovementioned computational burden of recomputing the data-dependent projector for every new data set, we present a new data-free dimension reduction method. The key idea is to control the KL divergence in expectation over the marginal density of data. We introduce an m-dimensional random vector

$$Y \sim \pi_{\text{data}}(y),$$

where π_{data} is the marginal density of data defined in (2). Note that the observed data y corresponds to a particular realization of Y. For a given projector P_r independent on the data, replacing y with Y in (9) and taking the expectation over Y yields

$$\mathbb{E}\big[D_{\text{KL}}(\pi_{\text{pos}}^{Y} || \pi_{\text{pos}}^{*,Y})\big] \leq \frac{\kappa}{2} \operatorname{trace}\left(\Gamma^{-1}(I_d - P_r^\top)\mathbb{E}[H(Y)](I_d - P_r)\right). \tag{12}$$

Here, the approximate posterior $\pi_{\text{pos}}^{*,Y}$ depends on Y via the optimal likelihood $\mathcal{L}^{*,Y}$. Similar to the data-dependent case, the leading generalized eigenvectors of the matrix pair $(\mathbb{E}[H(Y)], \Gamma)$ can be used to obtain a projector that minimizes the error bound. However, in this case, the matrix $\mathbb{E}[H(Y)]$ is the expectation of $H(y)$ over the marginal density of data, and thus it is independent of observed data. Let $(\lambda_i, v_i) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^d$ denotes the i-th eigenpair of $(\mathbb{E}[H(Y)], \Gamma)$ such that $\mathbb{E}[H(Y)]v_i = \lambda_i \Gamma v_i$, with $v_i^\top \Gamma v_j = \delta_{i,j}$ and $\lambda_i^y \geq \lambda_j^y$ for all $i \leq j$. The data-free projector P_r that minimizes the right-hand side of (12) is given by

$$\begin{aligned} \operatorname{Im}(P_r) &= \operatorname{span}\{v_1, \ldots, v_r\}, \\ \operatorname{Ker}(P_r) &= \operatorname{span}\{v_{r+1}, \ldots, v_d\}. \end{aligned} \tag{13}$$

When using this projector for defining the approximate posterior $\pi_{\text{pos}}^{*,Y}$, the expectation of the KL divergence $D_{\text{KL}}(\pi_{\text{pos}}^{Y}||\pi_{\text{pos}}^{*,Y})$ can be controlled as

$$\mathbb{E}\big[D_{\text{KL}}(\pi_{\text{pos}}^{Y}||\pi_{\text{pos}}^{*,Y})\big] \leq \frac{\kappa}{2}\sum_{i=r+1}^{d}\lambda_i. \tag{14}$$

Remark 3.4 (Bound in high probability). Inequality (14) gives a bound on $D_{\text{KL}}(\pi_{\text{pos}}^{Y}||\pi_{\text{pos}}^{*,Y})$ in expectation. In order to obtain a bound in high probability, let us use the Markov inequality $\mathbb{P}\{D_{\text{KL}}(\pi_{\text{pos}}^{Y}||\pi_{\text{pos}}^{*,Y}) \leq \varepsilon\} \geq 1-\varepsilon^{-1}\mathbb{E}[D_{\text{KL}}(\pi_{\text{pos}}^{Y}||\pi_{\text{pos}}^{*,Y})]$ for some $\varepsilon > 0$. Thus, for a given $0 < \eta \leq 1$, the condition $\frac{\kappa}{2}\sum_{i=r+1}^{d}\lambda_i \leq \frac{\varepsilon}{\eta}$ is sufficient to ensure that

$$D_{\text{KL}}(\pi_{\text{pos}}^{Y}||\pi_{\text{pos}}^{*,Y}) \leq \varepsilon,$$

holds with a probability greater than $1-\eta$.

Remark 3.5 (Coordinate selection). Similarly to Remark 3.3, instead of defining P_r as in (13), we can define a coordinate-aligned projector $P_r = \sum_{i\in\mathcal{I}} e_i e_i^\top$ by selecting an index set $\mathcal{I}$ corresponding to the r largest values of $(\Gamma^{-1})_{ii}\mathbb{E}[H(Y)]_{ii}$.

Now we show that the matrix $\mathbb{E}[H(Y)]$ admits a simple expression in terms of the Fisher information matrix associated with the likelihood function. This leads to a computationally convenient way to construct the data-free projector. Recall that the likelihood $\mathcal{L}^y(x)$, seen as a function of y, is the pdf of the data y conditionned on the parameter $x \in \mathbb{R}^d$. The Fisher information matrix associated with this family of pdf is

$$\mathcal{I}(x) = \int_{\mathbb{R}^m} \big(\nabla \log \mathcal{L}^y(x)\big)\big(\nabla \log \mathcal{L}^y(x)\big)^\top \mathcal{L}^y(x)\,\mathrm{d}y. \tag{15}$$

We can write

$$\begin{aligned}
\mathbb{E}[H(Y)] &= \int_{\mathbb{R}^m} H(y)\pi_{\text{data}}(y)\mathrm{d}y \\
&\overset{(10)}{=} \int_{\mathbb{R}^m}\left(\int_{\mathbb{R}^d}\big(\nabla \log \mathcal{L}^y(x)\big)\big(\nabla \log \mathcal{L}^y(x)\big)^\top \pi_{\text{pos}}^{y}(x)\mathrm{d}x\right)\pi_{\text{data}}(y)\mathrm{d}y \\
&\overset{(1)}{=} \int_{\mathbb{R}^m\times\mathbb{R}^d}\big(\nabla \log \mathcal{L}^y(x)\big)\big(\nabla \log \mathcal{L}^y(x)\big)^\top \frac{\mathcal{L}^y(x)\pi_{\text{pr}}(x)}{\int_{\mathbb{R}^d}\mathcal{L}^y(x')\pi_{\text{pr}}(x')\mathrm{d}x'}\,\pi_{\text{data}}(y)\mathrm{d}x\mathrm{d}y \\
&\overset{(2)}{=} \int_{\mathbb{R}^m\times\mathbb{R}^d}\big(\nabla \log \mathcal{L}^y(x)\big)\big(\nabla \log \mathcal{L}^y(x)\big)^\top \mathcal{L}^y(x)\pi_{\text{pr}}(x)\mathrm{d}x\mathrm{d}y \\
&\overset{(15)}{=} \int_{\mathbb{R}^d}\mathcal{I}(x)\pi_{\text{pr}}(x)\mathrm{d}x,
\end{aligned} \tag{16}$$

which shows that the matrix $\mathbb{E}[H(Y)]$ is the expectation of the Fisher information matrix over the prior. This expression does not involve any expectation over the posterior density, which is a major advantage compared to the expression (10) of the data-dependent matrix $H(y)$. The methodology presented here is summarized in Algorithm 2.

Algorithm 2: Data-free dimension reduction

Requires: π_{pr} satisfying Assumption 3.1, Fisher information matrix $\mathcal{I}(x)$ of $\mathcal{L}^y$, tolerance $\varepsilon > 0$, and maximal rank $r_{\max}$

Offline phase

Compute the matrix $H^I = \int_{\mathbb{R}^d} \mathcal{I}(x)\pi_{\text{pr}}(x)\mathrm{d}x$.

Compute the generalized eigendecomposition $H^I v_i = \lambda_i \Gamma v_i$.

Find the smallest r such that $\frac{\kappa}{2}\sum_{i=r+1}^{d}\lambda_i \leq \varepsilon$. If $r \geq r_{\max}$, set $r = r_{\max}$.

Return: Projector P_r defined by (13)

Online phase: given the data y **do:**

Define $\mathcal{L}^{*,y}$ as the conditional expectation defined in (7).

Return: Approximate posterior $\pi^{*,y}_{\text{pos}}(x) \propto \mathcal{L}^{*,y}(P_r x)\pi_{\text{pr}}(x)$

Example 3.6 (Gaussian likelihood)**.** Consider the parameter-to-data map is represented by a smooth forward model $G : \mathbb{R}^d \to \mathbb{R}^m$ and corrupted by an additive Gaussian noise $\xi_{\text{obs}} \sim \mathcal{N}(0, \Sigma_{\text{obs}})$ with non-singular covariance matrix $\Sigma_{\text{obs}} \in \mathbb{R}^{m\times m}$, i.e.,

$$y = G(x) + \xi_{\text{obs}}, \quad \text{where} \quad \xi_{\text{obs}} \sim \mathcal{N}(0, \Sigma_{\text{obs}}).$$

The likelihood function takes the form $\mathcal{L}^y(x) = Z^{-1}\exp(-\frac{1}{2}\|G(x) - y\|^2_{\Sigma^{-1}_{\text{obs}}})$, where $Z = \sqrt{(2\pi)^m \det(\Sigma_{\text{obs}})}$ is a normalizing constant. The Slepian-Bangs formula gives an explicit expression for the Fisher information matrix $\mathcal{I}(x) = \nabla G(x)^\top \Sigma^{-1}_{\text{obs}} \nabla G(x)$, where $\nabla G(x) \in \mathbb{R}^{m\times d}$ denotes the Jacobian of the forward model $G(x)$. By relation (16) we obtain

$$\mathbb{E}[H(Y)] = \int_{\mathbb{R}^d} \nabla G(x)^\top \Sigma^{-1}_{\text{obs}} \nabla G(x)\, \pi_{\text{pr}}(x)\mathrm{d}x. \tag{17}$$

A similar matrix was considered in [18] in the context of data-dependent dimension reduction. The major difference with (17) is that, in [18], the expectation is taken over the posterior density rather than over the prior.

4. Dimension reduction via parameter-reduced forward model

In the previous Section 3, the detection of the data-free informed subspace is based on an approximation of the likelihood function. In this section, we present an alternative strategy which, under Gaussian likelihood assumption, consist in approximating the forward model instead of the likelihood itself. This approach is similar to the vector-valued extension of the AS method [53] and still yields error bounds for the expected KL divergence.

As in Example 3.6, let us start with a Gaussian likelihood of the form

$$\mathcal{L}^y(x) = \frac{1}{Z}\exp\left(-\frac{1}{2}\|G(x) - y\|^2_{\Sigma^{-1}_{\text{obs}}}\right), \tag{18}$$

where $x \mapsto G(x)$ is a continuously differentiable forward model, $\Sigma_{\text{obs}} \in \mathbb{R}^{m\times m}$ is a non-singular covariance matrix and $Z = \sqrt{(2\pi)^m \det(\Sigma_{\text{obs}})}$ a normalizing constant. Our goal is to build a

low-dimensional approximation to the likelihood (18) by replacing the forward model with a ridge approximation $x \mapsto \widetilde{G}(P_r x)$. That is, we look for a likelihood approximation of the form

$$\widetilde{\mathcal{L}}^y(P_r x) = \frac{1}{Z} \exp\left(-\frac{1}{2}\|\widetilde{G}(P_r x) - y\|^2_{\Sigma_{\text{obs}}^{-1}}\right), \tag{19}$$

where P_r is a low-rank projector and where $\widetilde{G}$ is some parameter-reduced function defined over $\text{Ker}(P_r)$. In general, this approximate likelihood (19) is different than the previous one $\mathcal{L}^{*,y}$, see (7), and therefore $\widetilde{\mathcal{L}}^y$ might not be optimal with respect to the KL divergence as discussed in Section 3.1. The following proposition will guide the construction of the approximate forward model.

Proposition 4.1. Consider the posterior density $\pi^y_{\text{pos}}(x) \propto \mathcal{L}^y(x)\pi_{\text{pr}}(x)$ with a Gaussian likelihood as in (18). For any approximate forward model $\widehat{G} : \mathbb{R}^d \to \mathbb{R}^m$, the resulting approximate likelihood $\widehat{\mathcal{L}}^y(x) = \frac{1}{Z}\exp(-\frac{1}{2}\|\widehat{G}(x) - y\|^2_{\Sigma_{\text{obs}}^{-1}})$ defines an approximate posterior density $\widehat{\pi}^y_{\text{pos}}(x) \propto \widehat{\mathcal{L}}^y(x)\pi_{\text{pr}}(x)$ such that

$$\mathbb{E}\left[D_{\text{KL}}(\pi^Y_{\text{pos}}||\widehat{\pi}^Y_{\text{pos}})\right] + D_{\text{KL}}(\pi_{\text{data}}||\widehat{\pi}_{\text{data}}) = \frac{1}{2}\int_{\mathbb{R}^d} \|G(x) - \widehat{G}(x)\|^2_{\Sigma_{\text{obs}}^{-1}} \pi_{\text{pr}}(x)\text{d}x.$$

Here the expectation is taken over $Y \sim \pi_{\text{data}}$ and $\widehat{\pi}_{\text{data}}(y) = \int_{\mathbb{R}^d} \widehat{\mathcal{L}}^y(x)\pi_{\text{pr}}(x)\,\text{d}x$ is the approximate marginal density of data.

Proof. See Appendix A. □

Using an approximate forward model in the form of $\widehat{G}(x) = \widetilde{G}(P_r x)$, Proposition 4.1 ensures that the approximate posterior $\widetilde{\pi}^y_{\text{pos}}(x) \propto \widetilde{\mathcal{L}}^y(P_r x)\pi_{\text{pr}}(x)$ with $\widetilde{\mathcal{L}}^y(P_r x)$ as in (19) satisfies

$$\mathbb{E}[D_{\text{KL}}(\pi^Y_{\text{pos}}||\widetilde{\pi}^Y_{\text{pos}})] \leq \frac{1}{2}\int_{\mathbb{R}^d} \|G(x) - \widetilde{G}(P_r x)\|^2_{\Sigma_{\text{obs}}^{-1}} \pi_{\text{pr}}(x)\text{d}x, \tag{20}$$

This suggests to construct a ridge approximation $\widetilde{G}(P_r x)$ to $G(x)$ in the $L^2_{\pi_{\text{pr}}}$ sense. To accomplish this, we follow the methodology proposed in [53] for the approximation of multivariate function using gradient information. First, for any projector P_r, the optimal function $\widetilde{G}^*$ that minimizes the right-hand side of (20) is the conditional expectation

$$\widetilde{G}^*(x_r) = \int_{\text{Ker}(P_r)} G(x_r + x_\perp)\pi_{\text{pr}}(x_\perp|x_r)\,\text{d}x_\perp. \tag{21}$$

Then, similarly to Assumption 3.1, we assume that π_{pr} satisfies the following subspace Poincaré inequality.

Assumption 4.2 (Subspace Poincaré inequality)**.** There exists a symmetric positive definite matrix $\Gamma \in \mathbb{R}^{d\times d}$ and a scalar $\kappa > 0$ such that for any projector $P_r \in \mathbb{R}^{d\times d}$ and for any continuously differentiable function $h : \mathbb{R}^d \to \mathbb{R}$ the inequality

$$\int_{\mathbb{R}^d} \left(h(x) - h_{P_r}(x)\right)^2 \pi_{\text{pr}}(x)\text{d}x \leq \kappa \int_{\mathbb{R}^d} \|(I_d - P_r)^\top \nabla h(x)\|^2_{\Gamma^{-1}} \pi_{\text{pr}}(x)\text{d}x,$$

holds, where h_{P_r} is the conditional expectation of h defined by $h_{P_r}(x) = \int_{\text{Ker}(P_r)} h(P_r x + x'_\perp)\pi_{\text{pr}}(x'_\perp|P_r x)\text{d}x'_\perp$.

Assumption 4.2 is weaker than Assumption 3.1, in the sense that any distribution which satisfies the subspace logarithmic Sobolev inequality automatically satisfies the subspace Poincaré inequality with the same κ and the same Γ, see for instance [54, Corollary 2]. We refer to the recent contributions [4, 43, 49] for examples of probability distribution which satisfy (subspace) Poincaré inequality. As for the logarithmic-Sobolev constant, the Poincaré constant is hard to compute in practice, except the case of Gaussian prior. Using similar arguments as in the proof of Proposition 2.5 in [53], Assumption 4.2 allows to write

$$\int_{\mathbb{R}^d} \|G(x) - \widetilde{G}^*(P_r x)\|^2_{\Gamma_{\text{obs}}^{-1}} \pi_{\text{pr}}(x)\mathrm{d}x \leq \kappa \text{ trace}\left(\Gamma^{-1}(I_d - P_r^\top)H^G(I_d - P_r)\right), \tag{22}$$

holds for any projector P_r, where the matrix $H^G \in \mathbb{R}^{d\times d}$ is defined by

$$H^G = \int_{\mathbb{R}^d} \nabla G(x)^\top \Sigma_{\text{obs}}^{-1} \nabla G(x)\, \pi_{\text{pr}}(x)\mathrm{d}x, \tag{23}$$

with $\nabla G(x)$ the Jacobian matrix of $G(x)$ given by

$$\nabla G(x) = \begin{pmatrix} \frac{\partial G_1}{\partial x_1}(x) & \dots & \frac{\partial G_1}{\partial x_d}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial G_m}{\partial x_1}(x) & \dots & \frac{\partial G_m}{\partial x_d}(x) \end{pmatrix}.$$

Again, the projector P_r^G that minimizes the right-hand side of (22) can be constructed via the generalized eigenvalue problem $H^G v_i^G = \lambda_i^G \Gamma v_i^G$:

$$\begin{aligned} \text{Im}(P_r^G) &= \text{span}\{v_1^G, \dots, v_r^G\}, \\ \text{Ker}(P_r^G) &= \text{span}\{v_{r+1}^G, \dots, v_d^G\}. \end{aligned} \tag{24}$$

Using this projector to construct the approximate forward model $\widetilde{G}^*$ in (21) and the approximate likelihood as in (19), Proposition 4.1 and the inequality in (22) yield

$$\mathbb{E}\left[D_{\text{KL}}(\pi_{\text{pos}}^Y || \widetilde{\pi}_{\text{pos}}^Y)\right] \leq \frac{\kappa}{2} \sum_{i=r+1}^{d} \lambda_i^G. \tag{25}$$

The methodology is summarized in Algorithm 3.

The matrix H^G used in this case takes the same form as the matrix $\mathbb{E}[H(Y)]$ in Section 3.3 with the Gaussian likelihood (cf. Example 3.6), and hence results in the same data-free projector. However, the resulting approximate likelihood functions are not the same. Indeed in Section 3.3, the optimal approximate likelihood $\mathcal{L}^{*,y}$ is given as the conditional expectation of the likelihood function (cf. (7)), whereas here, $\widetilde{\mathcal{L}}^y$ is defined by the conditional expectation of the forward model $\widetilde{G}^*$ (cf. (21)). Using either the parameter-reduced likelihood in (7) or the parameter-reduced forward model in (21) results in the same parameter truncation error bound in terms of expected

KL divergence.

Algorithm 3: Data-free dimension reduction via forward model approximation

Requires: π_{pr} satisfying Assumption 3.1, Jacobian $\nabla G(x)$, tolerance $\varepsilon > 0$ and maximal rank $r = r_{\max}$.

Offline phase

Compute the matrix H^G defined in (23)
Compute the generalized eigendecomposition $H^G v_i^G = \lambda_i^G \Gamma v_i^G$
Find the smallest r such that $\frac{\kappa}{2}\sum_{i=r+1}^{d} \lambda_i \leq \varepsilon$. If $r \geq r_{\max}$, set $r = r_{\max}$
Assemble the projector P_r^G defined in (24)
Define $\widetilde{G}^*$ as the conditional expectation (21)

Return: Approximate forward model $x \mapsto \widetilde{G}^*(P_r^G x)$

Online phase: given the data y **do:**

Assemble $\widetilde{\mathcal{L}}^y(P_r^G x)$ as in (19)

Return: Approximate posterior $\widetilde{\pi}^y_{\text{pos}}(x) \propto \widetilde{\mathcal{L}}^y(P_r x)\pi_{\text{pr}}(x)$

Remark 4.3. Despite the similarity between the approximate likelihood functions given in (7) and (19), these two approaches offer different computational characteristics. Given the data-free informed subspace, the optimal parameter-reduced forward model $x_r \mapsto G^*(x_r)$ can be further replaced by a surrogate model $x_r \mapsto G^{\text{ROM}}(x_r)$ constructed in the offline phase. The surrogate model can be obtained using tensor methods [8, 42], the reduced basis method [44], polynomial techniques [35], etc., just to cite a few. All these approximation techniques do not scale well with the apparent parameter dimensions d, and thus parameter reduction can greatly improve the scalability of surrogate models.

In contrast, the conditional expectation of the likelihood function in (7) cannot be replaced with offline surrogate models because of the data-dependency of the likelihood.

5. Sampling the approximate posterior

Given a data-free informed subspace, the approximate posterior density has the factorized form

$$\widetilde{\pi}^y_{\text{pos}}(x) \propto \widetilde{\pi}^y_{\text{post}}(x_r)\pi_{\text{pr}}(x_\perp|x_r), \tag{26}$$

with either $\widetilde{\pi}^y_{\text{post}}(x_r) = \pi^y_{\text{post}}(x_r)$ in the optimal parameter-reduced likelihood approach of Section 3, or with $\widetilde{\pi}^y_{\text{post}}(x_r) = \widetilde{\mathcal{L}}^y(x_r)\pi^y_{\text{pr}}(x_r)$, $\widetilde{\mathcal{L}}^y(x_r) \propto \exp(-\frac{1}{2}\|\widetilde{G}^*(x_r) - y\|^2_{\Sigma_{\text{obs}}^{-1}})$ in the optimal parameter-reduced forward model approach of Section 4. The factorization (26) naturally suggests a dimension robust way to sampling the approximate posterior. The sampling method consists in first drawing samples $x_r^{(1)}, x_r^{(2)}, ..., x_r^{(K)}$ from the low-dimensional density $\widetilde{\pi}^y_{\text{pos}}(x_r)$ using either MCMC or SMC method. Then, for each sample $x_r^{(j)}$, we simulate a conditional prior sample $x_\perp^{(j)}$ from $\pi_{\text{pr}}(x_\perp|x_r^{(j)})$. In the end, $x^{(j)} = x_r^{(j)} + x_\perp^{(j)}$ are samples from the approximate posterior $\widetilde{\pi}^y_{\text{pos}}(x)$.

We emphasis here that the key is to be able to sample from the conditional prior $\pi_{\text{pr}}(x_\perp|x_r)$. This task is rather easy for Gaussian priors. We show in Section 7 how to sample from $\pi_{\text{pr}}(x_\perp|x_r)$ for non-Gaussian priors with a particular structure that can be exploited.

Remark 5.1. If the end goal is to compute expectation of some function h over of the approximate

posterior, the factorization (26) leads to

$$\begin{aligned}\int_{\mathbb{R}^d} h(x)\widetilde{\pi}_{\text{pos}}(x)\mathrm{d}x &= \int_{\text{Im}(P_r)} \left(\int_{\text{Ker}(P_r)} h(x_r + x_\perp)\pi_{\text{pr}}(x_\perp|x_r)\mathrm{d}x_\perp \right) \widetilde{\pi}_{\text{pos}}(x_r)\mathrm{d}x_r \\ &\approx \frac{1}{K}\sum_{j=1}^{K} \int_{\text{Ker}(P_r)} h(x_r^{(j)} + x_\perp)\pi_{\text{pr}}(x_\perp|x_r^{(j)})\mathrm{d}x_\perp,\end{aligned}$$

where $x_r^{(1)}, ..., x_r^{(K)}$ are samples from the approximate marginal posterior $\widetilde{\pi}^y_{\text{pos}}(x_r)$. This way, if the expectation over the conditional prior $\pi_{\text{pr}}(x_\perp|x_r^{(j)})$ can be carried out analytically, one can can simply avoid using conditional prior samples. Alternatively, the K conditional expectations $\int h(x_r^{(j)} + x_\perp)\pi_{\text{pr}}(x_\perp|x_r^{(j)})\mathrm{d}x_\perp$ can also be approximated via other accurate quadrature rule for $\pi_{\text{pr}}(x_\perp|x_r^{(j)})$. Either way, we assume that integration with respect to the conditional prior is tractable.

In Algorithm 4 we provide the details of an MCMC-based sampling procedure in which the approximate likelihood (defined by either optimal parameter-reduced likelihood or optimal parameter-reduced forward model) can be obtained as sample averages over the conditional prior $\pi_{\text{pr}}(x_\perp|x_r)$. To make these approximations generally applicable, we replace the conditional prior with the marginal prior $\pi_{\text{pr}}(x_\perp)$ in computing those conditional expectations in the Equations (27) and (28) in Algorithm 4. Note that the typical class of inverse problems equipped with a Gaussian prior $\pi_{\text{pr}} = \mathcal{N}(m_{\text{pr}}, \Sigma_{\text{pr}})$ is a special case. Since the projector P_r is orthogonal with respect to Σ_{pr}^{-1}, the marginal prior $\pi_{\text{pr}}(x_\perp)$ coincides with the conditional prior $\pi_{\text{pr}}(x_\perp|x_r)$.

A remaining question is how to choose the sample size N for computing the conditional expectations in (27) and (28). The following heuristic is developed based on the optimal parameter-reduced forward model. Consider the exact parameter-reduced forward model $\widetilde{G}^*(P_r x) = \widetilde{G}^*(x_r)$ and its sample-averaged approximation $\widetilde{G}_N(P_r x) = \widetilde{G}_N(x_r)$. The sample-averaged approximation defines an approximate posterior density

$$\widehat{\pi}^y_{\text{pos}}(x) \propto \exp\Big(-\frac{1}{2}\|\widetilde{G}_N(P_r x) - y\|^2_{\Sigma_{\text{obs}}^{-1}}\Big)\pi_{\text{pr}}(x),$$

that satisfies

$$\begin{aligned}\mathbb{E}[D_{\text{KL}}(\pi^Y_{\text{pos}}||\widehat{\pi}^Y_{\text{pos}})] &\overset{(20)}{\leq} \mathbb{E}\left[\frac{1}{2}\int_{\mathbb{R}^d} \|G(x) - \widetilde{G}_N(P_r x)\|^2_{\Sigma_{\text{obs}}^{-1}}\pi_{\text{pr}}(x)\mathrm{d}x\right] \\ &= \frac{1}{2}\,\mathbb{E}\left[\int_{\mathbb{R}^d} \|G(x) - \widetilde{G}^*(P_r x)\|^2_{\Sigma_{\text{obs}}^{-1}} + \|\widetilde{G}^*(P_r x) - \widetilde{G}_N(P_r x)\|^2_{\Sigma_{\text{obs}}^{-1}}\pi_{\text{pr}}(x)\mathrm{d}x\right] \\ &= \left(1 + \frac{1}{N}\right)\frac{1}{2}\int_{\mathbb{R}^d} \|G(x) - \widetilde{G}^*(P_r x)\|^2_{\Sigma_{\text{obs}}^{-1}}\pi_{\text{pr}}(x)\mathrm{d}x. \qquad (30)\end{aligned}$$

Here, the expectation is taken jointly over the data Y and the sample $\{x_\perp^{(i)}\}_{i=1}^N$. The above inequality directly follows from Proposition 4.1 and the fact that $\widetilde{G}^*(P_r x)$ is the conditional expectation of $G(x)$ over $\text{Ker}(P_r)$. We refer to Theorem 3.2 in [12] for more details on this derivation. Inequality (30) implies that the random approximate posterior $\widehat{\pi}^y_{\text{pos}}(x)$ can be used in place of $\widetilde{\pi}^y_{\text{pos}}(x)$, as the bounds on the expected Kullback-Leibler divergence in (20) and (30) are comparable. In addition, this suggests that the sample size N in (28) does not have to be large.

Algorithm 4: MCMC-based approach for sampling the approximate posterior.

Requires: A projector P_r, a sample size N for approximating the likelihood, a total posterior sample size K, and a proposal density $q(\cdot|x_r)$ on $\mathrm{Im}(P_r)$.

Sample-averaged likelihood approximation

Draw N i.i.d. samples $x_\perp^{(1)}, \dots, x_\perp^{(N)}$ from the marginal $\pi_{\mathrm{pr}}(x_\perp)$

if *optimal parameter-reduced likelihood is used* **then**

$$\widetilde{\mathcal{L}}_N^y(x_r) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}^y(x_r + x_\perp^{(i)}) \tag{27}$$

for any $x_r \in \mathrm{Im}(P_r)$

end

if *optimal parameter-reduced forward model is used* **then**

$$\widetilde{\mathcal{L}}_N^y(x_r) \propto \exp\Big(-\frac{1}{2}\|\widetilde{G}_N(x_r) - y\|^2_{\Sigma_{\mathrm{obs}}^{-1}}\Big), \widetilde{G}_N(x_r) = \frac{1}{N}\sum_{i=1}^{N} G(x_r + x_\perp^{(i)}) \tag{28}$$

for any $x_r \in \mathrm{Im}(P_r)$

end

Return: a sample-averaged approximate likelihood function $\widetilde{\mathcal{L}}_N^y(x_r)$

Subspace MCMC sampling

for $j = 1, 2, ..., K$ **do**

Given the Markov chain state $X_r^{(j-1)} = x_r$, propose a candidate $x_r^\dagger \sim q(\cdot|x_r)$

Evaluate the approximate likelihood function $\widetilde{\mathcal{L}}_N^y(x_r^\dagger)$

Compute the acceptance probability

$$\alpha(x_r^\dagger|x_r) = \min\left\{1\,,\ \frac{\widetilde{\mathcal{L}}_N^y(x_r^\dagger)\pi_{\mathrm{pr}}(x_r^\dagger)\, q(x_r|x_r^\dagger)}{\widetilde{\mathcal{L}}_N^y(x_r)\pi_{\mathrm{pr}}(x_r)\, q(x_r^\dagger|x_r)}\right\}. \tag{29}$$

With probability $\alpha(x_r^\dagger|x_r)$, **accept** $x_r^\dagger$ by setting $X_r^{(j)} = x_r^\dagger$, otherwise **reject** $x_r^\dagger$ by setting $X_r^{(j)} = x_r$.

end

Return: a Markov chain $X_r^{(1)}, X_r^{(2)}, ..., X_r^{(K)}$ with invariant density $\widetilde{\pi}_{\mathrm{pos}}^y(x_r)$

Approximate posterior sampling

for $j = 1, 2, ..., K$ **do**

Given the state $X_r^{(j)} = x_r^{(j)}$, draw a conditional prior sample $x_\perp^{(j)} \sim \pi_{\mathrm{pr}}(\cdot|x_r^{(j)})$

Compute the i-th approximate posterior sample $x^{(j)} = x_r^{(j)} + x_\perp^{(j)}$

end

Return: approximate marginal posterior samples $x^{(1)}, x^{(2)}, ..., x^{(K)}$

Even with $N = 1$, (20) and (30) differs only by a factor of 2. For the optimal parameter-reduced likelihood function, it is not obvious how to obtain a similar bound for the sampled-averaged conditional expectation in (27), see for instance the result [54, Proposition 5]. In this case, we adopt the identity (30) as a heuristic.

6. Sampling from the exact posterior

In this section, we present new strategies for sampling the exact posterior by adding minor modifications to Algorithm 4.

6.1. Pseudo-marginal for the optimal parameter-reduced likelihood

For the optimal parameter-reduced likelihood approach, Algorithm 4 replaces the optimal likelihood $\mathcal{L}^{*,y}(x_r)$ with the sample-average $\widetilde{\mathcal{L}}^y_N(x_r)$ defined by (27) using frozen (fixed) samples $\{x_\perp^{(i)}\}_{i=1}^N$. This way, Algorithm 4 produces samples from an estimation to the posterior approximation $\pi^{*,y}_{\text{pos}}(x) = \pi_{\text{pos}}(x_r)\pi_{\text{pr}}(x_\perp|x_r)$. In this section, we first show that replacing the frozen samples with freshly drawing samples $\{x_\perp^{(i)}\}_{i=1}^N$ at each MCMC iteration yields a pseudo-marginal MCMC [1] which samples exactly from $\pi^{*,y}_{\text{pos}}(x)$. In addition, we also show that an appropriate recycling of the data generated by this modified algorithm allows obtaining samples from the exact posterior $\pi^y_{\text{pos}}(x)$ itself.

We propose to modify Algorithm 4 by replacing the acceptance rate $\alpha_N(x_r^\dagger|x_r)$ in (29) with

$$\widehat{\alpha}_N(x_r^\dagger|x_r) = \min\left\{1\,,\ \frac{\pi_{\text{pr}}(x_r^\dagger)\left(\frac{1}{N}\sum_{i=1}^N \mathcal{L}^y(x_r^\dagger + x_\perp^{\dagger(i)})\right) q(x_r|x_r^\dagger)}{\pi_{\text{pr}}(x_r)\left(\frac{1}{N}\sum_{i=1}^N \mathcal{L}^y(x_r + x_\perp^{(i)})\right) q(x_r^\dagger|x_r)}\right\}. \tag{31}$$

Here, $\{x_\perp^{(i)}\}_{i=1}^N$ are i.i.d. samples from $\pi_{\text{pr}}(x_\perp|x_r)$ conditioned on the current state of the chain x_r and $\{x_\perp^{\dagger(i)}\}_{i=1}^N$ are i.i.d. samples from $\pi_{\text{pr}}(x_\perp|x_r^\dagger)$ conditioned on the proposed candidate $x_r^\dagger$. Compared to the previous acceptance rate (29) where $\{x_\perp^{(i)}\}_{i=1}^N = \{x_\perp^{\dagger(i)}\}_{i=1}^N$ where frozen, the new acceptance rate (31) requires to redraw fresh samples at each proposal candidate $x_r^\dagger$. This is summarized in Algorithm 5.

Algorithm 5: Pseudo-marginal MCMC for sampling the exact marginal posterior.

Requires: A projector P_r, a sample size N for approximating the likelihood, a total posterior sample size K, and a proposal density $q(\cdot|x_r)$ on $\text{Im}(P_r)$.

for $j = 1, 2, ..., K$ **do**

- Given the previous state of the Markov chain $X_r^{(j-1)} = x_r$ and the associated set of conditional prior samples $\{X_\perp^{(j-1,i)}\}_{i=1}^N = \{x_\perp^{(i)}\}_{i=1}^N$
- Propose a candidate $x_r^\dagger \sim q(\cdot|x_r)$
- Draw N independent samples $x_\perp^{\dagger(1)}, ..., x_\perp^{\dagger(N)} \sim \pi_{\text{pr}}(x_\perp|x_r^\dagger)$
- Compute the acceptance probability $\widehat{\alpha}(x_r^\dagger|x_r)$ as in (31)
- With probability $\widehat{\alpha}(x_r^\dagger|x_r)$, **accept** $X_r^{(j)} = x_r^\dagger$ and $\{X_\perp^{(j,i)}\}_{i=1}^N = \{x_\perp^{(\dagger,i)}\}_{i=1}^N$. Otherwise **reject** and set $X_r^{(j)} = x_r$ and $\{X_\perp^{(j,i)}\}_{i=1}^N = \{x_\perp^{(i)}\}_{i=1}^N$.

end

Return: the Markov chain $\{(X_r^{(j)}, \{X_\perp^{(j,i)}\}_{i=1}^N)\}_{j=1}^K$

In the next proposition we apply the analysis of pseudo-marginal MCMC [1] to show that $\pi^y_{\text{pos}}(x_r)$ is the invariant density of the Markov chain constructed by Algorithm 5. The key step is

to interpret Algorithm 5 as a classical Metropolis-Hastings algorithm that operates on the product space $\mathrm{Im}(P_r) \times \mathrm{Ker}(P_r)^N$.

Proposition 6.1. Algorithm 5 constructs an ergodic Markov chain $\{(X_r^{(j)}, \{X_\perp^{(j,i)}\}_{i=1}^N)\}_{j\geq 1}$ on the product space $\mathrm{Im}(P_r) \times \mathrm{Ker}(P_r)^N$ with invariant density

$$\pi_{\mathrm{tar}}^{y,N}(x_r, \{x_\perp^{(i)}\}_{i=1}^N) \propto \pi_{\mathrm{pr}}(x_r) \left(\sum_{i=1}^N \mathcal{L}^y(x_r + x_\perp^{(i)}) \right) \prod_{j=1}^N \pi_{\mathrm{pr}}(x_\perp^{(j)} | x_r). \tag{32}$$

The marginal of this target density satisfies $\pi_{\mathrm{tar}}^{y,N}(x_r) = \pi_{\mathrm{pos}}^y(x_r)$ so that the sequence $\{X_r^{(j)}\}_{j=1}^N$ is an ergodic Markov chain with the invariant density $\pi_{\mathrm{pos}}^y(x_r)$.

Proof. See Appendix B. □

Remark 6.2 (Choosing N in Algorithm 5). The statistical performance of pseudo-marginal methods depends on the variance of the sample-averaged estimate $\frac{1}{N}\sum_{i=1}^N \mathcal{L}^y(x_r + x_\perp^{(i)})$. This variance being inversely proportional to the sample size N, a larger N may result in better statistical efficiency of the MCMC chain. However, the computational cost per MCMC iteration increases linearly with N, while the improvement of the statistical efficiency will not follow the same rate. We refer the readers to [3, 23] for a detailed discussion on this topic and only provide an interpretation as follows. With $N \to \infty$, the Markov chain constructed by the pseudo-marginal MCMC converges to that of an idealized standard MCMC, which has the acceptance probability defined by the same proposal density and the exact evaluation of $\mathcal{L}^{*,y}(x_r)$. This way, even with a very large N, the statistical efficiency of the pseudo-marginal MCMC cannot be improved further beyond that of the idealized standard MCMC. As suggested by [23], the standard deviation of the logarithm of the parameter-reduced likelihood estimate, $\mathrm{var}[\log \widetilde{\mathcal{L}}_N^y]^{\frac{1}{2}}$, can be used to monitor the quality of the sample-averaged estimator.

It is remarkable to observe that, for $N = 1$, the target density (32) becomes the true posterior $\pi_{\mathrm{tar}}^{y,1}(x_r, x_\perp) = \pi_{\mathrm{pos}}^y(x_r + x_\perp)$. This means that Algorithm 5 actually produces samples $x = x_r + x_\perp$ from $\pi_{\mathrm{pos}}^y(x)$. For $N > 1$, we propose to recycle the Markov chain $\{X_\perp^{(1,i)}\}_{i=1}^N, \ldots, \{X_\perp^{(K,i)}\}_{i=1}^N$ produced by Algorithm 5 in order to generate samples from the exact posterior $\pi_{\mathrm{pos}}^y(x)$. This procedure is summarized in Algorithm 6 and a justification is provided in the following proposition.

Proposition 6.3. Let $\{(X_r^{(j)}, \{X_\perp^{(j,i)}\}_{i=1}^N)\}_{j\geq 1}$ be a Markov chain generated by Algorithm 5. For any $j \geq 1$ we randomly select $X_\perp^{(j)} \in \{X_\perp^{(j,i)}\}_{i=1}^N$ according to the probability

$$\mathbb{P}\left(X_\perp^{(j)} = X_\perp^{(j,k)} \middle| X_r^{(j)}, \{X_\perp^{(j,i)}\}_{i=1}^N \right) = \frac{\mathcal{L}^y(X_r^{(j)} + X_\perp^{(j,k)})}{\sum_{i=1}^N \mathcal{L}^y(X_r^{(j)} + X_\perp^{(j,i)})}, \quad 1 \leq k \leq N, \tag{33}$$

and we let $X^{(j)} = X_r^{(j)} + X_\perp^{(j)}$. Then $\{X^{(j)}\}_{j\geq 1}$ is a Markov chain with the exact posterior $\pi_{\mathrm{pos}}^y(x)$ as invariant density.

Proof. See Appendix C □

Algorithm 6: Recycling the Markov chain generated by Algorithm 5 to generate exact posterior samples

Requires: MCMC chain generated $\{(X_r^{(j)}, \{X_\perp^{(j,i)}\}_{i=1}^N)\}_{j=1}^K$ by Algorithm 5
for $j = 1, 2, ..., K$ **do**
  Subsample $X_\perp^{(j)} \in \{X_\perp^{(j,i)}\}_{i=1}^N$ according to the probability (33)
  Assemble $X^{(j)} = X_r^{(j)} + X_\perp^{(j)}$
end
Return: the Markov chain $\{X^{(j)}\}_{j=1}^K$ with invariant density $\pi_{\text{pos}}^y(x)$

6.2. Delayed acceptance for the optimal parameter-reduced forward model

For the optimal parameter-reduced forward model, the marginal density of the resulting approximate posterior does not coincide with that of the exact posterior in general. However, we can still modify the approximate inference algorithm 4 using the delayed acceptance technique [11, 38, 39] to explore the exact posterior. The delayed acceptance modifies Algorithm 4 by adding a second stage acceptance rejection within each MCMC iteration. Here we consider the sample-averaged likelihood $\widetilde{\mathcal{L}}_N^y(x_r)$ defined by either (27) (the optimal parameter-reduced likelihood) or (28) (the optimal parameter-reduced forward model), where the marginal prior sample set $\{x_\perp^{(i)}\}_{i=i}^N$ is prescribed. The following Proposition and Algorithm 7 detail this modification.

Proposition 6.4. Suppose we have a proposal distribution $q(\cdot|x_r)$ defined in the parameter reduced subspace $\text{Im}(P_r)$. We consider the following two stage Metropolis-Hastings method. In the first stage, we draw a proposal candidate $x_r^\dagger \sim q(\cdot|x_r)$. Then, with the probability

$$\alpha(x_r^\dagger|x_r) = \min\left\{1\,,\ \frac{\widetilde{\mathcal{L}}_N^y(x_r^\dagger)\pi_{\text{pr}}(x_r^\dagger)\,q(x_r|x_r^\dagger)}{\widetilde{\mathcal{L}}_N^y(x_r)\pi_{\text{pr}}(x_r)\,q(x_r^\dagger|x_r)}\right\}, \tag{34}$$

we move the proposal candidate $x_r^\dagger$ to the next stage. In the second stage, we draw a proposal candidate $\pi_{\text{pr}}(x_\perp^\dagger|x_r^\dagger)$ in the complement subspace $\text{Ker}(P_r)$ and then accept the pair of proposal candidates $(x_r^\dagger, x_\perp^\dagger)$ with the probability

$$\beta(x_r^\dagger, x_\perp^\dagger|x_r, x_\perp) = \min\left[1, \frac{\mathcal{L}^y(x_r^\dagger + x_\perp^\dagger)\,\widetilde{\mathcal{L}}_N^y(x_r)}{\mathcal{L}^y(x_r + x_\perp)\,\widetilde{\mathcal{L}}_N^y(x_r^\dagger)}\right]. \tag{35}$$

Then, the above procedure constructs an ergodic Markov chain with the full posterior $\pi_{\text{pos}}^y(x)$ as the invariant density.

Proof. This result can be derived from the standard delayed acceptance [11]. For completeness, we provide the proof in Appendix D. □

Remark 6.5. It worth to note that the delayed acceptance also opens the door to further accelerate the exact inference using surrogate models instead of the original forward model. The approximate likelihood $\widetilde{\mathcal{L}}_N^y(x_r)$ is deterministic and dimension reduced, which makes it possible to further approximate $\widetilde{\mathcal{L}}_N^y(x_r)$ using computationally fast surrogate models. In this case, the same delayed acceptance MCMC (Algorithm 7) can still produce ergodic Markov chains that converge to the

Algorithm 7: Delayed acceptance MCMC for sampling the exact posterior.

Requires: A projector P_r, a sample-averaged likelihood approximation defined in Algorithm 4, a total sample size K, and a proposal density $q(\cdot|x_r)$ on $\mathrm{Im}(P_r)$.

for $j = 1, 2, ..., K$ **do**

Given the Markov chain state $X^{(j-1)} = x_r + x_\perp$, propose a candidate $x_r^\dagger \sim q(\cdot|x_r)$

Compute the parameter-reduced likelihood $\widetilde{\mathcal{L}}_N^y(x_r^\dagger)$ using either using (27) or (28)

With probability $\alpha(x_r^\dagger|x_r)$ in (34) **move** $x_r^\dagger$ to the next stage as follows

Propose a candidate $x_\perp^\dagger \sim \pi_{\mathrm{pr}}(\cdot|x_r^\dagger)$

Compute the full likelihood $\mathcal{L}^y(x_r^\dagger + x_\perp^\dagger)$

With probability $\beta(x_r^\dagger, x_\perp^\dagger|x_r, x_\perp)$ in (35) **accept** $(x_r^\dagger, x^\dagger)$, otherwise **reject** $(x_r^\dagger, x^\dagger)$

Otherwise **reject** $x_r^\dagger$

end

Accept: set $X^j = x_r^\dagger + x_\perp^\dagger$

Reject: set $X^j = X^{(j-1)}$

Return: a Markov chain $X^{(1)}, X^{(2)}, ..., X^{(K)}$ with the invariant density $\pi_{\mathrm{pos}}^y(x)$

full posterior $\pi_{\mathrm{pos}}^y(x)$. In contrast, the pseudo-marginal method requires an unbiased Monte Carlo estimate of the exact marginal posterior $\pi_{\mathrm{pos}}^y(x_r)$ at every iteration, which is not straightforward to accelerate using surrogate models.

7. Non-Gaussian priors

The dimension reduction techniques presented in Sections 3 and 4 require one to evaluate the marginal prior density $\pi_{\mathrm{pr}}(x_r) = \int_{\mathrm{Ker}(P_r)} \pi_{\mathrm{pr}}(x_r + x_\perp)\mathrm{d}x_\perp$ and draw samples from the conditional prior $\pi_{\mathrm{pr}}(x_\perp|x_r) = \pi_{\mathrm{pr}}(x_r + x_\perp)/\pi_{\mathrm{pr}}(x_r)$. While these tasks are readily doable for Gaussian distributions, it might not be the case in general. In this section, we use Besov priors as an example to present strategies that can extend the proposed dimension reduction methods to some non-Gaussian priors.

7.1. Besov priors

Besov measure [21, 32, 34] naturally appears in image reconstruction problems in which the detection of edges and interfaces is important. Following [32, 34], we construct Besov priors using wavelet functions and consider functions on the one-dimensional torus $\mathbb{T} = (0, 1]$. Starting with a suitable compactly supported mother wavelet function $\psi_* \in \mathcal{L}_2(\mathbb{T})$, we can define an orthogonal basis

$$\psi_{j,k}(s) = 2^{\frac{j}{2}}\psi_*(2^j s - k), \quad j, k \in \mathbb{N}_{\geq 0}, \quad k \in [0, 2^j - 1].$$

This way, given a smoothness parameter $r > 0$ and integrability parameters $1 \leq p, q \leq \infty$, a function $f : s \mapsto f(s)$ in the Besov space $\mathcal{B}_{pq}^r(\mathbb{T})$ can be written as

$$f(s) = c_0 + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} 2^{-j(r+\frac{1}{2}-\frac{1}{p})} b_{j,k}\psi_{j,k}(s), \tag{36}$$

and satisfies

$$\|f\|_{\mathcal{B}^r_{pq}} := \left(|c_0|^q + \sum_{j=0}^{\infty} \left(\sum_{k=0}^{2^j-1} |b_{j,k}|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}} < \infty.$$

In a Bayesian setting, we can set $p = q$ and define the Besov-$\mathcal{B}^r_{pp}$ prior with the pdf‡

$$\pi_{\mathrm{pr}}(f) \propto \exp\Big(-\gamma \|f\|^p_{\mathcal{B}^r_{pp}} \Big), \tag{37}$$

where $\gamma > 0$ is a scale parameter. One can easily generalize the above definition of Besov priors to multidimensional settings by taking tensor products of the one-dimensional basis and associated coefficients.

We can discretize the Besov prior by truncating the infinite sum in (36) to the first D terms. This way, collecting all the coefficients into a parameter vector $x = (c_0, b_{0,0}, b_{0,1}, ..., b_{1,0}, b_{1,1}, ...) \in \mathbb{R}^d$, where $d = 2^{D+1}$, the discretized Besov-$\mathcal{B}^r_{pp}$ prior can be equivalently expressed as a product-form distribution over the parameter x with the pdf

$$\pi_{\mathrm{pr}}(x) = \prod_{i=1}^{d} \pi^{(i)}_{\mathrm{pr}}(x_i) \quad \text{with} \quad \pi^{(i)}_{\mathrm{pr}}(x_i) \propto \exp\left(-\gamma |x_i|^p\right). \tag{38}$$

7.2. Dimension reduction via coordinate selection

In general, we do not have closed form expressions for both the marginal $\pi_{\mathrm{pr}}(x_r)$ and the conditional $\pi_{\mathrm{pr}}(x_\perp|x_r)$, unless the projector P_r is aligned with the canonical basis. This leads to the construction of reduced subspace by selecting a subset of canonical basis. As discussed in Remarks 3.3 and 3.5, this task can be achieved by identifying an index set $\mathcal{I} \subset \{1, \dots, d\}$ with cardinality r such that $\mathcal{I}$ contains the indices of the r largest values of $i \mapsto (\Gamma^{-1})_{ii}\mathbb{E}[H(Y)]_{ii}$ in the data-dependent case or those of $i \mapsto (\Gamma^{-1})_{ii}\mathbb{E}[H(Y)]_{ii}$ in the data-free case. This leads to the projector $P_r = \sum_{i\in\mathcal{I}} e_i e_i^\top$, where $\{e_1, \dots, e_d\}$ is the canonical basis of $\mathbb{R}^d$. Thus, the product-form of (38) yields the marginal prior and the conditional prior

$$\pi_{\mathrm{pr}}(x_r) = \prod_{i\in\mathcal{I}} \pi^{(i)}_{\mathrm{pr}}(x_i) \quad \text{and} \quad \pi_{\mathrm{pr}}(x_\perp|x_r) = \prod_{i\notin\mathcal{I}} \pi^{(i)}_{\mathrm{pr}}(x_i),$$

respectively. In this formulation, evaluating the marginal prior density and drawing samples from the conditional prior become straightforward tasks.

Remark 7.1. For $1 \le q < 2$, the tails of $\pi^{(i)}_{\mathrm{pr}}(x_i)$ defined in (38) are heavier than Gaussian tails, and hence Assumptions 3.1 and 4.2 may not be satisfied. Nonetheless, one can still numerically apply the proposed dimension reduction methods without having the error bounds in (14) and (25). In this case, we set Γ to be the identity matrix in accordance with the fact that the prior components are independent and identically distributed.

Remark 7.2 (Other sparsity-inducing prior). There exist other shrinkage priors similar to Besov priors, in which the random function is expressed as a weighted linear combination of basis functions and the associated random weights follow other type of heavy tail distributions. For example, the horseshoe prior and the Student's t prior. See [10] for further discussions and references therein. The coordinate selection technique introduced here may also be applicable to those shrinkage priors.

‡ This pdf is used for demonstrating the intuition rather than a rigorous characterization, as it is defined with respect to the (non-existent) infinite-dimensional Lebesgue measure. However, the finite-dimensional discretization of the Besov measure, which is used in numerical simulations, has a pdf in this form with respect to Lebesgue measure.

7.3. Dimension reduction via prior normalization

Alternatively, we consider the case where the prior can be defined as the pushforward of the standard Gaussian measure with pdf $\mu(x) \propto \exp(-\frac{1}{2}\|x\|_2^2)$ under a C^1-diffeomorphism $T : \mathbb{R}^d \to \mathbb{R}^d$, which takes the form

$$\pi_{\text{pr}}(x) = T_\sharp \mu(x). \tag{39}$$

In other words, $\pi_{\text{pr}}(x)$ is the pdf of the random vector $X = T(Z)$ where $Z \sim \mu(z)$. For the Besov-$\mathcal{B}^r_{pp}$ prior defined in (38), the diffeomorphism T has a diagonal form $T(z) = (T_1(z_1), \ldots, T_d(z_d))$ with $T_i(z_i) = \Phi_i^{-1}(\Psi(z_i))$, where $\Phi_i(\cdot)$ is the cumulative density function (cdf) of $\pi_{\text{pr}}^{(i)}(x_i)$ defined in (38) and $\Psi(\cdot)$ is the cdf of the standard Gaussian. We provide details of the cdf $\Phi(\cdot)$ in Appendix E.

The invertibility of T allows us to reparametrize the Bayesian inverse problem in terms of the variable $z = T^{-1}(x)$, which is endowed with the Gaussian prior μ. With this change of variable, the likelihood function becomes $z \mapsto \mathcal{L}^y(T(z))$, and thus the matrix $H(y)$ used to reduce the dimension of z should be

$$H_z(y) = \int_{\mathbb{R}^d} \nabla T(z)^\top \left(\nabla \log \mathcal{L}^y(T(z))\right)\left(\nabla \log \mathcal{L}^y(T(z))\right)^\top \nabla T(z)\, \mu(z)\mathrm{d}z,$$

in the data-dependent case and $\mathbb{E}[H_z(Y)]$ in the data-dependent case. For the optimal parameter-reduced forward model in the Gaussian likelihood case (cf. Section 4), the forward model $x \mapsto G(x)$ is replaced by $z \mapsto G(T(z))$. This way, the matrix H^G should be replaced by

$$H_z^G = \int_{\mathbb{R}^d} \nabla T(z)^\top \nabla G(T(z))^\top \Sigma_{\text{obs}}^{-1} \nabla G(T(z)) \nabla T(z)\, \mu(z)\mathrm{d}z.$$

Using either of these matrices, we obtain a projector P_r to reduce the dimension in the variable $z = z_r + z_\perp$, where $z_r = P_r z$ and $z_\perp = (I_d - P_r)z$. In term of the original variable x, the dimension reduction method allows one to identify $x_r = T(P_r T^{-1}(x))$ with the observed data, while $x_\perp = T((I_d - P_r)T^{-1}(x))$ is informed by the prior only. Since x_r and $x_\perp$ are nonlinear with respect to x, the resulting method can be interpreted as a *nonlinear dimension reduction* method.

8. Example 1: elliptic PDE

We first validate our methods using an inverse problem of identifying the coefficient of a two-dimensional elliptic PDE from point observations of its solution.

8.1. Problem setup

Consider the problem domain $\Omega = [0,1]\times[0,1]$, with boundary $\partial\Omega$. We denote the spatial coordinate by $s = (s_1, s_2) \in \Omega$. We model the steady state potential solution field $p(s)$ for a given conductivity field $\kappa(s)$ and forcing function $f(s)$ using the Poisson's equation

$$-\nabla \cdot (\kappa(s)\nabla p(s)) = f(s), \quad s \in \Omega. \tag{40}$$

Let $\partial\Omega_{\text{n}} = \{s \in \partial\Omega \,|\, s_2 = 0\} \cup \{s \in \partial\Omega \,|\, s_2 = 1\}$ denote the top and bottom boundaries, and $\partial\Omega_{\text{d}} = \{s \in \partial\Omega \,|\, s_1 = 0\} \cup \{s \in \partial\Omega \,|\, s_1 = 1\}$ denote the left and right boundaries. We impose the mixed boundary condition:

$$p(s) = 0, \forall s \in \partial\Omega_{\text{d}}, \quad \text{and} \quad (\kappa(s)\nabla p(s)) \cdot \vec{n}(s) = 0, \forall x \in \partial\Omega_{\text{n}},$$

and let the forcing function take the form

$$f(s,t) = c\Big(\exp\big(-\frac{1}{2r^2}\|s-a\|^2\big) - \exp\big(-\frac{1}{2r^2}\|s-b\|^2\big)\Big), \forall t \geq 0,$$

with $r = 0.05$, which is the superposition of two Gaussian-shaped sink/source terms centered at $a = (0.5, 0.5)$ and $b = (2.5, 0.5)$, scaled by a constant $c = 6 \times 10^{-4}$. The conductivity field $\kappa(s)$ is endowed with a log-normal prior. That is, letting $x(s) = \log \kappa(s)$, the Gaussian process prior for $x(s)$ is defined by the stochastic PDE (see [37] and references therein):

$$-\triangle x(s) + \gamma x(s) = \mathcal{W}(s), \quad s \in \Omega, \tag{41}$$

where $\triangle$ is the Laplace operator and $\mathcal{W}(s)$ is the white noise process. We impose a no-flux boundary condition on the above SPDE and set $\gamma = 10$. Equations (40) and (41) are solved using the finite element method with bilinear basis functions. A mesh with 80×80 elements is used in this example. This leads to $n = 6400$ dimensional discretised parameters.

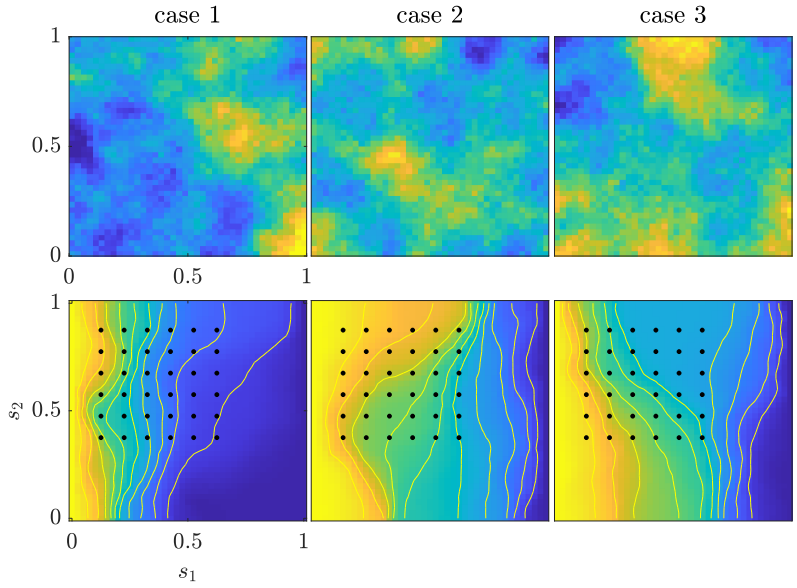


Figure 1: Setup of three test cases for the elliptic PDE example. The observation locations are shown as dots. Each column represent a test case, in which the top row shows the true conductivity fields and the bottom row shows the corresponding potential field.

We generate three "true" conductivity fields from the prior distribution and use them to simulate synthetic observed data sets. The true conductivity fields and the simulated potential fields are shown in Figure 1. Observations of the potential fields are measured at the $m = 36$ discrete locations shown as black dots in Figure 1. We set the standard derivation of the observation noise to $\sigma = 0.0415$, which corresponds to a signal-to-noise ratio of about 20.

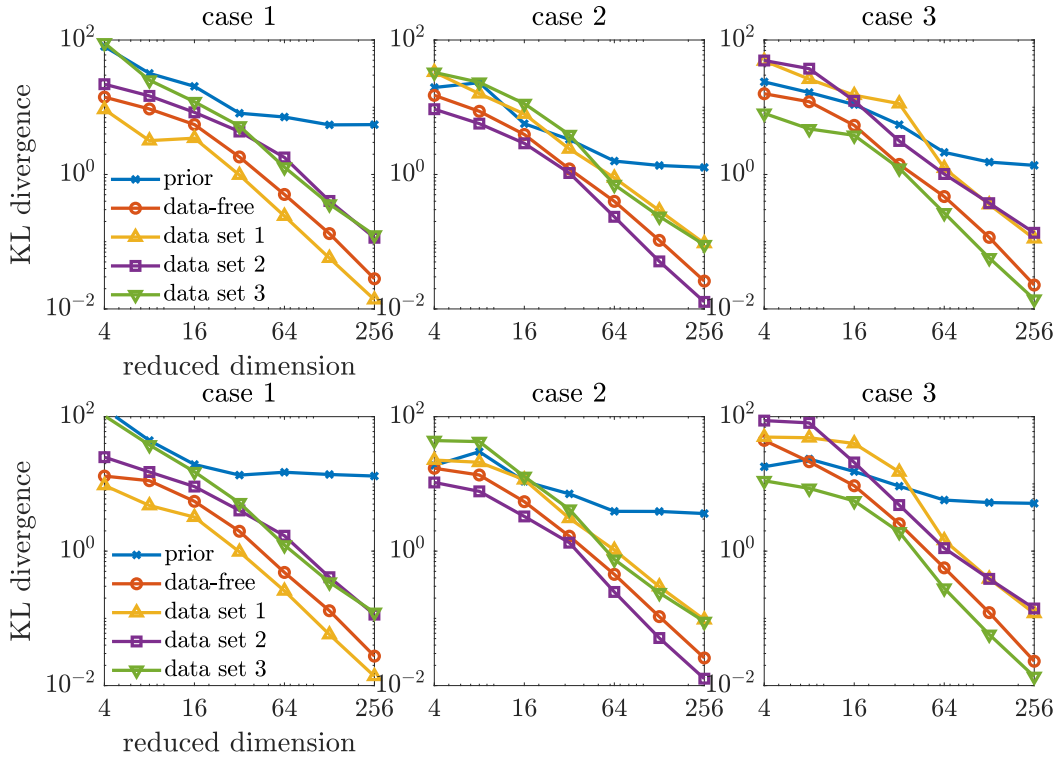


Figure 2: Elliptic PDE example. KL divergence of the full posterior densities from the approximate posterior densities defined by projectors with various ranks. Each column represents posteriors conditioned on a given data set. Top row: approximate posteriors are defined by the optimal parameter-reduced likelihood. Bottom: approximate posteriors are defined by the optimal parameter-reduced forward model.

8.2. Low-dimensional posterior approximations

We first compare the approximate posterior densities defined by the data-free dimension reduction with that of the data-dependent dimension reduction and that of the truncated Karhunen–Loéve expansion of the prior. We build five sets of projectors: the data-free projectors as detailed in Section 3.3, three sets of data-dependent projectors (see Section 3.2) that correspond to three synthetic data sets, and projectors defined by leading eigenvectors of the prior covariance (*i.e.* the truncated Karhunen–Loéve, referred to as the "prior-based projector" from hereinafter). For each of the data sets, the corresponding data-dependent projectors are constructed using the adaptive MCMC algorithm of [54]. Each set consists of projectors with ranks $r = 2^2, 2^4, ..., 2^8$. For each projector, we compute the KL divergences of the full posteriors from the approximated posterior densities defined by the optimal parameter-reduced likelihood (7). The results are shown in the top row of Figure 2. Using the same set of projectors, we also compare the KL divergences of the full posteriors from the resulting approximated posterior densities defined by the optimal parameter-reduced forward model (21). The results are shown in the bottom row of Figure 2.

In these experiments, we estimate the KL divergence using Monte Carlo integration with N posterior samples, which yields

$$D_{KL}(\pi^y_{\text{pos}}\|\tilde{\pi}^y_{\text{pos}}) \approx \frac{1}{N}\sum_{i=1}^{N}\left(\log \mathcal{L}^y(x^{(i)}) - \log \tilde{\mathcal{L}}^y(x^{(i)})\right) + \log\left(\frac{1}{N}\sum_{i=1}^{N}\exp\left(\log \tilde{\mathcal{L}}^y(x^{(i)}) - \log \mathcal{L}^y(x^{(i)})\right)\right),$$

where the second sample average accounts for the ratio between normalizing constants. For approximations that are close to the full posterior, using a reasonable number of (independent) posterior samples, e.g., 10^5 used here, make the standard deviations of the estimated KL divergence insignificant compared with the mean estimates in our numerical examples.

We observe that the optimal parameter-reduced likelihood and the optimal parameter-reduced forward model result in approximate posteriors with similar accuracy. For sufficiently large ranks ($r \geq 8$), the most accurate approximate posterior densities are obtained by the data-dependent projectors of the corresponding data set, followed by those obtained by the data-free projectors. We also observe that, for each data set, the data-dependent projectors constructed using other data sets result in less accurate approximations. By allowing a marginal loss of accuracy compared to the data-dependent construction, the data-free construction bypasses the computationally costly online dimension reduction process for every new data set. Compared with the prior-based dimension reduction, which is also an offline method, the data-free construction offers significantly more accurate approximations in this example.

For each of the data sets, we also compare the errors of the approximate posterior densities with the bounds defined in (14) and (25). Note that the right hand sides of (14) and (25) are the same up to the constant κ in this example. We plot the errors and the bounds (with $\kappa = 1$ for Gaussian prior) in Figure 3, in which all approximate posterior densities are defined by the data-free projectors. In this example, we observe that the errors of the approximate posterior densities follow the same trend as their corresponding error bounds. Note that both (14) and (25) give upper bounds on the expected KL divergence, and thus they may not bound the KL divergence for a realization of the data.

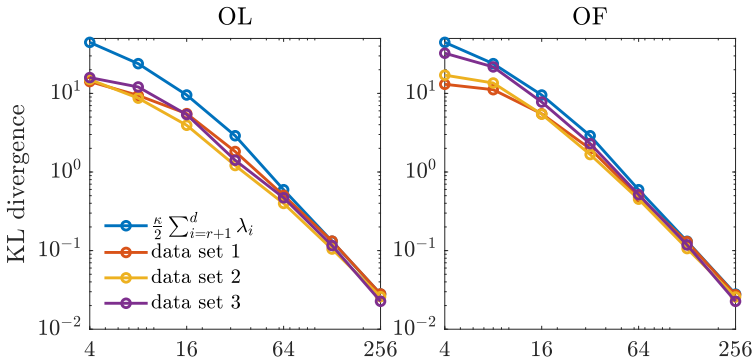


Figure 3: Elliptic PDE example. The bounds in (14) and (25) (with $\kappa = 1$) are compared to KL divergence of the full posterior densities from the approximate posterior densities defined by the data-free projectors with various ranks. Left: approximate posteriors are defined by the optimal parameter-reduced likelihood. Right: approximate posteriors are defined by the optimal parameter-reduced forward model.

8.3. Subspace accelerated sampling

Table 1: Acronyms of various inference algorithms used in numerical comparisons.

OL	approximate inference using Algorithm 4 and the optimal parameter-reduced likelihood function in (27)
PM	exact inference using the pseudo-marginal method (Algorithms 5 and 6)
OF	approximate inference using Algorithm 4 and the optimal parameter-reduced forward model in (28)
DA	exact inference using the delayed acceptance algorithm (Algorithm 7) with the approximation defined by the parameter-reduced forward model in (28)
H-MALA	exact inference using the Hessian preconditioned Langevin MCMC [45]
PCN	exact inference using the preconditioned Crank–Nicolson MCMC [7, 15]

We demonstrate the sampling performance of various approximate and exact inference algorithms introduced in Sections 5 and 6 using the posterior density conditioned on the first data set. All the methods used in the comparison and their acronyms are summarized in Table 1.

We use the Hessian-preconditioned Metropolis-Adjusted Langevin Algorithm (H-MALA) and the preconditioned Crank–Nicolson (PCN) MCMC as reference MCMC methods for sampling the full-dimensional posterior. Since H-MALA uses the low-rank decomposition of the Hessian matrix of the logarithm of the posterior density computed at the *maximum a posteriori* point to precondition MCMC, it can also be viewed as a data-dependent subspace-accelerated method. We refer to [17, 45] for a detailed discussion. In order to make a fair comparison with H-MALA, the MCMC algorithm we use on our data-free informed subspace is based on a Langevin proposal preconditioned by the same Hessian matrix used by H-MALA projected onto the data-free informed subspaces.

In Figure 4, the contours of the marginal posterior densities (marginalized onto the first two data-free LIS basis vectors) produced by approximate inference methods (with $r = 16$ and $r = 48$) are compared with those produced by their exact inference modifications (with $r = 48$). We can observe that the results produced by approximate inference methods approach those of their modifications as the rank of informed subspace increases.

To measure the efficiency of various MCMC methods, we use the average integrated autocorrelation times (IACTs) of parameters

$$\tau = \frac{1}{d}\sum_{i=1}^{d} \mathrm{IACT}(x_i),$$

where $\mathrm{IACT}(x_i)$ is the IACT of the i-th component of x. See [38, Section 12.7] for the definition of IACT. The data-free projectors with different ranks r and two different sample sizes $N = 2$ and $N = 5$ are used in this experiment. Here H-MALA and PCN are used as base cases to benchmark those MCMC methods accelerated by the informed subspace. All the methods are simulated for 5×10^5 iterations and repeated 10 times to report the mean and the standard deviation of τ. The initial state of all the simulations are randomly selected from a pre-computed Markov chain of posterior samples to avoid burn-in. The results are reported in Table 2.

For the approximate inference methods (OL and OF), the average IACTs consistently increase with the rank of the projectors, as the sampling performance of the Langevin proposal is expected

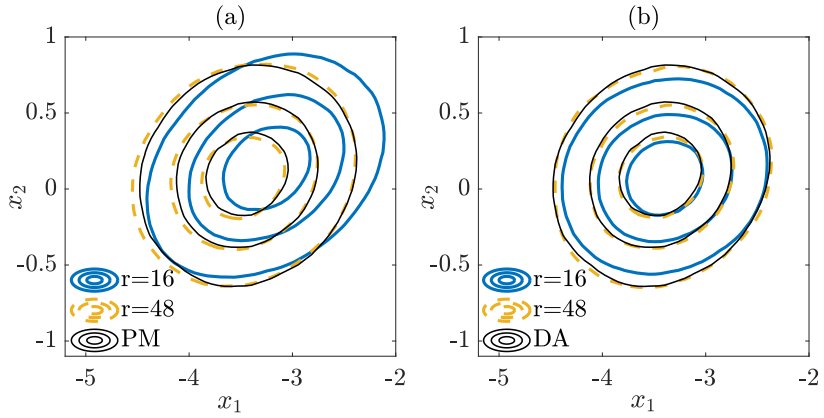


Figure 4: Elliptic PDE example. Contours of the marginal posterior densities computed by various inference algorithms using the data-free projector. (a): OL (with $r = 16$ and $r = 48$) and PM (with $r = 48$). (b): OF ($r = 16$ and $r = 48$) and DA (with $r = 48$). Here (x_1, x_2) represent the directions spanned by the first two data-free LIS basis vectors.

to decay with the underlying parameter dimension. Both OL and OF produce significantly smaller IACTs compared with the full-dimensional H-MALA.

Table 2: Elliptic PDE example. Average IACTs of parameters computed by various inference algorithms applied to the posterior conditioned on data set 1. Here the symbol - indicates poorly mixing Markov chains that do not have reliable estimate of the IACTs. All the data reported here are in the form of mean±standard derivation.

		IACT			IACT			IACT	
		OL	PM	$\sqrt{\text{var}}[\log \widetilde{\mathcal{L}}_N^y]$	OF	DA	$\mathbb{E}[\beta]$	HMALA	PCN
$N=2$	$r=16$	18.9±1.5	163±29	4.45±0.20	19.7±1.2	-	<0.1	164±17	1303±139
	$r=24$	34.9±1.1	106±13	2.65±0.19	35.7±2.1	-	<0.1		
	$r=32$	52.6±3.0	91.8±5.3	1.80±0.10	57.1±3.1	208±39	0.31±0.02		
	$r=40$	59.4±2.4	91.6±6.1	0.93±0.03	63.0±2.1	208±26	0.36±0.02		
	$r=48$	60.7±2.4	83.8±5.6	0.69±0.02	66.9±4.3	146±10	0.46±0.01		
$N=5$	$r=16$	18.7±1.0	102±8.2	2.28±0.10	19.3±1.3	-	<0.1		
	$r=24$	32.7±1.7	72.6±4.3	1.38±0.05	37.8±2.5	255±36	0.19±0.02		
	$r=32$	48.8±1.2	71.6±3.0	0.97±0.06	55.8±1.1	214±38	0.31±0.01		
	$r=40$	55.2±2.1	67.4±3.4	0.55±0.03	61.7±2.8	173±21	0.39±0.01		
	$r=48$	56.0±3.3	64.9±3.2	0.41±0.02	69.9±3.5	148±26	0.47±0.01		

Compared to the OL method, the PM method (the exact inference counterpart for OL) has a different behavior. Here we recall that the sample-averaged parameter-reduced likelihood, $\widetilde{\mathcal{L}}_N^y$, in the PM method is a random estimator, whereas $\widetilde{\mathcal{L}}_N^y$ in the OL method is deterministic because of the usage of prescribed samples. The standard deviation of the logarithm of $\widetilde{\mathcal{L}}_N^y$ in Table 2 confirms that low-rank projectors have rather large Monte Carlo errors as the approximation accuracy is

controlled by the rank truncation (cf. (14)). The exactness of the PM method comes at the cost of Monte Carlo error, which is controlled by the sample size N and the rank of the projector. We observe that increasing either the rank or the sample size can narrow the gap between the IACTs produced by PM and its OL counterpart. This experiment clearly suggests that PM needs to balance the sample size N and the rank of the projector to achieve the optimal performance.

Compared to the OF method, the DA method (the exact inference counterpart for OF) produces the largest IACTs among all subspace inference methods. This result is not surprising, as the second stage acceptance/rejection of DA necessarily deteriorates the statistical performance [11]. In Table 2, we observe that the second stage acceptance rates, $\mathbb{E}[\beta]$, increase with more accurate approximations obtained with higher projector ranks and larger sample sizes. As the result, the gaps between the IACTs produced by OF and DA are smaller for higher projector ranks and larger sample sizes.

Overall, approximate inference methods have better statistical performance compared to other methods in this example (cf. Table 2) and can obtain reasonably accurate results as shown in Figures 2 and Figure 4. With the additional cost that comes in the form of either Monte Carlo error (PM) or the second stage acceptance/rejection (DA), the exact inference modifications produce Markov chains with larger IACTs. Among all the exact inference methods, PM produces smaller IACTs compared with the full-dimensional H-MALA, PCN, and DA.

It is worth to mention that each iteration of the subspace MCMC method needs N number of forward model simulations, whereas H-MALA requires only one forward model simulation per iteration. In this example, approximate inference methods (OL and OF) with $N = 2$ still outperforms H-MALA in terms of IACTs per model evaluation. Exact inference methods, however, need more model evaluations than H-MALA to obtain the same number of effective samples (we will show in the subsection another example where H-MALA is outperformed by PM and DA). Notice that the forward model evaluations in each iteration can be embarrassingly parallelized: with parallel computing resources available, the subspace MCMC methods can still be more efficient than H-MALA in terms of the effective sample size per wall-clock time.

9. Example 2: PET with Poisson data

The second example is a two dimensional PET imaging problem, where we aim to reconstruct the density of the object from integer-valued Poisson observed data. We use here a Besov prior for which we access the coordinate selection technique and the prior normalization method presented in Sections 7.2 and 7.3.

9.1. Problem setup

In PET imaging, the goal is to identify an object of interest located inside a domain Ω subjected to gamma rays. The rays travel through Ω from multiple sources and the detectors count the number of incident photons (thus the data are integer-valued), see Figure 5a. The object of interest is described by its density of mass which is represented by $s \mapsto \exp(f(s))$, where $f : \Omega \to \mathbb{R}$ follows a Besov-$\mathcal{B}^2_{11}$ prior with the Haar wavelet, see Section 7.1. The change of intensity of a gamma ray along the path, $\ell_i(s), s \in \Omega$, can be modelled using Beer's law:

$$I_{d,i} = I_{s,i} \exp\Big(- \int_{\ell_i(s)} \exp\big(f(s)\big) ds \Big), \tag{42}$$

where $I_{d,i} \in \mathbb{R}_{\geq 0}$ and $I_{s,i} \in \mathbb{R}_{\geq 0}$ are the intensities at the detector and at the source, respectively. We assume that all the gamma ray sources have the same intensity, $I_{s,i} = I_s$ for $i = 1, ..., m$.

In this example, the domain Ω is discretized into a regular grid with d cells and the logarithm of the density is assumed to be piecewise constant. This yields the discretized parameter $x \in \mathbb{R}^d$. The line integrals in (42) are approximated by

$$\int_{\ell_i(s)} \exp(f(s))ds \approx \sum_{j=1}^{n} A_{ij} \exp(f_j),$$

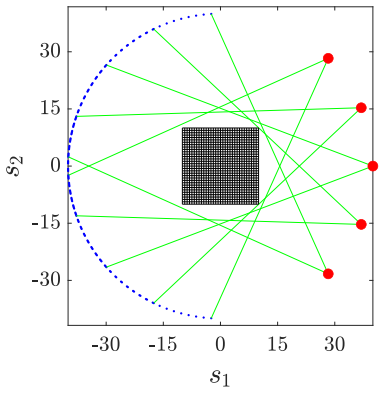
where $A_{ij} \in \mathbb{R}_{\geq 0}$ is the length of the intersection between line ℓ_i and cell j, and $\exp(f_j)$ is the discretized density in cell j. By discretizing the wavelet basis on the same grid and following the parametrization discussed in Section 7, we can write

$$f = Bx,$$

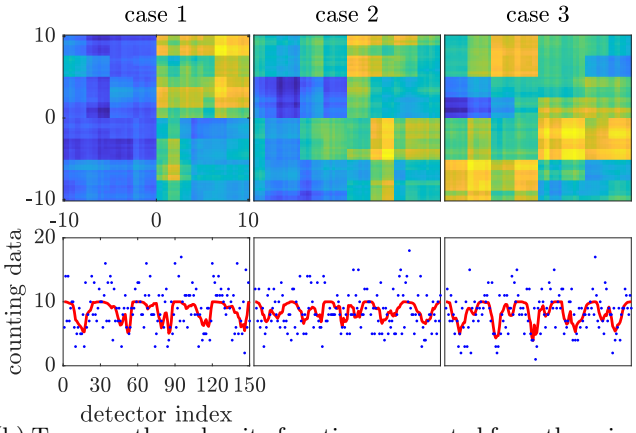
where $B \in \mathbb{R}^{d \times d}$ consists of discretized basis functions and x consists of associated coefficients. In this setting, x follows a product-form Laplace distribution given by (38) with $p = 1$ and with the scale parameter arbitrarily set to $\gamma = 1$. Suppose we have a total of m number of gamma ray paths and the corresponding matrix $A \in \mathbb{R}^{m \times d}$, the forward model $G : \mathbb{R}^d \to \mathbb{R}^m$ is defined as

$$G(x) = I_s \exp(-A \exp(Bx)). \tag{43}$$

We consider a PET setup shown in Figure 5a: the problem domain $\Omega = [-10, 10]^2$ is discretised into a $d = 64 \times 64$ regular grid, five radiation sources with intensity $I_s = 10$ are positioned with equal spaces on one side of a circle, spanning a 90° angle, and each source sends a fan of 30 gamma rays that are measured by detectors. This leads to $m = 150$ observations. The model setup is based on the code of [30].



(a) Discretised domain of interest $\Omega = [-10, 10]^2$ (mesh), radiation sources (red dots), and detectors (blue dots).



(b) Top row: three density functions generated from the prior. Bottom row: corresponding intensity function (solids lines) and measured counting data sets (dots).

Figure 5: The PET setup and three test cases.

We denote the observed data by $y \in \mathbb{N}^m$ where each element y_i is associated with the i-th gamma ray in the model. For the i-th gamma ray, recall that the intensity at the detector is computed by $G_i(x)$ for some input parameter x, and then the probability mass function of observing the counting data $Y_i = y_i$ is given by

$$\mathbb{P}(Y_i = y_i|x) = \frac{G_i(x)^{y_i} \exp(-G_i(x))}{y_i!}.$$

Suppose we can observe the counting data at all the detectors and assume the measurement processes are independent, we can write the likelihood function with the complete data as

$$\mathcal{L}^y(x) = \prod_{i=1}^{m} \mathbb{P}(Y_i = y_i|x) = \prod_{i=1}^{m} \frac{G_i(x)^{y_i} \exp(-G_i(x))}{y_i!}. \tag{44}$$

As shown in Appendix F, the Fisher information matrix of the above likelihood function takes the form

$$\mathcal{I}(x) = \nabla G(x)^\top M(x) \nabla G(x), \tag{45}$$

where $M(x)$ is a diagonal matrix with $M_{ii}(x) = G_i(x)^{-1}$ along its diagonal. We generate three "true" density functions from the prior distribution and use them to simulate synthetic data sets. The true density functions and the simulated data are shown in Figure 5b.

9.2. Numerical results using coordinate selection

We first present the results obtained by applying the coordinate selection method (cf. Section 7.2). Similar to the first example, here we will first compare the accuracy of approximate posterior densities defined by various approaches and projectors, and then benchmark the performance of MCMC methods. We adopt the same setup and acronyms as in Example 1 and Table 1.

For the approximate posterior densities, five sets of projectors built from selected coordinates, including the data-free projectors, three sets of data-dependent projectors (corresponding to three data sets), and the prior-based truncated wavelet basis, are considered. Each set consists of projectors with ranks $r = 2^3, 2^4, ..., 2^7$. The KL divergences of the full posteriors from the approximated posterior densities defined by the optimal parameter-reduced likelihood (7) are shown in the top row of Figure 6, while those of the optimal parameter-reduced forward model (21) are shown in the bottom row of Figure 6.

In this example, we observe similar results as the results of the elliptic PDE example. The optimal parameter-reduced likelihood and the optimal parameter-reduced forward model result in approximate posteriors with similar accuracy. The most accurate approximate posterior densities are obtained by the data-dependent projectors of the corresponding data set, followed by those obtained by the data-free projectors. For each data set, the data-dependent projectors constructed using other data sets result in less accurate approximations in general. However, the accuracy gaps between the data-free projectors and the data-dependent projectors (using other data sets) are not as significant as the elliptic PDE example. This can be caused by either the coordinate selection method or the rather large data size in this example. Compared with the prior-based dimension reduction, which is also an offline method, the data-free construction offers significantly more accurate approximations in this example. Overall, the data-free dimension reduction provides reasonably accurate posterior approximations for the Poisson observation process considered here.

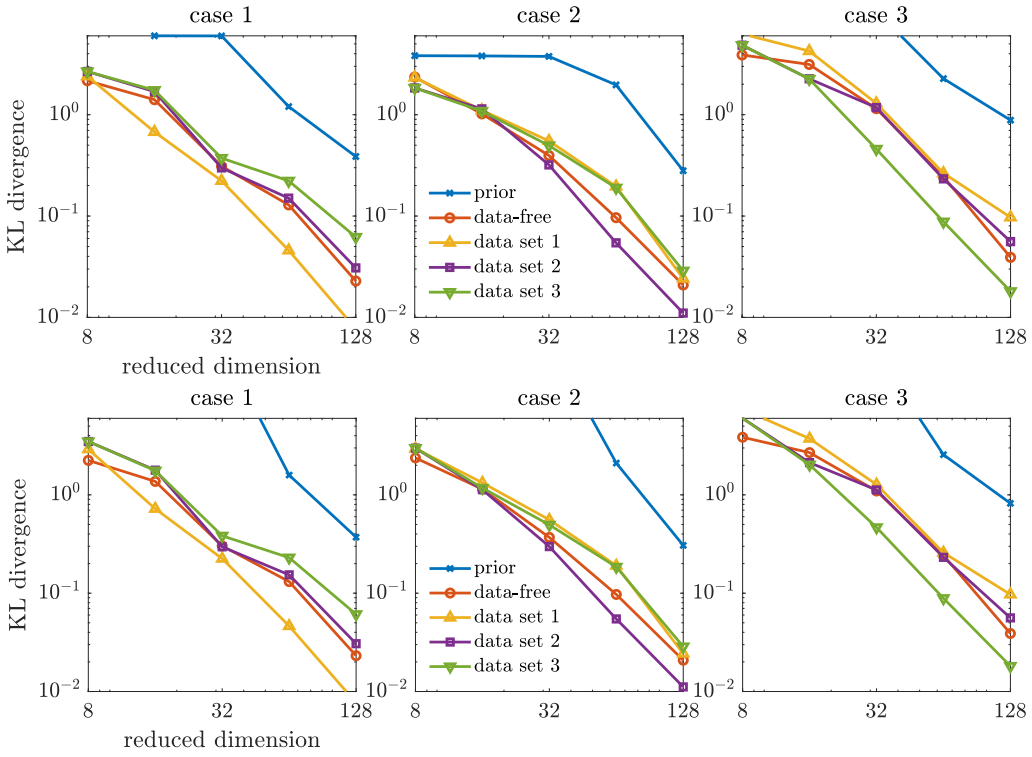


Figure 6: Same as Figure 2, but for the PET test case and where we used coordinate selection.

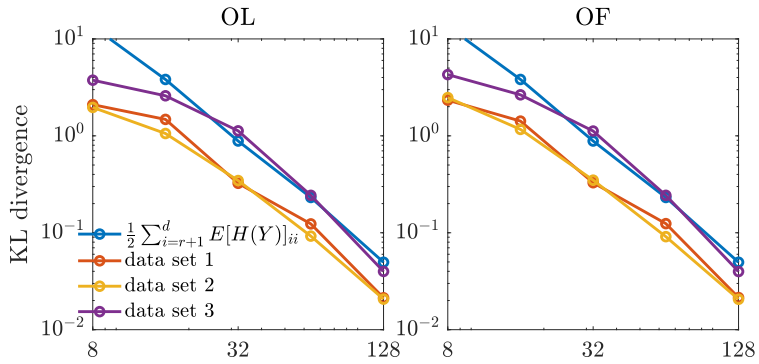


Figure 7: Same as Figure 3, but for the PET test case and where we used coordinate selection.

Although it remains an open question if the bounds in (14) and (25) can be applied for Besov priors, we provide a comparison of the errors of the approximate posterior densities (defined by the data-free projectors) with the bounds. The results are shown in Figure 7 with κ being replaced by 1. Interesting, we still observe that the errors of the approximate posterior densities follow the same trend as their corresponding bounds.

We then compare the performance of various subspace driven inference methods. In Figure 8, the contours of the the marginal posterior densities produced by approximate inference methods (with $r = 16$ and $r = 48$) are compared with those produced by their exact inference modifications (with $r = 48$). In this example, we observe that the contours produced by approximate inference methods are visually similar to those of exact inference methods. In addition, with increasing ranks, the contours produced by approximate inference methods approach those of the exact inference methods.

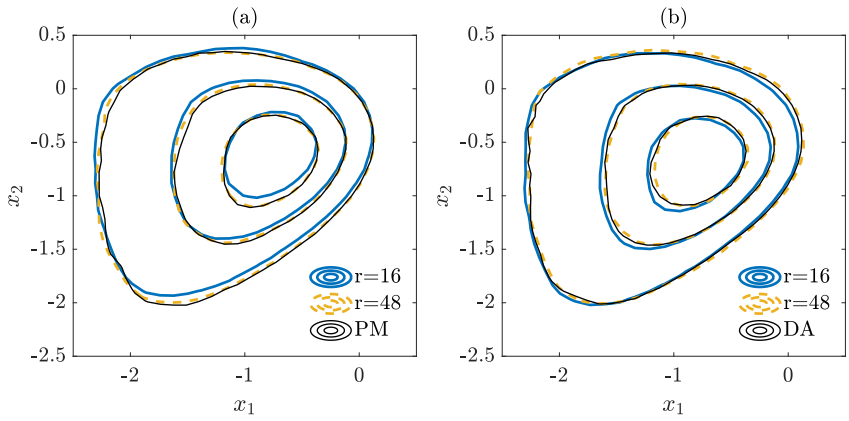


Figure 8: Same as Figure 4, but for PET and data set 1. Here $r = 16, 48$ are used in OL and OF, and $r = 48$ is used in PM and DA. Here (x_1, x_2) represent the first two coordinates selected by the data-free method.

Table 3: Same as Table 2, but for PET and data set 1. The coordinate selection is used in dimension reduction. Ranks of the subspaces are chosen to be 16, 32, and 48.

		IACT			IACT			IACT	
		OL	PM	$\sqrt{\text{var}}[\log \widetilde{\mathcal{L}}_N^y]$	OF	DA	$\mathbb{E}[\beta]$	HMALA	PCN
$N = 2$	$r = 16$	33.2±1.7	85.1±2.7	1.54±0.02	33.9±1.1	214±44	0.18±0.06	95.9±3.3	387±79
	$r = 32$	40.0±1.8	54.1±3.1	0.61±.007	41.0±2.2	87.8±6.5	0.55±0.01		
	$r = 48$	45.3±1.2	49.4±2.6	0.45±.002	46.0±2.2	73.5±5.8	0.66±0.01		
$N = 5$	$r = 16$	31.4±1.9	60.0±6.2	0.93±.006	31.8±1.4	220±65	0.22±0.04		
	$r = 32$	40.8±2.5	47.6±2.5	0.39±.004	42.8±2.4	88.0±6.8	0.56±0.01		
	$r = 48$	46.1±2.2	46.5±1.4	0.29±.001	46.3±1.9	69.5±4.0	0.67±0.01		

We use the average IACTs of the density function, $\tau = \frac{1}{d}\sum_{i=1}^{d} \text{IACT}(f_i)$, to measure the efficiency of various MCMC methods. The results are reported in Table 3. Here both PCN and H-MALA are implemented to sample the posterior in the transformed coordinate equipped with a Gaussian prior (cf. Section 7.3).

For the approximate inference methods (OL and OF), the average IACTs consistently increase with the rank of the projectors, as the sampling performance of the Langevin proposal is expected to decay with underlying the parameter dimension. Both OL and OF produce significantly smaller IACTs compared with the full-dimensional PCN and H-MALA method. Compared to the OL

method, the PM method, has a slightly higher IACTs in this example. This rather mild loss of performance (compared with the elliptic PDE example) is justified by the rather small values of $\widetilde{\mathcal{L}}_N^y$ (with $N = 2, 5$) in Table 3. Compared to the OF method, the DA method, again produces the largest IACTs among all subspace inference methods. However, the loss of performance here is not as severe as the elliptic PDE example, this is also justified by the improved second stage acceptance rates, $\mathbb{E}[\beta]$.

Overall, approximate inference methods have better statistical performance compared to other methods in this example and can obtain reasonably accurate results as shown in Figures 6 and Figure 8. With improved approximation errors, the exact inference methods also produces Markov chains with better mixing. Among all the exact inference methods, PM produces significantly smaller IACTs compared with other methods.

9.3. Numerical results using prior normalization

Then, we present the results obtained by applying the prior normalization method (cf. Section 7.3). The KL divergences of the full posteriors from the approximated posterior densities are shown in Figure 9. Here the result of prior-based dimension reduction is not presented, as the prior in the transformed space has an identity covariance matrix. We observe similar results as those obtained by the coordinate selection. We notice that the accuracy gaps between the data-free projectors and the data-dependent projectors (built using other data sets) are more significant compared with those obtained by the coordinate selection. The comparison of the errors of the approximate posterior densities (defined by the data-free projectors) with the bounds in (14) and (25) are provided in Figure 10. Here we have $\kappa = 1$ because the transformed coordinate is endowed with a Gaussian prior. We observe that the errors of the approximate posterior densities follow the same trend as their corresponding bounds. The IACTs of various MCMC methods are reported in Table 4. Again, the efficiency of subspace MCMC methods defined by the prior normalization is very close to that defined by the coordinate selection. Overall, both the coordinate selection and the prior normalization can be applied in this example to obtain accurate reduced-dimensional posterior approximations and derive efficient subspace MCMC methods.

Table 4: Same as Table 2, but for PET and data set 1. The prior normalization is used in dimension reduction. Ranks of the subspaces are chosen to be 16, 32, and 48.

		IACT			IACT			IACT	
		OL	PM	$\sqrt{\text{var}}[\log \widetilde{\mathcal{L}}_N^y]$	OF	DA	$\mathbb{E}[\beta]$	HMALA	PCN
$N = 2$	$r = 16$	35.8±1.7	81.4±6.0	1.48±0.02	33.5±1.8	168±23	0.25±0.02	95.9±3.3	387±79
	$r = 32$	42.8±2.0	55.1±2.9	0.64±.006	41.2±1.6	86.3±6.2	0.55±0.01		
	$r = 48$	45.0±2.4	51.8±2.0	0.46±.005	44.3±2.2	74.6±7.4	0.65±0.01		
$N = 5$	$r = 16$	35.1±1.9	54.1±4.1	0.88±0.01	32.8±2.8	151±21	0.26±0.02		
	$r = 32$	45.0±1.7	49.0±1.9	0.41±.003	42.0±2.6	83.1±5.1	0.55±0.01		
	$r = 48$	45.9±2.9	46.3±2.2	0.29±.003	44.4±0.8	70.6±3.7	0.66±0.01		

10. Conclusion

We present a new data-free strategy for reducing the dimensionality of large-scale statistical inverse problems. Compared to existing gradient-based dimension reduction technique, this new

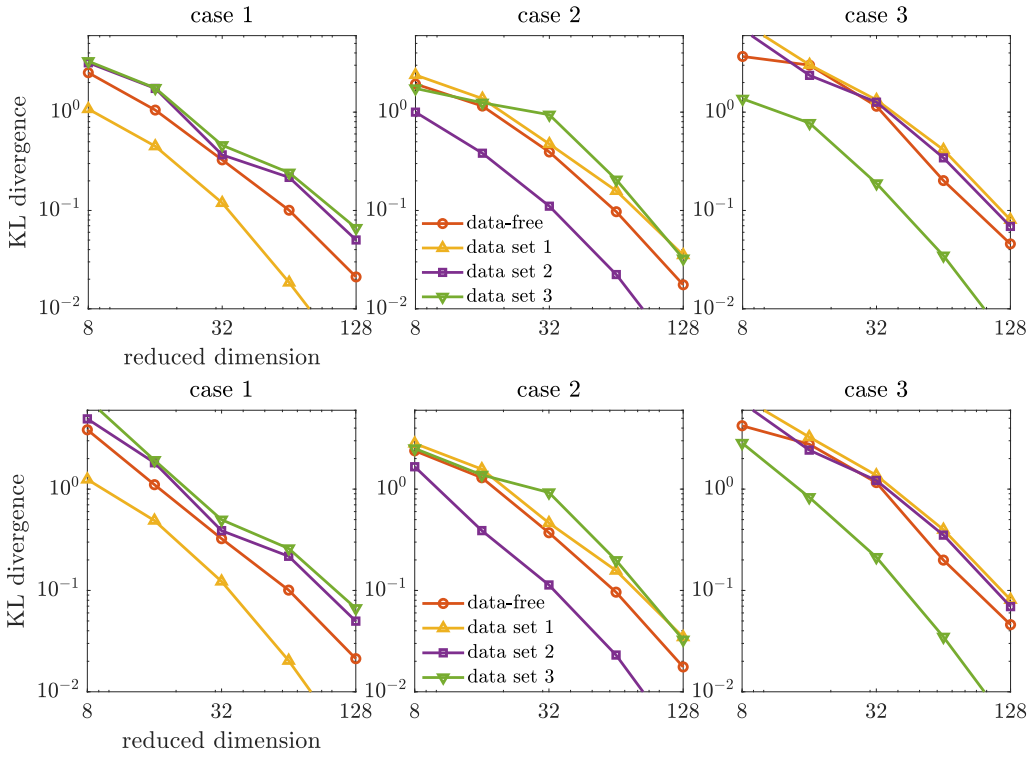


Figure 9: Same as Figure 2, but for PET. The prior normalization is used in dimension reduction.

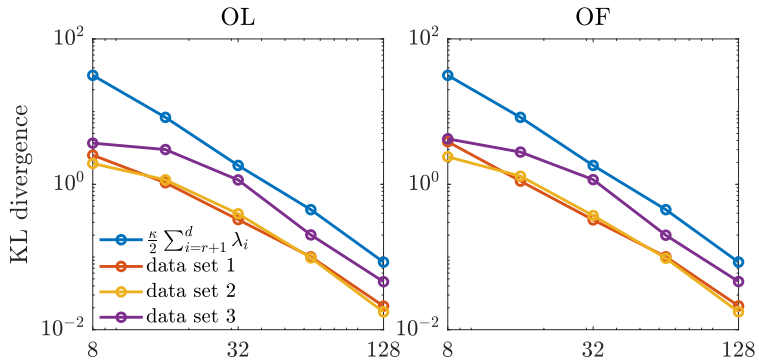


Figure 10: Same as Figure 3, but for PET. The prior normalization is used in dimension reduction.

approach identifies the computationally costly subspace construction in an offline phase. Our data-free dimension reduction is certified in the sense that its development is directly guided by factorizable posterior approximations and associated error bounds. The factorizable posterior approximations naturally offer dimension robust sampling methods for exploring the approximate posterior densities. More interestingly, by adding minor modifications to those approximate

inference algorithms, we further develop exact inference methods using the pseudo-marginal approach and the delayed acceptance approach. The resulting exact inference methods also scale well with parameter dimensionality, as the backbone of those methods is based on the dimension robust approximate inference methods. We also demonstrate the efficiency of our data-free dimensional reduction and various inference methods on two inverse problems involving a two-dimensional elliptic PDE with a Gaussian process prior and a PET problem with Poisson data and a Besov-$\mathcal{B}^2_{11}$ prior.

Acknowledgments

T. Cui and O. Zahm would like to acknowledge support from the INRIA associate team UNQUESTIONABLE. T. Cui also acknowledges support from the Australian Research Council.

Appendix A. Proof of Proposition 4.1

Recall $\pi^y_{\text{pos}}(x) = \frac{\mathcal{L}^y(x)\pi_{\text{pr}}(x)}{\pi_{\text{data}}(y)}$ and $\widehat{\pi}^y_{\text{pos}}(x) = \frac{\widehat{\mathcal{L}}^y(x)\pi_{\text{pr}}(x)}{\widehat{\pi}_{\text{data}}(y)}$. By definition of $\mathcal{L}^y(x)$ and $\widehat{\mathcal{L}}^y(x)$ we have

$$\begin{aligned} D_{\text{KL}}(\pi^y_{\text{pos}}||\widehat{\pi}^y_{\text{pos}}) &= \int_{\mathbb{R}^d} \log\left(\frac{\pi^y_{\text{pos}}(x)}{\widehat{\pi}^y_{\text{pos}}(x)}\right)\pi^y_{\text{pos}}(x)\mathrm{d}x \\ &= \log\frac{\widehat{\pi}_{\text{data}}(y)}{\pi_{\text{data}}(y)} + \int_{\mathbb{R}^d}\left(-\frac{1}{2}\|G(x)-y\|^2_{\Sigma^{-1}_{\text{obs}}} + \frac{1}{2}\|G^*(x)-y\|^2_{\Sigma^{-1}_{\text{obs}}}\right)\pi^y_{\text{pos}}(x)\mathrm{d}x \\ &= \log\frac{\widehat{\pi}_{\text{data}}(y)}{\pi_{\text{data}}(y)} + \int_{\mathbb{R}^d}\left(\frac{1}{2}\|e(x)\|^2_{\Sigma^{-1}_{\text{obs}}} - e(x)^\top\Sigma^{-1}_{\text{obs}}(G(x)-y)\right)\pi^y_{\text{pos}}(x)\mathrm{d}x \end{aligned} \tag{A.1}$$

where $e(x) = G(x) - G^*(x)$ is independent on y. Next we replace y by $Y \sim \pi_{\text{data}}$ and we take the expectation over Y. The first term in the above expression becomes

$$\mathbb{E}\left[\log\frac{\widehat{\pi}_{\text{data}}(Y)}{\pi_{\text{data}}(Y)}\right] = \int_{\mathbb{R}^m}\log\left(\frac{\widehat{\pi}_{\text{data}}(y)}{\pi_{\text{data}}(y)}\right)\pi_{\text{data}}(y)\mathrm{d}y = -D_{\text{KL}}(\pi_{\text{data}}||\widehat{\pi}_{\text{data}}). \tag{A.2}$$

Next, by definition of $\pi^y_{\text{pos}}(x)$, we have

$$\begin{aligned} \mathbb{E}\left[\int_{\mathbb{R}^d}\frac{1}{2}\|e(x)\|^2_{\Sigma^{-1}_{\text{obs}}}\pi^Y_{\text{pos}}(x)\mathrm{d}x\right] &= \frac{1}{2}\int_{\mathbb{R}^d\times\mathbb{R}^m}\|e(x)\|^2_{\Sigma^{-1}_{\text{obs}}}\pi^y_{\text{pos}}(x)\pi_{\text{data}}(y)\mathrm{d}x\mathrm{d}y \\ &= \frac{1}{2}\int_{\mathbb{R}^d\times\mathbb{R}^m}\|e(x)\|^2_{\Sigma^{-1}_{\text{obs}}}\mathcal{L}^y(x)\pi_{\text{pr}}(x)\mathrm{d}x\mathrm{d}y \\ &= \frac{1}{2}\int_{\mathbb{R}^d}\|G(x)-\widehat{G}(x)\|^2_{\Sigma^{-1}_{\text{obs}}}\pi_{\text{pr}}(x)\mathrm{d}x. \end{aligned} \tag{A.3}$$

For the last equality we used the fact that $y \mapsto \mathcal{L}^y(x)$ is a pdf so that $\int_{\mathbb{R}^m} \mathcal{L}^y(x)\mathrm{d}y = 1$. Using the same arguments, we have

$$\begin{aligned}
&\mathbb{E}\left[\int_{\mathbb{R}^d} e(x)^\top \Sigma_{\text{obs}}^{-1}\big(G(x) - Y\big)\pi_{\text{pos}}^Y(x)\mathrm{d}x\right] \\
&\quad= \int_{\mathbb{R}^d\times\mathbb{R}^m} e(x)^\top \Sigma_{\text{obs}}^{-1}\big(G(x) - y\big)\mathcal{L}^y(x)\pi_{\text{pr}}(x)\mathrm{d}x\mathrm{d}y \\
&\quad= \int_{\mathbb{R}^d} e(x)^\top \Sigma_{\text{obs}}^{-1} G(x)\pi_{\text{pr}}(x)\mathrm{d}x - \int_{\mathbb{R}^d} e(x)^\top \Sigma_{\text{obs}}^{-1}\left(\int_{\mathbb{R}^m} y\mathcal{L}^y(x)\mathrm{d}y\right)\pi_{\text{pr}}(x)\mathrm{d}x \\
&\quad= 0. \qquad\qquad (A.4)
\end{aligned}$$

The last equality is obtained by noting that the expectation of the data knowing the parameter x is $\int_{\mathbb{R}^m} y\mathcal{L}^y(x)\mathrm{d}y = G(x)$. Combining (A.2) (A.3) and (A.4), we obtain

$$\mathbb{E}\left[D_{\text{KL}}(\pi_{\text{pos}}^Y||\widehat{\pi}_{\text{pos}}^Y)\right] \overset{(A.1)}{=} -D_{\text{KL}}(\pi_{\text{data}}||\widehat{\pi}_{\text{data}}) + \frac{1}{2}\int_{\mathbb{R}^d} \|G(x) - \widehat{G}(x)\|^2_{\Sigma_{\text{obs}}^{-1}}\pi_{\text{pr}}(x)\mathrm{d}x,$$

which concludes the proof.

Appendix B. Proof of Proposition 6.1

Consider a Metropolis-Hastings algorithm which targets the pdf $\pi_{\text{tar}}^{y,N}$ defined by (32) using the following proposal density

$$q\left(x_r^\dagger, \{x_\perp^{\dagger(i)}\}_{i=1}^N \middle| x_r, \{x_\perp^{(i)}\}_{i=1}^N\right) = q(x_r^\dagger|x_r)\prod_{i=1}^N \pi_{\text{pr}}(x_\perp^{\dagger(i)}|x_r^\dagger), \qquad (B.1)$$

where $q(x_r^\dagger|x_r)$ is the same proposal density as the one used at step 1 of Algorithm (5). The acceptance probability of this Metropolis-Hastings algorithm is given by

$$\begin{aligned}
\alpha\left(x_r^\dagger, \{x_\perp^{\dagger(i)}\}_{i=1}^N \middle| x_r, \{x_\perp^{(i)}\}_{i=1}^N\right) &= \min\left\{1, \frac{\pi_{\text{tar}}^{y,N}(x_r^\dagger, \{x_\perp^{\dagger(i)}\}_{i=1}^N)\; q\left(x_r, \{x_\perp^{(i)}\}_{i=1}^N \middle| x_r^\dagger, \{x_\perp^{\dagger(i)}\}_{i=1}^N\right)}{\pi_{\text{tar}}^{y,N}(x_r, \{x_\perp^{(i)}\}_{i=1}^N)\; q\left(x_r^\dagger, \{x_\perp^{\dagger(i)}\}_{i=1}^N \middle| x_r, \{x_\perp^{(i)}\}_{i=1}^N\right)}\right\} \\
&= \min\left\{1, \frac{\pi_{\text{pr}}(x_r^\dagger)\left(\sum_{i=1}^N \mathcal{L}^y(x_r^\dagger + x_\perp^{\dagger(i)})\right) q(x_r|x_r^\dagger)}{\pi_{\text{pr}}(x_r)\left(\sum_{i=1}^N \mathcal{L}^y(x_r + x_\perp^{(i)})\right) q(x_r^\dagger|x_r)}\right\},
\end{aligned}$$

which is precisely $\widehat{\alpha}_N(x_r^\dagger|x_r)$ defined in (31). Note that the first two steps of Algorithm 5 consists of drawing a sample $(x_r^\dagger, \{x_\perp^{\dagger(i)}\}_{i=1}^N)$ from the proposal (B.1). This way, Algorithm 5 can be interpreted as a MCMC algorithm which targets $\pi_{\text{tar}}^{y,N}$. It remains to show that the marginal distribution

$\pi_{\text{tar}}^{y,N}(x_r)$ is the marginal posterior $\pi_{\text{pos}}^{y}(x_r)$. We can write

$$\begin{aligned}\pi_{\text{tar}}^{y,N}(x_r) &= \int_{\text{Ker}(P_r)^N} \pi_{\text{tar}}^{y,N}(x_r, x_\perp^{(1)}, ..., x_\perp^{(N)}) \mathrm{d}x_\perp^{(1)} ... \mathrm{d}x_\perp^{(N)} \\ &\propto \pi_{\text{pr}}(x_r) \sum_{i=1}^{N} \int_{\text{Ker}(P_r)} \mathcal{L}^y(x_r + x_\perp^{(i)}) \pi_{\text{pr}}(x_\perp^{(i)} | x_r) \mathrm{d}x_\perp^{(i)} \\ &\propto \int_{\text{Ker}(P_r)} \mathcal{L}^y(x_r + x_\perp^{(i)}) \pi_{\text{pr}}(x_r + x_\perp^{(i)}) \mathrm{d}x_\perp \propto \pi_{\text{pos}}^{y}(x_r)\end{aligned}$$

which concludes the proof.

Appendix C. Proof of Proposition 6.3

Recall that $\{(X_r^{(j)}, \{X_\perp^{(j,i)}\}_{i=1}^N)\}_{j\geq 1}$ admits $\pi_{\text{tar}}^{y,N}$ (32) as the invariant density, see Proposition 6.1. It remains to prove that $\{X^{(j)}\}_{j\geq 1}$ admits $\pi_{\text{pos}}^{y}(x)$ as the invariant density. For a given state $(X_r^{(j)}, \{X_\perp^{(j,i)}\} = (x_r, \{x_\perp^{(i)}\}_{i=1}^N)$, we have $X^{(j)} = x_r + x_\perp$ where $x_\perp \in \{x_\perp^{(i)}\}_{i=1}^N$ is selected with respect to the probability

$$\mathbb{P}\left(x_\perp = x_\perp^{(k)} \middle| x_r, \{x_\perp^{(i)}\}_{i=1}^N\right) = \frac{\mathcal{L}^y(x_r + x_\perp^{(k)})}{\sum_{i=1}^N \mathcal{L}^y(x_r + x_\perp^{(i)})}, \quad 1 \leq k \leq N. \tag{C.1}$$

Thus, we need to prove that the pdf $\pi(x)$ where $x = x_r + x_\perp$ is the posterior density $\pi(x) = \pi_{\text{pos}}^{y}(x)$. We can write

$$\begin{aligned}\pi(x) &= \pi(x_r, x_\perp) \\ &= \int_{\text{Ker}(P_r)^N} \pi(x_r, \{x_\perp^{(i)}\}_{i=1}^N, x_\perp) \, \mathrm{d}x_\perp^{(1)} ... \mathrm{d}x_\perp^{(N)} \\ &= \int_{\text{Ker}(P_r)^N} \pi\left(x_\perp \middle| x_r, \{x_\perp^{(i)}\}_{i=1}^N\right) \pi_{\text{tar}}^{y,N}\left(x_r, \{x_\perp^{(i)}\}_{i=1}^N\right) \mathrm{d}x_\perp^{(1)} ... \mathrm{d}x_\perp^{(N)},\end{aligned}$$

where $\pi(x_\perp | x_r, \{x_\perp^{(i)}\}_{i=1}^N)$ is the pdf of $x_\perp$ conditioned on $(x_r, \{x_\perp^{(i)}\}_{i=1}^N)$. By construction we have

$$\pi\left(x_\perp \middle| x_r, \{x_\perp^{(i)}\}_{i=1}^N\right) \overset{(C.1)}{=} \frac{\sum_{k=1}^N \delta_{x_\perp^{(k)}}(x_\perp) \mathcal{L}^y(x_r + x_\perp^{(k)})}{\sum_{i=1}^N \mathcal{L}^y(x_r + x_\perp^{(i)})}, \tag{C.2}$$

where $\delta_{x_\perp^{(k)}}$ denotes the Dirac mass function at point $x_\perp^{(k)}$. We can write

$$\begin{aligned}
\pi(x) &= \int_{\mathrm{Ker}(P_r)^N} \frac{\sum_{k=1}^N \delta_{x_\perp^{(k)}}(x_\perp)\mathcal{L}^y(x_r + x_\perp^{(k)})}{\sum_{i=1}^N \mathcal{L}^y(x_r + x_\perp^{(i)})} \pi_{\mathrm{tar}}^{y,N}\left(x_r, \{x_\perp^{(i)}\}_{i=1}^N\right) \mathrm{d}x_\perp^{(1)}...\mathrm{d}x_\perp^{(N)} \\
&\overset{(32)}{\propto} \sum_{k=1}^N \int_{\mathrm{Ker}(P_r)^N} \delta_{x_\perp^{(k)}}(x_\perp)\mathcal{L}^y(x_r + x_\perp^{(k)})\pi_{\mathrm{pr}}(x_r) \prod_{i=1}^N \pi_{\mathrm{pr}}(x_\perp^{(i)}|x_r)\, \mathrm{d}x_\perp^{(1)}...\mathrm{d}x_\perp^{(N)} \\
&\propto \sum_{k=1} \int_{\mathrm{Ker}(P_r)} \delta_{x_\perp^{(k)}}(x_\perp)\mathcal{L}^y(x_r + x_\perp^{(k)})\pi_{\mathrm{pr}}(x_r)\pi_{\mathrm{pr}}(x_\perp^{(k)}|x_r)\, \mathrm{d}x_\perp^{(k)} \\
&\propto \sum_{k=1} \mathcal{L}^y(x_r + x_\perp)\pi_{\mathrm{pr}}(x_r)\pi_{\mathrm{pr}}(x_\perp|x_r) \propto \pi_{\mathrm{pos}}^y(x_r + x_\perp),
\end{aligned}$$

which concludes the proof.

Appendix D. Proof of Proposition 6.4

To show the result of Proposition 6.4, we first interpret the first stage acceptance/rejection and the conditional prior sampling $\pi_{\mathrm{pr}}(x_\perp^\dagger|x_r^\dagger)$ as a joint proposal acting in the full parameter space $\mathrm{Im}(P_r)\oplus\mathrm{Ker}(P_r)$. The proposal $q(\cdot,|x_r)$ and the acceptance probability $\alpha(x_r^\dagger|x_r)$ defines an effective proposal distribution

$$\bar{q}(x_r^\dagger|x_r) = q(x_r^\dagger|x_r)\alpha(x_r^\dagger|x_r) + \left[1 - \int q(x_r^\dagger|x_r)\alpha(x_r'|x_r)dx_r'\right]\delta_{x_r}(x_r^\dagger),$$

where $\delta_{x_r}(\cdot)$ denotes the Dirac delta and the term in the bracket represents the probability of a proposal candidate being rejected. Then, we can define a joint proposal distribution

$$Q(x_r^\dagger, x_\perp^\dagger|x_r, x_\perp) := Q(x_r^\dagger, x_\perp^\dagger|x_r) = \pi_{\mathrm{pr}}(x_\perp^\dagger|x_r^\dagger)\,\bar{q}(x_r^\dagger|x_r),$$

for the MH to sample the full posterior.

Following the exactly same derivation in [11], one can show that accepting $(x_r^\dagger, x_\perp^\dagger) \sim Q(\cdot,\cdot|x_r, x_\perp)$ with the probability

$$\beta(x_r^\dagger, x_\perp^\dagger|x_r, x_\perp) = \min\left[1, \frac{\pi_{\mathrm{pos}}^y(x_r^\dagger + x_\perp^\dagger)\,Q(x_r, x_\perp|x_r^\dagger, x_\perp^\dagger)}{\pi_{\mathrm{pos}}^y(x_r + x_\perp)\,Q(x_r^\dagger, x_\perp^\dagger|x_r, x_\perp)}\right]$$

defines a Markov transition kernel with the full posterior $\pi_{\mathrm{pos}}^y(x_r + x_\perp)$ as its invariant distribution. Since the above acceptance probability is only used in the case where the first stage proposal candidate $x_r^\dagger$ is accepted, i.e., $x_r^\dagger \neq x_r$, we do not need to consider the Dirac delta term. This way, the above acceptance probability can be simplified as

$$\beta(x_r^\dagger, x_\perp^\dagger|x_r, x_\perp) = \min\left[1, \frac{\pi_{\mathrm{pos}}^y(x_r^\dagger + x_\perp^\dagger)\,\pi_{\mathrm{pr}}(x_\perp|x_r)\,q(x_r|x_r^\dagger)\,\alpha(x_r|x_r^\dagger)}{\pi_{\mathrm{pos}}^y(x_r + x_\perp)\,\pi_{\mathrm{pr}}(x_\perp^\dagger|x_r^\dagger)\,q(x_r^\dagger|x_r)\alpha(x_r^\dagger|x_r)}\right].$$

Substituting the identities

$$\frac{\alpha(x_r|x_r^\dagger)}{\alpha(x_r^\dagger|x_r)} = \frac{\widetilde{\mathcal{L}}_N^y(x_r)\pi_{\mathrm{pr}}(x_r)\,q(x_r^\dagger|x_r)}{\widetilde{\mathcal{L}}_N^y(x_r^\dagger)\pi_{\mathrm{pr}}(x_r^\dagger)\,q(x_r|x_r^\dagger)},$$

and

$$\frac{\pi^y_{\text{pos}}(x_r^\dagger + x_\perp^\dagger)\,\pi_{\text{pr}}(x_\perp|x_r)}{\pi^y_{\text{pos}}(x_r + x_\perp)\,\pi_{\text{pr}}(x_\perp^\dagger|x_r^\dagger)} = \frac{\mathcal{L}^y(x_r^\dagger + x_\perp^\dagger)\,\pi_{\text{pr}}(x_r^\dagger)}{\mathcal{L}^y(x_r + x_\perp)\,\pi_{\text{pr}}(x_r)},$$

into the above equation, we obtain

$$\beta(x_r^\dagger, x_\perp^\dagger|x_r, x_\perp) = \min\left[1, \frac{\mathcal{L}^y(x_r^\dagger + x_\perp^\dagger)\,\widetilde{\mathcal{L}}^y_N(x_r)}{\mathcal{L}^y(x_r + x_\perp)\,\widetilde{\mathcal{L}}^y_N(x_r^\dagger)}\right],$$

which is identical to the second stage acceptance probability in (35). Thus, the result follows.

Appendix E. Cumulative density function of $p(x) \propto \exp(-\gamma|x|^p)$

Given the pdf $p(x) = \frac{1}{c_{\gamma,p}}\exp(-\gamma|x|^p)$, $x \in \mathbb{R}$, we want to find its normalizing constant $c_{\gamma,p}$ and cdf. Using symmetry, the normalizing constant takes the form

$$c_{\gamma,p} = 2\int_0^\infty \exp(-\gamma x^p)dx,$$

and the cdf can be expressed as

$$\Phi(x) = \begin{cases} \dfrac{1}{2} + \dfrac{1}{c_{\gamma,p}}\displaystyle\int_0^x \exp(-\gamma t^p)dt & x \geq 0 \\ \dfrac{1}{2} - \dfrac{1}{c_{\gamma,p}}\displaystyle\int_0^{-x} \exp(-\gamma t^p)dt & x < 0 \end{cases}.$$

We first introduce the change-of-variable $z = \gamma t^p$ so that

$$t = \left(\frac{z}{\gamma}\right)^{\frac{1}{p}} \quad \text{and} \quad \frac{\mathrm{d}t}{\mathrm{d}z} = p^{-1}\gamma^{-\frac{1}{p}}\, z^{\frac{1}{p}-1}.$$

This yields

$$\int_0^x \exp(-\gamma t^p)dt = p^{-1}\gamma^{-\frac{1}{p}}\int_0^{\gamma x^p} z^{\frac{1}{p}-1}\exp(-z)dz := p^{-1}\gamma^{-\frac{1}{p}}\Gamma_{\text{lower}}(p^{-1}, \gamma x^p),$$

where $\Gamma_{\text{lower}}(\cdot,\cdot)$ is the lower incomplete gamma function. Following a similar derivation, we obtain $c_{\gamma,p} = 2\,p^{-1}\gamma^{-\frac{1}{p}}\,\Gamma(p^{-1})$, where $\Gamma(\cdot)$ is the Gamma function. This way, we have the cdf

$$\Phi(x) = \frac{1}{2} + \frac{\text{sign}(x)}{2\,\Gamma(p^{-1})}\Gamma_{\text{lower}}\Big(p^{-1}, \gamma\,\big(\text{sign}(x)\,x\big)^p\Big).$$

There are two notable special cases. The Gaussian distribution $\mathcal{N}(0, \sigma^2)$ can be specified using $\gamma = (2\sigma^2)^{-1}$ and $p = 2$, in which the cdf can be equivalently expressed using the error function. The Laplace distribution can be specified using $p = 1$, so that the cdf yields a simpler (but equivalent) expression in the form of

$$\Phi(x) = \frac{1}{2} + \frac{\text{sign}(x)}{2}\Big(1 - \exp\big(-\gamma\,\text{sign}(x)\,x\big)\Big).$$

Appendix F. Derivation of Fisher information matrices

Here we derive the Fisher information matrix for the Poisson likelihood case. Recall the Fisher information matrix

$$\mathcal{I}(x) = \int_{\mathbb{R}^m} \left(\nabla \log \mathcal{L}^y(x)\right)\left(\nabla \log \mathcal{L}^y(x)\right)^\top \mathcal{L}^y(x)\,\mathrm{d}y. \tag{F.1}$$

Defining the predata $\eta = G(x)$, we can express the gradient of the likelihood function as

$$\nabla_x \log \mathcal{L}^y(x) = \nabla G(x)^\top \nabla_\eta \log \mathcal{L}^y(\eta).$$

where

$$\mathcal{L}^y(\eta) = \prod_{i=1}^m \frac{\eta_i^{y_i}\,\exp(-\eta_i)}{y_i!}, \quad \text{subject to} \quad \eta = G(x).$$

This way, the Fisher information matrix can be rewritten as

$$\mathcal{I}(x) = \nabla G(x)^\top \left(\int_{\mathbb{R}^m} \left(\nabla \log \mathcal{L}^y(\eta)\right)\left(\nabla \log \mathcal{L}^y(\eta)\right)^\top \mathcal{L}^y(\eta)\,\mathrm{d}y \right) \nabla G(x), \tag{F.2}$$

subject to $\eta = G(x)$. The term in the brackets of the above equation is the Fisher information matrix of the Poisson distribution, which is a diagonal matrix

$$\left(\int_{\mathbb{R}^m} \left(\nabla \log \mathcal{L}^y(\eta)\right)\left(\nabla \log \mathcal{L}^y(\eta)\right)^\top \mathcal{L}^y(\eta)\,\mathrm{d}y \right)_{ii} = \frac{1}{\eta_i}.$$

Thus, the Fisher information matrix w.r.t. x is

$$\mathcal{I}(x) = \nabla G(x)^\top M(x) \nabla G(x), \tag{F.3}$$

where $M(x)$ is a diagonal matrix with $M_{ii}(x) = G_i(x)^{-1}$ along its diagonal.

References

[1] Christophe Andrieu, Gareth O Roberts, et al. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

[2] Christophe Andrieu, Matti Vihola, et al. Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *The Annals of Applied Probability*, 25(2):1030–1077, 2015.

[3] Christophe Andrieu, Matti Vihola, et al. Establishing some order amongst exact approximations of MCMCs. *The Annals of Applied Probability*, 26(5):2661–2696, 2016.

[4] Mario Bebendorf. A note on the Poincaré inequality for convex domains. *Zeitschrift für Analysis und ihre Anwendungen*, 22(4):751–756, 2003.

[5] Alexandros Beskos, Mark Girolami, Shiwei Lan, Patrick E Farrell, and Andrew M Stuart. Geometric MCMC for infinite-dimensional inverse problems. *Journal of Computational Physics*, 335:327–351, 2017.

[6] Alexandros Beskos, Ajay Jasra, Kody Law, Youssef Marzouk, and Yan Zhou. Multilevel sequential Monte Carlo with dimension-independent likelihood-informed proposals. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):762–786, 2018.

[7] Alexandros Beskos, Gareth Roberts, Andrew Stuart, and Jochen Voss. MCMC methods for diffusion bridges. *Stochastics and Dynamics*, 8(03):319–350, 2008.

[8] Marie Billaud-Friess, Anthony Nouy, and Olivier Zahm. A tensor approximation method based on ideal minimal residual formulations for the solution of high-dimensional problems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 48(6):1777–1806, 2014.

[9] S. Brooks, A. Gelman, G. Jones, and X. L. Meng, editors. *Handbook of Markov Chain Monte Carlo*. Taylor & Francis, 2011.

[10] Carlos M Carvalho, Nicholas G Polson, and James G Scott. Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80, 2009.

[11] J Andrés Christen and Colin Fox. Markov chain Monte Carlo using an approximation. *Journal of Computational and Graphical statistics*, 14(4):795–810, 2005.

[12] Paul G Constantine, Eric Dow, and Qiqi Wang. Active subspace methods in theory and practice: applications to kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4):A1500–A1524, 2014.

[13] Paul G Constantine, Carson Kent, and Tan Bui-Thanh. Accelerating Markov chain Monte Carlo with active subspaces. *SIAM Journal on Scientific Computing*, 38(5):A2779–A2805, 2016.

[14] Andrea F Cortesi, Paul G Constantine, Thierry E Magin, and Pietro M Congedo. Forward and backward uncertainty quantification with active subspaces: application to hypersonic flows around a cylinder. *Journal of Computational Physics*, 407:109079, 2020.

[15] Simon L Cotter, Gareth O Roberts, Andrew M Stuart, David White, et al. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28(3):424–446, 2013.

[16] Tiangang Cui, Gianluca Detommaso, and Robert Scheichl. Multilevel dimension-independent likelihood-informed MCMC for large-scale inverse problems. *arXiv preprint arXiv:1910.12431*, 2019.

[17] Tiangang Cui, Kody JH Law, and Youssef M Marzouk. Dimension-independent likelihood-informed MCMC. *Journal of Computational Physics*, 304:109–137, 2016.

[18] Tiangang Cui, James Martin, Youssef M Marzouk, Antti Solonen, and Alessio Spantini. Likelihood-informed dimension reduction for nonlinear inverse problems. *Inverse Problems*, 30(11):114015, 2014.

[19] Tiangang Cui and Xin T Tong. A unified performance analysis of likelihood-informed subspace methods. *arXiv preprint arXiv:2101.02417*, 2021.

[20] Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

[21] Masoumeh Dashti, Stephen Harris, and Andrew Stuart. Besov priors for Bayesian inverse problems. *Inverse Problems & Imaging*, 6(2):183, 2012.

[22] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

[23] Arnaud Doucet, Michael K Pitt, George Deligiannidis, and Robert Kohn. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313, 2015.

[24] Alain Durmus, Eric Moulines, et al. High-dimensional Bayesian inference via the unadjusted langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

[25] Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183):1–42, 2019.

[26] Ivan G Graham, Frances Y Kuo, James A Nichols, Robert Scheichl, Ch Schwab, and Ian H Sloan. Quasi-Monte Carlo finite element methods for elliptic PDEs with lognormal random coefficients. *Numerische Mathematik*, 131(2):329–368, 2015.

[27] Ivan G Graham, Frances Y Kuo, Dirk Nuyens, Robert Scheichl, and Ian H Sloan. Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications. *Journal of Computational Physics*, 230(10):3668–3694, 2011.

[28] Alice Guionnet and B Zegarlinksi. Lectures on logarithmic Sobolev inequalities. In *Séminaire de probabilités XXXVI*, pages 1–134. Springer, 2003.

[29] W. Hastings. Monte Carlo sampling using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[30] Jere Heikkinen. Statistical inversion theory in x-ray tomography. 2008.

[31] Jari Kaipio and Erkki Somersalo. *Statistical and computational inverse problems*, volume 160. Springer Science & Business Media, 2006.

[32] Ville Kolehmainen, Matti Lassas, Kati Niinimäki, and Samuli Siltanen. Sparsity-promoting Bayesian inversion. *Inverse Problems*, 28(2):025005, 2012.

[33] Shiwei Lan. Adaptive dimension reduction to accelerate infinite-dimensional geometric Markov chain Monte Carlo. *Journal of Computational Physics*, 392:71–95, 2019.

[34] Matti Lassas, Eero Saksman, and Samuli Siltanen. Discretization-invariant Bayesian inversion and Besov space priors. *Inverse problems and imaging*, 3(1):87–122, 2009.

[35] Olivier Le Maître and Omar M Knio. *Spectral methods for uncertainty quantification: with applications to computational fluid dynamics*. Springer Science & Business Media, 2010.

[36] Michel Ledoux. Logarithmic Sobolev inequalities for unbounded spin systems revisited. In *Séminaire de Probabilités XXXV*, pages 167–194. Springer, 2001.

[37] Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between gaussian fields and gaussian

Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

[38] Jun S Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2001.

[39] Jun S Liu and Rong Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American statistical association*, 93(443):1032–1044, 1998.

[40] Youssef M Marzouk and Habib N Najm. Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. *Journal of Computational Physics*, 228(6):1862–1902, 2009.

[41] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[42] Anthony Nouy. *Low-Rank Tensor Methods for Model Order Reduction*, pages 857–882. Springer International Publishing, Cham, 2017.

[43] Mario Teixeira Parente, Jonas Wallin, Barbara Wohlmuth, et al. Generalized bounds for active subspaces. *Electronic Journal of Statistics*, 14(1):917–943, 2020.

[44] Anthony T Patera, Gianluigi Rozza, et al. Reduced basis approximation and a posteriori error estimation for parametrized partial differential equations, 2007.

[45] Noemi Petra, James Martin, Georg Stadler, and Omar Ghattas. 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):A1525–A1555, 2014.

[46] Allan Pinkus. *Ridge functions*, volume 205. Cambridge University Press, 2015.

[47] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7:110–120, 1997.

[48] Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling for various metropolis-Hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

[49] Olivier Roustant, Franck Barthe, Bertrand Iooss, et al. Poincaré inequalities on intervals–application to sensitivity analysis. *Electronic journal of statistics*, 11(2):3081–3119, 2017.

[50] Alessio Spantini, Antti Solonen, Tiangang Cui, James Martin, Luis Tenorio, and Youssef Marzouk. Optimal low-rank approximations of Bayesian linear inverse problems. *SIAM Journal on Scientific Computing*, 37(6):A2451–A2487, 2015.

[51] A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

[52] Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*, volume 89. SIAM, 2005.

[53] Olivier Zahm, Paul G Constantine, Clementine Prieur, and Youssef M Marzouk. Gradient-based dimension reduction of multivariate vector-valued functions. *SIAM Journal on Scientific Computing*, 42(1):A534–A558, 2020.

[54] Olivier Zahm, Tiangang Cui, Kody Law, Alessio Spantini, and Youssef Marzouk. Certified dimension reduction in nonlinear Bayesian inverse problems. *arXiv preprint arXiv:1807.03712*, 2018.