

## **Low-rank Wasserstein polynomial chaos expansions in the framework of optimal transport**

Robert Gruhlke, Martin Eigel

submitted: March 17, 2022

Weierstrass Institute  
Mohrenstr. 39  
10117 Berlin  
Germany  
E-Mail: [robert.gruhlke@wias-berlin.de](mailto:robert.gruhlke@wias-berlin.de)  
[martin.eigel@wias-berlin.de](mailto:martin.eigel@wias-berlin.de)

No. 2927  
Berlin 2022



---

*2020 Mathematics Subject Classification.* 15A69, 35R13, 65N12, 65N22, 65J10, 97N50.

*Key words and phrases.* Tensor train format, Wasserstein metric, polynomial chaos expansion, alternating least squares, numerical upscaling, tensor chain, Sinkhorn divergence, optimal transport, multimodal distribution.

The authors acknowledge the support by the DFG SPP1886 “Polymorphic uncertainty modelling for the numerical design of structures”.

Edited by  
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)  
Leibniz-Institut im Forschungsverbund Berlin e. V.  
Mohrenstraße 39  
10117 Berlin  
Germany

Fax: +49 30 20372-303  
E-Mail: [preprint@wias-berlin.de](mailto:preprint@wias-berlin.de)  
World Wide Web: <http://www.wias-berlin.de/>

# Low-rank Wasserstein polynomial chaos expansions in the framework of optimal transport

Robert Gruhlke, Martin Eigel

## Abstract

A unsupervised learning approach for the computation of an explicit functional representation of a random vector  $Y$  is presented, which only relies on a finite set of samples with unknown distribution. Motivated by recent advances with computational optimal transport for estimating Wasserstein distances, we develop a new *Wasserstein multi-element polynomial chaos expansion* (WPCE). It relies on the minimization of a regularized empirical Wasserstein metric known as debiased Sinkhorn divergence.

As a requirement for an efficient polynomial basis expansion, a suitable (minimal) stochastic coordinate system  $X$  has to be determined with the aim to identify ideally independent random variables. This approach generalizes representations through diffeomorphic transport maps to the case of non-continuous and non-injective model classes  $\mathcal{M}$  with different input and output dimension, yielding the relation  $Y = \mathcal{M}(X)$  in distribution. Moreover, since the used PCE grows exponentially in the number of random coordinates of  $X$ , we introduce an appropriate low-rank format given as stacks of tensor trains, which alleviates the curse of dimensionality, leading to only linear dependence on the input dimension. By the choice of the model class  $\mathcal{M}$  and the smooth loss function, higher order optimization schemes become possible. It is shown that the relaxation to a discontinuous model class is necessary to explain multimodal distributions. Moreover, the proposed framework is applied to a numerical upscaling task, considering a computationally challenging microscopic random non-periodic composite material. This leads to tractable effective macroscopic random field in adopted stochastic coordinates.

## 1 Motivation

Measure transport has become a popular research topic in many scientific fields and is in particular of great interest in Uncertainty Quantification (UQ) and modern Machine Learning (ML). Our contribution provides a strategy to obtain a model representation optimized in distribution, approximating the law of the observed target  $Y$  with values in  $\mathbb{R}^N$ , for which only samples are at hand. This computed model is given explicitly in functional form as expansion in a compressed multi-element polynomial basis in a *stochastic reference coordinate system* determined by a random variable  $X$  with values in  $\mathbb{R}^M$ . We denote this model class by  $\mathcal{M} = \mathcal{M}(X)$  and optimize the representation by means of *computational optimal transport* [65] using measure fitting with a *debiased Sinkhorn loss* [39]. The choice of the reference coordinates  $X$ , consisting of independent random variables, ideally is adapted to the problem at hand<sup>1</sup>

---

<sup>1</sup>However, this interesting topic is not examined in this work and we assume a given reference system. An approach in this direction can be found in [73] using a log-loss to determine the needed input dimension and the associated numbers of degree of freedom.

This leads to a representation

$$Y \stackrel{d}{\approx} Y_{\text{PCE}} := \mathcal{M}(X) = \sum_{s=1}^S \sum_{\alpha \in \mathbb{N}_0^M} C_s[\alpha] P_\alpha^s(X), \quad (1)$$

which approximates the exact but unknown  $Y$  (in distribution), where  $S \in \mathbb{N}$  is the number of multi-elements and  $P_\alpha^s$  denotes the multi-element polynomial chaos indexed by  $\alpha$  on a decomposition into  $S$  subdomains.

In case that highly nonlinear maps are required to accurately represent the transformation of the reference to the target measure, the functional (polynomial chaos) representation easily leads to very high-dimensional representations. In particular, the complexity scales exponentially (“curse of dimensionality”) in the number of stochastic coordinates  $M$ , determining the size of the coefficient  $C_s$ . To make this approach feasible in practice, it is inevitable to compress the coefficient. We tackle this challenging task by introducing a low-rank tensor train ring (TTR) format. This is an extension of the popular tensor train (TT) format [64], which has been used successfully e.g. in UQ applications [29, 32, 25, 24], quantum physics models [82] and quantum chemistry [76]. In fact, the ring format is constructed by stacks of tensor trains and – opposite to the standard TT format – is able capable to represent vector valued output. It leads to an overall complexity that again scales only linearly in  $M$ .

The model design (1) can be seen as a generalization of the (much stricter) notion of optimal transport (OT), which in our method is conceptually carried out on each part of a decomposition of the preimage, mapping to a selection of samples determining the image. This approach allows for multimodal distribution representations in terms of random variables. While OT requires a diffeomorphism between two spaces of equal dimension, this assumption is relaxed in our approach, in particular allowing for non-continuous probability mass transport. Such a relaxation is inevitable if multimodalities should be represented with high accuracy. Moreover, we do not consider invertible mappings  $\mathcal{M}$  and use the notion of convergence in distribution for a generalized functional representation where input dimension  $M$  and output dimension  $N$  may differ.

Our methodological contribution combines different tools, ranging from optimal transport theory to polynomial chaos representations of random variables. It can be understood as an alternative model class to generative adversarial neuronal networks (GAN) and its extensions to Wasserstein distances denoted as Wasserstein GANs (WGAN), allowing for Riemannian optimization schemes. The performance of the technique is used to solve several challenging problems in uncertainty quantification such as the representation of multimodal distributions or random fields defined on different spatial scales, connected by the notion of “stochastic numerical upscaling”. Consequently, our contribution is related to different thematic fields, summarized in the following.

**Optimal transport** The notion of optimal transport [79] allows to compare probability measures in terms of required workload or costs to move one probability mass to another [52]. In the standard setting of the Monge problem [79] measurable sets  $X$  and  $Y$  with respective measures  $\mu$  and  $\nu$  such that  $\mu(X) = \nu(Y) < \infty$  are assumed. The task consists of finding an injective transport mapping  $T : X \rightarrow Y$  subject to some cost function  $c : X \times Y \rightarrow \mathbb{R}_+$ ,

$$T = \arg \min_{\tau} \left\{ \int_X c(x, \tau(x)) d\mu(x) : \tau_{\#} \mu = \nu \right\},$$

such that

$$\int_X g(T(x)) d\mu(x) = \int_Y g(y) dT_{\#} \mu(y),$$



for any  $\mu$ -measurable  $g : X \rightarrow \mathbb{R}$ . Let  $Y \sim \mu$  and  $X \sim \nu$ . If such a  $T$  exists, then

$$Y \stackrel{d}{=} T(X). \quad (2)$$

This concept of transport maps has been examined thoroughly in the context of Bayesian inverse problems [62, 7, 57, 13, 23]. As a matter of fact, these problems consist of determining a posterior measure given a prior measure conditioned on a set of observations, which resembles the transport problem. Consequently, the knowledge of the (approximate) transport map can help to significantly alleviate the high computational burden that e.g. is typical for Markov chain Monte Carlo methods. With this in mind, functional representations in a polynomial basis exploiting the beneficial structure of the Knothe-Rosenblatt transform were for instance developed in [74, 21, 20]. For the formulation of the variational problem, the Kullback-Leibler divergence is used. However, while this type of loss functional is appropriate in the setting of Bayesian inference due to the absolute continuity of the prior and posterior measures, the latter property does in general not hold true when optimizing with respect to measures.

Another drawback is that the existence of an injective transport map  $T$  is not ensured in general. A relaxation of this problem was formulated by Kantorovich, avoiding the missing guaranteed existence of a map  $T$  in the Monge problem. This concept is discussed in more detail in the first section of this work, including its computational challenges. By introducing a model class  $\mathcal{M}$  and a reference coordinate system represented by a random variable  $X$ , we obtain a representation motivated by (2) given by

$$Y \stackrel{d}{=} \mathcal{M}(X).$$

This type of representation generalizes the notion of transport in the following way. First if  $X$  and  $Y$  are random vectors with values in  $\mathbb{R}^M$  and  $\mathbb{R}^N$ ,  $N, M \in \mathbb{N}$ , then we allow that  $M \neq N$  while in a classical transport formulation  $M = N$  has to be satisfied. Second, we can drop the common injectivity or even diffeomorphism assumptions on  $T$ . Finally, we allow  $\mathcal{M}$  to be discontinuous, a feature that allows for a broader expressivity, which is essential for multi-modal random variables  $Y$ .

**Polynomial chaos & low-rank tensor formats** Polynomial chaos (PC) or multi-element PC [80] is ubiquitous in UQ [53, 72, 17] to represent a random variable  $Y : \Omega \rightarrow \mathbb{R}^N$  with finite variance in a coordinate system of  $M$  independent random variables  $X = (X_1, \dots, X_M)$ . These are used for an expansion in polynomials orthogonal with respect to given distribution of  $X$ . It is an explicit functional representation of the form

$$Y(\omega) = Y(X(\omega)) = \sum_{\alpha \in \mathbb{N}_0^M} C[\alpha] P_\alpha(X(\omega)), \quad \omega \in \Omega.$$

This surrogate includes all statistical information and allows for an computationally efficient evaluation of statistical quantities. In UQ, it has become extremely popular for the representation of data and solutions of random differential equations, especially PDEs [35, 72, 17]. Such polynomial chaos surrogates can be obtained by different means, the most common of which are stochastic Galerkin methods and empirical least squares projections, see [8, 11, 53, 30] and [16, 61, 34, 33], respectively. The convergence of the polynomial chaos expansion follows from the classical theory of Cameron and Martin [35].

When  $N = 1$ , low-rank tensor formats allow to compress the high-dimensional representation. These formats have initially been devised in the quantum chemistry and physics community and were later

reinvented in the field of computational mathematics. We refer the interested reader to the monograph [49] and the overview articles [63, 6]. The TT format was (re-)introduced in [64] and has become quite popular in computational mathematics, yielding a compressed representation of  $C$  for  $N = 1$  by

$$C[\alpha] = C[\alpha_1, \dots, \alpha_M] = C_1[\alpha_1] \cdot \dots \cdot C_M[\alpha_M] \quad (3)$$

with order 3 tensors  $C_m \in \mathbb{R}^{r_{m-1}, d_m, r_m}$ , tensor train ring rank  $r = (r_1, \dots, r_{m-1}) \in \mathbb{N}^{M-1}$ ,  $r_0 = r_m = 1$  and polynomial degree  $d_m \in \mathbb{N}$  in coordinate  $m$ . The number of degrees of freedom in this format is bounded by  $\mathcal{O}(Mdr^2)$  with  $r = \max\{r_1, \dots, r_M\}$ ,  $d = \max\{d_1, \dots, d_M\}$ . The complexity of this hierarchical representation thus is linear in the stochastic dimensions  $M$ . Moreover, these tensors form a differentiable manifold such that e.g. Riemannian optimization methods become feasible [75]. For  $N > 1$ , as mentioned above these formats cannot be used directly and thus require a modified design.

One idea would be to decompose  $C = C[i, \alpha]$  with  $i = 1, \dots, N$  denoting the  $i$ -th output component, as in (3) but using forth order tensor  $C_1 = C_1[i, \alpha_1]$  and third order tensors  $C_m = C_m[\alpha_m]$  for  $m = 2, \dots, M$ . However, in our application in practice this format results in an exponential growth of ranks with respect to the output dimension  $N$ . Consequently, we extend the standard TT format to the  $N \geq 1$  case, leading to a problem-adapted TTR or *stack of tensor trains* format, allowing each component tensor  $C_m$  to interact with any output component  $i = 1, \dots, N$ . The TTR format reads

$$C[i, \alpha] = C[i, \alpha_1, \dots, \alpha_M] = C_1[i, \alpha_1] \cdot \dots \cdot C_M[i, \alpha_M], \quad i = 1, \dots, N, \quad (4)$$

where all component tensors  $C_m \in \mathbb{R}^{N, r_{m-1}, d_m, r_m}$  are of order 4. For fixed  $i = 1, \dots, N$ , (4) can be interpreted as a classical tensor train format.

To get a bigger picture, it should be noted that the graph structure of tensor formats in some way resembles the topology (and expressiveness) of neural networks (NN) [18, 19, 1, 2, 5]. In fact, tensor networks can be seen as a subset of NNs with somewhat lesser expressivity<sup>2</sup> on the one hand, but much richer mathematical structure on the other.

**Relation to Generative Adversarial Neural Networks (GAN)** A very popular generative model represented by neural networks (NN) is the notion of GANs [46]. We introduce basic ideas since these methods pursue a similar goal as our approach but differ fundamentally in the way they achieve it. Classical Wasserstein GANs (WGANs), e.g. [55] are based on the dual formulation of  $\mathcal{W}_1$  optimal transport based on the difference of expectations of 1-Lipschitz functions given by

$$\min_{\theta} \max_{f: \text{Lip}(f) \leq 1} \mathbb{E}_{X \sim \mu} [f(X)] - \mathbb{E}_{X \sim \nu[\theta]} [f(X)].$$

The parametric measure  $\nu[\theta]$  is determined by some known distribution  $\nu$  represented as NN model  $g[\theta]$  subject to parameters  $\theta$ , cf. [4]. Accordingly, the following GAN optimization has to be solved,

$$\min_{\theta} \max_{f: \text{Lip}(f) \leq 1} \mathbb{E}_{X \sim \mu} [f(X)] - \mathbb{E}_{X \sim \nu} [f(g[\theta](X))]. \quad (5)$$

It has been observed that WGANs are difficult to train, in particular due to exhibited non-robustness. They may perform unconvincingly, e.g. when multi-modal distributions are approximated as in [55]. Since classical WGANs with  $\ell_1$  cost may not converge, several alternative cost functions were introduced such as the  $\ell_2$  cost in [66]. Smoothed WGANs based on a regularized Sinkhorn loss were introduced in [41] and applied for the pictures data sets MNIST and CIFAR-10 in [68]. Furthermore, the work in [9] considered the construction of invertible residual networks as generative models. In this

<sup>2</sup>mainly because they are a multilinear instead of a nonlinear representation

construction, the involved Lipschitz constraints to ensure invertibility remain a challenging task in the actual application.

Opposite to the GAN formulation, we do not use the dual characterization and also do not require a discriminator approximation or Lipschitz constraints. The used explicit polynomial chaos model class corresponds to the *generator* only within the GAN min-max formulation (5). This leads to a single model that needs to be trained. However, we underline the still present challenge of non-convex and non-linear optimization involving local minima, corresponding usually to inaccurate solutions. The effects already occur in simple scenarios to be discussed in examples in due course in the presentation of the approach.

We rely on second order optimization schemes based on automatic differentiation that become feasible because of the smooth parameter dependence of the debiased Sinkhorn loss. Additionally, we consider the UQ point of view and interpret this technique as a change of coordinates, which allows for possible stochastic dimensional reduction when  $M < N$ .

**Application context** This work is inspired considerably by the works in [73] and [69] where the derivation of effective random coarse grained fields is examined given only limited fine scale informations in terms of samples or observations. In [73] the construction of a random field representation in terms of a PCE was considered with coefficients defined on a Stiefel manifold associated to inherited correlation structures in the underlying Kosambi–Karhunen–Loève expansions. In order to make this approach tractable, a simplified loss function - namely a tensorized approximation of a maximum likelihood estimation - using one dimensional kernel density estimates to gain statistical information was considered. In contrast to this approach, here we consider a PCE obtained from optimizing (approximations of) Wasserstein losses. We denote this technique *Wasserstein Polynomial Chaos Expansions* (WPCE). It allows to capture the sample distribution accurately in an unsupervised manner without relying on kernel density estimations that converge slowly and lose statistical information in the estimation process.

The work [69] introduces a stochastic Bayesian upscaling framework to obtain PCEs of random variables. Here, the representation is obtained via a projection in terms of conditional expectations given as spectral Kalman filter. In its most elaborate form, this approach yields an assimilated random variable that matches the empirical distribution in the first and second moment only. Nevertheless, it should be noted that extensions to higher moments are mentioned. The practical realization however would be rather challenging.

A further strong motivation for the presented approach is provided by the applications we have in mind. First, an efficient generation of samples from a highly nonlinear (i.e. non-Gaussian) target measure is a challenging task, which can be found in many statistical applications. The access to e.g. multi-modal randomness in terms of density learning or standard Markov Chain Monte Carlo is a difficult task. Here we rely on such representation in terms of *functional approximations of random variables*. Such a functional representation exhibits two obvious advantages. On the one hand, it can be used in follow-up computations as a closed representation of randomness, relying on a known underlying reference coordinate system  $X$  and in turn can be used as a generator (or surrogate) to generate more samples of  $Y$ . On the other hand, we are interested in the *numerical upscaling* of random non-periodic microstructures as discussed in Section 4.3. In contrast to classical (random) homogenization with constant effective coefficient [10, 45, 44], we strive for an approach where the *macroscopic representation remains a random field*, hence also containing statistical information (of the finer scale) on the coarser scale. For practical applications, this provides much more useful information since e.g. the variance (or other statistical quantities of interest) of some system response can be determined. Upscaling of

stochastic non-periodic material is carried out in a stochastic pointwise sense, i.e. a sample of the fine scale field is upscaled numerically to obtain a sample of the related coarse scale random field for a given prescribed domain decomposition. Note that the homogenized material and the microscale material both have unknown or intractable distributions. In the homogenized framework, we allow for stochastic fluctuations without the need to (approximate) a stochastic homogenization problem defined on a unbounded domain.

The structure of this paper is as follows. The next Section 2 provides an overview of recent advances in computation optimal transport, building the framework for our approach. Section 3 introduces the structure of our model class based on a compressed multi-element PCE, its underlying reference coordinate system  $X$  and discusses challenges in the related optimization problem. Section 4 is devoted to the investigation of the numerical performance of the proposed method in the setting of multi-modal randomness and the representation of random fields on possible different scales in terms of stochastic upscaling. Eventually, a conclusion in Section 5 summarizes the main achievements and further directions.

## 2 Computational optimal transport

This section is concerned with the central goal of this work, namely the representation of a random variable  $Y$  with image in  $\mathbb{R}^N$ , which is unknown a priori and only finitely many samples are available or can be generated, typically involving high computational costs. As a standard tool, *kernel density estimation* (KDE) [15] is applied on an ensemble of samples to approximate a – possibly not existing – density of  $Y$ . Once a representation is obtained, further samples of  $Y$  can be drawn based on the density information with potentially negligible effort. However, the KDE suffers from the curse of dimensionality (CoD), i.e. the number of samples required for a reasonable approximation of the underlying density grows exponentially in the dimension  $N$ .

As an alternative strategy, we may aim for a functional representation

$$Y \stackrel{d}{=} \mathcal{M}(X) \tag{6}$$

based on some model class  $\mathcal{M}$  and some a priori chosen stochastic coordinate system encoded in a random variable  $X$ . We refer to Section 3 for a detailed discussion. This representation yields several advantages. First, it does not require the existence of an underlying density of  $Y$ . Second, samples of  $Y$  can be drawn with little effort if obtaining samples from  $X$  and the propagation through  $\mathcal{M}$  can be carried out efficiently. Finally, this functional representation can be advantageous to compute integral quantities in an analytical (sampling-free) manner. A common requirement is for instance the evaluation of quantities of interests like moments. This also is a central ingredient in stochastic Galerkin schemes [42, 58, 53] for the computation of inner products.

Given only sample information of  $Y$  the Wasserstein distance is a promising tool to compare distributions or their empirical counterparts even in the case when the measures do not share support. This property is useful when optimizing over  $\mathcal{M}$  based on a random or non-informed start value. In what follows we give an overview of results and challenges related to the Wasserstein distance and its computational advances based on regularization that in the end allow for the measure fitting application in mind.

## 2.1 Exact optimal transport: Wasserstein distance

Since only samples of  $Y$  are available,  $Y$  and  $\mathcal{M}(X)$  can only be compared in terms of samples or equivalently by determining the distance of the related empirical measures defined through the finite samples of  $Y$  and  $\mathcal{M}(X)$ . We intend to compute the distance of measures with the help of the *Wasserstein* or *Kantorovich–Rubinstein* metric [79], which are introduced in the following.

Let  $(\mathcal{V}, d)$  be a complete metric space,  $c: \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$  a symmetric continuous cost function and let  $\mathcal{D}(\mathcal{V})$  denote the set of probability measures on  $\mathcal{V}$ . The Kantorovich formulation [52] of optimal transport costs or Wasserstein costs between probability measures  $\mu, \nu \in \mathcal{D}(\mathcal{V})$  is defined as the minimal cost required to move each element of mass of  $\mu$  to each element of mass of  $\nu$  written as a linear problem over the set of transportation plans, which are probability measures on  $\mathcal{V} \times \mathcal{V}$ :

$$\mathcal{W}(\mu, \nu) := \mathcal{W}_c(\mu, \nu) := \inf_{\pi \in \mathcal{D}(\mathcal{V} \times \mathcal{V})} \{ \langle c, \pi \rangle : \pi_1 = \mu, \pi_2 = \nu \}, \quad (7)$$

where  $\pi_1 = \int_{y \in \mathcal{X}} d\pi(\cdot, y)$  and  $\pi_2 = \int_{x \in \mathcal{X}} d\pi(x, \cdot)$  are the marginals of the transportation plan  $\pi$ .

For  $p \in [1, \infty]$  let  $c(v_1, v_2) = d(v_1, v_2)^p$  and  $\mathcal{D}_p(\mathcal{V})$  denotes the set of measures on  $\mathcal{V}$  with finite moment of order  $p$ . Define the  $p$ -th *Wasserstein distance* by

$$\mathcal{W}_p(\mu, \nu) := \mathcal{W}_{c^p}(\mu, \nu)^{\frac{1}{p}}, \quad \mu, \nu \in \mathcal{D}_p(\mathcal{V}). \quad (8)$$

In practice, a measure  $\mu$  is unknown and only  $n \in \mathbb{N}$  *iid* samples of  $\mu$  or equivalently an empirical measure  $\mu_n$  are available. This raises the question of how well such an empirical measure explains the distribution  $\mu$  in terms of the Wasserstein distance (8). Since the Wasserstein metrizes the weak convergence of measures [79] and the empirical measure converges weakly to  $\mu$  [77], it follows for  $p \in [1, \infty)$  that

$$\mathcal{W}_p(\mu, \mu_n) \rightarrow 0, \quad \mu\text{-a.s. as } n \rightarrow \infty, \quad (9)$$

provided that  $X$  is compact and separable and  $\mu$  is a Borel measure. Unfortunately, the convergence rate of  $\mu_n$  to  $\mu$  in Wasserstein distance inevitably suffers from the curse of dimensionality. A negative result [26] showed that if  $\mu$  is absolutely continuous with respect to the Lebesgue measure on  $\mathbb{R}^N$  then for some  $C > 0$  it holds

$$\mathbb{E}[\mathcal{W}_1(\mu, \mu_n)] \geq Cn^{-1/N}.$$

In the opposite case, if  $N > 2$  and  $\mu$  are compactly supported on  $\mathbb{R}^N$  then

$$\mathbb{E}[\mathcal{W}_1(\mu, \mu_n)] \leq Cn^{-1/N}.$$

Similar negative results have been extended to the case  $p \in [1, \infty]$  in [81] based on the concept of upper and lower Wasserstein dimensions of the support of the underlying measure. In the special case of measures  $\mu$  with regular  $N$ -dimensional support [59] that is absolute continuous with respect to a Hausdorff measure, the same upper and lower asymptotical bounds hold true for  $\mathbb{E}[\mathcal{W}_p(\mu, \mu_n)]$  at least with  $p \in [1, N/2]$ .

A much more involved analysis in [81] shows that the rate of convergence of  $\mathcal{W}_p(\mu, \mu_n)$  depends on a notion of the intrinsic dimension of the measure  $\mu$ , which can be significantly smaller than the dimension of the metric space on which  $\mu$  is defined. Moreover, in the finite-sample regime, wildly different convergence rates may appear. In particular, they can enjoy much faster convergence for small  $n$ . An exemplary phenomenon is based on the different dimensional structure of measures at

different scales, i.e. the so called *multi-scale behaviour*. For illustration, let  $\mu$  be  $(m, \Delta)$ -clusterable, i.e.  $\text{supp}(\mu)$  lies in the union of  $m$  balls of radius at most  $\Delta$ . Then for all  $n \leq m(2\Delta)^{-2p}$ ,

$$\mathbb{E}[\mathcal{W}_p^p(\mu, \mu_n)] \leq (9^p + 3) \sqrt{\frac{m}{n}}.$$

Notably,  $\mu_n$  converges to  $\nu$  as  $n^{-1/2p}$  in the pre-asymptotic regime. This result applies for example to mixtures of Gaussian distributions. If the measure  $\mu$  with support in  $\mathbb{R}^N$  has approximately low-dimensional support for dimension  $\underline{N} < N$  then one obtains  $\mathbb{E}[\mathcal{W}_p^p(\mu, \mu_n)] \leq Cn^{-p/\underline{N}}$  in the finite sample range. Note that this finite range convergence may be much faster than  $n^{-p/N}$  provided  $\underline{N} \ll N$ .

So far we only discussed bounds on the expectation of  $\mathcal{W}_p(\mu, \mu_n)$ . This analysis is motivated by the so called *concentration around expectation* property. In particular, for  $n \in \mathbb{N}$  and  $p \in [1, \infty)$  the McDiarmid's inequality yields [81]

$$\mathbb{P}(\mathcal{W}_p^p(\mu, \mu_n) \geq \mathbb{E}[\mathcal{W}_p^p(\mu, \mu_n)] + t) \leq \exp(-2nt^2).$$

We may now consider two random variables  $Y \sim \mu$  and  $\mathcal{M}(X) \sim \nu$  with values in  $(\mathbb{R}^N, \|\cdot\|)$ , *iid* samples  $Y_1, \dots, Y_n \sim \mu$ ,  $\mathcal{M}(X_1), \dots, \mathcal{M}(X_m)$  and denote by  $\mu_n$  and  $\nu_m$  the empirical measures of  $\mu$  and  $\nu$ , respectively, for given  $n, m \in \mathbb{N}$  defined by

$$\mu_n = \sum_{i=1}^n a_i \delta_{Y_i}, \quad \nu_m = \sum_{k=1}^m b_k \delta_{\mathcal{M}(X_k)}.$$

Here,  $\delta_x$  denotes the delta distribution in point  $x$  and  $a_i \equiv 1/n$  and  $b_j \equiv 1/m$ . Let  $p \leq 0$  and  $C_p = [\|Y_i - \mathcal{M}(X_j)\|^p]_{ij} \in \mathbb{R}^{n,m}$  be the cost matrix and  $\mathbf{1}_n, \mathbf{1}_m$  be the  $n$  or  $m$ -dimensional vector of ones and define  $a = (a_i)$  and  $b = (b_j)$ . Then, the Wasserstein distance  $\mathcal{W}_p(\mu_n, \nu_m)$  is characterized by

$$\mathcal{W}_p^p(\mu_n, \nu_m) = \min_{\pi \in \mathcal{U}_{nm}} \langle \pi, C_p \rangle, \quad \mathcal{U}_{nm} = \{\pi \in \mathbb{R}_+^{n,m} : \pi \mathbf{1}_m = \mathbf{1}_n/m, \pi^T \mathbf{1}_n = \mathbf{1}_m/m\}. \quad (10)$$

This is a classical transport problem with the special case of uniform weights. It can thus be solved by standard solvers for min-cost flow problems. Unfortunately, the complexity of these depends at best cubically on the input data. Hence, when  $n$  or  $m$  get large, this approach becomes impractical. The optimal transport plan  $\pi^*$  is sparse [28] and lies on the boundary of the feasible region, which is a polytope [65]. Consequently, the solution process can become unstable due to ambiguities or discontinuous jumps of the cost functional and its orientation onto the spanned polytope of the admissible region, which is caused by the enforced constraints due to the admissible set  $\mathcal{U}_{nm}$ . Moreover, given an optimal transport plan  $\pi^*$  and a parametric cost function  $c = c(\theta)$ , which is for example induced by the parameters of the model class  $\mathcal{M}$ , then  $\theta \mapsto \pi^*(c(\theta))$  is not differentiable.

## 2.2 Inexact optimal transport: Entropic regularization

Motivated by the drawbacks of the exact optimal transport presented in Section 2.1, namely a lack of regularity and intractable computational costs for large sample sizes, a new era to computational optimal transport has started with [22]. It is based on *entropic regularization* of the Wasserstein distance, which can be traced back to the work of Schrödinger [71],

$$\mathcal{W}_{c,\epsilon}(\mu, \nu) = \min_{\pi \in \mathcal{D}(\mathcal{X} \times \mathcal{X})} \langle \pi, c \rangle + \epsilon \text{KL}(\pi, \mu \otimes \nu) \quad (11)$$

$$\text{subject to } \pi \geq 0, \quad \pi_1 = \mu \quad \pi_2 = \nu, \quad (12)$$

for positive values of “temperature”  $\epsilon > 0$  with *entropic penalty* (or *Kullback-Leibler divergence*)

$$\text{KL}(\pi, \mu \otimes \nu) = \left\langle \pi, \log \frac{d\pi}{d\mu \otimes \nu} \right\rangle - \langle \pi, 1 \rangle + \langle \mu \otimes \nu, 1 \rangle. \quad (13)$$

However, while the  $p$ -th Wasserstein distance satisfies the triangle inequality based on the positive definite loss function  $c(x, y) = \frac{1}{p} \|x - y\|^p$ , those geometric properties do not hold for the entropic loss  $\mathcal{W}_{c, \epsilon}$ . Therefore, minimizing an  $\mathcal{W}_{c, \epsilon}$  loss is not a sensible approach [38]. In general, if  $\nu$  is a given target measure, there exists a degenerate measure  $\mu$  such that  $\mathcal{W}_{c, \epsilon}(\mu, \nu) < \mathcal{W}_{c, \epsilon}(\nu, \nu)$ .

To circumvent these issues, the *debiased Sinkhorn divergence* was proposed in [67]. It is given by

$$\mathcal{S}_\epsilon(\mu, \nu) = \mathcal{W}_{c, \epsilon}(\mu, \nu) - \frac{1}{2} (\mathcal{W}_{c, \epsilon}(\mu, \mu) + \mathcal{W}_{c, \epsilon}(\nu, \nu)) \quad (14)$$

and has been adopted in the machine learning community e.g. to obtain generative models [41]. As being understood in [38], the debiased Sinkhorn divergence may metricizes the convergence in law and is thus suitable for measure-fitting applications. In particular we have the following result due to [38, 39].

**Theorem 2.1.** *Let  $\mathcal{V}$  be a compact metric space with a Lipschitz cost function  $c: \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}_+$  that induces a positive universal kernel  $k_\epsilon(\cdot, \cdot) = \exp(-c(\cdot, \cdot)/\epsilon)$  for  $\epsilon > 0$ . Then,  $\mathcal{S}_\epsilon$  defines a **symmetric, positive, definite, smooth** loss function that is **convex** in each of its input variables. It also metricizes the convergence in law: for all probability measures  $\mu, \nu \in \mathcal{D}(\mathcal{V})$ ,*

$$0 = \mathcal{S}_\epsilon(\nu, \nu) \leq \mathcal{S}_\epsilon(\mu, \nu), \quad (15)$$

$$\mu = \nu \Leftrightarrow \mathcal{S}_\epsilon(\mu, \nu) = 0, \quad (16)$$

$$\mu_n \rightarrow \mu \Leftrightarrow \mathcal{S}_\epsilon(\mu_n, \mu) \rightarrow 0. \quad (17)$$

*The results extend to measures with bounded support on an Euclidian space  $\mathcal{V} = \mathbb{R}^d$  endowed with ground cost functions  $c(x, y) = \|x - y\|$  or  $c(x, y) = \frac{1}{2} \|x - y\|^2$ , which induce exponential and Gaussian kernels, respectively.*

The smoothness of the debiased Sinkhorn divergence with explicit derivation of the gradient is already pointed out in [56] in the discrete measure setting. The debiased Sinkhorn divergence can be interpreted as an interpolation between the exact Wasserstein metric and kernel maximum mean discrepancies [67, 41]. This observation can be seen in the sampling complexity result obtained for the entropic regularized Wasserstein distance  $\mathcal{W}_{c, \epsilon}$  due to [40].

**Theorem 2.2.** *Let  $\mu, \nu$  be a probability on bounded subsets  $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^N$  and let  $\mathcal{W}_\epsilon$  be defined upon a smooth  $L$ -Lipschitz cost  $c$ . Then,*

$$\mathbb{E} [|\mathcal{W}_\epsilon(\mu, \nu) - \mathcal{W}_\epsilon(\mu_n, \nu_n)|] = \mathcal{O} \left( \frac{\exp(\kappa/\epsilon)}{\sqrt{n}} \left( 1 + \frac{1}{\epsilon^{\lceil N/2 \rceil}} \right) \right), \quad (18)$$

*with  $\kappa = 2L|\mathcal{X}| + \|c\|_\infty$  and the constants only depend on  $|\mathcal{X}|, |\mathcal{Y}|, d$  and  $\|c^{(k)}\|_\infty$  for  $k = 0, \dots, \lceil N/2 \rceil$ . In particular, we get the following asymptotic behaviour in  $\epsilon$ ,*

$$\mathbb{E} [|\mathcal{W}_\epsilon(\mu, \nu) - \mathcal{W}_\epsilon(\mu_n, \nu_n)|] = \mathcal{O} \left( \frac{\exp(\kappa/\epsilon)}{\epsilon^{\lceil N/2 \rceil} \sqrt{n}} \right), \quad \epsilon \rightarrow 0, \quad (19)$$

$$\mathbb{E} [|\mathcal{W}_\epsilon(\mu, \nu) - \mathcal{W}_\epsilon(\mu_n, \nu_n)|] = \mathcal{O} \left( \frac{1}{\sqrt{n}} \right), \quad \epsilon \rightarrow +\infty. \quad (20)$$

For  $\epsilon \rightarrow \infty$ , we obtain a rate of convergence independent on the dimension  $N$ , while for  $\epsilon \rightarrow 0$  the curse of dimensionality appears in terms of  $\epsilon$  itself. The question of the error introduced by the entropic regularization to the Wasserstein distance was answered in [40].

**Theorem 2.3.** *Let  $\mu$  and  $\nu$  be probability measures on  $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^N$  such that  $|\mathcal{X}| = |\mathcal{Y}| \leq D > 0$  and assume that the cost  $c$  is  $L$ -Lipschitz with respect to both arguments. Then it holds*

$$0 \leq \mathcal{W}_\epsilon(\mu, \nu) - \mathcal{W}(\mu, \nu) \leq 2\epsilon N \log \left( \frac{e^2 \cdot L \cdot D}{\sqrt{N} \cdot \epsilon} \right). \quad (21)$$

In particular, as  $\epsilon \rightarrow 0$ ,  $0 \leq \mathcal{W}_\epsilon(\mu, \nu) - \mathcal{W}(\mu, \nu) \sim 2\epsilon N \log(1/\epsilon)$ . (22)

In [56] an improved approximation for discrete measures is provided if the optimal transport plan is accessible. Concretely, for discrete measures  $\mu, \nu$ , let

$$\pi^* = \operatorname{argmin}_{\pi \in \mathcal{D}(\mathcal{V} \times \mathcal{V})} \langle \pi, c \rangle + \epsilon \operatorname{KL}(\pi, \mu \otimes \nu) \quad (23)$$

$$\text{subject to } \pi \geq 0, \quad \pi_1 = \mu \quad \pi_2 = \nu. \quad (24)$$

Then, for some  $C > 0$ ,

$$|\langle \pi^*, c \rangle - \mathcal{W}(\mu, \nu)| < C e^{-1/\epsilon}.$$

However, the solution of (10) in general lacks sparsity due to the regularization and one hence only obtains “blurred versions” of the sparse optimal transport plan of the Wasserstein distance [65].

Note that all results above extend to the debiased Sinkhorn divergence itself. Finally, Theorems 2.2 and 2.3 allow to define a “sweet spot choice” of  $\epsilon$  for the numerical evaluation of  $\mathcal{W}_\epsilon$  as discussed in the following section.

## 2.3 Computation of discrete OT

The following primal problem is the discrete counterpart to (10) for the entropic regularization

$$\mathcal{W}_{c,\epsilon}(\mu_n, \nu_m) = \min_{\pi \in \mathcal{U}_{nm}} \langle \pi, C \rangle + \epsilon \langle \pi, \log \pi - \mathbf{1}_{n,m} \rangle, \quad (25)$$

where  $\mathbf{1}_{n,m} \in \mathbb{R}^{n,m}$  is a matrix with all entries equal to one and component-wise application of the logarithm. Defining the kernel matrix  $K := \exp(-C/\epsilon)$  with component-wise application of the exponential, recalling the definition of weights  $a$  and  $b$  as in Section 2.1, the dual problem reads

$$\sup_{f \in \mathbb{R}^n, g \in \mathbb{R}^m} - \langle a, f \rangle - \langle b, g \rangle - \epsilon \langle \exp(-f/\epsilon), K \exp(-g/\epsilon) \rangle. \quad (26)$$

The primal and the dual formulation are linked via the optimality conditions

$$\begin{cases} \pi &= \operatorname{diag}(\exp\{-f/\epsilon\}) K \operatorname{diag}(\exp\{-g/\epsilon\}), \\ a &= \operatorname{diag}(\exp\{-f/\epsilon\}) K \exp\{-g/\epsilon\}, \\ b &= \operatorname{diag}(\exp\{-g/\epsilon\}) K^\top \exp\{-f/\epsilon\}. \end{cases} \quad (27)$$

These form the starting point for the numerical computation. The classical Sinkhorn–Knopp algorithm [22] relies on a fixed point iteration of the function

$$\Phi(u, v) = \begin{bmatrix} b \circledast K^T v \\ a \circledast K u \end{bmatrix} \in \mathbb{R}^{n+m},$$



where  $\odot$  denotes pointwise division of vectors. This scheme is related to the last two conditions in (27) with  $(u^*, v^*)$  denoting the fixed point with  $u^* = \exp(f^*/\epsilon)$ ,  $v^* = \exp(g^*/\epsilon)$  and the unique solution to (26)  $f^*, g^*$ . Let  $\delta$  be a required tolerance for the iteration error. Then the Sinkhorn algorithm outputs an approximated transportation plan in  $\mathcal{O}(\delta^{-2}n^2\|C\|_\infty^2 \ln n)$  iterations [3, 28]. Alternatively, (25) or (26) might be solved with other solvers such as an “Adaptive Primal-Dual Accelerated Gradient Descent” [28] in  $\mathcal{O}\left(\min\left\{\delta^{-1}n^{9/4}\sqrt{\|C\|_\infty \ln n}, \delta^{-2}n^2R\|C\|_\infty \ln n\right\}\right)$  or with a Newton scheme [12] which results in local quadratic convergence.

Since the regularization parameter  $\epsilon$  causes numerical instabilities in the iterative process as  $\epsilon \rightarrow 0$ , a stabilized Sinkhorn algorithm variant has been formulated in log space by several authors [28, 70].

With a large number of point data  $n, m \gg 1$ , these classic formulations may still become cumbersome or infeasible. Hence, modern state-of-the-art computations of regularized optimal transport or Sinkhorn divergences rely on  $\epsilon$ -scaling heuristics, adaptive kernel truncation or multiscale schemes [70], e.g. enhanced by  $k$ -mean clustering [38]. A very promising recent approach relies on streamed sample data in a so-called online Sinkhorn scheme [60] to better approximate the original continuous regularized optimal transport problem with a sampling complexity arbitrarily close to  $\mathcal{O}(1/N)$ . From the vast amount of codes for computational transport, we only refer to `GeomLoss` by Jean Feydy [39], with embarrassingly parallel GPU implementations and a reference online implementation with linear memory footprint. The present work relies on this library for the numerical experiments, which allows to compute measure fitting applications with efficient computation of the gradients with sample data  $n > 1 \times 10^6$  within seconds.

### 3 Model class and stochastic reference coordinate system

The following section is devoted to the design of the model class  $\mathcal{M}$  and the underlying stochastic reference coordinate system  $X$ , defining the parameter dependent random variable  $\mathcal{M}(X)$  that we aim to fit close to  $Y$  in distribution. The first section is devoted to the actual dimension of the random variable  $X$ . Then we will introduce our model class based on compressed multi-element generalized polynomial chaos. Finally, we formulate the optimization problem and discuss involved challenges.

#### 3.1 Dimensionality

It is inevitable to choose a suitable reference coordinate system for an efficient approximation of the unknown  $Y$  in distribution. Given a sufficient amount of samples of  $Y$  such that empirical marginals can be inferred with some confidence, the question is if we can make an informed guess about the dimensions of the reference coordinate system  $X$ . It turns out that at first glance, the choice is very flexible as stated next.

**Theorem 3.1.** *Let  $X, Y$  be a continuous random vectors with images  $\text{img } X \subset \mathbb{R}^M$  and  $\text{img } Y \subset \mathbb{R}^N$  for  $1 \leq N, M < \infty$  such that there exists an invertible transformation  $T_Y$  with uniformly distributed  $T_Y(Y)$  on  $[0, 1]^M$ . Then there exists  $f: \mathbb{R}^M \rightarrow \mathbb{R}^N$ , such that  $Y \stackrel{d}{=} f(X)$ .*

*Proof.* Consider transformations  $T_X$  and  $T_Y$  such that  $T_X(X)$  and  $T_Y(Y)$  are uniformly distributed over  $[0, 1]^M$  and  $[0, 1]^N$ , respectively. Those transformations may be constructed via the marginal cumulative distribution functions as in the Rosenblatt transformation [54]. Since  $[0, 1]^M$  and  $[0, 1]^N$  have same cardinality, there exists a bijection  $\Phi: [0, 1]^M \rightarrow [0, 1]^N$ . Define the random variable  $Z := \Phi(T_X(X))$  with image  $\text{img}(Z) \subset [0, 1]^N$ . Furthermore, let  $T_Z$  be a transformation such that  $T_Z(Z)$  is uniformly distributed on  $[0, 1]^N$ . Then by construction,

$$Y \stackrel{d}{=} f(X), \quad f := T_Y^{-1} \circ T_Z \circ \Phi \circ T_X. \quad \square \quad (28)$$

We hence found a representation of a continuous  $N$ -dimensional random vector based on an  $M$ -dimensional stochastic reference coordinate system  $X$ . Note that  $f$  in general may not be bijective, highly non-linear and not necessarily continuously differentiable. Moreover, the choice of  $f$  is not unique.

In practice, we may look for the components of  $Y$  that are highly correlated and choose the dimension of  $M$  dependent on  $N$  and the number of highly correlated components. This is motivated by the fact that the image of highly correlated components might be close to a lower-dimensional structure and thus easier to be approximated in distribution in a lower-dimensional reference coordinate system.

As the other extreme,  $Y$  may consist of fully uncorrelated components but has a simple one-dimensional dependency to be explored. This fact is discussed in the following example.

**Example 3.2.** *Let  $X \sim \mathcal{U}(-\pi, \pi)$ . Define  $Y := (\sin(X), \cos(X), \sin(2X), \dots, \cos(nX))$  with images in  $\mathbb{R}^{2n}$  for  $n \in \mathbb{N}$ . Then all components of  $Y$  are uncorrelated, in particular  $\mathbb{E}[Y_i, Y_j] = c_{ij} \delta_{ij}$  for some constant  $c_{ij} \in \mathbb{R}$ .*

#### 3.2 Polynomial chaos expansion and tensor train ring compression

Let  $X \sim \mu = \mu_0 \otimes \dots \otimes \mu_0$  be an  $M$ -dimensional finite reference coordinate system with  $X = (X_1, \dots, X_M)$  and independent reference random variables  $X_i \sim \mu_0, i = 1, \dots, M$  such that  $\mu_0$

has finite arbitrary moments. Let  $\{p_i, i \in \mathbb{N}_0\}$  with  $p_0 \equiv 1$  be a set of orthogonal polynomials w.r.t. to  $\mu_0$ , that is  $\mathbb{E}[p_i(X_1)p_j(X_1)] := \delta_{ij}$ . We define the set of tensorized stochastic polynomials

$$\{P_\alpha, \alpha \in \mathbb{N}_0^M\}, \quad P_\alpha(X) = \prod_{i=1}^M p_{\alpha_i}(X_i). \quad (29)$$

These polynomials are orthogonal w.r.t.  $\mu$ , in particular  $\mathbb{E}[P_\alpha(X)P_\beta(X)] = \delta_{\alpha\beta}$  for  $\alpha, \beta \in \mathbb{N}_0^M$ . Such type of polynomial construction typically arise in the context of stochastic Galerkin schemes with convergence in  $L^p(\Omega, \mu, \mathcal{V})$  for some separable Banach space  $\mathcal{V}$  [43, 83, 72, 35, 27] even for the case  $M = \infty$ . Given an unknown random variable  $Y$  taking values in  $\mathbb{R}^N$ , we look for an approximation in distribution of the form

$$Y \stackrel{d}{\approx} Y_{\text{PCE}} := \sum_{\alpha \in \mathbb{N}_0^M} C[\alpha] P_\alpha(X), \quad C: \mathbb{N}_0^M \rightarrow \mathbb{R}^N. \quad (30)$$

Practically, we aim for a truncated version of the polynomial chaos expansion. Consider a finite index set  $\Lambda \subset \mathbb{N}_0^M$ ,  $0 \in \Lambda$  with  $0 \in \mathbb{N}_0^M$  and  $|\Lambda| < \infty$ . We define the truncated polynomial chaos expansion representation  $Y_\Lambda \stackrel{d}{\approx} Y_{\text{PCE}}$  by

$$Y_\Lambda := \sum_{\alpha \in \Lambda} C_\Lambda[\alpha] P_\alpha(X), \quad C_\Lambda: \Lambda \rightarrow \mathbb{R}^N. \quad (31)$$

The representation (31) may suffer from the curse of dimensionality if the set  $\Lambda$  grows exponential in  $M$ . In order to see this, consider the tensorized index set

$$\Lambda = \bigtimes_{m=1}^M \Lambda_m, \quad \Lambda_i \subset \mathbb{N}_0. \quad (32)$$

If  $|\Lambda_m| = d \in \mathbb{N}$  for  $m = 1, \dots, M$ , then the tensor  $C_\Lambda$  has  $d^M$  inputs, resulting in a total of  $Nd^M$  degrees of freedom in this approximation class. A technique to circumvent such exponential complexity can be found with tensor compressions by so-called *low-rank formats*. Popular tensor formats in the literature are tensor trains (TT) [64] or hierarchical tensors (HT) [49] for  $N = 1$ .

In the numerical investigations we observe an unfavourable exponential growth of ranks with growing  $N$ . To alleviate this, we consider a *tensor network format* as our compression tool, namely the *tensor train ring* (TTR). To make this concrete, define  $r_0 = r_M = 1$  and let  $r_m \in \mathbb{N}$  for  $m = 1, \dots, M-1$  and  $d_m = |\Lambda_m|$ . Moreover, choose  $C_\Lambda$  of the form

$$C_\Lambda[\alpha] = C_{\Lambda_1}[\alpha_1] \bullet \dots \bullet C_{\Lambda_M}[\alpha_M], \quad C_{\Lambda_j}[\alpha_j]: \mathbb{N}^{r_{j-1}, r_j} \rightarrow \mathbb{R}^N. \quad (33)$$

Here,  $\bullet$  denotes the contraction between adjacent indices and the Hadamard product w.r.t. to the output dimension  $N$ . Specifically, for  $A: \mathbb{N}^{q,r} \rightarrow \mathbb{R}^N$ ,  $B: \mathbb{N}^{r,s} \rightarrow \mathbb{R}^N$  and with Hadamard product  $\odot$ ,

$$A \bullet B = \sum_{k=1}^r A[:, k] \odot B[k, :] : \mathbb{N}^{q,s} \rightarrow \mathbb{R}^N. \quad (34)$$

We may identify the output dimension  $N$  with an additional index  $i$  as a tensor representation. With some abuse of notation, we let  $C_\Lambda[i, \alpha] = (C_\Lambda[\alpha])_i$  such that (33) can be rewritten as

$$C_\Lambda[i, \alpha] = C_{\Lambda_1}[i, \alpha_1] \cdot \dots \cdot C_{\Lambda_M}[i, \alpha_M] \in \mathbb{R}, \quad i = 1, \dots, N. \quad (35)$$

Thus, the components  $C_{\Lambda_1}$  define tensors of order 4 and we compactly use the notation and ordering of indices as

$$C_{\Lambda_m} = (C_{\Lambda_m}[k_{m-1}, i, \alpha_m, k_m]) \in \mathbb{R}^{r_{m-1}, N, d_m, r_m}, \quad (36)$$

where  $i = 1, \dots, N$  resorts to the second and  $\alpha_m \in \Lambda_m$ ,  $|\Lambda_m| = d_m$  resorts to the third index, respectively. In summary, we seek a tensor ring representation of rank  $\mathbf{r} = (r_0, \dots, r_M)$  with  $k_0 = k_M = 1$  of the form

$$Y_\Lambda = \sum_{k_1=1}^{r_1} \dots \sum_{k_{m-1}=1}^{r_{m-1}} \bigodot_{m=1}^M C_{\Lambda_m}[k_{m-1}, :, \alpha_m, k_m] p_{\alpha_m}(X_m), \quad (37)$$

where  $:$  denotes the vector representation of output dimension  $N$ . Note that while the representation in (37) looks complicated at first glance, it is only a contraction of tensors of order 4, which can be realized efficiently in practice using the Einstein summation convention. Once such a representation (37) is available, we can conveniently compute statistics such as moments in a sampling-free manner. This is due to the separated structure in the dependence of the input reference coordinate system  $X = (X_1, \dots, X_m)$  and the required high-dimensional quadrature then corresponds to index contractions. Let  $r_m = r$  for  $m = 1, \dots, M-1$  and  $d_m = d$  for  $m = 1, \dots, M$ . Then the proposed tensor train ring format has  $NMdr^2$  degrees of freedom, meaning a linear dependence on the input dimension  $M$  related to  $X$  and effectively circumventing the curse of dimensionality. We remark at this point that the tensor train ring does not yield a closed format with respect to the Frobenius norm. Here, we consider approximations in distribution only based on the Sinkhorn divergence, which introduces a weaker topology acting on the tensor train ring.

For a more specific ‘‘dependence pattern’’ of  $Y_\Lambda$ , we can set entries in the components  $C_{\Lambda_m}$  denoted as *degrees of freedom* to zero and thus specify the dependence of the output. As an example, for given  $i \in \{1, \dots, N\}$ , setting

$$C_{\Lambda_m}[k_{m-1}, i, \alpha_m, k_m] = 0, \quad k_{m-1} = 1, \dots, r_{m-1}, \alpha_m = 1, \dots, d_m, k_m = 1, \dots, r_m,$$

corresponds to no dependence of  $(Y_\Lambda)_i$  on  $X_m$ . In addition to the practical meaning of the model, this also reduces the number of degrees of freedom in the optimization process. Special cases involve that each component in  $Y_\Lambda$  is modeled independently with a univariate polynomial chaos expansion or a triangular parameter dependence structure as in Knothe-Rosenblatt transport maps in the case of  $N = M$ , see [84] for a recent result.

### 3.2.1 Multi-element polynomial chaos

A straight-forward generalization of the polynomial chaos as model class is the multi-element generalized polynomial chaos [80] described in this section. Let  $X = (X_1, \dots, X_M)$  denote a reference coordinate system with image  $\mathcal{X} = \text{img}(X) \subset \mathbb{R}^M$  and density  $f_X$ . Moreover, let  $\mathcal{X}^s$  denote a disjoint partition of  $\mathcal{X}$  for  $s = 1, \dots, S \in \mathbb{R}$ . Assume that the underlying distribution  $\mu$  of  $X$  has arbitrary finite moments, then one can define a set of functions  $\{P_\alpha^s : \alpha \in \mathbb{N}_0^M\}$  that are piecewise polynomials on  $\{\mathcal{X}^s\}$  with support on  $\mathcal{X}^s$  and also orthonormal with respect to the expectation. In particular, it holds for  $\alpha, \beta \in \mathbb{N}_0^M$ ,  $s, s' \in \mathbb{N}$  that

$$\mathbb{E}[P_\alpha^s(X)P_\beta^{s'}(X)] = \int_{\mathcal{X}^s \cup \mathcal{X}^{s'}} \chi_{\mathcal{X}^s \cap \mathcal{X}^{s'}}(x) P_\alpha^s(x) P_\beta^{s'}(x) f_X(x) d\lambda(x) = \delta_{\alpha, \beta} \delta_{s, s'},$$

where  $\chi$  denotes the indicator function.

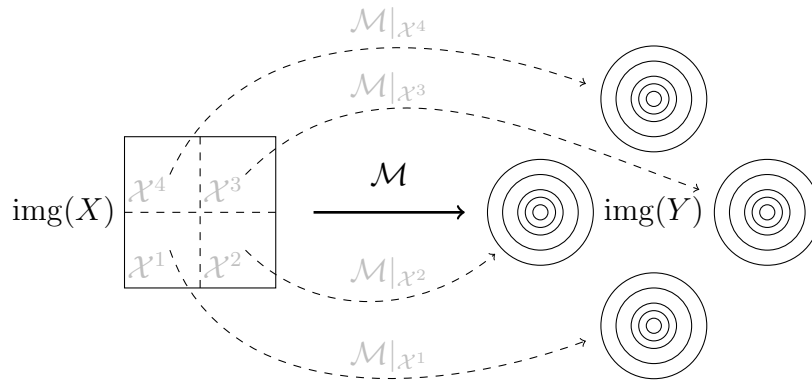


Figure 1: Sketch of a separated propagation of masses through a model class based on piecewise continuous representations, *e.g.* realized via multi-element polynomial chaos.

On the one hand, due to the locality of the function, a model class based on this notion is able to move masses of probability separately subject to the partition, which is illustrated in Figure 1. On the other hand, if  $\mathcal{X}$  is compact and  $Y$  is a multimodal random variable such that each accumulation of mass is completely disjoint, there exists no continuous model class  $\mathcal{M}$  such that  $Y \stackrel{d}{=} \mathcal{M}(X)$  since probability mass would have to be split. This in turn motivates the usage of piecewise continuous approximation classes.

In practice, to identify multimodal behaviour, one may utilize methods based on  $k$ -means clustering to determine the number of modes  $k$ , *e.g.* using the G-means algorithm of [50]. This can then be used to define an adequate partitioning of  $S$ . Note that the definition of the partition and thus the structure of non-continuity can be chosen arbitrarily as long as the resulting generation of reference coordinate samples is well balanced across the subdomains. One should specifically avoid situations where only few samples are in some  $\mathcal{X}^s$ , resulting in limited available information in the measure fitting application.

### 3.3 The optimization problem

In the following the optimization procedure is described that allows to obtain a transport representation as described before. Let  $\theta$  denote the degrees of freedom of the parametric model class  $\mathcal{M}(\cdot) = \mathcal{M}[\theta](\cdot)$ , which *e.g.* could be a tensor train ring format with underlying (multi-element) polynomial chaos as feature functions. Here,  $\theta$  relates to the sub-parts of the tensor train ring decomposition of the coefficient tensor as in (36). For a given stochastic coordinate system  $X$  and a batch of realizations  $\mathbf{Y} \in \mathbb{R}^{n,N}$  of the random variable  $Y$ , compute  $m$  samples of  $X$  represented as a batch of samples  $\mathbf{X} \in \mathbb{R}^{m,M}$ . Then, for fixed  $\theta$ ,  $\mathcal{M}[\theta][\mathbf{X}] \in \mathbb{R}^{m,N}$  defines a batch of samples that can be compared to  $\mathbf{Y}$  in terms of empirical measures  $\mu_n$  and  $\nu_m = \nu_m[\theta]$  and the Sinkhorn divergence. The respective optimization problem becomes

$$\min_{\theta} \mathcal{S}_{\epsilon}(\mu_n, \nu_m[\theta]). \quad (38)$$

Based on the dependence structure of  $\mathcal{M}$  on  $\theta$ , (38) defines a non-convex non-linear optimization problem with an almost everywhere differentiable function. Note that all  $\theta$  such that  $\mathcal{M}[\theta] \equiv 0$  define non-smooth points. We demonstrate the non-convexity and non-linearity for a very simple model class with linear parameter dependence in Example 3.3.

**Example 3.3.** Let  $X \sim \mathcal{U}(-1, 1)$  be uniformly distributed and denote by  $L_k$  the  $k$ -th Legendre polynomial defined on  $[-1, 1]$ . Moreover, let  $Y = L_2(X)$  and the model class  $\mathcal{M}(x) = c_0 + c_2 L_2(x)$ . Let  $\theta = [c_0, c_2]$ , then  $\theta^* = [0, -1]$  and  $\theta^* = [0, 1]$  define local minima of the function  $\theta \rightarrow \mathcal{S}_\epsilon(\mu_n, \nu_m[\theta])$  up to statistical noise introduced by the finite number of samples  $n, m \in \mathbb{N}$ . The left plot in Figure 2 illustrates this example in log-scale.

**Proposition 3.4.** Then  $f(X_1) = f(X_2)$  in distribution. Moreover, if  $f$  is odd, it holds that  $f(X_1) = -f(X_2)$  in distribution.

*Proof.* Since  $f$  is odd it holds that  $f(-x) = -f(x)$  for  $x \in \Xi$ . Given that  $X_1 = X_2$  in distribution and  $X_1 = -X_1$  in distribution since  $\mathcal{D}(\mathbb{R}^n)$  is symmetric, it follows  $f(X_1) = f(-X_1) = -f(X_1) = -f(X_2)$ .  $\square$

As a consequence, given an orthogonal model class with respect to the reference coordinate system  $X \sim \mathcal{D}(\mathbb{R}^n)$  of a symmetric distribution  $\mathcal{D}(\mathbb{R}^n)$ , we can decompose like

$$\mathcal{M}(x) = \sum_{\alpha \in \text{ODD}} C[\alpha] P_\alpha(x) + \sum_{\alpha \notin \text{ODD}} C[\alpha] P_\alpha(x). \quad (39)$$

Here  $\text{ODD} \subset \Lambda$  denotes the set of indices corresponding to odd polynomials. Due to Proposition 3.4, we can then choose  $C[\alpha] \geq 0$  for  $\alpha \in \text{ODD}$ . This may reduce the number of local minima significantly, in particular for  $d$  getting large. Typical examples for symmetric distributions are the uniform distribution on  $(-1, 1)^d$  and the standard normal distribution on  $\mathbb{R}^d$ .

**Example 3.5.** Let again  $X \sim \mathcal{U}(-1, 1)$  and  $L_k$  denote the  $k$ -th Legendre polynomial defined on  $[-1, 1]$ . Assume  $Y = L_1(X) + L_2(X)$  and the model class  $\mathcal{M}$  be given as  $\mathcal{M}(x) = c_1 L_1(x) + c_2 L_2(x)$ . The right plot in Figure 2 illustrates the example with multiple local minima in log-scale, including the mentioned reduction effect.

**Proposition 3.6.** Let  $\mathcal{M}$  be a model class as in Section 3.2 with orthonormal basis of increasing degree encoded in  $\alpha \in \mathbb{N}_0^d$  w.r.t. the underlying stochastic coordinate system  $X$ . Then,  $C[:, 0] = \mathbb{E}[\mathcal{M}(X)]$ .

*Proof.* The proof follows by construction, since the basis is orthonormal w.r.t. the associated measure  $\mu$  of  $X$ .  $\square$

Consequently, given  $N$  samples of the unknown random variable  $Y$ , we can compute the empirical mean and bound  $C[:, 0]$  around the empirical mean value in the optimization routine. In particular, in the numerical experiments the obtained optimized parameter always leads the model class to match the empirical mean exactly. This observation is in agreement with the best possible sample rate of  $1/\sqrt{n}$  obtained for large  $\epsilon$  since it matches the convergence rate of the empirical mean to the exact mean.

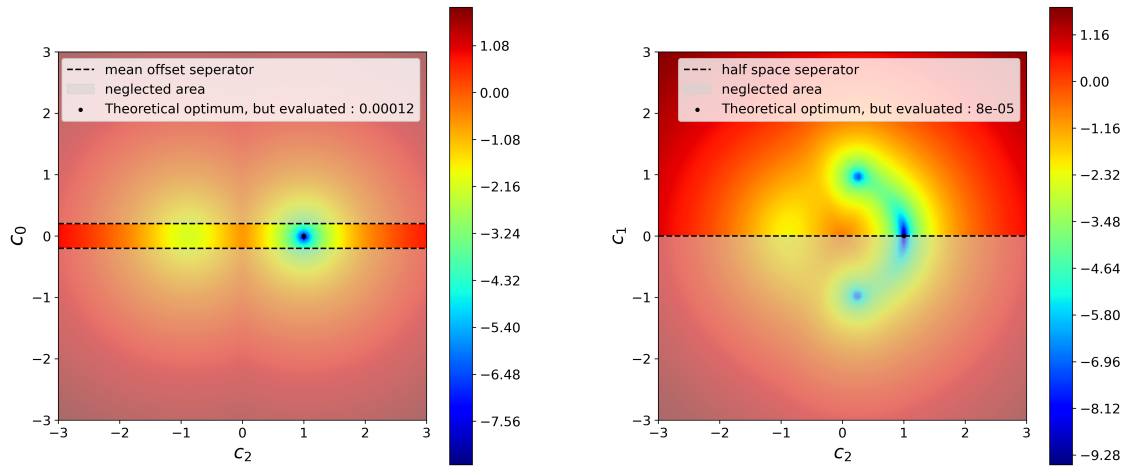


Figure 2: Plots of  $\theta \rightarrow \mathcal{S}_\epsilon(\mu_n, \nu_m[\theta])$  in log-scale for Example 3.3 (left) and Example 3.5 (right) showing the presence of multiple local minima and the effect of neglecting areas in the parameter domain motivated by Proposition 3.4 and 3.6.

## 4 Numerical experiments

In this section we illustrate the performance of the approach presented above with some high-dimensional applications. Model classes based on polynomial chaos, its multi-element extension and a compression based on the tensor train ring format are used.

First, we examine the multimodal random variables, where we point out the necessity of the non-continuous model class to represent isolated accumulation of probability masses. Second, a simple random field discretization is carried out as a toy example to represent random fields with non-linear but smooth dependence of the stochastic input variable. Finally, from a non-periodic random micro-material, an effective upscaled piecewise constant random macro-material is constructed as porosity coefficient of a diffusion problem.

For the numerical realization, several open source software packages are used. The python package `geomloss`[39] provides the calculation of the debiased Sinkhorn loss and gradient information. The minimization with automatic differentiation is carried out with `pytorch-minimize`. The compressed model class is implemented with our python package `TensorTrain`[48] while multi-element polynomial chaos and sample generation is realized with `Alea`[31]. For the numerical upscaling experiment, the composite structures are generated with our library `Bubbles`[47] and the numerical solution process of the corresponding partial differential equations is realized using `Fenics`[37]. Finally, `seaborn`[36] is used throughout this work for generating plots of the obtained numerical results.

### 4.1 Multimodal distribution

Multimodal distributions especially pose a challenge when representing random variables explicitly. In this section, we consider multimodal random variables  $Y$  modeled as a mixture of Gaussian distributions in  $N = 1$  and  $N = 2$  dimensions. We consider two cases, in the following denoted as *weakly* or *strongly disconnected*. By weakly disconnected we mean multimodal probability masses, such that

**Algorithm 1** Fitting of  $Y \stackrel{d}{\approx} \mathcal{M}(X)$ 


---

<b>Input:</b>	$\left\{ \begin{array}{l} \mu_n \text{ or } (Y^i)_{i=1}^n \subset \mathbb{R}^N, N \in \mathbb{N}, \\ X \sim \mathcal{D}(\mathbb{R}^M), M \in \mathbb{N}, \\ m \in \mathbb{N}, \\ \mathcal{M}(X) = \mathcal{M}[\theta](X) \text{ based on } \{P_\alpha^s\}, \\ (\mu, \nu) \mapsto \mathcal{S}_\epsilon(\mu, \nu) \in \mathbb{R}_+, \\ (\theta_0, \ell) \mapsto \theta^* = \mathcal{O}(\theta_0, \ell), \end{array} \right.$	$\triangleright$ <i>data samples</i> , $\triangleright$ <i>input stochastic coordinate system</i> , $\triangleright$ <i>number of samples drawn of <math>X</math></i> , $\triangleright$ <i>model class</i> , $\triangleright$ <i>Sinkhorn loss operator</i> , $\triangleright$ <i>local optimizer</i> .
---------------	--	---

**Output:** Trained model class with parameter  $\theta^*$  such that  $\mathcal{M}[\theta^*](X) \stackrel{d}{\approx} Y$ .

---

Compute base samples  $(X^j)_{j=1}^m \subset \mathbb{R}^M$ .  
Precompute evaluations  $P_\alpha^s(X)$  for  $s = 1, \dots, S, \alpha \in \Lambda$ .  
Define  $\theta \rightarrow \mathcal{M}[\theta](X)$  which defines  $\theta \rightarrow \nu_m[\theta]$ .  
Set up the loss functional  $\theta \rightarrow \ell(\theta) := \mathcal{S}_\epsilon(\mu_n, \nu_m[\theta])$ .  
Minimize  $\ell(\theta)$  using restarted BFGS and obtain  $\theta^*$ .

---

the mass in the transition area between them is not negligible, see Figure 3 for an illustration. Strongly disconnected then refers to isolated accumulations of multi-modal distributions, see the top right box in Figure 4 for an example.

#### 4.1.1 Weakly disconnected multimodality

In the first validation experiment, we consider the case of a weak disconnected multimodal behavior, where the term weak indicates that the mass between two modes is not negligible, i.e. the modes are not fully disconnected. Let  $T$  be a any diffeomorphic transport map with Lipschitz constant  $L > 0$  such that  $Y = T(X)$  for some random variable  $X$  with connected compact image. Note that as modes separate further and the connection becomes “thinner”, the Lipschitz constant of the map grows larger unboundedly.

Here we consider a one dimensional random variable  $Y$  as a mixture of the Gaussian distributions  $\mathcal{N}(-2, 1)$  and  $\mathcal{N}(1, 0.5)$ , i.e.  $N = 1$ . Assume the reference coordinate system  $X \sim \mathcal{U}([-1, 1]^2)$  and recall the tensor ring format from Section 3.2 with global Legendre polynomials of various degree as model class. In this setting,  $M = 2$  and the model class does not coincide with (an approximation of) a classical transport map. We motivate the choice of  $M > N$  to allow for a compressed representation of the functional representation  $\mathcal{M}(X)$ . In particular in the computation using  $M = N$ , a polynomial degree of 81 was required to obtain similar results as in the right plot of Figure 3. This has to be compared to the degrees of freedom in the compressed format given as  $M(d + 1)r$ , where  $d$  denotes the polynomial degree and  $r$  denotes the rank of the tensor train ring.

Figure 3 illustrates the resulting fitted random variable with respect to the debiased Sinkhorn loss  $\mathcal{S}_\epsilon$  with  $\epsilon = 0.05$ . We note that the model class is challenged by accurately representing the “valley” between the modes. It can be observed that the approximation quality increases only slightly in this area as the polynomial degree increases.



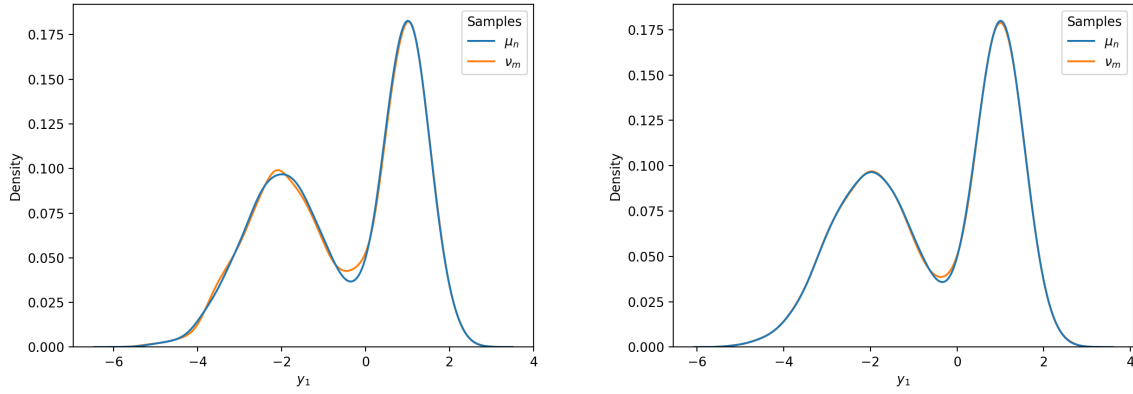


Figure 3: Tensor train ring fitting of a multimodal distribution. Reference coordinate system  $X = (U_1, U_2)$   $U_1, U_2 \sim \mathcal{U}(-1, 1)$  *iid* with tensorized Legendre chaos with tensor train ring rank  $r = (1, 3, 1)$  based on  $K = 1 \times 10^5$  samples. **Left:** Polynomial degree 4 with a minimized debiased Sinkhorn loss of  $2.0 \times 10^{-4}$ . **Right:** Polynomial degree 8 with a minimized debiased Sinkhorn loss of  $3.6 \times 10^{-5}$ .

#### 4.1.2 Strongly disconnected multimodality

When considering multimodal densities, for disjoint multimodalities with a stochastic input coordinate system  $X$  and connected image, the “transport map” inevitably exhibits discontinuities to split the separated probability mass, as mentioned in Section 3.2.1

As a model problem for  $Y$  we again consider a mixture of Gaussian distributions. For a number  $B \in \mathbb{N}$  of modes, let

$$\Theta[B] = \left\{ 0, \dots, \frac{2\pi B}{B-1} \right\}.$$

Then  $Y$  is defined as mixture of

$$\mathcal{N}(\mu_{S,\theta}, \sigma_B^2 I_2), \quad \mu_{S,\theta} = (S \cos(\theta), S \sin(\theta))^T, \quad \theta \in \Theta[B],$$

with shift parameter  $S > 0$  and variance  $\sigma_B^2 > 0$  depending on  $B$  to ensuring disconnection.

As a first setup we consider  $B = 4$  with  $S = 6$  and  $\sigma_B^2 = 1$  based on a partition of  $[-1, 1]^2$  into  $2 \times 2$  squares. The numerical results are shown in Figure 4. The second setup involves  $B = 8$  modes using a partition of  $[-1, 1]^2$  into  $2 \times 4$  squares with shift  $S = 6$  and smaller variance  $\sigma_B^2 = 0.1$ .

The results are visualized by means of the kernel density estimation in the `seaborn` package. Motivated by Proposition 3.6, the degrees of freedom associated to the (local) mean values are set as start value within the iterative minimizing process, while the remaining values are initialized randomly with standard normal distribution. In particular, we set the degrees of freedom associated with polynomial coefficients for the constant contributions to the means  $\mu_{S,\theta}$ . All other coefficients are randomly initialized *iid* Gaussian with mean 0 and variance 0.01.

## 4.2 Random field with tensor train rings

We consider the random field

$$\kappa(x, \omega) = \cos(U_1(\omega)x + U_2(\omega)) + U_3(\omega), \quad x \in [0, 1] \quad (40)$$

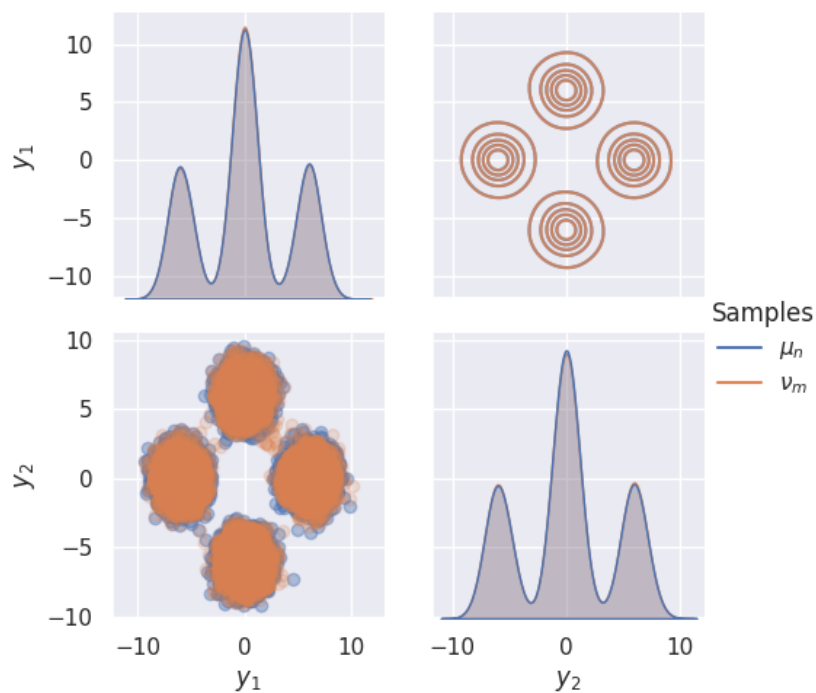


Figure 4: Verification of a multi-element polynomial chaos model class  $\mathcal{M}$  with stochastic reference coordinate system  $X = \mathcal{U}([-1, 1]^2)$  with  $2 \times 2$  elements. The empirical measures are based on  $n = m = 1.6 \times 10^4$  samples. Each local polynomial is a tensor product of orthonormal polynomials of degree 7. The final debiased Sinkhorn loss  $\mathcal{S}_\epsilon$  with  $\epsilon = 0.05$  is given by  $\mathcal{S}_\epsilon(\mu_n, \nu_m) \approx 6 \times 10^{-3}$ .

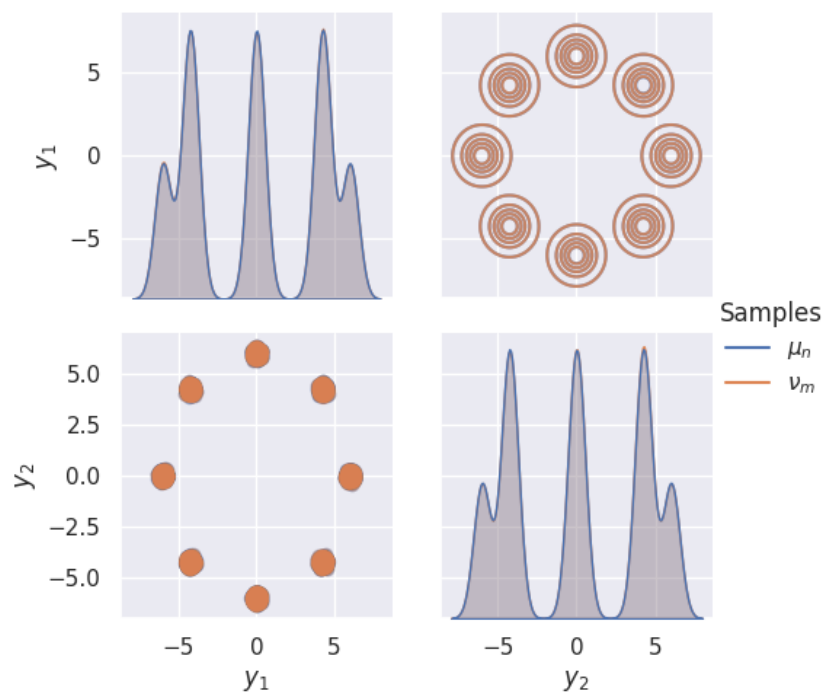


Figure 5: Verification of a multi-element polynomial chaos model class  $\mathcal{M}$  with stochastic reference coordinate system  $X = \mathcal{U}([-1, 1]^2)$  with  $2 \times 4$  elements. The empirical measures are based on  $n = m = 3.2 \times 10^4$  samples. Each local polynomial is a tensor product of orthonormal polynomials of degree 7. The final debiased Sinkhorn loss  $\mathcal{S}_\epsilon$  with  $\epsilon = 0.05$  is given by  $\mathcal{S}_\epsilon(\mu_n, \nu_m) \approx 2 \times 10^{-3}$ .

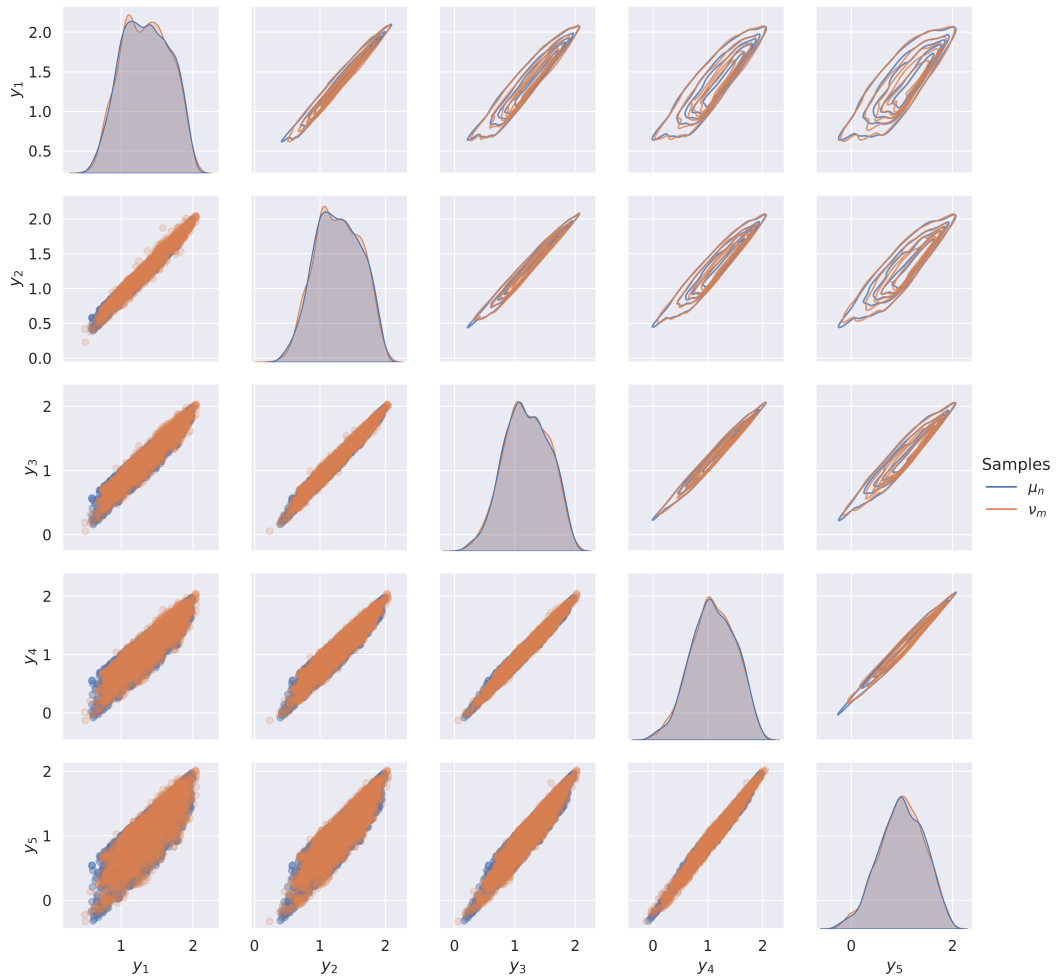


Figure 6: Marginals of the  $N = 5$  discretization points of the random field from (40) with a tensor train ring of rank  $r = (2, 2)$  using tensorized Legendre polynomials of degree  $(9, 9, 1)$  for  $X \sim \mathcal{U}([-1, 1]^3)$ , i.e.  $M = 3$  based on  $n = m = 2 \times 10^3$  samples with  $\epsilon = 0.05$  for the debiased Sinkhorn loss  $\mathcal{S}_\epsilon(\mu_n, \nu_m) \approx 7.8 \times 10^{-4}$ .

with *iid* random variables  $U_1, U_2, U_3 \sim \mathcal{U}(-1, 1)$ . For  $N \in \mathbb{N}$ , let  $x_i = (i - 1)/(N - 1)$ ,  $i = 1, \dots, N$  be a discretization of  $[0, 1]$  and define the  $\mathbb{R}^N$ -valued random variable  $Y = (Y_i)_i$  by

$$Y_i = \kappa(x_i, \omega), \quad i = 1, \dots, N.$$

Figures 6 and 7 depict the correlation of the marginals for  $N = 5$  and  $N = 10$ , respectively. The plots show the different levels of correlation intensity, which corresponds to the distance of the discretization points in  $[0, 1]^2$ . The model class successfully captures these details with very small tensor train ring rank of  $(2, 2)$  with a debiased Sinkhorn loss of  $7.8 \times 10^{-4}$ .

### 4.3 Application to numerical upscaling

The motivation for the following experiment is the simulation with materials exhibiting random non-periodic microstructures as shown in Figure 8. Such computations pose significant challenges since the large number of inclusions results in a large set of stochastic dimensions. This is due to the assumed parametrization, which determines the location as well as the shape of these inclusions.

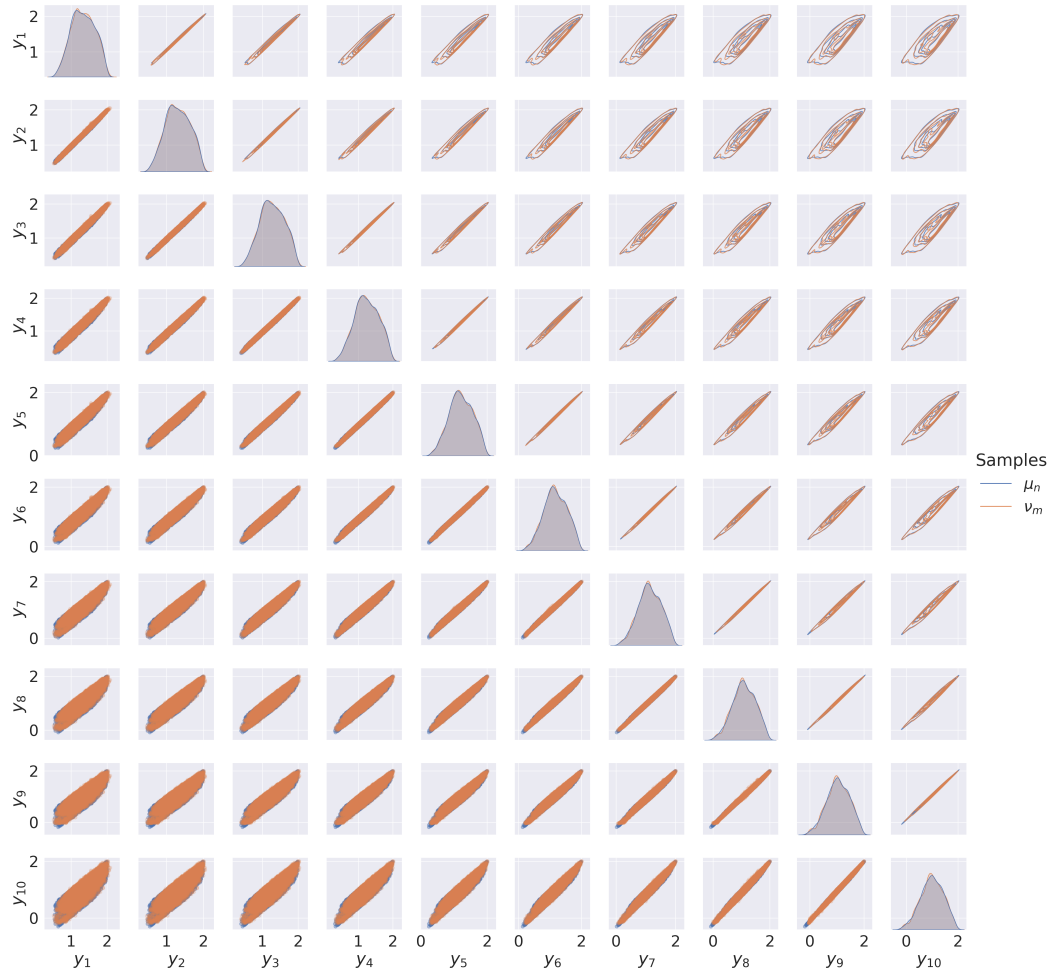


Figure 7: Marginals of the  $N = 10$  discretization points of the random field from (40) with a tensor train ring of rank  $r = (2, 2)$  using tensorized Legendre polynomials of degree  $(9, 9, 1)$  for  $X \sim \mathcal{U}([-1, 1]^3)$ , i.e.  $M = 3$  based on  $n = m = 5 \times 10^3$  samples with  $\epsilon = 0.05$  for the debiased Sinkhorn loss  $\mathcal{S}_\epsilon(\mu_n, \nu_m) \approx 6.3 \times 10^{-4}$ .

For a statistical analysis of problems with such data, one would have to consider a large number of realizations for accurate results, which becomes computationally expensive due to the microscopic structures that have to be resolved.

An alternative way we pursue here is to determine an effective macroscopic (“upscaled”) random field which leads to the same statistical system response for some quantity of interest. In particular, we use an approach of “numerical upscaling” to obtain (samples of) a macroscopic material description. As a result, we derive a functional representation of a *piecewise constant random field* as upscaled stochastic information. Note that this is in contrast to constant effective material descriptions, which are usually obtained from (stochastic) homogenization methods where information about stochastic fluctuations is lost for subsequent numerical computations.

The idea of functional representation of upscaled material is inspired by and extends the works in [73] based on an optimization with maximum likelihood estimators (MLE) and [69] where a functional representation is obtained by means of Kalman filters in a Bayesian context. Here we obtain a functional representation by minimizing the debiased Sinkhorn divergence that metricizes the space of probability distributions. As a result, based on the chosen model class we are able to fit random variables that are close to the observed samples in the sense of distribution. This is fundamentally different to the functional representation obtained in [73], which uses a tensorized approximation of the MLE, relying on an approximation of uncorrelated random variables, and the representation in [69], where the resulting random variable has matching mean and variance only.

We consider a Lipschitz domain  $D \subset \mathbb{R}^2$  on which a random composite field  $\kappa \in L^2(\Omega, \sigma, \mathbb{P}; L^\infty(D))$  is defined. This random field represents a model for random composite materials. Figure 8 depicts example realizations. The random material is described as 2-phase matrix composite material consisting of a matrix and inclusions. Each inclusion is a star-shaped domain, *i.e.* for  $\kappa_0, \kappa_1 \in \mathbb{R}_+$

$$\kappa(x, \omega) = \begin{cases} \kappa_0 & = x \in D_{\text{incl}}(\omega), \\ \kappa_1 & = x \in D \setminus D(\omega), \end{cases}$$

for a random domain  $D_{\text{incl}}(\omega) \subset D$  to be specified below. A sample of the random domain  $D_{\text{incl}}$  is drawn as a set of star-shaped random inclusions, which are parametrized by their boundary description and correlated by geometric constraints. Specifically, we assume that the set of inclusions is separated in the sense that the set of convex hulls of the polygonal approximations of the inclusions does not collide for a given prescribed discretization.

Each realization of a random inclusion  $\mathcal{I}(\omega)$  is given with an interface  $\partial\mathcal{I}(\omega)$  parametrized as

$$\partial\mathcal{I}(\omega) = \{P(\omega) + \rho(\theta, \omega)(\cos \theta, \sin \theta)^T, \theta \in [0, 2\pi]\} \subset \mathbb{R}^2$$

with center point  $P \sim \mathcal{U}(D)$  and radius  $\rho$  being a random field given as

$$\rho(\theta, \omega) = \rho_0(\omega) \exp \left( \sum_{\ell=1}^5 a_\ell(\omega) \sin(\ell\theta) + b_\ell(\omega) \cos(\ell\theta) \right)$$

with  $\rho_0 \sim \mathcal{U}(0.01, 0.1)$  such that for the realization  $\rho_0(\omega)$  we choose *i.i.d.*  $a_\ell, b_\ell \sim \mathcal{U}(-\rho_0(\omega), \rho_0(\omega))$  for  $\ell = 1, \dots, 5$ . Then a realization of  $\kappa$  is obtained by successively adding or rejecting of up to 40 inclusions that satisfy the geometric constraint.

Subsequently, we consider a numerical upscaling scheme as used in [78]. Let  $\delta D = \Gamma_D \dot{\cup} \Gamma_N$  with  $|\Gamma_D| > 0$  with disjoint Dirichlet boundary  $\Gamma_D$  and Neumann boundary segment  $\Gamma_N$ , respectively. We

consider a random partial differential equation given as

$$\begin{cases} -\operatorname{div} \mathcal{A}(x, \omega) \nabla \mathcal{U}(x, \omega) = f(x) & \text{a.s. in } D \times \Omega, \\ \mathcal{U}(x, \omega) = g(x) & \text{a.s. in } \Gamma_D \times \Omega, \\ \mathcal{A}(x, \omega) \partial_n [\mathcal{U}](x, \omega) = h(x) & \text{a.s. in } \Gamma_N \times \Omega, \end{cases} \quad (41)$$

with deterministic sufficiently regular data  $f, g, h$  such that there exists a unique weak solution  $\mathcal{U}(\cdot, \omega)$   $\mathbb{P}$ -almost everywhere for a given realization of a random field  $\mathcal{A}(\cdot, \omega)$  to be specified. Let  $N_s \in \mathbb{N}$  and consider a disjoint decomposition  $D = \bigcup_{s=1}^{N_s} D_s$ . For  $\omega \in \Omega$  and on each  $D_s$  choosing  $\mathcal{A} = \kappa$ , we solve the following auxiliary Dirichlet problems on the micro scale,

$$\begin{cases} -\operatorname{div} \kappa(x, \omega) \nabla u_j(x, \omega) = 0 & \text{a.s. in } D_s \times \Omega, \\ u_j(x, \omega) = x_j & \text{a.s. in } \partial D_s \times \Omega, \end{cases} \quad (42)$$

for  $j = 1, 2$ . We refer to Figure 8 for an illustration of a partition for different realizations of the microscopic random field  $\kappa$ . The effective macroscopic random permeability tensor field  $K$  is assumed to be piecewise constant in  $x$  with respect to the partition  $\{D_s\}$ , given by

$$[K(x, \omega)|_{D_s}]_{ij} \equiv K_{ij}^s(\omega) := \frac{1}{|D_s|} \int_{D_s} \kappa(x, \omega) \frac{\partial u_i}{\partial x_j} dx, \quad i, j = 1, 2.$$

Assuming that  $K$  is symmetric and possibly anisotropic on  $D_s$  for each  $s = 1, \dots, N_s$ , we can encode  $K$  as a random vector  $Y$  with values in  $\mathbb{R}^{3N_s}$  and realization given by

$$Y(\omega) = [K_{11}^1, K_{22}^1, K_{12}^1, \dots, K_{11}^{N_s}, K_{22}^{N_s}, K_{12}^{N_s}](\omega)^T. \quad (43)$$

While we can draw samples of  $Y$  representing the upscaled random tensor field  $K$  with the above concept, the actual distribution is unknown. This motivates the functional representation of  $Y$  which then enables to draw new samples very efficiently.

For the numerical investigation we considered two cases with  $N_s = 1$  and  $N_s = 4$  with  $D = [0, 1]^2$  and partition as illustrated in Figure 8. The corresponding output random variable  $Y$  from (43) representing the anisotropic behaviour thus is  $N = 3$  and  $N = 12$  dimensional, respectively.

In the case of  $N_s = 1$  with  $\operatorname{img}(Y) \subset \mathbb{R}^3$ , we use a  $M = 2$  dimensional reference coordinate system based on *iid* uniform random variables  $X_1, X_2 \sim \mathcal{U}(-1, 1)$  defining  $X = (X_1, X_2)$ . The model class  $\mathcal{M}$  is defined as a tensorized Legendre polynomial chaos of degree 9 in each stochastic mode. The associated coefficient tensor is represented as a tensor train ring with rank 6. The result of the fitting procedure with debiased Sinkhorn loss of  $2.4 \times 10^{-5}$  is shown in Figure 9 as corner plots.

For the second case of  $N_s = 4$  with  $\operatorname{img}(Y) \subset \mathbb{R}^{12}$ , we use a stochastic reference coordinate system  $X$  with  $M = 3$ . Figure 10 depicts corner plots of the data samples of  $Y$ , showing strong correlation structure only locally that is associated with the diagonal entries of the upscaled anisotropic tensor. To bridge the dimensional gap between  $N = 12$  and  $M = 3$ , we allow flexible non-linear behaviour in terms of a large polynomial degree of 15 and a tensor train ring rank of (9, 9). The resulting final debiased Sinkhorn loss is about  $9 \times 10^{-3}$ . In contrast to the  $N_s = 1$  case, we observe a slightly larger deviation between the data and the model class distribution fit for the resulting marginals.

Based on the obtained upscaled macroscopic random field  $K$ , we investigate the quality of the model class in terms of propagation and generalization. For this, we consider the differential equation (41) with  $\mathcal{A} = K$ . Let  $\mathcal{U}(\omega)$  denote the weak solution of (41) for a realization of the random field  $K(\omega)$ . Moreover, define the quantity of interest  $q: H^1(D) \rightarrow \mathbb{R}$  as the spatial mean value

$$q(u) := \frac{1}{|D|} \int_D v dx, \quad \forall u \in H^1(D).$$

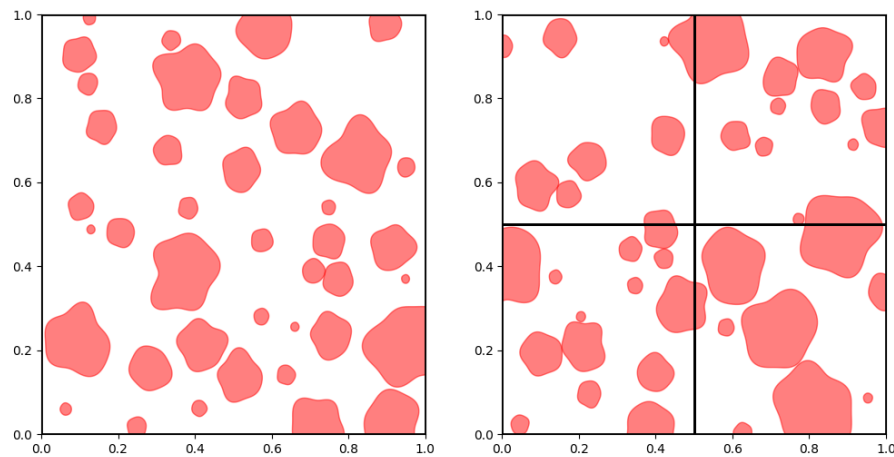


Figure 8: **Left:** Random composite sample with  $N_s = 1$  partition for the upscaling process. **Right:** Random composite sample with  $N_s = 4$  partition (black grid) used in the upscaling scheme.

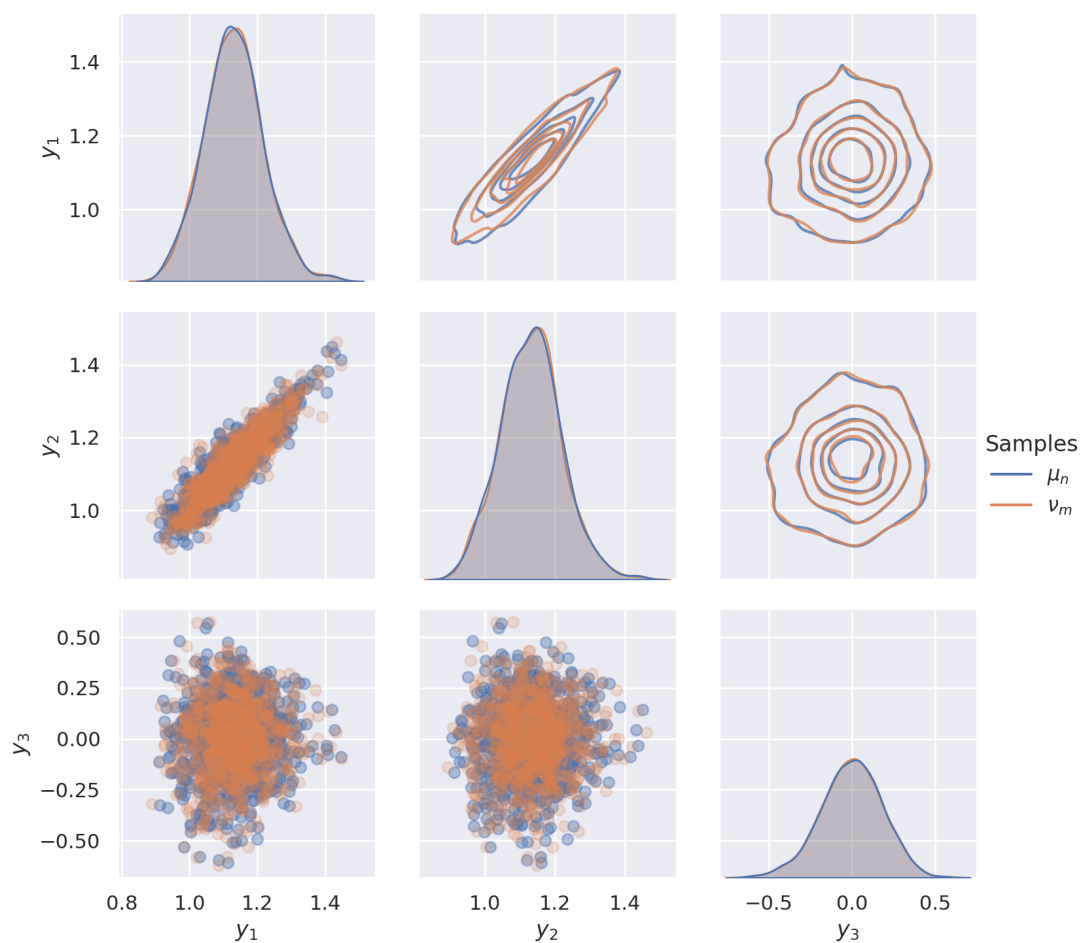


Figure 9: Upscaling experiment with  $N_s = 1$ ,  $N = 3$ ,  $M = 2$ , polynomial degree 9, and tensor train ring rank 6 with a resulting debiased Sinkhorn divergence of  $2.4 \times 10^{-5}$ . Here,  $Y_1$  and  $Y_2$  are scaled with 0.1 to improve the optimization performance.



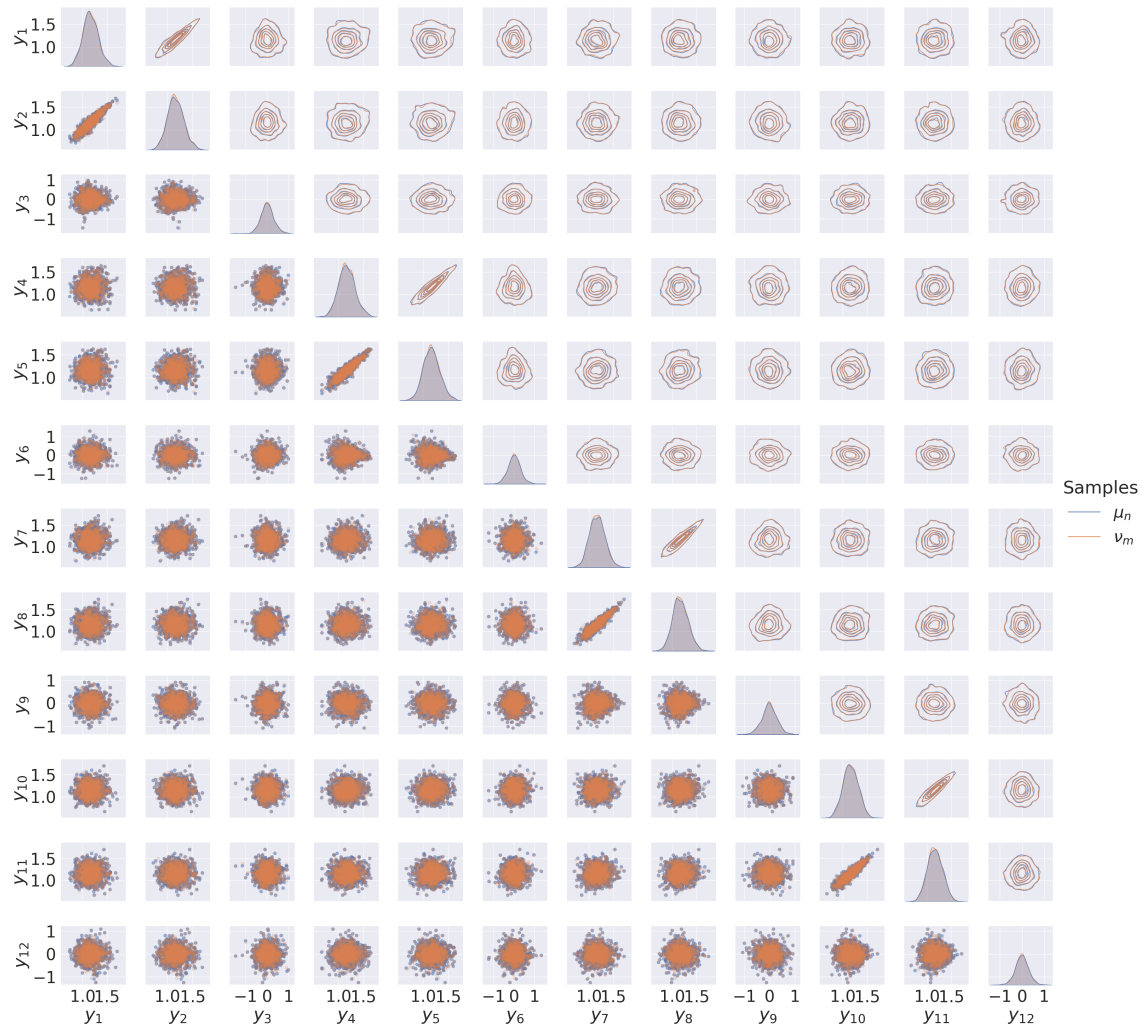


Figure 10: Upscaling experiment with  $N_s = 4$ ,  $N = 12$ ,  $M = 3$ , polynomial degree 15, and tensor train ring rank  $(9, 9)$  with a debiased Sinkhorn divergence of  $9 \times 10^{-3}$ .  $Y_k$  is scaled by 0.1 for  $k \bmod 3 \equiv 1$  or  $k \bmod 3 \equiv 2$ .

This in turn defines the random variable  $\bar{u}(\omega) := q(\mathcal{U}(\omega))$ .

We now propagate three different types of samples of  $K$  through the Darcy problem (1) for one subdomain  $N_s = 1$  yielding  $N = 3$ , and (2) for a decomposition into 4 subdomains  $N_s = 4$  yielding  $N = 12$ , respectively. First, the data samples of  $Y$  are used to define samples of  $K$  and thus samples of  $\bar{u}$ . We refer to these samples as *reference samples* propagated through  $q$ . Second, the underlying samples of  $X$  used to fit the model class  $\mathcal{M}$  are propagated through  $\mathcal{M}[\theta^*]$ , yielding approximate samples of  $Y$ . These are referred to as *model fitted samples* propagated through  $q$ . Third, new samples of the underlying reference coordinate system  $X$  are drawn, mapped through the fit in the model class  $\mathcal{M}[\theta^*]$  and eventually get propagated through  $q$ . Figure 11 shows the results of these experiments.

On the one hand, we observe that the samples used to optimize the surrogate in the model class in turn produce results that are very close to the propagated data sample distribution for both scenarios with  $N_s = 1$  and  $N_s = 4$ , which is to be expected. On the other hand, the propagation of newly generated samples results in a deviation of the resulting distribution for  $N_s = 4$ , while there is only sample noise deviation for  $N_s = 1$ . Table 1 and 2 depict the impact and strength of the deviation in terms of mean and variance. A plausible explanation for this deviation is the very small number of samples used and the insufficient expressiveness of the model class  $\mathcal{M}$  to “explain” the data sample distribution for  $N_s = 4$  with a resulting debiased Sinkhorn loss of  $2.9 \times 10^{-3}$  only, compared to the  $2.4 \times 10^{-5}$  loss value for  $N_s = 1$ .

Table 1: First and second order statistics of the 3 resulting distributions displayed in Figure 11. Case of one subdomain  $N_s = 1$ .







	samples		
			
mean	$7.41 \times 10^{-2}$	$7.41 \times 10^{-2}$	$7.43 \times 10^{-2}$
variance	$3.24 \times 10^{-5}$	$3.18 \times 10^{-5}$	$3.28 \times 10^{-5}$

Table 2: First and second order statistics of the 3 resulting distributions displayed in Figure 11. Case of 4 subdomains  $N_s = 4$ .

	samples		
			
mean	$7.26 \times 10^{-2}$	$7.25 \times 10^{-2}$	$7.55 \times 10^{-2}$
variance	$3.75 \times 10^{-5}$	$3.53 \times 10^{-5}$	$6.11 \times 10^{-5}$

## 5 Conclusion

In this work we consider a very flexible computational method to obtain functional representations of random variables that are determined based on samples only without any knowledge of the underlying distribution. The random variables (or fields) are expanded in a polynomial basis and optimized in an unsupervised way by means of debiased Sinkhorn losses that leads to a fitting of the discrete source and target measures. In order to represent the uncertainty in the approximation, we use a

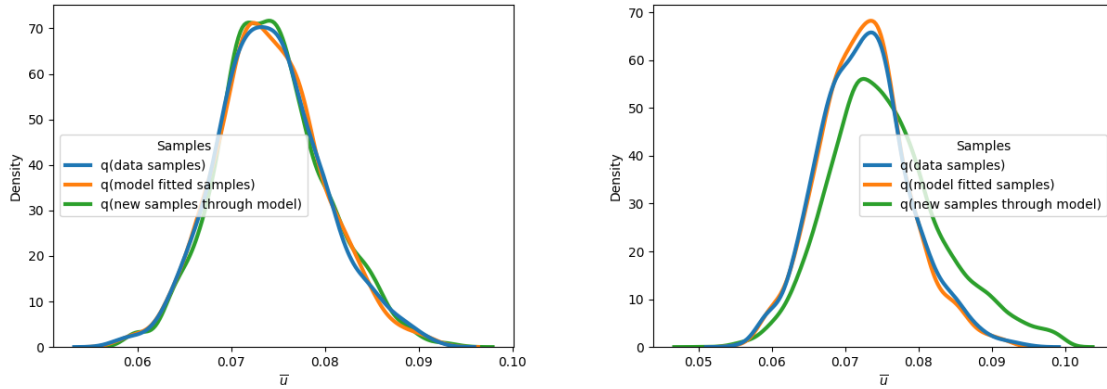


Figure 11: KDE plots of  $q(\mathcal{Y})$  for three sets of samples of  $\mathcal{Y}$ :  $1 \times 10^4$  data samples of  $Y$ , the  $1 \times 10^4$  propagated base samples through  $\mathcal{M}$  and  $1 \times 10^5$  new samples drawn from  $X$  again propagated through  $\mathcal{M}$  are used to obtain samples from  $\bar{u}$ . **Left:** Case of one subdomain  $N_s = 1$ . **Right:** Case of 4 subdomains  $N_s = 4$ .

multi-element polynomial chaos expansion as model class  $\mathcal{M}$ . This high-dimensional representation is compressed with a possibly very efficient low-rank compression with a tensor train ring format. The definition of (polynomial chaos) basis functions relies on the choice of the underlying stochastic input or reference coordinate system denoted as  $X$ , which is an  $M$ -dimensional random vector. Since the target random vector  $Y$  is  $N$ -dimensional, the model class  $\mathcal{M}$  representation can be seen a relaxation of classical transport structures for which  $N = M$  and  $\mathcal{M}$  has to be a diffeomorphism. We show that the relaxation in particular to discontinuous probability mass transport is necessary to obtain accurate functional representations of multi-modal measures.

While the optimization problem is non-convex and non-linear, the involved smoothness and high-speed computation of the debiased Sinkhorn loss allows for fast computation of accurate measure fittings through the parameter dependent model class in a distributional sense, based on second order schemes like BFGS or Hessian supported Newton schemes.

The developed technique is applied to several common challenging tasks in UQ. This includes the representation of a multi-modal distribution and an example of a smooth random field. Furthermore, a numerical stochastic upscaling scheme with functional representation of the macroscale random field is developed. This can be understood as a new approach within the framework laid out in [73], [69] and [14], where functional representations are obtained based on tensorized maximum likelihood losses, Kalman filtering in a Bayesian framework and optimization of Kullback-Leibner losses, respectively. It extends these works in the sense that the new technique allows for a fitting in distribution, whereas only finite moment fits are obtained in [69] or simplified approximations of maximum likelihood estimators of uncorrelated random variables are performed in [73], thus possibly ignoring any correlation structure between components of the samples.

As an outlook, the proposed unsupervised functional representation scheme can be applied and extended in various directions. A first extension concerns the used metric  $\bar{d}$  in the definition of the Wasserstein metric and in turn of the debiased Sinkhorn divergence. While in this work we considered the Euclidean norm case, choosing  $\bar{d}$  to be a Sobolev norm allows to learn random fields arising in random partial differential equations in distribution. This idea extends the method presented in [34] based on empirical expectation losses.

A second extensions is related to Bayesian inference. Here, we aim to obtain posterior information

encoded in a functional approximation of a random variable as in [69], which corresponds to  $N = M$  in our setting. This representation allows the incorporation of posterior information in follow-up computations based on samples or sampling free methods such as stochastic Galerkin schemes.

Third, we recall that the optimization subject to debiased Sinkhorn loss leads to a fit in the model class to the data sample distribution. However, there might be circumstances where more statistical information of the image  $Y$  such as moments is available. This for instance arises with affine representations of random fields  $\kappa$  in terms of Kosambi–Karhunen–Loève expansions

$$\kappa(x, \omega) = \kappa_0(x) + \sum_{i=1}^N Y_i(\omega) \kappa_i(x).$$

Here, the random vector  $Y = (Y_1, \dots, Y_N)$  has unknown distributions but is known to be centralized with uncorrelated  $Y_i$  with unit variance, yielding a Stiefel manifold setting as in [73]. With this the optimization problem (38) can easily be extended to a constrained formulation such that  $\mathcal{M}[\theta]$  exhibits mean  $0 \in \mathbb{R}^N$  and covariance  $I \in \mathbb{R}^{N,N}$ . In the performed numerical investigations this constraint was introduced by penalty terms to circumvent difficulties that would arise when modifying the used tensor train ring format. We believe that this type of additional information should improve the generalization error for small numbers of data samples for training in the model class.

Finally, the underlying debiased Sinkhorn divergence still is a dimension-dependent metric in terms of the sample complexity. It would thus be worthwhile to extend our view to classes of dimension-free metrics such as the ones in the recent work [51].

## References

- [1] Mazen Ali and Anthony Nouy. Approximation with tensor networks. part ii: Approximation rates for smoothness classes. *arXiv preprint arXiv:2007.00128*, 2020.
- [2] Mazen Ali and Anthony Nouy. Approximation with tensor networks. part iii: Multivariate approximation. *arXiv preprint arXiv:2101.11932*, 2021.
- [3] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *arXiv preprint arXiv:1705.09634*, 2017.
- [4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.
- [5] Markus Bachmayr, Anthony Nouy, and Reinhold Schneider. Approximation by tree tensor networks in high dimensions: Sobolev and compositional functions. *arXiv preprint arXiv:2112.01474*, 2021.
- [6] Markus Bachmayr, Reinhold Schneider, and André Uschmajew. Tensor networks and hierarchical tensors for the solution of high-dimensional partial differential equations. *Foundations of Computational Mathematics*, 16(6):1423–1472, 2016.
- [7] Ricardo Baptista, Olivier Zahm, and Youssef Marzouk. An adaptive transport framework for joint and conditional density estimation. *arXiv preprint arXiv:2009.10303*, 2020.

- [8] Joakim Beck, Fabio Nobile, Lorenzo Tamellini, and Raúl Tempone. Convergence of quasi-optimal stochastic galerkin methods for a class of pdes with random coefficients. *Computers & Mathematics with Applications*, 67(4):732–751, 2014.
- [9] Jens Behrmann, Will Grathwohl, Ricky TQ Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *International Conference on Machine Learning*, pages 573–582. PMLR, 2019.
- [10] Xavier Blanc, Claude Le Bris, and P-L Lions. Stochastic homogenization and random lattices. *Journal de mathématiques pures et appliquées*, 88(1):34–63, 2007.
- [11] Géraud Blatman and Bruno Sudret. Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach. *Comptes Rendus Mécanique*, 336(6):518–523, 2008.
- [12] Christoph Brauer, Christian Clason, Dirk Lorenz, and Benedikt Wirth. A sinkhorn-newton method for entropic optimal transport. *arXiv preprint arXiv:1710.06635*, 2017.
- [13] Michael Brennan, Daniele Bigoni, Olivier Zahm, Alessio Spantini, and Youssef Marzouk. Greedy inference with structure-exploiting lazy maps. *Advances in Neural Information Processing Systems*, 33:8330–8342, 2020.
- [14] Michael Brennan, Daniele Bigoni, Olivier Zahm, Alessio Spantini, and Youssef Marzouk. Greedy inference with structure-exploiting lazy maps. *Advances in Neural Information Processing Systems*, 33:8330–8342, 2020.
- [15] Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, 2017.
- [16] Abdellah Chkifa, Albert Cohen, Giovanni Migliorati, Fabio Nobile, and Raul Tempone. Discrete least squares polynomial approximation with random evaluations- application to parametric and stochastic elliptic pdes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(3):815–837, 2015.
- [17] Albert Cohen and Ronald DeVore. Approximation of high-dimensional parametric pdes. *Acta Numerica*, 24:1–159, 2015.
- [18] Nadav Cohen, Or Sharir, and Amnon Shashua. On the expressive power of deep learning: A tensor analysis. In *Conference on learning theory*, pages 698–728. PMLR, 2016.
- [19] Nadav Cohen and Amnon Shashua. Convolutional rectifier networks as generalized tensor decompositions. In *International Conference on Machine Learning*, pages 955–963. PMLR, 2016.
- [20] Tiangang Cui and Sergey Dolgov. Deep composition of tensor-trains using squared inverse rosenblatt transports. *Foundations of Computational Mathematics*, pages 1–60, 2021.
- [21] Tiangang Cui, Sergey Dolgov, and Olivier Zahm. Conditional deep inverse rosenblatt transports. *arXiv preprint arXiv:2106.04170*, 2021.
- [22] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

- [23] Gianluca Detommaso, Tiangang Cui, Youssef Marzouk, Alessio Spantini, and Robert Scheichl. A stein variational newton method. *Advances in Neural Information Processing Systems*, 31, 2018.
- [24] Sergey Dolgov, Karim Anaya-Izquierdo, Colin Fox, and Robert Scheichl. Approximation and sampling of multivariate probability distributions in the tensor train decomposition. *Statistics and Computing*, 30(3):603–625, 11 2019.
- [25] Sergey Dolgov, Boris N Khoromskij, Alexander Litvinenko, and Hermann G Matthies. Polynomial chaos expansion of random coefficients and the solution of stochastic partial differential equations in the tensor train format. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):1109–1135, 2015.
- [26] Richard Mansfield Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.
- [27] Charles F Dunkl and Yuan Xu. *Orthogonal Polynomials of Several Variables*, volume 155. Cambridge University Press, 2014.
- [28] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn’s algorithm. In *International conference on machine learning*, pages 1367–1376. PMLR, 2018.
- [29] M. Eigel, J. Neumann, R. Schneider, and Sebastian Wolf. Non-intrusive tensor reconstruction for high-dimensional random pdes. *Computational Methods in Applied Mathematics*, 19:39 – 53, 2019.
- [30] Martin Eigel, Claude Jeffrey Gittelsohn, Christoph Schwab, and Elmar Zander. Adaptive stochastic galerkin fem. *Computer Methods in Applied Mechanics and Engineering*, 270:247–269, 2014.
- [31] Martin Eigel, Robert Gruhlke, Manuel Marschall, Philipp Trunschke, and Elmar Zander. ALEA - A Python Framework for Spectral Methods and Low-Rank Approximations in Uncertainty Quantification.
- [32] Martin Eigel, Manuel Marschall, Max Pfeffer, and Reinhold Schneider. Adaptive stochastic galerkin FEM for lognormal coefficients in hierarchical tensor representations. *Numerische Mathematik*, 145(3):655–692, 6 2020.
- [33] Martin Eigel, Johannes Neumann, Reinhold Schneider, and Sebastian Wolf. Non-intrusive tensor reconstruction for high-dimensional random pdes. *Computational Methods in Applied Mathematics*, 19(1):39–53, 2019.
- [34] Martin Eigel, Reinhold Schneider, Philipp Trunschke, and Sebastian Wolf. Variational monte carlo—bridging concepts of machine learning and high-dimensional partial differential equations. *Advances in Computational Mathematics*, 45(5):2503–2532, 2019.
- [35] Oliver G Ernst, Antje Mugler, Hans-Jörg Starkloff, and Elisabeth Ullmann. On the convergence of generalized polynomial chaos expansions. *ESAIM: Mathematical Modelling and Numerical Analysis*, 46(2):317–339, 2012.
- [36] Michael Waskom et al. seaborn: statistical data visualization. *python package*, 2012-2021.
- [37] FEniCS Project - Automated solution of Differential Equations by the Finite Element Method.

- [38] Jean Feydy. *Geometric data analysis, beyond convolutions*. PhD thesis, PhD thesis, Université Paris-Saclay, 2020.
- [39] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019.
- [40] Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583. PMLR, 2019.
- [41] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.
- [42] Roger G. Ghanem and Pol D. Spanos. *Stochastic finite elements: a spectral approach*. Springer-Verlag, New York, 1991.
- [43] Roger G Ghanem and Pol D Spanos. *Stochastic finite elements: a spectral approach*. Courier Corporation, 2003.
- [44] Antoine Gloria, Stefan Neukamm, and Felix Otto. An optimal quantitative two-scale expansion in stochastic homogenization of discrete elliptic equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 48(2):325–346, 2014.
- [45] Antoine Gloria and Felix Otto. An optimal error estimate in stochastic homogenization of discrete elliptic equations. *The annals of applied probability*, pages 1–28, 2012.
- [46] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [47] Robert Gruhlke and Till Schäfer. Bubbles - A Python Framework for composite modelling.
- [48] Robert Gruhlke and David Sommer. TensorTrain - A Python Framework for Tensor Train approximations with PyTorch and NumPy backend.
- [49] Wolfgang Hackbusch. *Tensor spaces and numerical tensor calculus*, volume 42. Springer, 2012.
- [50] Greg Hamerly and Charles Elkan. Learning the k in k-means. *Advances in neural information processing systems*, 16:281–288, 2004.
- [51] Jiequn Han, Ruimeng Hu, and Jihao Long. A class of dimensionality-free metrics for the convergence of empirical measures. *arXiv preprint arXiv:2104.12036*, 2021.
- [52] Leonid Kantorovich. On the transfer of masses (in russian). In *Doklady Akademii Nauk*, volume 37, pages 227–229, 1942.
- [53] Olivier Le Maître and Omar M Knio. *Spectral methods for uncertainty quantification: with applications to computational fluid dynamics*. Springer Science & Business Media, 2010.
- [54] Régis Lebrun and Anne Dutfoy. Do rosenblatt and nataf isoprobabilistic transformations really differ? *Probabilistic Engineering Mechanics*, 24(4):577–584, 2009.

- [55] Huidong Liu, Xianfeng Gu, and Dimitris Samaras. Wasserstein gan with quadratic transport cost. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4832–4841, 2019.
- [56] Giulia Luise, Alessandro Rudi, Massimiliano Pontil, and Carlo Ciliberto. Differential properties of sinkhorn approximation for learning with wasserstein distance. *arXiv preprint arXiv:1805.11897*, 2018.
- [57] Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini. An introduction to sampling via measure transport. *arXiv preprint arXiv:1602.05023*, 2016.
- [58] Hermann G Matthies. Stochastic finite elements: Computational approaches to stochastic partial differential equations. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik: Applied Mathematics and Mechanics*, 88(11):849–873, 2008.
- [59] Pertti Mattila. *Geometry of sets and measures in Euclidean spaces: fractals and rectifiability*, volume 44. Cambridge university press, 1999.
- [60] Arthur Mensch and Gabriel Peyré. Online sinkhorn: Optimal transport distances from sample streams. *arXiv preprint arXiv:2003.01415*, 2020.
- [61] Giovanni Migliorati, Fabio Nobile, Erik von Schwerin, and Raúl Tempone. Approximation of quantities of interest in stochastic pdes by the random discrete  $l^2$  projection on polynomial spaces. *SIAM Journal on Scientific Computing*, 35(3):A1440–A1460, 2013.
- [62] Rebecca Morrison, Ricardo Baptista, and Youssef Marzouk. Beyond normality: Learning sparse probabilistic graphical models in the non-gaussian setting. *Advances in neural information processing systems*, 30, 2017.
- [63] Anthony Nouy. Low-rank tensor methods for model order reduction. *arXiv preprint arXiv:1511.01555*, 2015.
- [64] Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.
- [65] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.
- [66] Thomas Pinetz, Daniel Soukup, and Thomas Pock. What is optimized in wasserstein gans? In *Proceedings of the 23rd Computer Vision Winter Workshop*, 2018.
- [67] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.
- [68] Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D Lee. On the convergence and robustness of training gans with regularized optimal transport. *Advances in Neural Information Processing Systems*, 31, 2018.
- [69] Sadiq M Sarfaraz, Bojana V Rosić, Hermann G Matthies, and Adnan Ibrahimbegović. Stochastic upscaling via linear bayesian updating. In *Multiscale Modeling of Heterogeneous Structures*, pages 163–181. Springer, 2018.



- [70] Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3):A1443–A1481, 2019.
- [71] Erwin Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. In *Annales de l'institut Henri Poincaré*, volume 2, pages 269–310, 1932.
- [72] Christoph Schwab and Claude Jeffrey Gittelson. Sparse tensor discretizations of high-dimensional parametric and stochastic pdes. *Acta Numerica*, 20:291–467, 2011.
- [73] Christian Soize. Identification of high-dimension polynomial chaos expansions with random coefficients for non-gaussian tensor-valued random fields using partial and limited experimental data. *Computer methods in applied mechanics and engineering*, 199(33-36):2150–2164, 2010.
- [74] Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. Inference via low-dimensional couplings. *The Journal of Machine Learning Research*, 19(1):2639–2709, 2018.
- [75] Michael Maximilian Steinlechner. Riemannian optimization for solving high-dimensional problems with low-rank tensor structure. Technical report, EPFL, 2016.
- [76] Szilárd Szalay, Max Pfeffer, Valentin Murg, Gergely Barcza, Frank Verstraete, Reinhold Schneider, and Örs Legeza. Tensor product methods and entanglement optimization for ab initio quantum chemistry. *International Journal of Quantum Chemistry*, 115(19):1342–1391, 2015.
- [77] Veeravalli S Varadarajan. On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 19(1/2):23–26, 1958.
- [78] Maria Vasilyeva and Aleksey Tyrylgin. Machine learning for accelerating macroscopic parameters prediction for poroelasticity problem in stochastic media. *Computers & Mathematics with Applications*, 84:185–202, 2021.
- [79] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.
- [80] Xiaoliang Wan and George Em Karniadakis. Multi-element generalized polynomial chaos for arbitrary probability measures. *SIAM Journal on Scientific Computing*, 28(3):901–928, 2006.
- [81] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.
- [82] Albert H Werner, Daniel Jaschke, Pietro Silvi, Martin Kliesch, Tommaso Calarco, Jens Eisert, and Simone Montangero. Positive tensor network approach for simulating open quantum many-body systems. *Physical review letters*, 116(23):237201, 2016.
- [83] Dongbin Xiu and George Em Karniadakis. The wiener–askey polynomial chaos for stochastic differential equations. *SIAM journal on scientific computing*, 24(2):619–644, 2002.
- [84] Jakob Zech and Youssef Marzouk. Sparse approximation of triangular transports. part ii: the infinite dimensional case. *arXiv preprint arXiv:2107.13422*, 2021.