Statistics and Computing (2022) 32:27
https://doi.org/10.1007/s11222-022-10087-1

Low-rank tensor reconstruction of concentrated densities with application to Bayesian inversion

Martin Eigel[1] · Robert Gruhlke[1] · Manuel Marschall[2]

Received: 10 August 2020 / Accepted: 4 February 2022 / Published online: 12 March 2022

Abstract

This paper presents a novel method for the accurate functional approximation of possibly highly concentrated probability densities. It is based on the combination of several modern techniques such as transport maps and low-rank approximations via a nonintrusive tensor train reconstruction. The central idea is to carry out computations for statistical quantities of interest such as moments based on a convenient representation of a reference density for which accurate numerical methods can be employed. Since the transport from target to reference can usually not be determined exactly, one has to cope with a perturbed reference density due to a numerically approximated transport map. By the introduction of a layered approximation and appropriate coordinate transformations, the problem is split into a set of independent approximations in seperately chosen orthonormal basis functions, combining the notions h- and p-refinement (i.e. "mesh size" and polynomial degree). An efficient low-rank representation of the perturbed reference density is achieved via the Variational Monte Carlo method. This nonintrusive regression technique reconstructs the map in the tensor train format. An a priori convergence analysis with respect to the error terms introduced by the different (deterministic and statistical) approximations in the Hellinger distance and the Kullback–Leibler divergence is derived. Important applications are presented and in particular the context of Bayesian inverse problems is illuminated which is a main motivation for the developed approach. Several numerical examples illustrate the efficacy with densities of different complexity and degrees of perturbation of the transport to the reference density. The (superior) convergence is demonstrated in comparison to Monte Carlo and Markov Chain Monte Carlo methods.

Keywords Tensor train · Uncertainty quantification · VMC · Low-rank · Bayesian inversion · Partial differential equations with random coefficients

Mathematics Subject Classification 62F15 · 62G07 · 35R60 · 60H35 · 65C20 · 65N12 · 65N22 · 65J10

1 Overview

We derive a novel numerical method for the functional representation of highly concentrated probability density functions whose probability mass concentrates along low-dimensional structures. Such situations may for instance arise in computational Bayesian inference with highly informed data. The difficult task of obtaining samples from the posterior distribution is usually attacked with Markov Chain Monte Carlo (MCMC) methods. Despite their popularity, the convergence rate of these methods is ultimately limited by the underlying Monte Carlo sampling technique, see e.g. Dodwell et al. (2019) for recent multilevel techniques in this context. Moreover, practical issues e.g. regarding the initial number of samples (burn-in) or a specific convergence assessment arise, requiring profound experience.

M. Eigel and R. Gruhlke acknowledge the support by the DFG SPP1886 "Polymorphic uncertainty modelling for the numerical design of structures".

✉ Robert Gruhlke
gruhlke@wias-berlin.de

Martin Eigel
eigel@wias-berlin.de

Manuel Marschall
manuel.marschall@ptb.de

[1] Weierstrass Institute, Mohrenstrasse 39, 10117 Berlin, Germany

[2] Physikalisch-Technische Bundesanstalt, Braunschweig and Berlin, Germany

In this work, we propose a new approach based on *function space representations with efficient surrogate models* in several instances. This is motivated by our previous work on adaptive low-rank approximations of solutions of parametric random PDEs with Adaptive Stochastic Galerkin FEM (ASGFEM, see e.g. Eigel et al. 2017, 2020) and in particular the sampling-free Bayesian inversion presented in Eigel et al. (2018) where the setting of uniform random variables was examined. A generalization to the important case of Gaussian random variables turns out to be non-trivial from a computational point of view due to the difficulties caused by the representation of highly concentrated densities in a compressing tensor format which is required to cope with the high dimensionality of the problem. As a consequence, we develop a discretization approach which takes into account the potentially difficult to represent structure of the probability density at hand by a combination of several transformations and approximations that can be chosen adaptively to counteract the interplay of the employed numerical approximations. With the computed functional representation of the density, the evaluation of moments and statistical quantities of interest can be carried out efficiently and with high accuracy. Additionally, we point out that the generated surrgate can be used for the fast generation of samples from the posterior distribution.

The central idea of the method is to construct a map which transports the target density to some convenient reference density and to employ low-rank regression techniques to obtain a functional representation, for which accurate numerical methods are available. Transport maps for probability densities are a classical topic in mathematics. They are under active research in particular in the area of optimal transport (Villani 2008; Santambrogio 2015) and also have become popular in current machine learning research (Tran et al. 2019; Rezende and Mohamed 2015; Detommaso et al. 2019). A main application we have in mind is Bayesian inversion where, given a prior density and some observations of the forward model, a posterior density should be determined. In this context, the rescaling approaches in Schillings and Schwab (2016) and Schillings et al. (2020) based on the Laplace approximation can be considered as transport maps of a certain (affine) form. More general transport maps have been examined extensively in El Moselhy and Marzouk (2012) and Parno and Marzouk (2018) and other works of the research group. Obtaining a transport map is in general realized by minimizing a certain loss functional, e.g. the Kullback–Leibler distance, between the target and the push-forward of a reference density. This process has been analyzed and improved using iterative maps (Brennan et al. 2020) or multi-scale approaches (Parno et al. 2016). However, the optimization, the loss functional and the chosen model class for the transport map yield only an approximation to an *exact* transport. We hence suppose that, in general, only an inexact transport is available. Considering the pullback of the target, this can be interpreted as starting from a slightly or severly perturbed reference density. One then has to cope with the degree of the perturbation in subsequent approximation steps to enable an accurate explicit representation of this new reference density.

Finding a suitable approximation relies on concepts from adaptive finite element methods (FEM). In addition to the selection of (local) approximation spaces of a certain degree (in the spirit of “p-refinement”), we introduce a spatial decomposition of the density representation into layers (similar to “h-refinement”) around some center of mass of the considered density. This enables to exploit the (assumed) decay behavior of the approximated density. Overall, this “hp-refinement” allows to balance inaccuracies and hence to compensate perturbations of the reference density by putting more computational effort into the discretization part. Consequently, one enjoys the freedom to decide whether more effort should be invested into computing an exact transport map or into a more elaborate discretization of the perturbed reference density.

For eventual computations with the devised (possibly high-dimensional) functional density representation, an efficient representation format is required. In our context, hierarchical tensors and in particular tensor trains (TT) prove to be advantageous, cf. Bachmayr et al. (2016) and Oseledets (2011). These low-rank formats enable to alleviate the curse of dimensionality under suitable conditions and allow for efficient evaluations of high-dimensional objects. For each layer of the discretization we aim to obtain a low-rank tensor representation of the respective perturbed reference density. In certain ideal cases such as transporting to the standard Gaussian density a rank-one representation is sufficient. A discussion on the low-rank approximation of perturbations of Gaussian densities is carried out in Rohrbach et al. (2020). In more general cases, a low-rank representability may be observed numerically. To allow for tensor methods to be applicable, the desired discretization layers have to be tensor domains. Therefore, the underlying perturbed reference density is transformed to an alternative coordinate system which benefits the representation and enables to exploit the regularity and decay behavior of the density. To generate a tensor train representation (coupled with a function basis which is then also called extended or functional TT format Gorodetsky et al. 2015), the Variational Monte Carlo (VMC) method (Eigel et al. 2019b) is employed. It basically is a tensor regression approach based on function samples for which a convergence analysis is available. Notably, depending on the chosen loss functional, it leads to the best approximation in the respective model space. It has previously been examined in the context of random PDEs in Eigel et al. (2019b) as an alternative nonintrusive numerical approach to Stochastic Galerkin FEM in the TT format (Eigel et al. 2017, 2020). The approximation of Eigel et al. (2020) is used in one of the presented examples for Bayesian inversion with the random Darcy equation with lognormal coefficient. We note that surrogate models of the forward model have been used in the context of MCMC e.g. in Li and Marzouk (2014) and tensor representations (obtained by cross approximation) were used in Dolgov et al. (2020) to improve the efficiency of MCMC sampling.

The derivation of our method is supported by an a priori convergence analysis with respect to the Hellinger distance and the Kullback–Leibler divergence. In the analysis, different error sources have to be considered, in particular a layer truncation error depending on decay properties of the density, a low-rank truncation error and model space approximations are introduced. Moreover, the VMC error analysis (Eigel et al. 2019b) comprising statistical estimation and numerical approximation errors is adjusted to be applicable to the devised approach. While not usable for an a posteriori error control in its current initial form, the derived analysis leads the way to more elaborate results for this promising method in future research.

With the constructed functional density surrogate, sampling-free computations of statistical quantities of interest such as moments or marginals become feasible by fast tensor contractions.

While several assumptions have to be satisfied for this method to work most efficiently, the approach is rather general and can be further adapted to the problem at hand. Moreover, it should be emphasized that by constructing a functional representation, structural properties of the density at hand (in particular smoothness, sparsity, low-rank approximability and decay behavior in different parameters) can be exploited in a much more extensive way than what is possible with sampling based methods such as MCMC, leading to more accurate statistical computations and better convergence rates. We note that the perturbed posterior surrogate can be used to efficiently generate samples by rejection sampling or within an MCMC scheme. Since the perturbed transport can be seen as a preconditioner, the sample generation can be based on the perturbed prior. These samples can then be transported to the posterior by the determined push-forward. As a prospective extension, the constructed posterior density could directly be used in a Stochastic Galerkin FEM based on the integral structure, closing the loop of forward and inverse problem, resulting in the inferred forward problem with model data determined by Bayesian inversion from the observed data.

The structure of the paper is as follows. Section 2 is concerned with the representation of probability densities and introduces a relation between a target and a reference density. Such a transport map can be determined numerically by approximation in a chosen class of functions and with an assumed structure, leading to the concept of perturbed reference densities. To counteract the perturbation, a layered truncated discretization is introduced. An efficient low-rank representation of the mappings is described in Sect. 3 where the tensor train format is discussed. In order to obtain this nonintrusively, the Variational Monte Carlo (VMC) tensor reconstruction is reviewed. A priori convergence results with respect to the Hellinger distance and Kullback–Leibler divergence are derived in Sect. 4. For practical purposes, the proposed method is described in terms of an algorithm in Sect. 5. Possible applications we have in mind are examined in Sect. 6. In particular, the setting of Bayesian inverse problems is recalled. Moreover, the computation of moments and marginals is scrutinized. Section 7 illustrates the performance of the proposed method. In addition to an examination of the numerical sensitivity of the accuracy with respect to the perturbation of the transport maps, a typical model problem from Uncertainty Quantification (UQ) is depicted, namely the identification of a parametrization for the random Darcy equation with lognormal coefficient given as solution of a stochastic Galerkin FEM.

2 Density representation

The aim of this section is to introduce the central ideas of the proposed approximation of densities. For this task, two established concepts are reviewed, namely *transport maps* (El Moselhy and Marzouk 2012; Marzouk et al. 2016), which are closely related to the notion of optimal transport (Villani 2008; Santambrogio 2015), and *hierarchical low-rank tensor representations* (Oseledets 2011; Hackbusch 2012; Bachmayr et al. 2016). By the combination of these techniques, assuming the access to a suitable transformation, the developed approach yields a functional representation of the density in a format which is suited for computations with high-dimensional functions. In particular, it becomes feasible to accurately handle highly concentrated posterior densities. While transport maps on their own in principle enable the generation of samples of some target distribution, the combination with a functional low-rank representation allows for integral quantities such as (centered) moments to become computable cheaply. Given an approximate transport map, the low-rank representation can be seen as a further approximation step (compensating for the inaccuracy of the used transport) to gain direct access to the target density.

Consider a target measure π with Radon-Nikodym derivative with respect to the Lebesgue measure λ denoted as f with support in $\mathbb{R}^d$, $d < \infty$, i.e.

$$f(y) := \frac{\mathrm{d}\pi}{\mathrm{d}\lambda}(y), \quad y \in Y := \mathbb{R}^d. \tag{1}$$

In the following we assume that point evaluations of f are available up to a multiplicative constant, motivated by the framework of Bayesian posterior density representation with unknown normalization constant. Furthermore, let π_0 be some reference measure exhibiting a Radon-Nikodym derivative with respect to to the Lebesgue measure denoted as f_0. This is motivated by the prior measure and density in the context of Bayesian inference.

In the upcoming sections we relate π and π_0 with the help of a transport map T. If the exact T is replaced by some $\tilde{T}$, in Sect. 2.2 we alternatively can describe π with the help of a so-called associated *auxiliary measure* $\tilde{\pi}_0$ with *auxiliary density* $\tilde{f}_0$. As in the Bayesian context, we also may refer to it as a *perturbed prior measure* with *perturbed prior density*. This concept introduces a possible workload balancing between an approximation of T and the associated perturbed prior.

2.1 Transport maps

The notion of density transport is classical and has become a popular research area with *optimal transport*, see e.g. Villani (2008) and Santambrogio (2015). From a practical point of view, it has been employed to improve numerical approaches for Bayesian inverse problems for instance in El Moselhy and Marzouk (2012), Brennan et al. (2020) and Dolgov et al. (2020). Similar approaches are discussed in terms of sample transport e.g. for Stein's method (Liu and Wang 2016; Detommaso et al. 2018) or multi-layer maps (Brennan et al. 2020). We review the properties required for our approach in what follows. Note that since our target application is Bayesian inversion, we usually use the terms "prior" and "posterior" instead of the more general "reference" and "target" densities.

Let $X := \mathbb{R}^d$ and assume that there exists an exact transport map $T\colon X \to Y$, which is a diffeomorphism[1] relating π and π_0 via

$$f_0(x) = f(T(x))|\mathcal{J}_T(x)|, \quad x \in X. \tag{2}$$

This change of variables formula allows computations to be carried out in terms of the measure π_0, which is commonly assumed to be of a simpler structure. In particular, for any function $Q\colon Y \to \mathbb{R}$ integrable with respect to π, it holds that

$$\int_Y Q(y)\mathrm{d}\pi(y) = \int_X Q(T(x))f_0(x)\mathrm{d}\lambda(x). \tag{3}$$

Note that the computation of the right-hand side in (3) may still be a challenging task depending on the actual structure of $Q \circ T$. In "Appendix A" we list several examples of transport maps with an exploitable structure.

2.2 Inexact transport and the perturbed prior

In general, the transport map T is unknown or difficult to determine and hence has to be approximated by some $\tilde{T}\colon X \to Y$, e.g. using a polynomial representation with respect to π_0 (El Moselhy and Marzouk 2012) or with a composition of simple maps in a reduced space such as in Brennan et al. (2020). As a consequence of the approximation, it holds

$$\int_Y Q(y)\mathrm{d}\pi(y) \approx \int_X Q(\tilde{T}(x))\mathrm{d}\pi_0(x), \tag{4}$$

subject to the accuracy of the involved approximation of T. The approximate transport $\tilde{T}$ relates a measure $\tilde{\pi}_0$ with density $\tilde{f}_0$ to the target measure π, whereas

$$\tilde{f}_0(x) = f(\tilde{T}(x))|\mathcal{J}_{\tilde{T}}(x)|. \tag{5}$$

We henceforth refer to (5) as the auxiliary reference or *perturbed prior density*. Using this construction, the moment computation reads

$$\int_Y Q(y)\mathrm{d}\pi(y) = \int_X Q(\tilde{T}(x))\tilde{f}_0(x)\mathrm{d}\lambda(x). \tag{6}$$

If one would know $\tilde{f}_0$, by (5) and (6) one would also have access to the exact posterior.

Equation (6) is the starting point of the proposed method by approximating $\tilde{f}_0$ in another coordinate system which is better adapted to the structure of the approximate (perturbed) prior. For this, consider a (fixed) diffeomorphism

$$\Phi\colon \hat{X} \subset \mathbb{R}^d \to X, \quad \hat{x} \mapsto x = \Phi(\hat{x}) \tag{7}$$

with Jacobian $\hat{x} \mapsto |\mathcal{J}_\Phi(\hat{x})|$ and define the *transformed perturbed prior*

$$\hat{f}_0\colon \hat{X} \mapsto \mathbb{R}_+, \quad \hat{x} \mapsto \hat{f}_0(\hat{x}) := \tilde{f}_0(\Phi(\hat{x})). \tag{8}$$

If (8) can be approximated accurately by some function $\hat{f}_0^h$ then

$$\int_Y Q(y)\mathrm{d}\pi(y) \approx \int_{\hat{X}} Q(\tilde{T}(\Phi(\hat{x})))\hat{f}_0^h(\hat{x})|\mathcal{J}_\Phi(\hat{x})|\mathrm{d}\lambda(\hat{x}) \tag{9}$$

with accuracy determined only by the approximation quality of $\hat{f}_0^h$. Thus, (9) enables to shift the complexity of a transport map approximation $\tilde{T} \approx T$ to a functional approximation $\hat{f}_0^h \approx \hat{f}_0$ in a new coordinate system $\hat{X}$ and allows for a

[1] The requirements on T can be weakened, e.g. to local Lipschitz continuity.

balancing of both effects. The construction of $\tilde{T}$ and a suitable map in (7) may be used to obtain a convenient transformed perturbed prior given in (8). An approximation thereof can be significantly simpler compared to a possibly complicated target density f or the computation of the exact transport T. We give some example scenarios and more details on the choice of Φ in "Appendix B".

2.3 Layer based representation

To further refine and motivate the notion of an adapted coordinate system, let $L \in \mathbb{N}$ and $(X^\ell)_{\ell=1}^L$ be pairwise disjoint domains in X such that

$$K := \bigcup_{\ell=1}^{L} \overline{X^\ell} \tag{10}$$

is simply connected and compact and define $X^{L+1} := X \setminus K$. Then, for given $L \in \mathbb{N}$ we may decompose the perturbed prior $\tilde{f}_0$ as

$$\tilde{f}_0(x) = \sum_{\ell=1}^{L+1} \tilde{f}_0^\ell(x) \quad \text{with} \quad \tilde{f}_0^\ell := \chi_\ell \tilde{f}_0, \tag{11}$$

where χ_ℓ denotes the indicator function on X^ℓ. Moreover, for any tensor set $\hat{X}^\ell := \times_{i=1}^d \hat{X}_i^\ell$ and diffeomorphism $\Phi^\ell : \hat{X}^\ell \to X^\ell, 1 \le \ell \le L+1$, we may represent the *localized perturbed prior* $\tilde{f}_0^\ell$ as a pull-back function

$$\tilde{f}_0^\ell = \hat{f}_0^\ell \circ \Phi^{\ell^{-1}}, \tag{12}$$

where $\hat{f}_0^\ell$ is a map defined on $\hat{X}^\ell$ as in (8). This layer based coordinate change enables a representation of the density on bounded domains. Even though the remainder layer is unbounded, we assume that K is sufficiently large to cover all probability mass of $\tilde{f}_0$ except for a negligible higher-order error.

Up to this point, the choice of transformation Φ^ℓ, $\ell = 1, \ldots, L+1$, is fairly general. However, for the further development of the method we assume the following property.

Definition 1 (*rank 1 stability*) Let $\mathcal{X}, \hat{\mathcal{X}} = \times_{i=1}^d \hat{\mathcal{X}}_i \subset \mathbb{R}^d$ be open and bounded sets. A diffeomorphism $\Phi : \hat{\mathcal{X}} \mapsto \mathcal{X}$ is called *rank 1 stable* if Φ and the Jacobian $|\mathcal{J}_\Phi|$ have rank 1, i.e. there exist univariate functions $\Phi_i : \hat{\mathcal{X}}_i \to \mathcal{X}$, $h_i : \hat{\mathcal{X}} \to \mathbb{R}, i = 1, \ldots, d$, such that for $\hat{x} \in \hat{\mathcal{X}}$

$$\Phi(\hat{x}) = \bigodot_{i=1}^d \Phi_i(\hat{x}_i), \quad |\mathcal{J}_\Phi(\hat{x})| = \prod_{i=1}^d h_i(\hat{x}_i), \tag{13}$$

where $\bigodot$ denotes the Hadamard product of vectors.

Due to the notion of rank 1 stable transformations, the map $\hat{x} \mapsto T(\Phi(\hat{x}))$ in (9) inherits the rank structure of T, see Sect. 3. Furthermore, since the Jacobian $\hat{x} \mapsto |\mathcal{J}_\Phi(\hat{x})|$ is rank 1, we can construct tensorized orthonormal basis functions which may be used to approximate the transformed perturbed prior in (8).

Remark 1 The described concept can be extended to any rank $r \in \mathbb{N}$ Jacobian of Φ, i.e.

$$|\mathcal{J}_\Phi(\hat{x})| = \sum_{k=1}^r \prod_{i=1}^d h_{i,k}(\hat{x}_i). \tag{14}$$

Motivated by the right-hand side in (9), one may use different approximations of the perturbed transformed prior $\tilde{f}_0 \circ \Phi$ in r distinct tensorized spaces, each associated to the rank 1 weight $\prod_{i=1}^d h_{i,k}$.

2.4 Layer truncation

This paragraph is devoted to the treatment of the last (remainder or truncation) layer X^{L+1} introduced in (11) with the aim to suggest some approximation choices.

If $\tilde{f}_0$ is represented in the layer format (11), it is convenient to simply extend the function to zero after layer $L \in \mathbb{N}$. By this, the remaining (possibly small) probability mass is neglected. Such a procedure is typically employed in numerical applications and does not impose any computational issues since events on the outer truncated domain are usually exponentially unlikely for a truncation value chosen sufficiently large. Nevertheless, in order to present a rigorous treatment, we require properties like absolute continuity, which would be lost by using a cut-off function. Inspired by Schillings et al. (2020) regarding the information limit of unimodal posterior densities,[2] we suggest a Gaussian approximation for the last layer $L+1$ on the unbounded domain X^{L+1}, i.e. for some s.p.d. $\Sigma \in \mathbb{R}^{d,d}$ and $\mu \in \mathbb{R}^d$ we define the density

$$\tilde{f}_0^{\mathrm{Trun}}(x) := C_L \begin{cases} \tilde{f}_0^\ell(x), & x \in X^\ell, \ell = 1, \ldots, L, \\ f_{\Sigma,\mu}(x), & x \in X^{L+1}, \end{cases} \tag{15}$$

with $C_L = (C_L^< + C_L^>)^{-1}$, where

$$C_L^< := \int\limits_{X \setminus K} f_{\Sigma,\mu}(x) \, \mathrm{d}\lambda(x), \tag{16}$$

[2] A result of Schillings et al. (2020) is that under suitable conditions the posterior distribution converges to a Gaussian in the limit of zero noise and infinite measurements.

$$C_L^> := \sum_{\ell=1}^{L} \int_{X^\ell} \tilde{f}_0^\ell(x)\, \mathrm{d}\lambda(x), \tag{17}$$

and $f_{\Sigma,\mu}$ denotes the Gaussian probability density function with mean μ and covariance matrix Σ.

Remark 2 A particular choice for μ and Σ would be the mean and covariance of the normalized version of $\tilde{f}_0|_K$

or the MAP and the corresponding square root of the Hessian of $\tilde{f}_0$.

Note that the constant $C_L^<$ in (16) may exhibit an analytic form whereas computing $C_L^>$ suffers from the curse of dimensionality and is in general not available. To circumvent this issue and render further use of the representation (15) feasible, we now aim for an approximation model to adequately represent the localized transformed perturbed prior maps $\hat{f}_0^\ell = \tilde{f}_0^\ell \circ \Phi^\ell$ from (12).

3 Approximation model

The computation of high-dimensional integrals and the efficient construction of surrogates is a challenging task with a multitude of approaches. Some of these techniques are sparse grid methods (Chen and Schwab 2016; Garcke and Griebel 2012), collocation (Ernst et al. 2019; Nobile et al. 2008; Foo and Karniadakis 2010) or modern sampling techniques (Gilks et al. 1995; Rudolf and Sprungk 2017; Neal 2001).

A promising class of approximation model relies on the concept of compression in terms of low-rank formats (Grasedyck et al. 2013). Those can be seen as generalization of the singular value decomposition of functions with high-dimensional input. We give a short introduction of the low-rank format considered in Sect. 3.1 and present a numerical scheme to construct such an approximation model by the Variational Monte Carlo (VMC) method in Sect. 3.2. In order to enhance readability, in the first two sections we use the notation $\hat{g}$ to underline the connection to the transformed perturbed prior $\hat{f}_0$ and its localizations, which is part of the total approximation model in Sect. 3.3.

3.1 Low-rank tensor train format

Let $\hat{\mathcal{X}} = \bigotimes_{i=1}^d \hat{\mathcal{X}}_i \subset \mathbb{R}^d$, $\hat{\mathcal{X}}_i$, $i \in [d] := \{1,\ldots,d\}$, and consider a map $\hat{g}\colon \hat{\mathcal{X}} \to \mathbb{R}$. The function $\hat{g}$ can be represented in the TT format if there exists a *rank vector* $\boldsymbol{r} = (r_1,\ldots,r_{d-1}) \in \mathbb{N}^{d-1}$ and univariate functions $\hat{g}^i[k_{i-1},k_i]\colon \hat{\mathcal{X}}_i \to \mathbb{R}$ for $k_i \in [r_i]$, $i \in [d]$, such that for all $\hat{x} \in \hat{\mathcal{X}}$ and $k_0 = k_d = 1$ there holds

$$\hat{g}(\hat{x}) = \sum_{\boldsymbol{k}=\boldsymbol{1}}^{\boldsymbol{r}} \prod_{i=1}^{d} \hat{g}^i[k_{i-1},k_i](\hat{x}_i), \quad \boldsymbol{k} := (k_1,\ldots,k_{d-1}). \tag{18}$$

In the forthcoming sections we consider weighted tensorized Lebesgue spaces in which the perturbed prior segments are defined. In particular, for a non-negative weight function $w\colon \hat{\mathcal{X}} \to \mathbb{R}$ with $w = \bigotimes_{i=1}^d w_i$, $w \in L^1(\hat{\mathcal{X}})$, define

$$\mathcal{V}(\hat{\mathcal{X}}) := L^2(\hat{\mathcal{X}}, w) \\ = \left\{ \hat{g}\colon \hat{\mathcal{X}} \to \mathbb{R} \mid \|\hat{g}\|_{\mathcal{V}}^2 := \int_{\hat{\mathcal{X}}} \hat{g}(\hat{x})^2 w(\hat{x})\, \mathrm{d}\lambda(\hat{x}) < \infty \right\}. \tag{19}$$

This space may be identified by its tensorization $L^2(\hat{\mathcal{X}}, w) = \bigotimes_{i=1}^d L^2(\hat{\mathcal{X}}_i, w_i)$.

We assume that there exists a complete orthonormal basis $\{P_k^i : k \in \mathbb{N}\}$ in $L^2(\hat{\mathcal{X}}_i, w_i)$ for every $i \in [d]$ which is known a priori. For discretization purposes, we introduce the finite dimensional subspaces

$$\mathcal{V}_{i,n_i} := \overline{\operatorname{span}\left\{P_1^i,\ldots,P_{n_i}^i\right\}} \subseteq L^2(\hat{\mathcal{X}}_i, w_i), \tag{20}$$

for $i = 1,\ldots,d$, and $n_i \in \mathbb{N}$. On these we formulate the *extended tensor train format* in terms of the coefficient tensors

$$G^i\colon \left[r_{i-1}\right] \times [n_i] \times [r_i] \to \mathbb{R}, \\ (k_{i-1}, j, k_i) \mapsto G^i[k_{i-1}, j, k_i], \quad i \in [d], \tag{21}$$

such that every univariate function $\hat{g}^i \in \mathcal{V}_{i,n_i}$ can be written as

$$\hat{g}^i[k_{i-1},k_i](\hat{x}_i) = \sum_{j=1}^{n_i} G^i[k_{i-1},j,k_i] P_j^i(\hat{x}_i) \quad \text{for } \hat{x} \in \hat{\mathcal{X}}_i. \tag{22}$$

Any function

$$\hat{g} \in \mathcal{V}_\Lambda := \bigotimes_{i=1}^{d} \mathcal{V}_{i,n_i} \subseteq \mathcal{V}(\hat{\mathcal{X}}) \tag{23}$$

can be expressed in the full tensor format by a high dimensional algebraic tensor $G\colon \Lambda := \bigtimes_{i=1}^d [n_i] \to \mathbb{R}$ and tensorized functions $P_\alpha := \bigotimes_{i=1}^d P_{\alpha_i}$ for multiindices $\alpha =$

$(\alpha_1,\dots,\alpha_d)\in\Lambda$ such that

$$\hat{g}(\hat{x}) = \sum_{\boldsymbol{\alpha}\in\Lambda} G[\alpha_1,\dots,\alpha_d]\prod_{i=1}^{d} P_{\alpha_i}(\hat{x}_i). \tag{24}$$

In contrast to this, the format given by (18) and (22) admits a linear structure in the dimension. More precisely, the memory complexity of $\mathcal{O}(\max\{n_1,\dots,n_d\}^d)$ in (24) reduces to

$$\mathcal{O}(\max\{r_1,\dots,r_{d-1}\}^2\cdot d\cdot\max\{n_1,\dots,n_d\}). \tag{25}$$

This observation raises the question of expressibility for certain classes of functions and the existence of a representation rank $\boldsymbol{r}$ where $\max\{r_1,\dots,r_{d-1}\}$ stays sufficiently small for practical computations. This issue is e.g. addressed in Schneider and Uschmajew (2014), Bachmayr et al. (2017) and Griebel and Harbrecht (2013) under certain assumptions on the regularity and in Espig et al. (2009), Oseledets (2011), Ballani et al. (2013) and Eigel et al. (2019b) explicit (algorithmic) constructions of the format are discussed even in case that $\hat{g}$ has no analytic representation.

For later reference we define the finite dimensional low-rank manifold of rank $\boldsymbol{r}$ tensor trains by

$$\mathcal{M}_{\boldsymbol{r}}(\hat{\mathcal{X}}) := \{\hat{g}\in\mathcal{V}(\hat{\mathcal{X}}) \mid \hat{g} \text{ as in (18)}, \ \hat{g}^i \text{ as in (22)}\}. \tag{26}$$

This is an embedded manifold in the finite full tensor space $\mathcal{V}_\Lambda$ from (23). We also require the concept of the algebraic (full) tensor space

$$\mathbb{T} := \left\{G\colon \mathbb{N}^d\to\mathbb{R}\right\} \tag{27}$$

and the corresponding low-rank form for given $\boldsymbol{r}\in\mathbb{N}^{d-1}$ defined by

$$\mathbb{TT}_{\boldsymbol{r}} := \left\{G\colon \Lambda\to\mathbb{R} \mid G[\alpha] = \sum_{\boldsymbol{k}=\boldsymbol{1}}^{\boldsymbol{r}}\prod_{i=1}^{d} G[k_{i-1},\alpha_i,k_i]\right\}. \tag{28}$$

In the following, a method is reviewed that can be used to construct surrogates in the presented extended tensor train format using only samples of the sought high-dimensional function.

3.2 Tensor train regression by Variational Monte Carlo

We review the sampling-based VMC method presented in Eigel et al. (2019b) which is employed here to construct TT approximations of the transformed local perturbed priors $\hat{f}_0^\ell$ as defined in (12). The approach generalizes the concept of randomized tensor completion (Eigel et al. 2019a). Its analysis relies on the theory of statistical learning, leading to a priori convergence results. It can also be seen as a generalized tensor least squares technique. In principle, the algorithm for the construction of the surrogate is interchangeable and e.g. an alternative cross-interpolation method for probability densities is presented in Dolgov et al. (2020).

For the VMC framework, consider the *model class*

$$\mathcal{M} := \mathcal{M}_{\boldsymbol{r}}(\hat{\mathcal{X}},\underline{c},\overline{c}) \subset \mathcal{M}_{\boldsymbol{r}}(\hat{\mathcal{X}}) \tag{29}$$

of truncated rank $\boldsymbol{r}\in\mathbb{R}^{d-1}$ tensor trains which is given by

$$\mathcal{M}_{\boldsymbol{r}}(\hat{\mathcal{X}},\underline{c},\overline{c}) := \left\{\hat{g}\in\mathcal{M}_{\boldsymbol{r}} \mid \underline{c}\le\hat{g}\le\overline{c} \quad \text{a.e. in } \hat{\mathcal{X}}\right\}$$

for $0\le\underline{c}<\overline{c}\le\infty$. The model class $\mathcal{M}$ is a finite subset of the truncated nonlinear space $\mathcal{V}(\hat{\mathcal{X}},\underline{c},\overline{c})\subseteq\mathcal{V}(\hat{\mathcal{X}})$ defined by

$$\mathcal{V}(\hat{\mathcal{X}},\underline{c},\overline{c}) := \{\hat{g}\in L^2(\hat{\mathcal{X}},w) \mid \underline{c}\le\hat{g}\le\overline{c} \quad \text{a.e. in } \hat{\mathcal{X}}\}. \tag{30}$$

This space is equipped with the metric

$$d_{\mathcal{V}(\hat{\mathcal{X}},\underline{c},\overline{c})}(\hat{g}_1,\hat{g}_2) := \|\hat{g}_1-\hat{g}_2\|_{\mathcal{V}(\hat{\mathcal{X}})}.$$

Moreover, note that due to the orthonormality of $\{P_\alpha\}_{\alpha\in\mathbb{N}^d}$ in $\mathcal{V}(\hat{\mathcal{X}})$, for every $\hat{g}\in\mathcal{V}(\hat{\mathcal{X}})$ it holds

$$\|\hat{g}\|_{\mathcal{V}} = \|G\|_{\ell^2(\mathbb{T})} \quad \text{with} \quad \hat{g} = \sum_{\alpha} G[\alpha]P_\alpha \in\mathcal{V}(\hat{\mathcal{X}}). \tag{31}$$

Additionally, we define a *loss function* $\iota\colon\mathcal{V}(\hat{\mathcal{X}},\underline{c},\overline{c})\times\hat{\mathcal{X}}\to\mathbb{R}$ such that $\iota(\cdot,\hat{x})$ is continuous for almost all $\hat{x}\in\hat{\mathcal{X}}$ and $\iota(\hat{g},\cdot)$ is integrable with respect to the weight function w of $\mathcal{V}(\hat{\mathcal{X}})$ for every $\hat{g}\in\mathcal{V}(\hat{\mathcal{X}},\underline{c},\overline{c})$ and the *cost functional* $\mathcal{J}:\mathcal{V}(\hat{\mathcal{X}},\underline{c},\overline{c})\to\mathbb{R}$ given by

$$\mathcal{J}(\hat{g}) := \int_{\hat{\mathcal{X}}}\iota(\hat{g},\hat{x})w(\hat{x})\mathrm{d}\lambda(\hat{x}). \tag{32}$$

Then, the objective of the method is to find a minimizer

$$\hat{g}^* \in \operatorname{argmin}_{\hat{g}\in\mathcal{V}(\hat{\mathcal{X}},\underline{c},\overline{c})}\mathcal{J}(\hat{g}). \tag{33}$$

Due to the infinite dimensional setting, we confine the minimization problem in (33) to our model class $\mathcal{M}$. This yields the minimization problem

$$\text{find } \hat{g}^*_{\mathcal{M}} \in \operatorname{argmin}_{\hat{g}\in\mathcal{M}}\mathcal{J}(\hat{g}). \tag{34}$$

Subsequently, instead of $\mathscr{J}$ a tractable empirical functional is considered, namely

$$\mathscr{J}_N(\hat g) := \frac{1}{N}\sum_{k=1}^{N} \iota(\hat g, \hat x^k), \tag{35}$$

with independent samples $\{\hat x^k\}_{k\le N}$ distributed according to the measure $w\lambda$ with a (possibly rescaled) weight function w with respect to the Lebesgue measure λ. The corresponding empirical optimization problem reads

$$\text{find } \hat g^*_{\mathcal{M},N} \in \operatorname{argmin}_{\hat g\in\mathcal{M}} \mathscr{J}_N(\hat g). \tag{36}$$

In our application the loss functional relates to the Kullback–Leibler divergence or the L^2 error for a given function $\hat g$ and target $\hat f_0$. In order to emphasize this dependence, we may write $\iota(\hat g, \hat x;\, \hat f_0)$.

3.3 The total approximation model

We employ the VMC approach from Sect. 3.2 to build an approximation of $\hat f_0^\ell = \tilde f_0^\ell \circ \Phi^\ell$ in TT format for each layer $\hat X^\ell = \times_{i=1}^d \hat X_i^\ell$. In particular, we choose $\hat{\mathcal{X}} = \hat X^\ell$ and $w = w^\ell = |\mathcal{J}_{\Phi^\ell}|$, for $\ell = 1,\ldots,L$. For a given number $N_\ell \in \mathbb{N}$ of samples and bounds $0 \le \underline{c}_\ell < \overline{c}_\ell < \infty$, we denote this approximation as $\hat f_0^{\ell,\mathrm{TT},N_\ell} \in \mathcal{M}^\ell := \mathcal{M}(\hat X^\ell, \underline{c}_\ell, \overline{c}_\ell)$. It is a solution of

$$\hat f_0^{\ell,\mathrm{TT},N_\ell} \in \operatorname{argmin}_{v\in\mathcal{M}^\ell} \frac{1}{N_\ell}\sum_{k=1}^{N_\ell} \iota(v, \hat x^k;\, \hat f_0), \tag{37}$$

with samples $\{\hat x^k\}_{k=1}^{N_\ell}$ drawn from the (possibly rescaled) finite measure $w_\ell\lambda$. Finally, for a given transport $\tilde T$ and a choice of transformations $\Phi^\ell\colon \hat X^\ell \to X^\ell$, according to (15) our approximate of the perturbed prior $\tilde f_0$ denoted by $\tilde f_0^{\mathrm{Trun,TT}}$ is defined by

$$\tilde f_0^{\mathrm{Trun,TT}}(x) := C_L^{\mathrm{TT}} \begin{cases} \tilde f_0^{\ell,\mathrm{TT}}(x), & x \in X^\ell,\ \ell = 1,\ldots,L, \\ f_{\Sigma,\mu}(x), & x \in X^{L+1}, \end{cases} \tag{38}$$

with $\tilde f_0^{\ell,\mathrm{TT}} = \hat f_0^{\ell,\mathrm{TT},N_\ell} \circ \left(\Phi^\ell\right)^{-1}$. Here, $C_L^{\mathrm{TT}} := (C_L^< + C_L^{>,\mathrm{TT}})^{-1}$ with $C_L^<$ from (16) and

$$C_L^{>,\mathrm{TT}} := \sum_{\ell=1}^{L} \int_{X^\ell} \tilde f_0^{\ell,\mathrm{TT}}(x)\,\mathrm{d}\lambda(x). \tag{39}$$

We refer to Fig. 1 for a visual presentation of the involved objects, approximations and transformations.

4 Error estimates

We now discuss the accuracy of the actual approximation of the target density f given by

$$\tilde f^{TT} := \tilde f_0^{\mathrm{Trun,TT}} \circ \tilde T^{-1} \otimes |\mathcal{J}_{\tilde T^{-1}}|. \tag{40}$$

Since our approach is based on several components like transport, truncation, low-rank compression and the VMC method, these components are examined separately in the upcoming sections. Our main result is stated in Sect. 4.5.

4.1 Transport invariant measures of discrepancy

In this section we derive a relation property such that the error of the approximation $\tilde f_0^{\mathrm{Trun,TT}}$ to the the perturbed prior transfers directly to the discrepancy between $\tilde f^{\mathrm{TT}}$ and f. Note that this property is canonical since passing to the image space of some measurable function is fundamental in probability theory. Ideally such a relation is an equivalence of the form

$$\mathrm{d}\left(Y;\, f,\, \tilde f^{TT}\right) = \mathrm{d}\left(X;\, \tilde f_0,\, \tilde f_0^{\mathrm{Trun,TT}}\right). \tag{41}$$

Prominent measures of discrepancy for two absolutely continuous Lebesgue probability density functions h_1 and h_2 on some measurable space Z are the squared Hellinger distance

$$\mathrm{d}^2_{\mathrm{Hell}}(Z, h_1, h_2) := \frac{1}{2}\int_Z \left(\sqrt{h_1}(z) - \sqrt{h_2(z)}\right)^2 \mathrm{d}\lambda(z), \tag{42}$$

and the Kullback–Leibler divergence

$$\mathrm{d}_{\mathrm{KL}}(Z, h_1, h_2) := \int_Z \log\left(\frac{h_1(z)}{h_2(z)}\right) h_1(z)\,\mathrm{d}\lambda(z). \tag{43}$$

For the Hellinger distance, the absolute continuity assumption can be dropped from an analytical point of view. We observe that both $\mathrm{d}_{\mathrm{Hell}}$ and d_{KL} satisfy (41).

Lemma 1 *Let* $\sharp \in \{\mathrm{Hell}, \mathrm{KL}\}$. *It then holds*

$$\mathrm{d}_\sharp(Y;\, f,\, \tilde f^{TT}) = \mathrm{d}_\sharp(X;\, \tilde f_0,\, \tilde f_0^{\mathrm{Trun,TT}}). \tag{44}$$

Proof We only show (44) for $\sharp = \mathrm{KL}$ since $\sharp = \mathrm{Hell}$ follows by similar arguments. By definition

$$\mathrm{d}_{\mathrm{KL}}(Y;\, f,\, \tilde f^{\mathrm{TT}}) = \int_Y \log\left(\frac{f(y)}{\tilde f^{\mathrm{TT}}(y)}\right) f(y)\,\mathrm{d}\lambda(y), \tag{45}$$

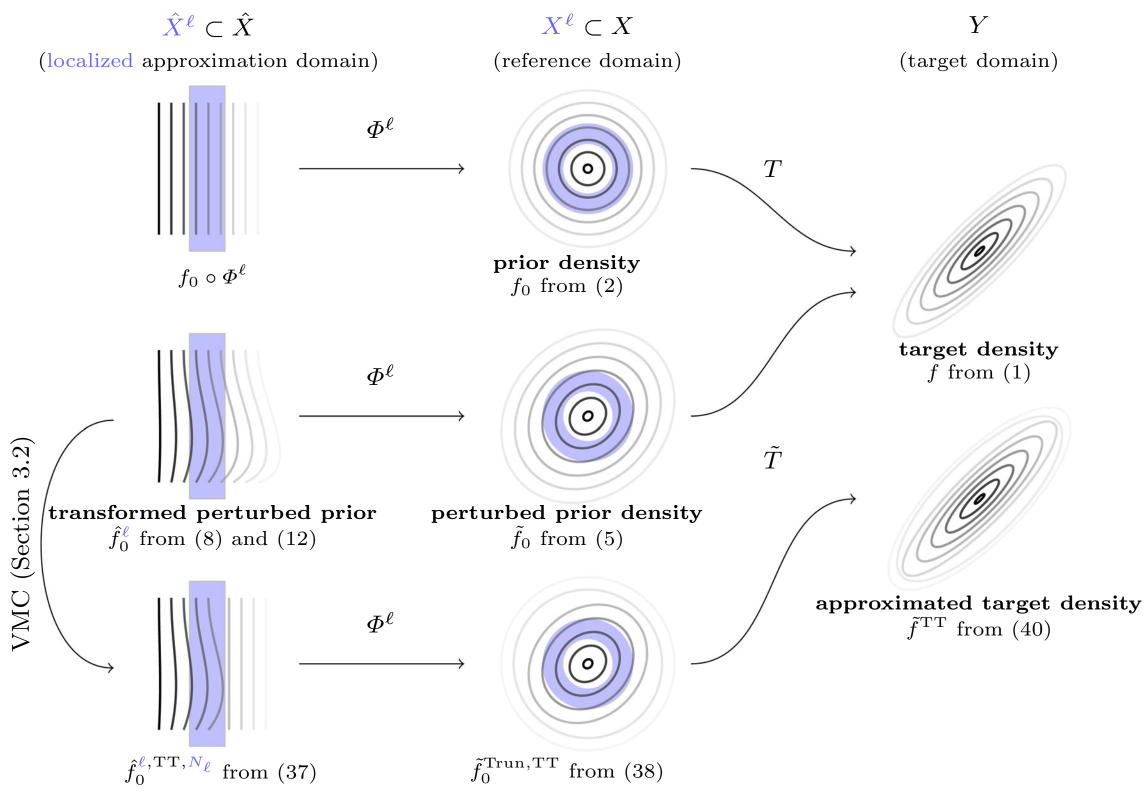

$\hat{X}^\ell \subset \hat{X}$
(localized approximation domain)
$X^\ell \subset X$
(reference domain)
Y
(target domain)
Φ^ℓ
T
$f_0 \circ \Phi^\ell$
prior density
f_0 from (2)
target density
f from (1)
Φ^ℓ
$\tilde{T}$
transformed perturbed prior
$\hat{f}_0^\ell$ from (8) and (12)
perturbed prior density
$\tilde{f}_0$ from (5)
approximated target density
$\tilde{f}^{\mathrm{TT}}$ from (40)
VMC (Section 3.2)
Φ^ℓ
$\hat{f}_0^{\ell,\mathrm{TT},N_\ell}$ from (37)
$\tilde{f}_0^{\mathrm{Trun},\mathrm{TT}}$ from (38)

Fig. 1 Overview of the presented method sketching the different involved transformations and approximations with references to the respective equations

and the introduction of the transport map $\tilde{T}$ yields the claim

$$\int_X \log\left(\frac{f \circ \tilde{T}(x)}{\tilde{f}^{\mathrm{TT}} \circ \tilde{T}(x)} \cdot \frac{|\mathcal{J}_{\tilde{T}}(x)|}{|\mathcal{J}_{\tilde{T}}(x)|}\right) \tilde{f}_0(x)\, \mathrm{d}\lambda(x)$$

$$= \mathrm{d}_{\mathrm{KL}}(X;\, \tilde{f}_0,\, \tilde{f}_0^{\mathrm{Trun},\mathrm{TT}}). \tag{46}$$

□

4.2 Truncation error

Since our approximation scheme relies on the truncation of the density, we introduce a convenient type of decay on the outer layer of the perturbed prior.

Definition 2 (*outer polynomial exponential decay*) A function $\tilde{f}_0\colon X \to \mathbb{R}^+$ has outer polynomial exponential decay if there exists a simply connected compact set $K \subset X$ with a polynomial π^+ which is positive on $X \setminus K$ and some $C > 0$ such that

$$\tilde{f}_0(x) \leq C \exp\left(-\pi^+(x)\right), \quad x \in X \setminus K. \tag{47}$$

The error introduced by the Gaussian extension is estimated in the next lemma.

Lemma 2 (truncation error) *For $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d,d}$, let $\tilde{f}_0$ have outer polynomial exponential decay with positive polynomial $\tilde{\pi}^+$ and $\tilde{C} > 0$ with $K = \overline{B_R(\mu)}$ for some $R > 0$. Then, for $C_\Sigma = 1/2\lambda_{\min}(\Sigma^{-1})$ there exists a $C = C(\tilde{C}, \Sigma, d, C_\Sigma) > 0$ such that*

$$\|\tilde{f}_0 - \tilde{f}_0^{\mathrm{Trun}}\|_{L^1(X\setminus K)} \lesssim \|\exp\left(-\tilde{\pi}^+\right)\|_{L^1(X\setminus K)} + \Gamma\left(d/2, C_\Sigma R^2\right)$$

and

$$\left|\int_{X\setminus K} \log\left(\frac{\tilde{f}_0}{f_{\Sigma,\mu}}\right) \tilde{f}_0 \,\mathrm{d}x\right| \leq \int_{X\setminus K} \left(\frac{1}{2}\|x\|^2_{\Sigma^{-1}} + \tilde{\pi}^+(x)\right) e^{-\tilde{\pi}^+(x)}\, \mathrm{d}\lambda(x)$$

with the incomplete Gamma function Γ.

Proof The proof follows immediately from the definition of $\tilde{f}_0^{\mathrm{Trun}}$. □

In the case that the perturbed prior is close to a Gaussian standard normal distribution, it holds $C \approx 1$.

4.3 Low-rank compression error

In this section we discuss the error introduced by compressing a full algebraic tensor into a tensor in low-rank tensor train format. The higher order singular value decomposition (HOSVD) (Oseledets and Tyrtyshnikov 2010) is based on successive unfoldings of the full tensor into matrices, which are orthogonalized and possibly truncated by a singular value decomposition. This algorithm leads to the following result.

Lemma 3 (Theorem 2.2 Oseledets and Tyrtyshnikov 2010) *For any $g \in \mathcal{V}_\Lambda$ and $\boldsymbol{r} \in \mathbb{R}^{d-1}$ there exists an extended low-rank tensor train $g_{\boldsymbol{r}} \in \mathcal{M}_{\boldsymbol{r}}$ such that*

$$\|g - g_{\boldsymbol{r}}\|^2_{\mathcal{V}(\hat{X})} \le \sum_{i=1}^{d-1} \sigma_i^2, \tag{48}$$

where σ_i is the distance of the i-th unfolding matrix of the coefficient tensor of g in the HOSVD to its best rank r_i approximation in the Frobenius norm.

Remark 3 Estimate (48) is rather unspecific as the σ_i cannot be quantified a priori. In the special case of Gaussian densities we refer to Rohrbach et al. (2020) for an examination of the low-rank representation depending on the covariance structure. When the transport $\tilde{T}$ maps the considered density only "closely" to a standard Gaussian, the results can be applied immediately to our setting and more precise estimates are possible.

4.4 VMC error analysis

To examine the VMC convergence in our setting, we recall the analysis of Eigel et al. (2019b) in a slightly more general manner. Analogously to Sect. 3.1, we use the notation $\hat{\mathcal{X}}$ and w as a placeholder for any layer $\hat{X}^\ell$ and weight w_ℓ for $\ell = 1, \ldots, L$. Here we assume that $L^2(\hat{\mathcal{X}}, w)$ is continuously embedded in $L^1(\hat{\mathcal{X}}, w)$.

Recall the cost functional $\mathscr{J}$ from (32) defined by a loss function ι depending on the transformed perturbed prior $\hat{f}_0$ as in Sect. 3.3. As a first step we show compatibility conditions of two specific types of loss functions corresponding to the Kullback–Leibler divergence and the L^2-norm.

Lemma 4 (KL loss compatibility) *Let $\hat{f}_0 \in \mathcal{V}(\hat{\mathcal{X}}, 0, c^*)$ for $c^* < \infty$ and $0 < \underline{c} < \overline{c} < \infty$. Then*

$$\begin{aligned}\mathcal{V}(\hat{X}, \underline{c}, \overline{c}) \ni \hat{g} \mapsto \iota(\hat{g}, \hat{x}) &= \iota(\hat{g}, \hat{x}, \hat{f}_0) \\ &:= -\log(\hat{g}(\hat{x}))\hat{f}_0(\hat{x})\end{aligned} \tag{49}$$

is uniformly bounded and Lipschitz continuous on the model class $\mathcal{M} = \mathcal{M}_{\boldsymbol{r}}(\hat{\mathcal{X}}, \underline{c}, \overline{c})$ if $P_\alpha \in L^\infty(\hat{\mathcal{X}})$ for every $\alpha \in \Lambda$. Furthermore, $\mathscr{J}$ is globally Lipschitz continuous on the metric space $(\mathcal{V}(\hat{\mathcal{X}}, \underline{c}, \overline{c}), d_{\mathcal{V}(\hat{\mathcal{X}}, \underline{c}, \overline{c})})$.

Proof The loss ι is bounded on $\mathcal{M}_{\boldsymbol{r}}(\hat{\mathcal{X}}, \underline{c}, \overline{c})$ since $0 < \underline{c} < \overline{c} < \infty$. Let $\hat{g}_1, \hat{g}_2 \in \mathcal{V}(\hat{\mathcal{X}}, \underline{c}, \overline{c})$. Then

$$|\iota(\hat{g}_1, \hat{x}) - \iota(\hat{g}_2, \hat{x})| \le \underbrace{\frac{1}{\underline{c}} \sup_{\hat{x} \in \hat{\mathcal{X}}} \{\hat{f}_0(\hat{x})\}}_{:= C^* < \infty} |\hat{g}_1(\hat{x}) - \hat{g}_2(\hat{x})|. \tag{50}$$

The global Lipschitz continuity of $\mathscr{J}$ follows by using (50) and

$$\begin{aligned}|\mathscr{J}(\hat{g}_1) - \mathscr{J}(\hat{g}_2)| &\le C^* \|\hat{g}_1 - \hat{g}_2\|_{L^1(\hat{\mathcal{X}}, w)} \\ &\le C C^* d_{\mathcal{V}(\hat{\mathcal{X}}, \underline{c}, \overline{c})}(\hat{g}_1, \hat{g}_2),\end{aligned} \tag{51}$$

with a constant C related to the embedding of $L^2(\hat{\mathcal{X}}, w)$ into $L^1(\hat{\mathcal{X}}, w)$. If $\hat{g}_1, \hat{g}_2$ are in $\mathcal{M}_{\boldsymbol{r}}(\hat{\mathcal{X}}, \underline{c}, \overline{c})$ with coefficient tensors G_1 and $G_2 \in \mathbb{TT}_{\boldsymbol{r}}$ then by Parseval's identity and the finite dimensionality of $\mathcal{M}_{\boldsymbol{r}}(\hat{\mathcal{X}}, \underline{c}, \overline{c})$ there exists $c = c\left(\sup_{\alpha \in \Lambda} \|P_\alpha\|_{L^\infty(\hat{X})}\right) > 0$ such that

$$\begin{aligned}|\hat{g}_1(x) - \hat{g}_2(x)| \le c\|G_1 - G_2\|_{\ell^2(\mathbb{T})} &= c\|\hat{g}_1 - \hat{g}_2\|_{\mathcal{V}} \\ &= c\, d_{\mathcal{V}(\hat{\mathcal{X}}, \underline{c}, \overline{c})}(\hat{g}_1, \hat{g}_2),\end{aligned} \tag{52}$$

which yields the Lipschitz continuity on $\mathcal{M}_{\boldsymbol{r}}(\hat{\mathcal{X}}, \underline{c}, \overline{c})$. □

Lemma 5 (L^2-loss compatibility) *Let $\hat{f}_0 \in \mathcal{V}(\hat{X}, 0, c^*)$ for $c^* < \infty$ and let $0 \le \underline{c} < \overline{c} < \infty$. Then*

$$\mathcal{V}(\hat{\mathcal{X}}, \underline{c}, \overline{c}) \ni \hat{g} \mapsto \iota(\hat{g}, \hat{x}) = \iota(\hat{g}, \hat{x}, \hat{f}_0) := |\hat{g}(\hat{x}) - \hat{f}_0(\hat{x})|^2 \tag{53}$$

is uniformly bounded and Lipschitz continuous on $\mathcal{M}_{\boldsymbol{r}}(\hat{\mathcal{X}}, \underline{c}, \overline{c})$ provided $P_\alpha \in L^\infty(\hat{\mathcal{X}})$ for every $\alpha \in \Lambda$.

Proof Let $\hat{g}_1, \hat{g}_2 \in \mathcal{V}(\hat{\mathcal{X}}, \underline{c}, \overline{c})$. Then

$$\begin{aligned}|\iota(\hat{g}_1, \hat{x}) - \iota(\hat{g}_2, \hat{x})| \le\ & |\hat{g}_1(\hat{x}) - \hat{g}_2(\hat{x})| \cdot |\hat{g}_2(\hat{x}) + \hat{g}_2(\hat{x})| \\ & + 2|\hat{g}_1(\hat{x}) - \hat{g}_2(\hat{x})|\hat{f}_0(\hat{x}).\end{aligned} \tag{54}$$

Due to $\overline{c} < \infty$, the Lipschitz property follows as in the proof of Lemma 4 if $\hat{g}_1, \hat{g}_2 \in \mathcal{M}_{\boldsymbol{r}}(\hat{\mathcal{X}}, \underline{c}, \overline{c})$. □

Let $\hat{g}_{\mathcal{M}}$ and $\hat{g}_{\mathcal{M},N}$ be as in (34) and (36). The analysis examines different error components with respect to $\hat{f}_0 \in \mathcal{V}(\hat{\mathcal{X}}, 0, c^*)$ for some $0 < c^* < \infty$ defined by

$$\mathcal{E} := \left|\mathscr{J}(\hat{f}_0) - \mathscr{J}\left(\hat{g}^*_{\mathcal{M},N}\right)\right|, \tag{55}$$

$$\mathcal{E}_{\text{app}} := \left|\mathscr{J}(\hat{f}_0) - \mathscr{J}\left(\hat{g}^*_{\mathcal{M}}\right)\right|, \tag{56}$$

$$\mathcal{E}_{\text{gen}} := \left|\mathscr{J}\left(\hat{g}^*_{\mathcal{M}}\right) - \mathscr{J}\left(\hat{g}^*_{\mathcal{M},N}\right)\right|, \tag{57}$$

denoting the VMC, approximation and generalization error, respectively. By a simple splitting, the VMC error can be bounded by the approximation and the generalization error,[3] namely

$$\mathcal{E} \leq \mathcal{E}_{\mathrm{app}} + \mathcal{E}_{\mathrm{gen}}. \tag{58}$$

Due to the global Lipschitz property on $\mathcal{V}(\hat{\mathcal{X}}, \underline{c}, \overline{c})$ with $\underline{c} > 0$ in the setting of (49) or $\underline{c} \geq 0$ as in (53), the approximation error can be bounded by the best approximation in $\mathcal{M}$. In particular there exists $C > 0$ such that

$$\mathcal{E}_{\mathrm{app}} \leq C \inf_{v \in \mathcal{M}} \|h^* - v\|^2_{\mathcal{V}(\hat{\mathcal{X}})}. \tag{59}$$

We note that such a bound by the best approximation in $\mathcal{M}$ with respect to the $\mathcal{V}(\hat{\mathcal{X}})$-norm may not be required when using the Kullback–Leibler divergence if one is interested directly in the best approximation in this divergence. Then the assumption $\underline{c} > 0$ can be relaxed in the construction of $\mathcal{V}(\hat{\mathcal{X}}, \underline{c}, \overline{c})$ since no global Lipschitz continuity of $\mathcal{J}$ in Lemma 4 is necessary. Thus the more natural subspace of $\mathcal{V}(\hat{\mathcal{X}}, 0, \overline{c})$ of absolutely continuous functions with respect to $\hat{f}_0$ may be considered instead.

It remains to bound the statistical generalization error $\mathcal{E}_{\mathrm{gen}}$. For this the notion of covering numbers is required. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be an abstract probability space.

Definition 3 (*covering number*) Let $\epsilon > 0$. The covering number $\nu(\mathcal{M}, \epsilon)$ denotes the minimal number of open balls of radius ϵ with respect to the metric $d_{\mathcal{V}(\hat{\mathcal{X}}, \underline{c}, \overline{c})}$ needed to cover $\mathcal{M}$.

Lemma 6 *Let ι be defined as in (49) or (53). Then there exist $C_1, C_2 > 0$ only depending on the uniform bound and the Lipschitz constant of $\mathcal{M}$ given in Lemmas 4 and 5, respectively, such that for $\epsilon > 0$ and $N \in \mathbb{N}$ denoting the number of samples in the empirical cost functional in (35) it holds*

$$\mathbb{P}[\mathcal{E}_{\mathrm{gen}} > \epsilon] \leq 2\nu(\mathcal{M}, C_2^{-1}\epsilon)\delta(1/4\epsilon, N), \tag{60}$$

with $\delta(\epsilon, N) \leq 2\exp(-2\epsilon^2 N/C_1^2)$.

Proof The claim follows immediately from Lemmas 4 and 5, respectively, and (Thm. 4.12, Cor. 4.19 Eigel et al. 2019b). □

Remark 4 *(choice of $\underline{c}, \overline{c}$ and $\hat{\mathcal{X}}$)* Due to the layer based representation in (11) and (15) on each layer $\hat{X}^\ell = \Phi^{-1}(X^\ell)$ we have the freedom to choose $\underline{c}$ separately. In particular, assuming that the perturbed prior $\tilde{f}_0$ decays per layer, we can choose $\underline{c}$ according to the decay and with this control the constant in (50).

4.5 A priori estimate

In this section we state our main convergence result.

Assumption 1 For a target density $f : Y \to \mathbb{R}_+$ and a transport map $\tilde{T} : X \to Y$, there exists a simply connected compact domain K such that $\tilde{f}_0 = (f \circ T) \otimes |\mathcal{J}_T| \in L^2(K)$ has outer polynomial exponential decay with polynomial π^+ on $X \setminus K$. Consider the symmetric positive definite matrix $\Sigma \in \mathbb{R}^{d,d}$ and $\mu \in \mathbb{R}^d$ as the covariance and mean for the outer approximation $f_{\Sigma,\mu}$. Furthermore, let $K = \overline{\bigcup_{\ell=1}^L X^\ell}$ where X^ℓ is the image of a rank-1 stable diffeomorphism $\Phi^\ell : \hat{X}^\ell \to X^\ell$ such that there exists $0 < c_\ell^* < \infty$ with $\hat{f}_0^\ell(\hat{x}) \leq c_\ell^*$ for $\hat{x} \in \hat{X}_\ell$ for $\ell = 1, \ldots, L$.

We can now formulate the main theorem of this section regarding the convergence of the developed approximation with respect to the Hellinger distance and the KL divergence.

Theorem 1 (A priori convergence) *Let Assumption 1 hold and let a sequence of sample sizes $(N^\ell)_{\ell=1}^L \subset \mathbb{N}$ be given. For every $\ell = 1, \ldots, L$, consider bounds $0 < \underline{c}^\ell < \overline{c}^\ell < \infty$ and let $\tilde{f}^{\mathrm{TT}}$ be defined as in (40). Then there exist constants $C, C_\Sigma, C^\ell, C_\iota^\ell > 0$, $\ell = 1, \ldots, L$, such that for $\sharp \in \{\mathrm{KL}, \mathrm{Hell}\}$ and $p_{\mathrm{Hell}} = 2$ and $p_{\mathrm{KL}} = 1$*

$$\mathrm{d}_\sharp^{p_\sharp}(Y, f, \tilde{f}^{\mathrm{TT}}) \leq C \left(\sum_{\ell=1}^L \left(\mathcal{E}^\ell_{\mathrm{best}} + \mathcal{E}^\ell_{\mathrm{sing}} + \mathcal{E}^\ell_{\mathrm{gen}} \right) + \mathcal{E}^\sharp_{\mathrm{trun}} \right). \tag{61}$$

Here, $\mathcal{E}^\ell_{\mathrm{best}}$ denotes the error of the best approximation $\hat{g}_\Lambda^{\ell,*}$ to $\hat{f}_0^\ell$ in the full truncated space $\mathcal{V}_\Lambda^\ell(\underline{c}^\ell, \overline{c}^\ell) = \mathcal{V}_\Lambda^\ell \cap \mathcal{V}(\hat{X}^\ell, \underline{c}^\ell, \overline{c}^\ell)$ given by

$$\mathcal{E}^\ell_{\mathrm{best}} := \|\hat{f}_0^\ell - \hat{g}_\Lambda^{\ell,*}\|_{\mathcal{V}(\hat{X}^\ell)} = \inf_{\hat{g}^\ell \in \mathcal{V}_\Lambda^\ell(\underline{c}^\ell, \overline{c}^\ell)} \|\hat{f}_0^\ell - \hat{g}^\ell\|_{\mathcal{V}(\hat{X}^\ell)},$$

$\mathcal{E}^\ell_{\mathrm{sing}}$ is the low-rank approximation error of the algebraic tensor $G : \Lambda \to \mathbb{R}$ associated to $\hat{g}_\Lambda^{\ell,*}$ and the truncation error $\mathcal{E}^\#_{\mathrm{trun}}$ is given by

$$\left(\mathcal{E}^{\mathrm{Hell}}_{\mathrm{trun}}\right)^2 := \|\exp(-\pi^+)\|_{L^1(X \setminus K)} + \Gamma\left(d/2, C_\Sigma R^2\right),$$
$$\mathcal{E}^{\mathrm{KL}}_{\mathrm{trun}} := \int_{X \setminus K} \left(\frac{1}{2}\|x\|^2_{\Sigma^{-1}} + \tilde{\pi}^+(x) \right) e^{-\tilde{\pi}^+(x)} \,\mathrm{d}\lambda(x).$$

Furthermore, for any $(\epsilon^\ell)_{\ell=1}^L \subset \mathbb{R}_+$ the generalization errors $\mathcal{E}^\ell_{\mathrm{gen}}$ can be bounded in probability by

$$\mathbb{P}(\mathcal{E}^\ell_{\mathrm{gen}} > \epsilon^\ell) \leq 2\nu(\mathcal{M}^\ell, C^\ell \epsilon^\ell)\delta^\ell(1/4\epsilon^\ell, N^\ell),$$

with ν denoting the covering number from Definition 3 and $\delta^\ell(\epsilon, N) \leq 2\exp(-2\epsilon^2 N/C_\iota^\ell)$.

[3] Neglecting an intractable optimization error.

Proof We first prove (61) for $\sharp = \mathrm{Hell}$. Note that $|\sqrt{a} - \sqrt{b}| \leq \sqrt{|a-b|}$ for $a, b \geq 0$ and with Lemma 1 it holds

$$\begin{aligned} \mathrm{d}^2_{\mathrm{Hell}}(Y; f, \tilde{f}^{\mathrm{TT}}) &= \mathrm{d}^2_{\mathrm{Hell}}(X; \tilde{f}_0, \tilde{f}_0^{\mathrm{Trun,TT}}) \\ &\leq 1/2 \| \tilde{f}_0 - \tilde{f}_0^{\mathrm{Trun,TT}} \|_{L^1(K)} \\ &\quad + 1/2 \| \tilde{f}_0 - \tilde{f}_0^{\mathrm{Trun,TT}} \|_{L^1(X \setminus K)}. \end{aligned}$$

Since $K = \cup_{\ell=1}^L X^\ell$ and X^ℓ are bounded, there exist constants $C(X^\ell) > 0$, $\ell = 1, \dots, L$, such that

$$\begin{aligned} \| \tilde{f}_0 - \tilde{f}_0^{\mathrm{Trun,TT}} \|_{L^1(K)} &= \sum_{\ell=1}^{L} \| \tilde{f}_0 - \tilde{f}_0^{\mathrm{Trun,TT}} \|_{L^1(X^\ell)} \\ &\leq \sum_{\ell=1}^{L} C(X_\ell) \| \tilde{f}_0 - \tilde{f}_0^{\mathrm{Trun,TT}} \|_{L^2(X^\ell)}. \end{aligned}$$

Moreover, by construction

$$\| \tilde{f}_0 - \tilde{f}_0^{\mathrm{Trun,TT}} \|_{L^2(X^\ell)} = \| \hat{f}_0^\ell - \hat{f}_0^{\ell,\mathrm{TT},N_\ell} \|_{\mathcal{V}(\hat{X}^\ell)}. \tag{62}$$

The claim follows by application of Lemmas 2, 3 and 6 together with (58).

To show (61) for $\sharp = \mathrm{KL}$, note that by Lemma 1 and the construction (38) it holds

$$\begin{aligned} \mathrm{d}_{\mathrm{KL}}(Y; f, \tilde{f}^{\mathrm{TT}}) = \sum_{\ell=1}^{L} \int_{X^\ell} \log \frac{\tilde{f}_0}{\tilde{f}_0^{\ell,\mathrm{TT}}} \tilde{f}_0 \mathrm{d}\lambda(x) \\ + \int_{X \setminus K} \log \frac{\tilde{f}_0}{f_{\Sigma,\mu}} \tilde{f}_0 \mathrm{d}\lambda(x). \end{aligned} \tag{63}$$

Using Lemma 2 we can bound the integral over $X \setminus K$ by the truncation error $\mathcal{E}_{\mathrm{trun}}$. Employing the loss function and cost functional of Lemma 4 yields

$$\int_{X^\ell} \log \frac{\tilde{f}_0}{\tilde{f}_0^{\ell,\mathrm{TT}}} \tilde{f}_0 \mathrm{d}\lambda(x) \leq \mathcal{E}^\ell_{\mathrm{app}} + \mathcal{E}^\ell_{\mathrm{gen}}. \tag{64}$$

The claim follows by application of Lemmas 3 and 6 together with (58). □

4.6 Polynomial approximation in weighted L^2 spaces

In order to make the error bound (61) in Theorem 1 more explicit with respect to $\mathcal{E}_{\mathrm{best}}$, we consider the case of a smooth density function with analytic extension. The analysis follows the presentation in Babuška et al. (2010) and leads to exponential convergence rates by an iterative interpolation argument based on univariate best approximation bounds by interpolation. An analogous analysis for more general regularity classes is possible but not in the scope of this article.

Let $\hat{\mathcal{X}} = \bigotimes_{i=1}^d \hat{\mathcal{X}}_i \subset \mathbb{R}^d$ be bounded and $w = \otimes_{i=1}^d w_i \in L^\infty(\hat{\mathcal{X}})$ a non-negative weight such that $\mathcal{C}(\hat{\mathcal{X}}) \subset \mathcal{V} := L^2(\hat{\mathcal{X}}, w) = \bigotimes_{i=1}^d L^2(\hat{\mathcal{X}}_i, w_i)$.

For a Hilbert space H, a bounded set $I \subset \mathbb{R}$ and a function $f \in \mathcal{C}(I; H) \subset L^2(I, w; H)$ with weight $w \colon I \to \mathbb{R}$, let $\mathcal{I}_n \colon \mathcal{C}(I; H) \to L^2(I, w; H)$ denote the continuous Lagrange interpolation operator.

Assume that $f \in \mathcal{C}(I; H)$ admits an analytic extension in the region of the complex plane $\Sigma(I; \tau) := \{z \in \mathbb{C} | \operatorname{dist}(z, I) \leq \tau\}$ for some $\tau > 0$. Then, referring to Babuška et al. (2010),

$$\| f - \mathcal{I}_n f \|_{L^2(I,w;H)} \lesssim \sigma(n, \tau) \max_{z \in \Sigma(I;\tau)} \| f(z) \|_H, \tag{65}$$

with $\sigma(n, \tau) := 2(\rho - 1)^{-1} \exp(-n \log(\rho))$ and $\rho := 2\tau/|I| + \sqrt{1 + 4\tau^2/|I|^2} > 1$. Using an iterative argument over d dimensions, a convergence rate for the interpolation of $f \in \mathcal{C}(\hat{\mathcal{X}}; \mathbb{R}) \subset L^2(\hat{\mathcal{X}}, w; \mathbb{R})$ can be derived from the 1-dimensional convergence. More specifically, let $\mathcal{I}_\Lambda : \mathcal{C}(\hat{\mathcal{X}}) \mapsto L^2(\hat{\mathcal{X}}, w)$ denote the continuous interpolation operator $\mathcal{I}_\Lambda := \mathcal{I}^1_{n_1} \circ \mathcal{I}^{2:d}_{n_2:n_d}$ written as composition of a 1-dimensional and a $d-1$-dimensional interpolation with continuous

$$\mathcal{I}^1_{n_1} : \mathcal{C}(\hat{\mathcal{X}}_1) \to L^2\left(\bigtimes_{i=2}^d \hat{\mathcal{X}}_i, \otimes_{i=2}^d w_i \right)$$

and

$$\mathcal{I}^{2,\dots,d}_{n_2,\dots,n_d} : \mathcal{C}\left(\bigtimes_{i=2}^d \hat{\mathcal{X}}_i \right) \to H$$

with $H = L^2(\bigtimes_{i=2}^d \hat{\mathcal{X}}_i, \otimes_{i=2}^d w_i)$. Then, for $f \in \mathcal{C}(\hat{\mathcal{X}})$ and some $C > 0$ it follows

$$\begin{aligned} \| f - \mathcal{I}_\Lambda f \| &\leq \| f - \mathcal{I}^1_{n_1} f \| + \| \mathcal{I}^1_{n_1} (f - \mathcal{I}^{2,\dots,d}_{n_2,\dots,n_d} f) \| \\ &\lesssim \| f - \mathcal{I}^1_{n_1} f \| \\ &\quad + \sup_{\hat{x}_1 \in \hat{\mathcal{X}}_1} \| f(x_1) - \mathcal{I}^{2,\dots,d}_{n_2,\dots,n_d} f(x_1) \|_H. \end{aligned}$$

The second term of the last bound is a $d-1$-dimensional interpolation and can hence be bounded uniformly over $\hat{x}_1$ by a similar iterative argument. We summarize the convergence result for $\mathcal{E}^\ell_{\mathrm{best}}$ in the spirit of (Theorem 4.1 Babuška et al. 2010).

Lemma 7 *Let $\hat{f}_0 \in \mathcal{C}(\hat{X}^\ell) \subset L^2(\hat{X}^\ell, w)$ admit an analytic extension in the region*

$$\Sigma(\hat{X}^\ell, (\tau_i^\ell)_{i=1}^d) = \bigtimes_{i=1}^d \Sigma(\hat{X}_i^\ell, \tau_i^\ell)$$

for some $\tau_i^\ell > 0$, $\ell = 1, \ldots, L$, $i = 1, \ldots, d$. Then, with σ from (65)*,*

$$\inf_{v \in \mathcal{V}_\Lambda} \|\hat{f}_0 - v\|_{L^2(\hat{X}^\ell, w)} \lesssim \sum_{i=1}^{d} \sigma(n_i, \tau_i).$$

In case that $\underline{c} \leq \hat{f}_0(\hat{x}), \hat{g}^*(\hat{x}) \leq \overline{c}$ is satisfied for $\hat{g}^* := \operatorname{argmin}_{\hat{g} \in \mathcal{V}_\Lambda} \|f - \hat{g}\|_{L^2(\hat{X}^\ell, w)}$, the decay rate carries over onto the space $\mathcal{V}_\Lambda^\ell(\underline{c}^\ell, \overline{c}^\ell)$. If only $\underline{c} \leq \hat{f}_0(\hat{x}) \leq \overline{c}$ holds, the image of $\hat{g}^*$ can be restricted to $[\underline{c}, \overline{c}]$, see e.g. Cohen and Migliorati (2017). This approximation in fact admits a smaller error than $\hat{g}^*$.

Remark 5 The interpolation argument on polynomial discrete spaces could be expanded to other orthonormal systems such as trigonometric polynomial, admitting well-known Lebesgue constants as in Da Fies and Vianello (2013).

Remark 6 Explicit best approximation bounds for appropriate smooth weights w as in the case of spherical coordinates can be obtained using partial integration techniques as in Mead and Delves (1973). There, the regularity class of $\hat{f}_0$ uses high-order weighted Sobolev spaces based on derivatives of w as in the case of classical polynomials.

5 Algorithm

Since a variety of techniques is employed in the proposed density discretization, this section provides an exemplary algorithmic workflow to illustrate the required steps in practical applications (see also Fig. 7 for a sketch of the components of the method). The general method to obtain a representation of the density (1) by its auxiliary reference (5) is summarized in Algorithm 1. Based on this, the computation of possible quantities of interest such as moments (6) or marginals are considered in Sects. 6.2.1 and 6.2.2, respectively. In the following we briefly describe the involved algorithmic procedures.

Computing the transformation Obtaining a suitable transport map is a current research topic and examined e.g. in Papamakarios et al. (2021), Parno and Marzouk (2018), Tran et al. (2019) and Marzouk et al. (2016). In Sect. 2.1, two naive options are introduced. In the numerical applications, we employ an affine transport and also illustrate the capabilities of a quadratic transport in a two-dimensional example. For the affine linear transport we utilize a semi-Newton optimizer to obtain the maximum value of f and an approximation of the Hessian at the optimal value, see Sect. A.1. For the construction of a quadratic transport we rely on the library `TransportMaps` (Baptista et al. 2015-2018). The task to provide the (possibly inexact) transport map is summarized in the function

$$\tilde{T} \leftarrow \text{ComputeTransport}[f]. \tag{66}$$

In the following paragraphs we assume Φ^ℓ to be the multivariate polar transformation as in "Appendix B.1", defined on the corresponding hyperspherical shells $\hat{X}^\ell$. We refer to $\hat{X}_1^\ell$ as the *radial dimension* and $\hat{X}_i^\ell$ as the *angular dimensions* for $1 < i \leq d$. The computations on each shell $\hat{X}^\ell$, $\ell = 1, \ldots, L$, are fully decoupled and suitable for parallelization. Note that the proposed method is easily adapted to other transformations Φ^ℓ.

Generating an orthonormal basis To obtain suitable finite dimensional subspaces, one has to introduce spanning sets that allow for an efficient computation of e.g. moments (3) and the optimization of the functional (32). Given a fixed dimension vector $\boldsymbol{n}^\ell \in \mathbb{N}^d$ for the current $\hat{X}^\ell$, $\ell = 1, \ldots, L$, and introducing the weight w^ℓ given by the jacobian of the chosen transformation Φ^ℓ, the function

$$\mathcal{P}^\ell = \{\mathcal{P}_i^\ell\}_{i=1}^d \leftarrow \text{GenerateONB}[\hat{X}^\ell, \boldsymbol{n}^\ell, w^\ell, \tau_{\text{GS}}] \tag{67}$$

can be split into three distinct algorithmic parts as follows.

- *1st coordinate* $\hat{x}_1$: The computation of an orthonormal polynomial basis $\{P_{1,\alpha}^\ell\}_\alpha$ with respect to the weight $w_1^\ell(\hat{x}_1) = \hat{x}_1^{d-1}$ in the radial dimension by a stabilized Gram-Schmidt method. This is numerically unstable since the involved summations cause cancellation. As a remedy, we define *arbitrary precision polynomials* with a significant digit length τ_{mant} to represent polynomial coefficients. By this, point evaluations of the orthonormal polynomials and computations of integrals of the form

$$\int_{\hat{X}_1^\ell} \hat{x}_1^m P_{1,\alpha}^\ell(\hat{x}_1) \hat{x}_1^{d-1} \mathrm{d}\lambda(\hat{x}_1), \quad m \in \mathbb{N}, \tag{68}$$

  e.g. required for computing moments with polynomial transport, can be realized with high precision. The length τ_{mant} is set to 100 in the numerical examples and the additional run-time is negligible as the respective calculations can be precomputed.
- *2nd coordinate* $\hat{x}_2$: Since $\hat{X}_2^\ell = [0, 2\pi]$ and to preserve periodicity, we employ trigonometric polynomials given by

$$P_{2,j}^\ell(\hat{x}_2) = \begin{cases} \frac{1}{\sqrt{2\pi}}, & j = 1 \\ \frac{\sin(\frac{j}{2}\hat{x}_2)}{\sqrt{\pi}}, & j \text{ even} \\ \frac{\cos(\frac{j-1}{2}\hat{x}_2)}{\sqrt{\pi}}, & j > 1 \text{ odd.} \end{cases} \tag{69}$$

Note that here the weight function is constant, i.e. $w_2^\ell(\hat{x}_2) \equiv 1$, and the defined trigonometric polynomials are orthonormal in $L^2(\hat{X}_2^\ell)$.

- *coordinate* $\hat{x}_3, \ldots, \hat{x}_d$: On the remaining angular dimensions $i = 3, \ldots, d$, we employ the usual Gram-Schmidt orthogonalization algorithm on $[0, \pi]$ with weight function $w_i^\ell(\hat{x}_i) = \sin^i(\hat{x}_i)$, based on polynomials.

Fortunately, the basis for dimensions $1 < i \leq d$ coincides on every layer $\ell = 1, \ldots, L$. It hence can be computed just once and passed to the individual process handling the current layer. Only the basis in the radial dimension needs to be adjusted to $\hat{X}^\ell$. The parameter τ_{GS} collects all tolerance parameters for the applied numerical quadrature and the significant digit length τ_{mant}.

Generation of Samples We utilize Monte Carlo samples to approximate the L^2 integral for the empirical minimization formulation (37). To generate such samples on $\hat{X}^\ell$, we employ the inverse transform sampling based on the inverse cumulative distribution function of the normalized version of the weight function ω^ℓ. The generation process of $N \in \mathbb{N}$ samples is denoted as the function

$$\mathcal{S}^\ell := \left\{\left(\hat{x}^s, \hat{f}_0^\ell(\hat{x}^s)\right)\right\}_{s=1}^N \\ \uparrow \mathrm{GenerateSamples}[\hat{f}_0^\ell, \hat{X}^\ell, w^\ell, N]. \tag{70}$$

Reconstruction of a Tensor Train surrogate The VMC tensor reconstruction of Sect. 3 is summarized in the function

$$\left\{\hat{F}_{0,i}^{\ell,\mathrm{TT}}\right\}_{i=1}^d \leftarrow \mathrm{ReconstructTT}[\mathcal{S}^\ell, \mathcal{P}^\ell, \boldsymbol{r}^\ell, \tau_{\mathrm{Recon}}]. \tag{71}$$

The tensor components $\hat{F}_{0,i}^{\ell,\mathrm{TT}}$ are associated with the corresponding basis $\mathcal{P}_i^\ell$ to form a rank $\boldsymbol{r}^\ell$ extended tensor train as defined in (18) and (22). The additional parameter τ_{Recon} collects all parameters that determine the VMC algorithm.

The method basically involves the optimization of a loss functional over the set of tensor trains with rank (at most) $\boldsymbol{r}^\ell$. In the presented numerical computations we consider a mean-squared loss and the respective empirical approximation based on a current sample set $\mathcal{S}^\ell$. The tensor optimization—based on a rank adaptive, alternating direction fitting (ADF) algorithm—is implemented in the `xerus` library (Huber and Wolf 2014-2017) and wrapped in the `ALEA` framework (Eigel et al.). Additionally, the machine learning framework `PyTorch` (Paszke et al. 2017) can be utilized in `ALEA` to minimize the empirical cost functional from (35) by a wide range of state-of-the-art stochastic optimizers. The latter enable stochastic gradient methods to compute the tensor coefficients as known from machine learning applications. With this setting in mind, the actual meaning of the parameter τ_{Recon} depends on the chosen optimizer. In this article we focus on the ADF implementation and initialize e.g. the starting rank, the number of iteration of the ADF and a target residual norm.

6 Applications

In the preceding sections the creation of surrogate models of quite generic probability density functions are developed. Based on this, in the following we focus on actual applications where such a representation is beneficial. We start with the framework of Bayesian inverse problems with target density (1) corresponding to the Lebesgue posterior density. Subsequently, we cover the computation of moments and marginals.

6.1 Bayesian inversion

This section is devoted to a brief review of the Bayesian paradigm. We assume that the reader is familiar with the concept of statistical inverse problems and hence focus on the general formalism and highlight the notation with the setup of Sect. 2 in mind. We closely follow the presentation in Eigel et al. (2018) and refer to Stuart (2010), Dashti and Stuart (2016) and Kaipio and Somersalo (2006) for a comprehensive overview.

Let Y and $\mathcal{Y}$ denote separable Hilbert spaces equipped with inner products $\langle\cdot,\cdot\rangle_H$ for $H \in \{Y, \mathcal{Y}\}$. Assume there exists a *parameter to observation* map $\mathcal{G}\colon Y \to \mathcal{Y}$ such that for some given observation $\delta \in \mathcal{Y}$ the relation $\delta = \mathcal{G}(y) + \eta$ holds for some $y \in Y$ and noise $\eta \in \mathcal{Y}$. For instance in the Darcy model we take y as parameter determining the permeability of some porous medium and let $\mathcal{G}$ describe the observation of water pressure at some location in the domain. By taking η as a random variable with law $\mathcal{N}(0, C_0)$ for some symmetric positive definite covariance operator C_0 on $\mathcal{Y}$, the inference of y which explains the observation δ becomes a statistical inverse problem. Consequently, the quantities y and δ become random variables over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in Y and $\mathcal{Y}$, respectively. In Stuart (2010) mild conditions on the parameter to observation map are derived to show a continuous version of Bayes formula which yields the existence and uniqueness of the Radon-Nikodym derivative of the (posterior) measure π_δ of the conditional random variable $y|\delta$ with respect to a prior measure π_0 of y. More precisely, by assuming η to be Gaussian and independent with respect to y, both measures π_0 and π_δ on Y are related by the *Bayesian potential*

$$\Psi(y, \delta) := \frac{1}{2}\langle C_0^{-1}(\delta - \mathcal{G}(y)), \delta - \mathcal{G}(y)\rangle_{\mathcal{Y}}. \tag{72}$$

The posterior density is given by

$$\mathrm{d}\pi_\delta(y) = Z^{-1}\exp\left(-\Psi(y, \delta)\right)\mathrm{d}\pi_0(y). \tag{73}$$

Algorithm 1 Tensor train surrogate creation of perturbed prior

```
Input:   Lebesgue target density $f\colon \mathbb{R}^d \to \mathbb{R}_+$                                   (1)
         tensor spaces $\left\{\hat{X}^\ell\right\}_{\ell=1}^L$, with $\hat{X}^\ell = \bigtimes_{i=1}^d \hat{X}_i^\ell$                 (12)
         coordinate transformations $\Phi^\ell\colon \hat{X}^\ell \to X^\ell \subset \mathbb{R}^d$                (13)
            with rank-1 Jacobians $w^\ell := |\mathcal{J}_{\Phi^\ell}|\colon \hat{X}^\ell \to \mathbb{R}$
         basis dimensions $(\boldsymbol{n}^1, \ldots, \boldsymbol{n}^L)$, $\boldsymbol{n}^\ell \in \mathbb{N}^d$ for $\ell = 1, \ldots, L$        (22)
         sample size $N_\ell \in \mathbb{N}$, $\ell = 1, \ldots, L$ for level-wise reconstruction
         tensor train ranks $(\boldsymbol{r}^1, \ldots, \boldsymbol{r}^L)$, $\boldsymbol{r}^\ell \in \mathbb{N}^{d-1}$, for $\ell = 1, \ldots, L$     (18)
         Gram-Schmidt tolerance parameter $\tau_{\mathrm{GS}}$
         tensor reconstruction parameter $\tau_{\mathrm{Recon}}$
Output:  Level-wise low-rank approximation of perturbed prior

  Diffeomorphism $\tilde{T}$   $\leftarrow$   ComputeTransport[$f$]

  for $\ell = 1, \ldots, L$, (in parallel) do
    • Set transformed perturbed prior $\hat{f}_0^\ell(\hat{x}) := \left(f \circ \tilde{T} \otimes |\mathcal{J}_{\tilde{T}}|\right) \circ \Phi^\ell(\hat{x}),\ \hat{x} \in \hat{X}^\ell$
    • Build one-dimensional ONB $\mathcal{P}_i^\ell$ of $\mathcal{V}_{i,n_i^\ell} \subseteq L^2(\hat{X}_i^\ell, w_i^\ell)$ for $i = 1, \ldots, d$

          $\mathcal{P}^\ell = \{\mathcal{P}_i^\ell\}_{i=1}^d$   $\leftarrow$   GenerateONB[$\hat{X}^\ell, \boldsymbol{n}^\ell, w^\ell, \tau_{\mathrm{GS}}$]

    • Generate samples with respect to the weight $w^\ell$

      $\mathcal{S}^\ell := \left\{\left(\hat{x}^s, \hat{f}_0^\ell(\hat{x}^s)\right)\right\}_{s=1}^N$   $\leftarrow$   GenerateSamples[$\hat{f}_0^\ell, \hat{X}^\ell, w^\ell, N$]

    • Reconstruct TT surrogate $\tilde{f}_0^{\ell,\mathrm{TT}}\colon \hat{X}^\ell \to \mathbb{R}$

            $\left\{\tilde{F}_{0,i}^{\ell,\mathrm{TT}}\right\}_{i=1}^d$   $\leftarrow$   ReconstructTT[$\mathcal{S}^\ell, \mathcal{P}^\ell, \boldsymbol{r}^\ell, \tau_{\mathrm{Recon}}$]

    • Equip tensor components with basis

                  $\hat{f}_0^{\ell,\mathrm{TT}}(\hat{x})$   $:=$   $\sum_{\boldsymbol{k}}^{r^\ell} \prod_{i=1}^d \hat{f}_{0,i}^{\ell,\mathrm{TT}}[k_{i-1}, k_i](\hat{x}_i)$
      where $\hat{f}_{0,i}^{\ell,\mathrm{TT}}[k_{i-1}, k_i](\hat{x}_i)$   $:=$   $\sum_{j=1}^{n_j^\ell} \hat{F}_{0,i}^{\ell,\mathrm{TT}}[k_{i-1}, \mu_i, k_i] P_{i,j}^\ell(\hat{x}_i)$
  end for
  return $\left\{\tilde{f}_\ell\right\}_{l=1}^L$
```

We assume it exists and denote the normalization constant $Z := \mathbb{E}_{\pi_0}\left[\exp\left(-\Psi(y, \delta)\right)\right]$. Note that we interchangeably write y as an element of Y and the corresponding random variable with values in Y.

This simplified version of a *Bayesian inverse problem* can be cast into the framework of this manuscript by using the notation f for the posterior density (73) and $f_0 = \mathrm{d}\pi_0$ for the prior density.

6.2 Statistical quantities of interest

A common task is the efficient computation of the expectation of some quantity of interest (QoI) $Q\colon Y \to \mathbb{R}$,

$$\mathbb{E}\left[Q\right] = \int_Y Q(y) f(y) \mathrm{d}\lambda(y). \tag{74}$$

We discuss the special case of moment computations in Sect. 6.2.1 and the basis representations of marginals in Sect. 6.2.2. In those cases the structure of Q allows for direct computations of the integrals via tensor contractions. For more involved choices of the QoI, we suggest a universal sampling approach by repeated evaluation of the low-rank surrogate. More precisely, by application of the integral transformation we can approximate

$$\mathbb{E}\left[Q\right] \approx \sum_{\ell=1}^L \int_{\hat{X}^\ell} Q \circ \tilde{T} \circ \Phi^\ell(\hat{x}) \tilde{f}_0^{\ell,\mathrm{TT}}(\hat{x}) |\mathcal{J}_{\Phi^\ell}(\hat{x})| \mathrm{d}\lambda(\hat{x}) \tag{75}$$

and replace the integrals over $\hat{X}^\ell$ by Monte Carlo estimates with samples according to the (normalized) weight $|\mathcal{J}_{\Phi^\ell}|$. Those samples can be obtained by uniform sampling on the tensor spaces $\hat{X}^\ell$ and the inverse transform approach as mentioned in the paragraph **Generating Samples** of Sect. 5. Alternatively, efficient MCMC sampling by marginalization can be employed (Weare 2007).

6.2.1 Moment computation

In this section we discuss the computation of moments for the presented layer-based format with low-rank tensor train approximations. In particular, we are interested in an efficient generation of the moment map

$$\boldsymbol{\alpha} \mapsto \int\limits_{Y} y^{\boldsymbol{\alpha}} f(y) \mathrm{d}\lambda(y), \quad \boldsymbol{\alpha} = (\alpha_k)_k \in \mathbb{N}_0^d. \tag{76}$$

Given some transport $\tilde{T} \colon X \to Y$ with an associated perturbed prior $\tilde{f}_0 = (f \circ \tilde{T}) \otimes |\mathcal{J}_{\tilde{T}}|$, an integral transformation yields

$$\int\limits_{Y} y^{\boldsymbol{\alpha}} f(y) \mathrm{d}\lambda(y) = \int\limits_{X} \tilde{T}(x)^{\boldsymbol{\alpha}} \tilde{f}_0(x) \mathrm{d}\lambda(x). \tag{77}$$

We fix $1 \leq \ell \leq L$ and assume tensor spaces $\hat{X}^\ell$, X^ℓ such that a layer-based splitting can be employed to obtain integrals over X^ℓ of the form

$$\int\limits_{Y} y^{\boldsymbol{\alpha}} f(y) \mathrm{d}\lambda(y) = \sum_{\ell=1}^{L} \int\limits_{X^\ell} \tilde{T}(x)^{\boldsymbol{\alpha}} \tilde{f}_0(x) \mathrm{d}x. \tag{78}$$

Note that we neglect the remaining unbounded layer X^{L+1} since for moderate $|\alpha|$ and $\mathrm{vol}(\bigcup_{\ell=1}^{L} X^\ell)$ sufficiently large, the contribution to the considered moment does not have a significant influence on the overall approximation. Additionally, a rank-1 stable diffeomorphism $\Phi^\ell \colon \hat{X}^\ell \mapsto X^\ell$ is assumed for which there exist univariate functions $\Phi^\ell_{,j} \colon \hat{X}^\ell_j \to X^\ell$ with $\Phi^\ell_{,j} = (\Phi^\ell_{i,j})_{i=1}^d$ and $h_j \colon \hat{X}^\ell_j \to \mathbb{R}$ for every $j = 1, \ldots, d$, such that

$$\Phi^\ell(\hat{x}) = \prod_{j=1}^{d} \Phi^\ell_{,j}(\hat{x}_j) \quad \text{and} \quad |\mathcal{J}_{\Phi^\ell}|(\hat{x}) = \prod_{j=1}^{d} h_j(\hat{x}_j). \tag{79}$$

Moments under affine transport Let

$$H = [h_{ki}]_{k,i=1}^d = [h_1, h_2, \ldots, h_d] \in \mathbb{R}^{d,d}$$

be a symmetric positive definite matrix and $M = (M_i)_{i=1}^d \in \mathbb{R}^d$ such that the considered transport map takes the form

$$\tilde{T}(\cdot) = H \cdot + M. \tag{80}$$

With the multinomial coefficient for $j \in \mathbb{N}$, $\boldsymbol{\beta} \in \mathbb{N}_0^d$ for $j = |\boldsymbol{\beta}|$ given by

$$\binom{j}{\boldsymbol{\beta}} := \frac{j!}{\beta_1! \cdot \ldots \cdot \beta_d!},$$

the computation of moments corresponds to the multinomial theorem as seen in the next lemma.

Lemma 8 *Let $k \in \mathbb{N}$ with $1 \leq k \leq d$ and $\alpha_k \in \mathbb{N}_0$. It holds*

$$[H\Phi^\ell(\hat{x}) + M)]_k^{\alpha_k} = \sum_{j_k=0}^{\alpha_k} \sum_{|\boldsymbol{\beta}_k|=j_k} C_k^H[j_k, \alpha_k, \boldsymbol{\beta}_k] \times \prod_{j=1}^{d} \boldsymbol{\Phi}_j^{\boldsymbol{\beta}_k}(\hat{x}_j), \tag{81}$$

where the high-dimensional coefficient C_k^H is given by

$$C_k^H[j_k, \alpha_k, \boldsymbol{\beta}_k] := \binom{\alpha_k}{j_k} c_k^{\alpha_k - j_k} \binom{j_k}{\boldsymbol{\beta}_k} h_k^{\boldsymbol{\beta}_k}, \tag{82}$$

with $c_k := \sum_{i=1}^{d} h_{ki} M_i$ and

$$\boldsymbol{\Phi}_j^{\boldsymbol{\beta}_k} := [\Phi^\ell_{1,j}(\hat{x}_j), \ldots, \Phi^\ell_{d,j}(\hat{x}_j)]^{\boldsymbol{\beta}_k}. \tag{83}$$

Proof Note that

$$[H\Phi^\ell(\hat{x}) + M)]_k^{\alpha_k} = \sum_{j_k=0}^{\alpha_k} \binom{\alpha_k}{j_k} c_k^{\alpha_k - j_k} \times \left(\sum_{i=1}^{d} h_{ki} \prod_{j=1}^{d} \Phi^\ell_{ij}(\hat{x}_j) \right)^{j_k}.$$

The statement follows by the multinomial theorem since

$$\left(\sum_{i=1}^{d} h_{ki} \prod_{j=1}^{d} \Phi^\ell_{ij}(\hat{x}_j) \right)^{j_k} = \sum_{|\boldsymbol{\beta}_k|=j_k} \binom{j_k}{\boldsymbol{\beta}_k} \left(\prod_{i=1}^{d} h_{ki}^{(\boldsymbol{\beta}_k)_i} \right) \times \left(\prod_{j=1}^{d} \prod_{i=1}^{d} \Phi^\ell_{ij}(\hat{x}_j)^{(\boldsymbol{\beta}_k)_i} \right).$$

Generalizing Lemma 8 to multiindices $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ yields

$$[H\Phi^\ell(\hat{x}) + M)]^{\boldsymbol{\alpha}} = \sum_{\boldsymbol{j}=0}^{\boldsymbol{\alpha}} \sum_{(|\boldsymbol{\beta}_k|)_k=\boldsymbol{j}} \left(\prod_{k=1}^{d} C_k^H[j_k, \alpha_k, \boldsymbol{\beta}_k] \right) \times \prod_{j=1}^{d} \boldsymbol{\Phi}_j^{\sum_{k=1}^{d} \boldsymbol{\beta}_k}(\hat{x}_j), \tag{84}$$

where $\sum_{(|\boldsymbol{\beta}_k|)_k=\boldsymbol{j}} := \sum_{|\boldsymbol{\beta}_1|=j_1} \ldots \sum_{|\boldsymbol{\beta}_d|=j_d}$ is used.

Exploiting the layer-wise TT representation of $\hat{f}_\ell$ from (38) and using the rank-1 stable map (79), the high-dimensional integral over X^ℓ simplifies to

$$\begin{aligned}
&\int\limits_{X_\ell} \tilde{T}(x)^{\boldsymbol{\alpha}} \tilde{f}_0(x)\mathrm{d}\lambda(x) \\
&\quad = \sum_{\boldsymbol{j}=0}^{\boldsymbol{\alpha}} \sum_{(|\boldsymbol{\beta}_k|)_k=\boldsymbol{j}} \left(\prod_{k=1}^{d} C_k^H[j_k, \alpha_k, \boldsymbol{\beta}_k] \right) \\
&\qquad \times \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{r}_\ell} \prod_{i=1}^{d} \int\limits_{\hat{X}_i} \left[\hat{f}_{\ell,i}[k_{i-1}, k_i] \otimes \boldsymbol{\Phi}_i^{\sum_{k=1}^{d} \boldsymbol{\beta}_k} \otimes h_i \right](\hat{x}_i)\, \mathrm{d}\hat{x}_i .
\end{aligned} \tag{85}$$

Note that the right-hand side is composed via decoupled one dimensional integrals only. We point out that while the structure is simplified, the definition of $\boldsymbol{\Phi}_j$ in (83) a priori results in several integrals (indexed by $\sum_{k=1}^{d} \boldsymbol{\beta}_k$). These integrals, whose number depends on the cardinality of $\boldsymbol{\alpha}$, have to be computed. In several cases this simplifies further, e.g. when Φ^ℓ transforms the spherical coordinate system to Cartesian coordinates.

Moment computation using spherical coordinates In the special case that Φ^ℓ is the multivariate polar transformation of "Appendix B.1", the number of distinct computation of integrals from (85) reduces significantly. Recall that $\hat{x}_1 = \rho$, $\hat{x}_{2:d} = \boldsymbol{\theta} = (\theta_0, \ldots, \theta_{d-2})$ and let $\beta_i^k := (\boldsymbol{\beta}_k)_i$ be the i-th entry of $\boldsymbol{\beta}_k$. We find that

$$\boldsymbol{\Phi}_1^{\sum_{k=1}^{d} \boldsymbol{\beta}_k}(\rho) = \rho^{|\boldsymbol{j}|}, \tag{86}$$

$$\boldsymbol{\Phi}_2^{\sum_{k=1}^{d} \boldsymbol{\beta}_k}(\theta_0) = \cos^{\left(\sum_{k=1}^{d} \beta_1^k\right)}(\theta_0) \sin^{\left(\sum_{k=1}^{d} \beta_2^k\right)}(\theta_0), \tag{87}$$

$$\boldsymbol{\Phi}_{i+2}^{\sum_{k=1}^{d} \boldsymbol{\beta}_k}(\theta_i) = \sin^{\left(\sum_{l=1}^{i} \sum_{k=1}^{d} \beta_l^k\right)}(\theta_i) \cos^{\left(\sum_{k=1}^{d} \beta_{i+1}^k\right)}(\theta_i). \tag{88}$$

for $1 \le i \le d-2$.

The exponential complexity due to the indexing by $\sum_{k=1}^{d} \boldsymbol{\beta}_k$ reduces to linear complexity in $|\boldsymbol{\alpha}|$. More precisely, the amount of exponents in (86) - (88) is linear in the dimensions since the sums only depend on $|\boldsymbol{\alpha}|$, leading to $\mathcal{O}(|\boldsymbol{\alpha}|d)$ different integrals that may be precomputed for each tuple (k_{i-1}, k_i). This exponential complexity in the rank vanishes in the presence of an approximation basis associated with each coordinate dimension as defined in Sect. 3.

6.2.2 Computation of marginals

In probability theory and statistics, marginal distributions and especially marginal probability density functions provide insights into an underlying joint density by means of lower dimensional functions that can be visualized. The computation of marginal densities is a frequent problem encountered e.g. in parameter estimation and when using sampling techniques since histograms and corner plots provide easy access to (in general high-dimensional) integral quantities.

In contrast to the Markov Chain Monte Carlo algorithm, the previously presented method of a layer based surrogate for the Lebesgue density function $f : Y = \mathbb{R}^d \to \mathbb{R}$ allows for a functional representation and approximation of marginal densities without additional evaluations of f.

For $y \in Y$ and $i = 1, \ldots, d$, define

$$y_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots y_d)$$

as the marginalized variable where the i-th component is left out and $f(y_{-i}, y_i) := f(y)$. Then, for given $i = 1, \ldots, d$, the i-th marginal density reads

$$\mathrm{d}f_i(y_i) := \int_{\mathbb{R}^{d-1}} f(y_{-i}, y_i)\mathrm{d}\lambda(y_{-i}). \tag{89}$$

Computing this high-dimensional integral by quadrature or sampling is usually infeasible and the transport map approach as given by (3) fails since the map $T : X \to Y$ cannot be used directly in (89). Alternatively, we can represent $\mathrm{d}f_i : \mathbb{R} \to \mathbb{R}$ in a given orthonormal basis $\{\varphi_j\}_{j=1}^{N_\varphi}$ and consider

$$\mathrm{d}f_i(y_i) = \sum_{j=1}^{N_\varphi} \beta_j \varphi_j(y_i), \tag{90}$$

where β_j, $j = 1, \ldots, N_\varphi$ denotes the $L^2(\mathbb{R})$ projection coefficient

$$\beta_j := \int_{\mathbb{R}} \varphi_j(y_i)\mathrm{d}f_i(y_i)\mathrm{d}\lambda(y_i). \tag{91}$$

With this, the marginalization can be carried out similarly to the computations in Sect. 6.2.1 by replacing f with $\tilde{f}^{TT}$.

A convenient basis is given by monomials since (91) then simplifies to

$$\beta_j = \int_{\mathbb{R}^d} y_i^j f(y)\mathrm{d}\lambda(y). \tag{92}$$

This is the moment corresponding to the multiindex $\alpha = (\alpha_k)_{k=1}^d \in \mathbb{N}^d$ with $\alpha_k = \delta_{k,j}$. Alternatively, indicator functions may be considered in the spirit of histograms.

6.3 Stochastic Galerkin finite element method

Consider the case when the target density $f : Y \to \mathbb{R}_+$ with $Y = \mathbb{R}^d$ should be used in a stochastic Galerkin Finite Element method (SGFEM, see e.g. Eigel et al. 2017, 2020). Note that f and its approximation $\tilde{f}^{\mathrm{TT}}$ in general is not of product type. If there exists $c > 0$ and a norm $\|\cdot\|_Y$ on Y such that

$$\int_Y e^{c\|y\|_Y} \tilde{f}^{\mathrm{TT}}(y)\mathrm{d}\lambda(y) < \infty$$

then there exists a complete orthonormal basis of polynomials (Dunkl and Xu 2014). The moment computation from Sect. 6.2.1 can then be used to obtain the polynomials numerically based on a Gram-Schmidt orthonormalization process. This approach is out of the scope of the present work and will be discussed elsewhere.

6.4 Samples from the posterior

In case one is interested to draw samples from f, we propose using standard sample generation schemes like rejection sampling or MCMC on the approximated perturbed prior $\tilde{f}_0^{\mathrm{TT}}$. Such a scheme can be implemented efficiently based on the presented surrogate since fast point evaluations are possible. Once samples are obtained from $\tilde{f}_0^{\mathrm{TT}}$, mapping them through $\tilde{T}$ yields samples from $\tilde{f}^{\mathrm{TT}}$. Note that we cannot draw samples in the reference space $\hat{X}$ and map them via the transformations $\{\Phi_\ell\}$ to obtain samples from $\tilde{f}_0^{\mathrm{TT}}$ since the transformation is not a transport.

7 Numerical validation and applications

This section is devoted to a numerical validation of the proposed Algorithm 1 using various types of transformations T while applying the scheme to practical problems. We focus on three example settings. The first consists of an artificial Gaussian posterior density which could be translated to a linear forward model and Gaussian prior assumptions in the Bayesian setting. Second, we study the approximation under inexact transport and conclude as a third setting with an actual Bayesian inversion application governed by the log-normal Darcy flow problem. All considered examples satisfy Assumption 1 and are in fact analytic. This is obvious for the Gaussian and the transformed Gaussian in Sects. 7.1 and 7.2. In the log-normal case we refer to Babuška et al. (2007) and Hoang and Schwab (2014).

Remark 7 When dealing with probability density functions, a key requirement for a valid approximation is non-negativity. Some comments are in order since this property cannot be guaranteed when applying usual concepts to obtain tensor train approximiations. Nevertheless, we would like to point out that we did not experience any numerical difficulties when computing the presented experiments and the following two approaches have not been implemented.

Note that the chosen function space of bounded functions $\mathcal{V}(\hat{\mathcal{X}}, \underline{c}, \overline{c})$ and the model class $\mathcal{M} = \mathcal{M}_{\boldsymbol{r}}(\hat{\mathcal{X}}, \underline{c}, \overline{c}) \subset \mathcal{V}(\hat{\mathcal{X}}, \underline{c}, \overline{c})$ implies the non-negativity of our approximation in (38). Numerically, those spaces are not immediately accessible in the context of tensor regression. We mention two possible methods that can be implemented to handle the numerical optimization in $\mathcal{M}$:

- *Constrained optimization*: The boundedness in the model class can be translated into constraints on the tensor components of the tensor train model.
- *Square root trick*: One can obtain a tensor train approximation $\sqrt{\hat{g}}$ of $\sqrt{\hat{f}_0}$, as *e.g.* applied in Cui and Dolgov (2021). Subsequently squaring the result yields the desired non-negative tensor train approximation $\hat{g}$ of $\hat{f}_0$. However, taking the square of an extended tensor train increases the rank and the number of basis elements in the representation. Therefore, careful adjustments have to be made to ensure that the analysis of Sect. 4 is still valid. Due to the upper bound of $\hat{f}_0$ in Assumption 1 and the fact that for any $\hat{g} \in \mathcal{M}$ it holds for some $C > 0$ that

$$\|\hat{f}_0 - \hat{g}\|^2_{\mathcal{V}(\hat{X})} \leq C\|\sqrt{\hat{f}_0} - \sqrt{\hat{g}}\|^2_{\mathcal{V}(\hat{X})}.$$

Then, the error analysis can be carried out as presented with adapted model classes that take care of the increasing rank.

7.1 Validation experiment 1: Gaussian density

In this experiment we confirm the theoretical results of Sect. 4 and verify the numerical algorithm. Even though the examined approximation of a Gaussian density is not a challenging task for the proposed algorithm, it can be seen as the most fundamental illustration, revealing the possible rank-1 structure of the perturbed prior under optimal transport.

We consider the posterior density determined by a Gaussian density with covariance matrix $\Sigma \in \mathbb{R}^{d,d}$ and mean $\mu \in \mathbb{R}^d$ given by

$$\frac{\mathrm{d}\pi}{\mathrm{d}\lambda}(x) = f(x) = C\exp\left(-\frac{1}{2}\|x-\mu\|^2_{\Sigma^{-1}}\right), \qquad (93)$$

where $C = (2\pi)^{-d/2}\det\Sigma^{-1/2}$ is the normalizing factor of the multivariate Gaussian. We set the covariance operator such that the Gaussian density belongs to uncorrelated random variables, i.e. Σ exhibits a diagonal structure, and it holds for some $0 < \sigma \ll 1$ that $\Sigma = \sigma^2 I$. This Gaussian setting has several benefits when used as validation. On the one hand, we have explicit access to the quantities that are usually of interest in Bayesian inference like the mean, covariance, normalization constant and marginals. On the other hand, the optimal transport to a standard normal density

$$f_0(x) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\|x\|^2\right) \tag{94}$$

is given by an affine linear function, defined via mean μ and covariance Σ as proposed in Remark 2. We subsequently employ the multivariate polar transformation from Example B.1 and expect a rank-1 structure in the reconstruction of the local approximations of the (perturbed) prior.

The remainder of this section considers different choices of $\sigma \in \mathbb{R}$ and $d \in \mathbb{N}$ and highlights the stability of our method under decreasing variance (i.e. with higher density concentration) and increasing dimension. The approximations are compared with their exact counterparts. More specifically, the error of the normalization constant is observed, namely

$$\mathrm{err}_Z := |1 - Z_h|, \tag{95}$$

and the relative error of the mean and covariance in the Euclidean and Frobenius norms $|\cdot|_2$ and $|\cdot|_\mathrm{F}$,

$$\mathrm{err}_\mu := |\mu - \mu_h|_2 |\mu|_2^{-1}, \quad \mathrm{err}_\Sigma := |\Sigma - \Sigma_h|_\mathrm{F} |\Sigma|_\mathrm{F}^{-1}. \tag{96}$$

Additionally the convergence in terms of the Kullback–Leibler divergence (43) and the Hellinger distance (42) are analyzed. Computing the Kullback–Leibler divergence is accomplished by Monte Carlo samples $(x_i)_{i=1}^{N_\mathrm{KL}}$ of the posterior (i.e. in this case the multivariate Gaussian posterior) to compute the empirical approximation

$$\begin{aligned} \mathrm{d}_\mathrm{KL}(\pi, \pi_h) &= \int_{\mathbb{R}^d} \log\left(\frac{f(x)}{f_h(x)}\right) f(x) \mathrm{d}\lambda(x) \\ &\approx \frac{1}{N_\mathrm{KL}} \sum_{i=1}^{N_\mathrm{KL}} \log\left(\frac{f(x_i)}{f_h(x_i)}\right). \end{aligned} \tag{97}$$

The index h generically denotes the employed approximation (38). A similar computation can be carried out for the Hellinger distance such that

$$\mathrm{d}_\mathrm{Hell}^2(\pi, \pi_h) \approx \frac{1}{2N_\mathrm{Hell}} \sum_{i=1}^{N_\mathrm{Hell}} \left(\sqrt{\frac{f_h(x_i)}{f(x_i)}} - 1\right)^2.$$

In the numerical experiments, the convergence of these error measures is depicted with respect to the number of calls to the forward model (i.e. the Gaussian posterior density), the discretization of the radial component $\rho \in [0, \infty)$ in the polar coordinate system and the number of samples on each layer X^ℓ, $\ell = 1, \ldots, L$, for fixed $L \in \mathbb{N}$.

In Table 1 err_Z is depicted for different choices of σ and d. The experiment comprises radial discretizations $0 = \rho_0 < \rho_1 < \ldots < \rho_L = 10$ with $L = 19$ equidistantly chosen layers and 1000 samples of f_0 on each resulting subdomain X^ℓ. The generated basis (67) contains polynomials of maximal degree 7 in ρ_ℓ, $\ell = 0, \ldots, L$, and constant functions in every angular direction. The choice of constant functions relies on the assumption that the perturbed prior that has to be approximated corresponds to the polar transformation of (94), which is a function in ρ only. Additional numerical test show that even much fewer samples and a larger basis lead to the assumed rank-1 structure. It can be observed that the approximation quality of Z is invariant under the choice of σ and fairly robust with the dimension d, which is expected since the transformation is exact and the function to reconstruct is a rank-1 object.

In Fig. 2 we show the convergence behavior of the developed method in terms of the number of calls of the posterior density f. Here, the presented low-rank surrogate is constructed on an increasing number of layers. Taking 100 samples on each layer starting from only one, i.e. from $L = 1$ up to $L = 50$ on [0, 10], the VMC tensor reconstruction is carried out. This increase in the number of layers translates directly to an increase in posterior calls. Figure 2 (left) shows the convergence in terms of the Kullback–Leibler divergence and the Hellinger distance. In accordance with Theorem 1, the Kullback–Leibler divergence converges faster. For the computation, $N_\mathrm{KL} = N_\mathrm{Hell} = 1000$ samples of the posterior f are drawn and we show the mean and 90% quantile of 30 repeated evaluations, where each run includes the recomputation of the VMC approximation. For a quantitive comparison in terms of posterior statistics, in Fig. 2 (right) we show the relative error err_Σ for the presented algorithm (TT, blue), a reference Monte Carlo sampling using samples from the posterior f (MC, green) and a Hamiltonian Monte Carlo sampler (HMC, red) as it comes out of the box with `Matlab`. HMC is a sensible choice as a reference method since the tuning phase of HMC involves the computation of the MAP and Hessian to improve the exploration phase of the sample chains, which is in some sense similar to employed affine linear transport in our algorithm. The HMC tuning phase usually took more than 1000 posterior evaluations, which are not counted in Fig. 2. As a result, we observe fast convergence of our method which takes advantage of the regularity of f in comparison to the sampling techniques, which are inherently limited by the $1/\sqrt{N}$ Monte Carlo convergence. HMC performs as well as MC, which comes from the fact that the estimated MAP and Hessian push the sampler into the posterior region. To emphasize the stability of the results, we repeated the VMC approximation 30 times and again show the mean and 90% quantile. For the sampling methods, we repeated HMC and MC 1000 times each and show the mean of the relative covariance approximation only. Concerning the stagnation of err_Σ we suspect a precision problem in the computation, which is confirmed by the small variance.

Table 1 Numerical approximation of Z in the Gaussian example

Dimension	$\sigma^2 = 10^{-2}$	$\sigma^2 = 10^{-4}$	$\sigma^2 = 10^{-6}$	$\sigma^2 = 10^{-8}$
2	$5.24 \cdot 10^{-11}$	$1.09 \cdot 10^{-10}$	$2.8 \cdot 10^{-11}$	$9.3 \cdot 10^{-11}$
4	$2.21 \cdot 10^{-10}$	$4.57 \cdot 10^{-10}$	$5.48 \cdot 10^{-10}$	$3.4 \cdot 10^{-10}$
6	$5.01 \cdot 10^{-11}$	$9.5 \cdot 10^{-11}$	$7.49 \cdot 10^{-11}$	$6.19 \cdot 10^{-10}$
8	$1.48 \cdot 10^{-11}$	$8.21 \cdot 10^{-10}$	$2.99 \cdot 10^{-10}$	$2.1 \cdot 10^{-10}$
10	$2.91 \cdot 10^{-9}$	$9.61 \cdot 10^{-10}$	$4.43 \cdot 10^{-11}$	$2.46 \cdot 10^{-9}$

Error of the normalization constant computed via a TT surrogate to $Z = 1$

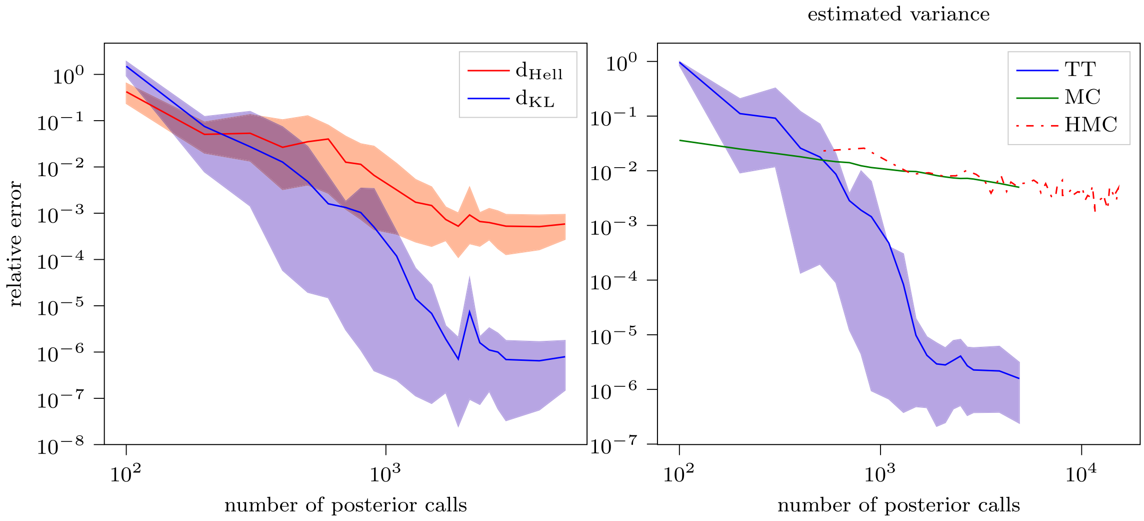


Fig. 2 Gaussian density example with $d = 10$, mean $\mu = \mathbf{1}$ and noise level $\sigma = 10^{-7}$. Tensor reconstructions are repeated with 30 random sample sets to show quantile range from 5–95% (pastel) to the mean (bold). Hellinger distance and Kullback–Leibler divergence are shown (left). In (right) the relative covariance error err_Σ is shown for the tensor reconstruction algorithm (blue) for a plain MC sampling of the posterior (green) and a tuned Hamiltonian MC sampler (red). (Color figure online)

Nevertheless, an approximation of around four magnitudes smaller than MC and HMC for the covariance is achieved.

7.2 Validation experiment 2: Perturbation of exact transport

In the following experiment we consider a so-called "banana example" as posterior density, see e.g. Marzouk et al. (2016). Let f_0 be the density of a standard normal Gaussian measure and let T_Σ be the affine transport of $\mathcal{N}(0, I)$ to the Gaussian measure $\mathcal{N}(0, \Sigma)$. Furthermore, set

$$T_2(x) = \begin{pmatrix} x_1 \\ x_2 - (x_1^2 + 1) \end{pmatrix}. \tag{98}$$

The exact transport T from $\mathcal{N}(0, I)$ to the curved and concentrated banana distribution with density f is then given by

$$T(x) = T_2 \circ T_\Sigma(x), \quad \Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}. \tag{99}$$

Note that the employed density can be transformed into a Gaussian using a quadratic transport function. For this experiment, we employ transport maps $\tilde{T}$ of varying accuracy for the pull-back of the posterior density to a standard Gaussian. In particular, we use an approximation $\tilde{T}_1$ (obtained with Baptista et al. 2015-2018) of the optimal affine transport T_1 and the quadratic transport T to build an approximation $\tilde{T}$ given as convex combination

$$\tilde{T}(x) = (1 - t)\, \tilde{T}_1(x) + t\, T(x), \quad t \in [0, 1]. \tag{100}$$

For $t = 1$, the transport map is optimal since it generates the desired reference density. For $0 \leq t < 1$, a perturbed prior density is obtained with perturbation strength determined by t. The impact of the perturbed transport is visualized in Fig. 3.

It can be observed that the transformed perturbed prior is not of rank-1 when the transformation is inexact. Furthermore, the difference between the target prior and the perturbed prior is imminent, which implies that e.g. a Laplace approximation to the considered banana density would neglect possible important features of the distribution.

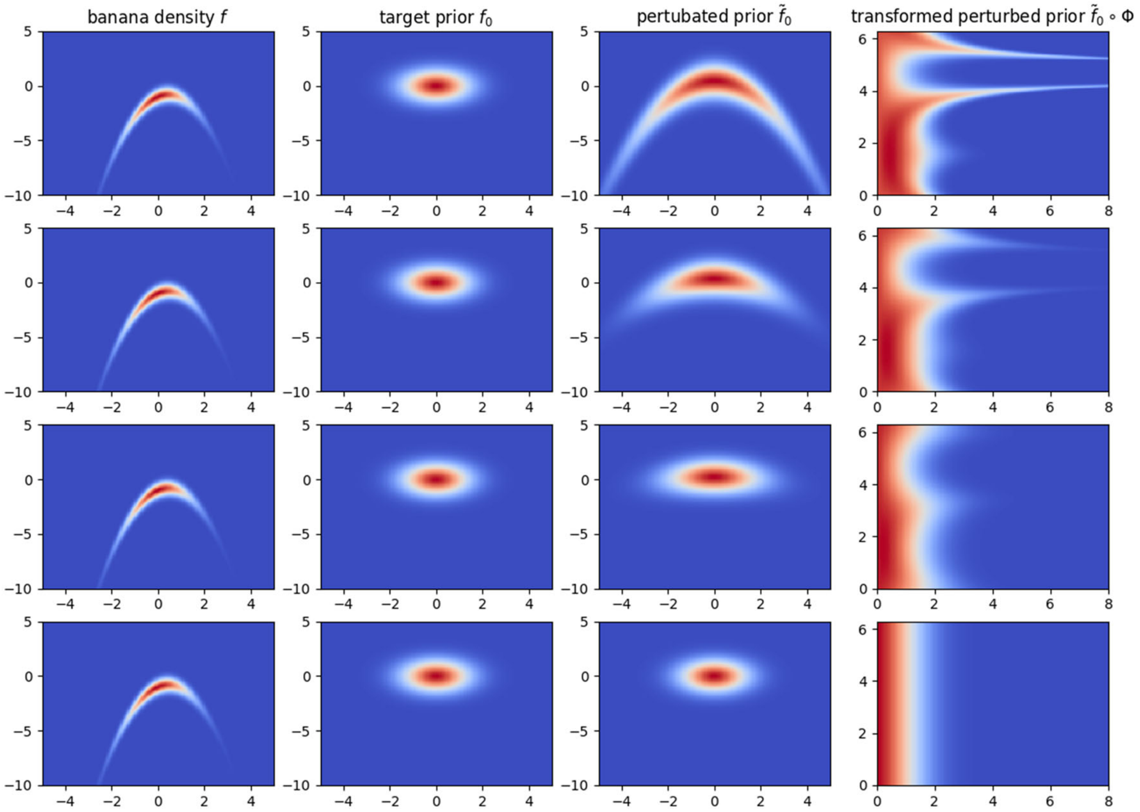


Fig. 3 Illustration of the effect of different transports in (100) for $t = 0, 0.25, 0.5, 1$. (top to bottom)

In Fig. 4 we illustrate the impact of an inexact transport on the approximation results in terms of err_μ and err_Σ. For the considered target density, mean and covariance are known analytically and hence no reference sampling has to be carried out. We additionally employ an MCMC sampling to show the improvement due to the additional low-rank reconstruction. For the optimal transport map, one observes that the surrogate reconstruction reduces to the approximation of a rank-1 Gaussian density, which can be done efficiently with few evaluations of f. If the transport is only linear and inaccurate, results comparable to MCMC are achieved. For a (somewhat) more accurate transport, the low-rank reconstruction leads to (already) drastically improved estimates.

7.3 Bayesian inversion with log-normal Darcy model

Random partial differential equations (PDEs), i.e. PDEs with correlated random data, play an important role in the popular field of Uncertainty Quantification (UQ). A prominent benchmark example is the (stationary) ground water flow model, also called the Darcy problem, as e.g. examined in Eigel et al. (2014, 2017, 2020). This linear second order PDE model on some domain $D \subset \mathbb{R}^d$, $d = 1, 2, 3$ is determined by a forcing term $g \in L^2(D)$ and the random quantity $a(y) \in L^\infty(D)$, which for almost every $y \in Y$ models a conductivity or permeability coefficient. The physical system is described by

$$-\operatorname{div}(a(y)\nabla q(y)) = g \quad \text{in } D, \quad q(y)|_{\partial D} = 0, \tag{101}$$

and the solution $q(y) \in V := H_0^1(D)$ corresponds to the system response. Pointwise solvability of (101) for almost every $y \in Y$ is guaranteed by a Lax–Milgram argument. For details we refer to Schwab and Gittelson (2011).

For the applications in this article, we employ a truncated log-normal coefficient field

$$a(y) = \exp\left(\sum_{k=1}^{d} a_k y_k\right), \tag{102}$$

for some fixed $(a_k)_{k=1}^d$ with $a_k \in L^2(D)$ denoting planar Fourier cosine modes and the image of some random variable with law $\mathcal{N}(0, I)$ denoted by $y = (y_k)_{k=1}^d \in Y$. A detailed description and an adaptive Galerkin approach to solve this problem can be found in Eigel et al. (2020).

For the inverse problem, the observation operator is modeled by 144 equidistantly distributed observations in $D = [0, 1]^2$ of the solution $q(y^*) \in H_0^1(D)$ for some $y^* \in Y = \mathbb{R}^d$, which is drawn from a standard normal distribution. Additionally, the observations are perturbed by a centered

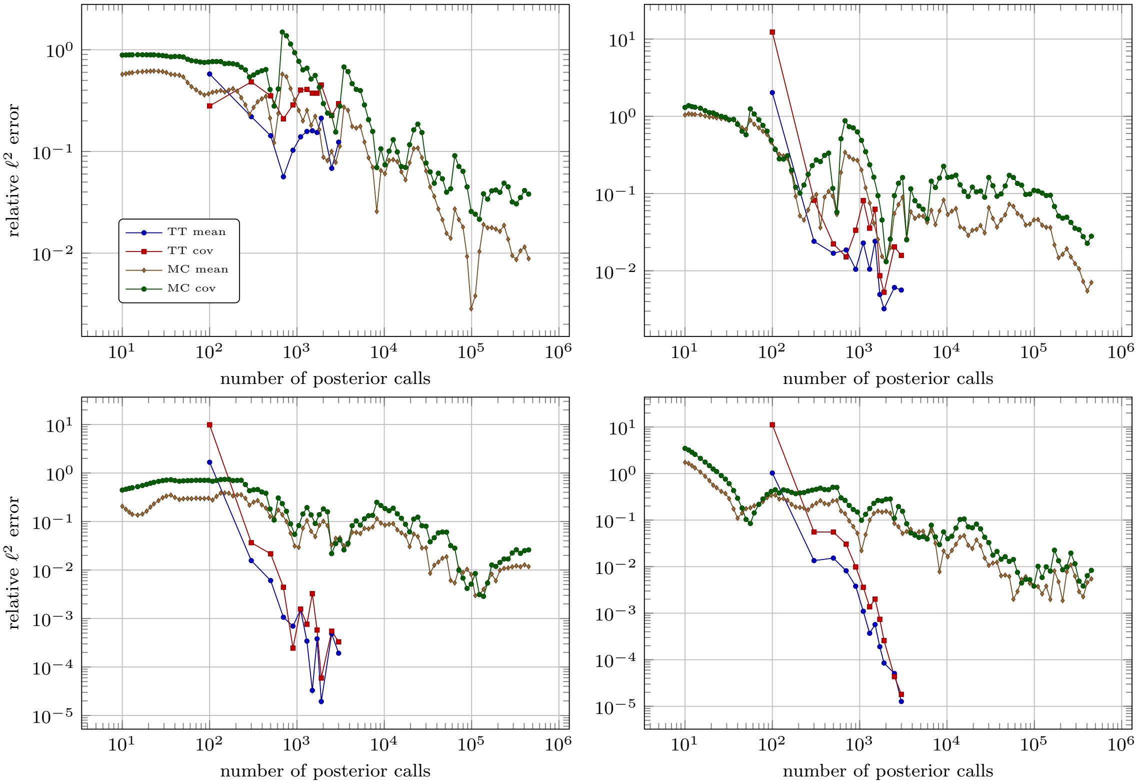


Fig. 4 Convex combination of affine and quadratic transport for the banana posterior. Affine linear map ($t = 0$ top left), transport with $t = 0.25$ (top right), $t = 0.5$ (bottom left) and exact quadratic map ($t = 1$, bottom right). Error quantities err_μ and err_Σ for the employed tensor train surrogate and a MCMC approximation in terms of the number of calls to the posterior function. The surrogate is reconstructed from 100 samples per layer, yielding a tensor with radial basis up to polynomial degree 9 and Fourier modes up to degree 20

Gaussian noise with covariance σI with $\sigma = 10^{-7}$. We consider the Bayesian posterior density (73) and set

$$f(y) = Z^{-1} \mathrm{d}\pi_\delta(y) \mathrm{d}\pi_0(y) \tag{103}$$

as the target density of the measure π on Y according to (1). This target density is now approximated by the developed algorithm.

To obtain the desired relative error quantities as above, we employ reference computations that involve adaptive quadrature for the two dimensional example in Fig. 5 and Markov Chain Monte Carlo integration with 10^6 steps of the chain and a burn-in time of 1000 samples for the experiment in Fig. 6. We point out that the MCMC chains are started in an area defined by the numerically obtained MAP point used for the employed transport map. This is in some sense similar to the HMC algorithm considered in the previous example. For the reconstruction algorithm, an affine linear transport is estimated by Hessian information of the log-likelihood and on every layer we employ 100 samples. As above, this corresponds to taking initially one layer resulting in 100 posterior calls. We increase the number of layers up to $L = 25$. The respective relative errors are displayed in Fig. 6.

The stagnation of the graphs in Fig. 5 is on the one hand governed by the observation noise and on the other hand explicable by a non-optimal reference solution since the TT approximation yields results equivalent to an adaptive quadrature when taking $L = 5$ layers of refinement and thus a total of 500 samples.

The improvement of the mean and covariance estimate by the low-rank approach can already be observed for a low sample number. We note that the Monte Carlo estimate does not allow for an adequate computation of the empirical covariance, which therefore is left out of the comparison.

8 Conclusion

We propose a novel approach to approximate probability densities with high accuracy, combining the notion of *transport maps* and *low-rank functional representations* of auxiliary (perturbed) reference densities. Based on a suitable class of transformations, an approximation with respect to a finite tensorized basis can be carried out in extended hierarchical tensor formats. This yields a compressed representation for an efficient computation of statistical quantities of inter-

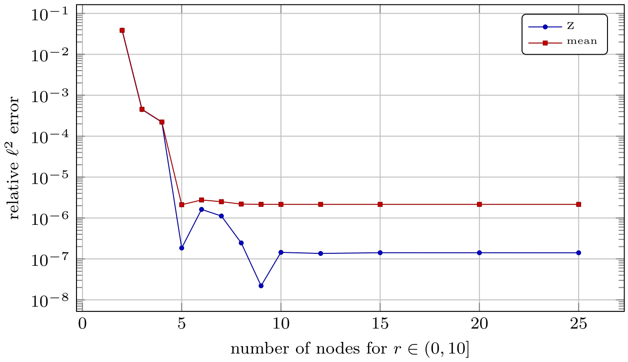


Fig. 5 Comparison of the computed reference and the low-rank surrogate of (1) normalization constant (err_Z), and (2) mean (err_μ). For the Darcy setting with $d = 2$ we observe 144 nodes in the physical domain. The measurements are perturbed by Gaussian noise with deviation $\eta = 1e - 7$. We employ an adaptive quadrature in the two dimensional space to obtain the reference quantities. The stagnation of the graphs are due to non-optimal reference solutions. More precisely, the TT approximation yields equivalent results to adaptive quadrature when taking 5 nodes of refinement

est (e.g. moments or marginals) in a sampling free manner. In this work, the multivariate polar transformation is used as a particular rank-1 stable transformation. The method requires point evaluations of the perturbed reference density (up to a multiplicative constant). The approach can hence be applied to not normalized posterior densities in the context of Bayesian inversion. An a priori convergence analysis is examined and we illustrate the performance of the method via an inverse problem with a log-normal Darcy forward model. A comparison with classical MCMC illustrates the superior convergence in terms of the moment accuracy relative to the number of posterior evaluations. Future research will be concerned with

- application: usage of the approximated densities for subsequent computations e.g. with SGFEM,
- analysis: Given a function $\tilde{f}_0$ it has to be examined which rank-1 stable transformations Φ lead to a low-rank function $\tilde{f}_0 \circ \Phi$.

Acknowledgements The authors would like to thank Reinhold Schneider for fruitful discussions. Furthermore, the authors would like to thank the anonymous reviewers for comments that helped to improve the manuscript.

Author Contributions All authors contributed equally to the manuscript.

Data Availability All data is stored in a data repository at BitBucket.

Code Availability The code is stored in a data repository at BitBucket: https://bitbucket.org/aleadev/alea.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

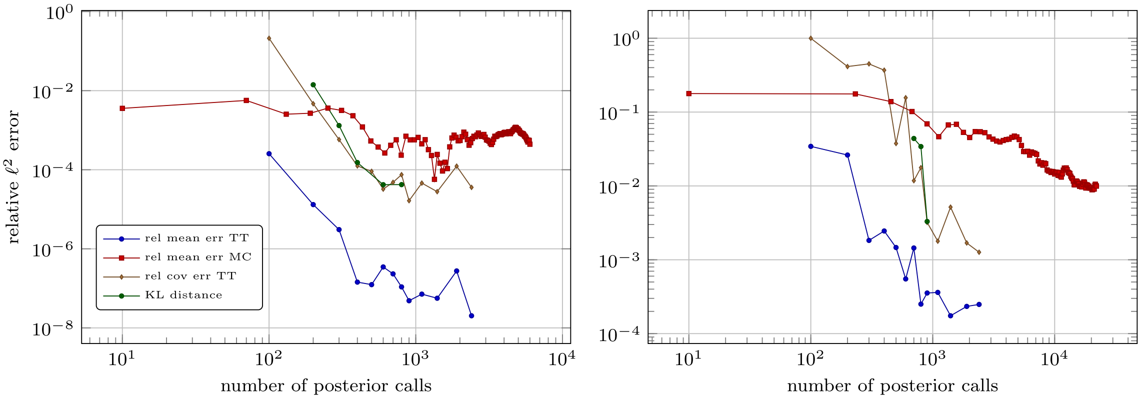


Fig. 6 Darcy example with $d = 2$ (left) and $d = 10$ (right). Comparison of an MCMC method and the low-rank surrogate for computing the mean error (err_μ) with respect to the number of calls to the solution of the forward problem. The reference mean is computed with 10^6 MCMC samples. Additionally, the KL divergence is shown, which is computed using empirical integration

A Typical transport maps

A.1 Affine transport

In Schillings and Schwab (2016) and Schillings et al. (2020) the authors employ an affine linear preconditioning for acceleration of MCMC or sparse-grid integration in the context of highly informative and concentrated Bayesian posterior densities, using a s.p.d. matrix $H \in \mathbb{R}^{d,d}$ and $M \in \mathbb{R}^d$. In the mentioned articles, up to a multiplicative constant, H corresponds to the inverse square root of the Hessian at the MAP (maximum a posteriori probability) M, i.e. the location of the local optimum of an Laplace approximation of the posterior density. This rather simple construction, under the assumption of an unimodal density, leads to stable numerical algorithms for the computation of quantities of interest as the posterior mass concentrates. When considering the push-forward of a reference density f_0 to a target density f this concept coincides with an affine transport

$$y = T(x) = Hx + M, \quad x \in X.$$

In the transport setting H and M may be computed for instance via some minimization of the Kullback–Leibler divergence as in El Moselhy and Marzouk (2012). Note that H and M do not necessarily have to be the inverse square root of the Hessian or the MAP. Figure 7 illustrates the concept of an affine transport.

A.2 Quadratic transport

A more general class of polynomial transport exhibits the form

$$T(x) = \frac{1}{2} x : \boldsymbol{A} : x + Hx + M, \quad x \in X, \tag{104}$$

with $\boldsymbol{A} = (\boldsymbol{A}_{ijk})_{i,j,k=1}^d \in \mathbb{R}^{d,d,d}$, $H \in \mathbb{R}^{d,d}$, $M \in \mathbb{R}^d$ and $(x : \boldsymbol{A} : x)_k = \sum_{i,j=1}^d x_i A_{ijk} x_j$. Such a quadratic transport may be used for simple nonlinear transformations as depicted in Fig. 8.

A.3 More general transport maps

The parametrization of transport maps can be chosen quite liberally as long as certain criteria are satisfied, which are either directly imposed in the ansatz space $\mathcal{T}$ of the maps or added as constraints during optimization. In particular, the approximate transport map has to be invertible, which can be ensured by requiring a positive Jacobian. A commonly used measure for transport optimization is the Kullback–Leibler divergence[4] leading to the optimization problem

$$\min_{T \in \mathcal{T}} \mathrm{d}_{\mathrm{KL}}(Y; T\pi_0, \pi) \quad \text{s.t.} \quad \det \mathcal{J}_T > 0 \qquad \pi\text{-a.e.} \tag{105}$$

Several suggestions regarding simplifications and special choices of function spaces $\mathcal{T}$ such as smooth triangular maps based on higher-order polynomials or radial basis functions can for instance be found in the review article (El Moselhy and Marzouk 2012). An interesting idea is to subdivide the task into the iterative computation of simple correction maps which are then composed as proposed in Brennan et al. (2020). We again emphasize that while an accurate transport map is desirable, any approximation of such a map can in principle be used with the proposed method. In fact one can decide whether it is beneficial to spend more effort on the approximation of the perturbed density or on a better representation of the transport.

B Coordinate transformations and example scenarios

Balancing the complexity of a transport map $\tilde{T}$ and the approximation of the perturbed transformed prior can be challenging but allows for some flexibility. Especially the change of variables induced by a sensible choice of the transformation Φ (7) can lead to simplified structures. In the following we list some scenarios in which a coordinate transform leads to advantageous represntations.

B.1 Gaussian perturbed prior density

Let f_0 be defined on $\mathbb{R}^d$ and $\tilde{T}$ maps f to $\tilde{f}_0$ which, in some sense, is close to a standard Gaussian density. In this case, Φ from (7) may be chosen as the d-dimensional spherical transformation. This shifts the exponential decay of $\tilde{f}_0$ to the one dimensional radial parameter. The accuracy of an approximation can then be improved easily by additional h-refinements, as described in Sect. 2.3. For the d-dimensional spherical coordinate system, a simple layer layout is given in terms of hyperspherical shells. In particular, for $\ell = 1, \ldots, L+1 < \infty$, with $0 = \rho_1 < \rho_2 < \ldots < \rho_{L+1} < \rho_{L+2} = \infty$, let

$$\hat{X}^\ell := [\rho_\ell, \rho_{\ell+1}] \times [0, 2\pi] \times \bigtimes_{i=2}^{d-2} [0, \pi],$$

$$X^\ell := B_{\rho_{\ell+1}}(0) \setminus B_{\rho_\ell}(0) \subset X,$$

i.e. $\hat{X}^\ell$ and X^ℓ denote the corresponding adopted (transformed) and the original parameter space, respectively. Then,

[4] Although in machine learning Wasserstein or Sinkhorn distances have become very popular when so-called normalizing flows are computed.

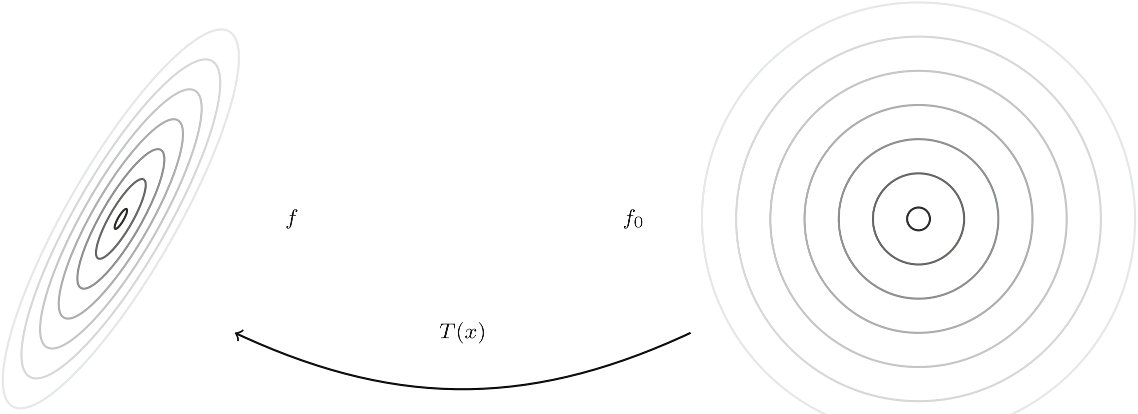
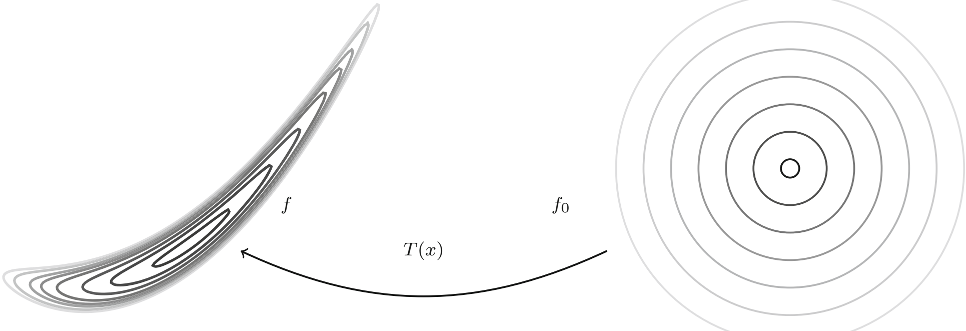


Fig. 7 Illustration of affine transport: translation, rotation and rescaling

Fig. 8 Illustration of quadratic transport: affine properties and bending



for $\hat{x} = (\rho, \theta_0, \boldsymbol{\theta}) \in \hat{X}$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{d-2})$, the d-dimensional spherical transformation $\Phi^\ell \colon \hat{X}^\ell \to X^\ell$ reads

$$\Phi^\ell(\hat{x}) = \rho \begin{bmatrix} \cos\theta_0 \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{d-3} \sin\theta_{d-2} \\ \sin\theta_0 \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{d-3} \sin\theta_{d-2} \\ \cos\theta_1 \sin\theta_2 \cdots \sin\theta_{d-3} \sin\theta_{d-2} \\ \cos\theta_2 \cdots \sin\theta_{d-3} \sin\theta_{d-2} \\ \vdots \\ \cos\theta_{d-3} \sin\theta_{d-2} \\ \cos\theta_{d-2} \end{bmatrix}.$$

Moreover, the Jacobian is given by

$$\det \mathcal{J}_{\Phi^\ell}(\rho, \theta_0, \boldsymbol{\theta}) = \rho^{d-1} \prod_{i=1}^{d-2} \sin^i \theta_i .$$

The structure of the transformation and its jacobian leads to the following result.

Proposition 1 *The multivariate spherical coordinate transformation is rank-1 stable.*

B.2 Mixed bounded and unbounded domains

Let $\hat{I} := \bigtimes_{i=1}^n [\hat{a}_i, \hat{b}_i]$ and $I := \bigtimes_{i=1}^n [a_i, b_i]$ with $-\infty < \hat{a}_i < \hat{b}_i < \infty$, $-\infty < a_i < b_i < \infty$ for $i = 1, \ldots, n$ and let a diffeomorphism $\Phi^1 \colon \hat{I} \to I$ with $I = \Phi^1(\hat{I})$ of the form

$$\Phi^1(\hat{x}^1) = [\Phi^1_1(\hat{x}^1_1), \ldots, \Phi^1_n(\hat{x}^1_n)]^T. \tag{106}$$

In applications, $\Phi^1_i \colon [\hat{a}_i, \hat{b}_i] \to [a_i, b_i]$ might be chosen as a "blow up function" as illustrated in Fig. 9 or as identity.

Then the *generalized hypercylindrical transformation* $\Phi \colon \hat{I} \times \mathbb{R}^d \to I \times \mathbb{R}^d$ is given as

$$\Phi(\hat{x}^1, \hat{x}^2) = [\Phi^1(\hat{x}^1), \Phi^2(\hat{x}^2)],$$

where $\Phi^2 \colon \mathbb{R}^d \to \mathbb{R}^d$ denotes the d-dimensional spherical transformation from "Appendix B.1". This construction leads to the following proposition.

Proposition 2 *The generalized hypercylindrical coordinate transformation is rank-1 stable.*

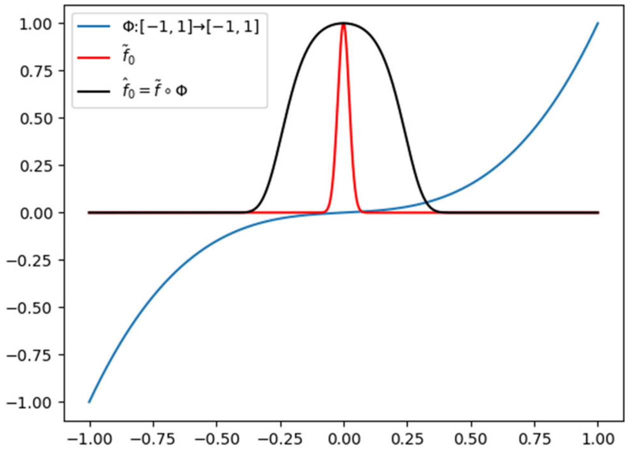


Fig. 9 Illustration of one-dimensional blow-up of a local feature

References

Babuška, I., Nobile, F., Tempone, R.: A stochastic collocation method for elliptic partial differential equations with random input data. SIAM J. Numer. Anal. **45**(3), 1005–1034 (2007)

Babuška, I., Nobile, F., Tempone, R.: A stochastic collocation method for elliptic partial differential equations with random input data. SIAM Rev. **52**(2), 317–355 (2010)

Bachmayr, M., Schneider, R., Uschmajew, A.: Tensor networks and hierarchical tensors for the solution of high-dimensional partial differential equations. Found. Comput. Math. **16**(6), 1423–1472 (2016)

Bachmayr, M., Cohen, A., Dahmen, W.: Parametric PDEs: sparse or low-rank approximations? IMA J. Numer. Anal. **38**(4), 1661–1708 (2017)

Ballani, J., Grasedyck, L., Kluge, M.: Black box approximation of tensors in hierarchical tucker format. Linear Algebra Appl. **438**(2), 639–657 (2013)

Baptista, R.M., Bigoni, D., Morrison, R., Spantini, A.: `TransportMaps`, (MIT Uncertainty Quantification Group , 2015–2018). http://transportmaps.mit.edu/docs/

Brennan, M., Bigoni, D., Zahm, O., Spantini, A., Marzouk, Y.: Greedy inference with structure-exploiting lazy maps. Adv. Neural Inf. Process. Syst. **33**, 8330–8342 (2020)

Chen, P., Schwab, C.: Sparse-grid, reduced-basis Bayesian inversion: nonaffine-parametric nonlinear equations. J. Comput. Phys. **316**, 470–503 (2016)

Cohen, A., Migliorati, G.: Optimal weighted least-squares methods. SMAI J. Comput. Math. **3**, 181–203 (2017). https://doi.org/10.5802/smai-jcm.24

Cui, T., Dolgov, S.: Deep composition of Tensor-Trains using squared inverse Rosenblatt transports. Found. Comput. Math., pp. 1–60 (2021)

Da Fies, G., Vianello, M.: On the Lebesgue constant of subperiodic trigonometric interpolation. J. Approx. Theory **167**, 59–64 (2013)

Dashti, M., Stuart, A.M.: The Bayesian approach to inverse problems. Handbook of uncertainty quantification, pp. 1–118 (2016)

Detommaso, G., Cui, T., Marzouk, Y., Spantini, A., Scheichl, R.: A Stein variational Newton method. In: Advances in Neural Information Processing Systems, pp. 9169–9179 (2018)

Detommaso, G., Kruse, J., Ardizzone, L., Rother, C., Köthe, U., Scheichl, R.: Hint: Hierarchical Invertible Neural Transport for General and Sequential Bayesian inference. arXiv preprint arXiv:1905.10687 (2019)

Dodwell, T., Ketelsen, C., Scheichl, R., Teckentrup, A.: Multilevel Markov chain Monte Carlo. SIAM Rev. **61**(3), 509–545 (2019)

Dolgov, S., Anaya-Izquierdo, K., Fox, C., Scheichl, R.: Approximation and sampling of multivariate probability distributions in the tensor train decomposition. Stat. Comput. **30**(3), 603–625 (2020)

Dunkl, C.F., Xu, Y.: Orthogonal Polynomials of Several Variables, vol. 155. Cambridge University Press, Cambridge (2014)

Eigel, M., Gruhlke, R., Marschall, M., Zander, E.: `alea`—a python framework for spectral methods and low-rank approximations in uncertainty quantification. https://bitbucket.org/aleadev/alea

Eigel, M., Gittelson, C.J., Schwab, C., Zander, E.: Adaptive stochastic Galerkin FEM. Comput. Methods Appl. Mech. Eng. **270**, 247–269 (2014)

Eigel, M., Pfeffer, M., Schneider, R.: Adaptive stochastic Galerkin FEM with hierarchical tensor representations. Numer. Math. **136**(3), 765–803 (2017)

Eigel, M., Marschall, M., Schneider, R.: Sampling-free Bayesian inversion with adaptive hierarchical tensor representations. Inverse Prob. **34**(3), 035010 (2018)

Eigel, M., Neumann, J., Schneider, R., Wolf, S.: Non-intrusive tensor reconstruction for high-dimensional random PDEs. Comput. Methods Appl. Math. **19**(1), 39–53 (2019a)

Eigel, M., Schneider, R., Trunschke, P., Wolf, S.: Variational Monte Carlo—bridging concepts of machine learning and high-dimensional partial differential equations. Adv. Comput. Math. (2019b). https://doi.org/10.1007/s10444-019-09723-8

Eigel, M., Marschall, M., Pfeffer, M., Schneider, R.: Adaptive stochastic Galerkin FEM for lognormal coefficients in hierarchical tensor representations. Numer. Math. **145**(3), 655–692 (2020)

El Moselhy, T.A., Marzouk, Y.M.: Bayesian inference with optimal maps. J. Comput. Phys. **231**(23), 7815–7850 (2012)

Ernst, O.G., Sprungk, B., Tamellini, L.: On expansions and nodes for sparse grid collocation of lognormal elliptic PDEs. arXiv preprint arXiv:1906.01252 (2019)

Espig, M., Grasedyck, L., Hackbusch, W.: Black box low tensor-rank approximation using fiber-crosses. Constr. Approx. **30**(3), 557 (2009)

Foo, J., Karniadakis, G.E.: Multi-element probabilistic collocation method in high dimensions. J. Comput. Phys. **229**(5), 1536–1557 (2010)

Garcke, J., Griebel, M.: Sparse Grids and Applications, vol. 88. Springer, Berlin (2012)

Gilks, W.R., Richardson, S., Spiegelhalter, D.: Markov Chain Monte Carlo in Practice. Chapman and Hall/CRC, London (1995)

Gorodetsky, A.A., Karaman, S., Marzouk, Y.M.: Function-Train: a continuous analogue of the tensor-train decomposition. arXiv preprint arXiv:1510.09088 (2015)

Grasedyck, L., Kressner, D., Tobler, C.: A literature survey of low-rank tensor approximation techniques. GAMM Mitteilungen **36**(1), 53–78 (2013)

Griebel, M., Harbrecht, H.: On the construction of sparse tensor product spaces. Math. Comput. **82**(282), 975–994 (2013)

Hackbusch, W.: Tensor Spaces and Numerical Tensor Calculus, vol. 42. Springer, Berlin (2012)

Hoang, V.H., Schwab, C.: N-term Wiener chaos approximation rates for elliptic PDEs with lognormal Gaussian random inputs. Math. Models Methods Appl. Sci. **24**(04), 797–826 (2014)

Huber, B., Wolf, S.: Xerus—a general purpose tensor library. https://libxerus.org/ (2014–2017)

Kaipio, J., Somersalo, E.: Statistical and Computational Inverse Problems, vol. 160. Springer, Berlin (2006)

Li, J., Marzouk, Y.M.: Adaptive construction of surrogates for the Bayesian solution of inverse problems. SIAM J. Sci. Comput. **36**(3), A1163–A1186 (2014)

Liu, Q., Wang, D.: Stein variational gradient descent: a general purpose Bayesian inference algorithm. In: Advances in neural information processing systems, pp. 2378–2386 (2016)

Marzouk, Y., Moselhy, T., Parno, M., Spantini, A.: An introduction to sampling via measure transport. arXiv preprint arXiv:1602.05023 (2016)

Mead, K., Delves, L.: On the convergence rate of generalized Fourier expansions. IMA J. Appl. Math. **12**(3), 247–259 (1973)

Neal, R.M.: Annealed importance sampling. Stat. Comput. **11**(2), 125–139 (2001)

Nobile, F., Tempone, R., Webster, C.G.: A sparse grid stochastic collocation method for partial differential equations with random input data. SIAM J. Numer. Anal. **46**(5), 2309–2345 (2008)

Oseledets, I.V.: Tensor-train decomposition. SIAM J. Sci. Comput. **33**(5), 2295–2317 (2011)

Oseledets, I., Tyrtyshnikov, E.: TT-cross approximation for multidimensional arrays. Linear Algebra Appl. **432**(1), 70–88 (2010)

Papamakarios, G., Nalisnick, E., Rezende, D.J., Mohamed, S., Lakshminarayanan, B.: Normalizing flows for probabilistic modeling and inference. J. Mach. Learn. Res. **22**(57), 1–64 (2021)

Parno, M.D., Marzouk, Y.M.: Transport map accelerated Markov chain Monte Carlo. SIAM/ASA J. Uncertain. Quantif. **6**(2), 645–682 (2018)

Parno, M., Moselhy, T., Marzouk, Y.: A multiscale strategy for Bayesian inference using transport maps. SIAM/ASA J. Uncertain. Quantif. **4**(1), 1160–1190 (2016)

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: NIPS-W (2017)

Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International Conference on Machine Learning. PMLR, pp. 1530–1538 (2015)

Rohrbach, P.B., Dolgov, S., Grasedyck, L., Scheichl, R.: Rank bounds for approximating Gaussian densities in the tensor-train format. arXiv preprint arXiv:2001.08187 (2020)

Rudolf, D., Sprungk, B.: Metropolis-Hastings importance sampling estimator. PAMM **17**(1), 731–734 (2017)

Santambrogio, F.: Optimal transport for applied mathematicians. Birkäuser NY **55**, 58–63 (2015)

Schillings, C., Schwab, C.: Scaling limits in computational Bayesian inversion. ESAIM Math. Model. Numer. Anal. **50**(6), 1825–1856 (2016)

Schillings, C., Sprungk, B., Wacker, P.: On the convergence of the Laplace approximation and noise-level-robustness of Laplace-based Monte Carlo methods for Bayesian inverse problems. Numer. Math. **145**(4), 915–971 (2020)

Schneider, R., Uschmajew, A.: Approximation rates for the hierarchical tensor format in periodic Sobolev spaces. J. Complex. **30**(2), 56–71 (2014)

Schwab, C., Gittelson, C.J.: Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. Acta Numer. **20**, 291–467 (2011)

Stuart, A.M.: Inverse problems: a Bayesian perspective. Acta Numer. **19**, 451–559 (2010)

Tran, D., Vafa, K., Agrawal, K., Dinh, L., Poole, B.: Discrete flows: invertible generative models of discrete data. Adv. Neural. Inf. Process. Syst. **32**, 14719–14728 (2019)

Villani, C.: Optimal Transport: Old and New, vol. 338. Springer, Berlin (2008)

Weare, J.: Efficient Monte Carlo sampling by parallel marginalization. Proc. Natl. Acad. Sci. **104**(31), 12657–12662 (2007)