On the representation and learning of monotone triangular transport maps

Ricardo Baptista, Youssef Marzouk, Olivier Zahm

On the representation and learning of monotone triangular transport maps

Ricardo Baptista[1], **Youssef Marzouk**[1], **and Olivier Zahm**[2]

[1] *Massachusetts Institute of Technology*
Cambridge, MA 02139 USA
rsb@mit.edu, ymarz@mit.edu,

[2] *Université Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK*
38000, Grenoble, France
olivier.zahm@inria.fr

Abstract: Transportation of measure provides a versatile approach for modeling complex probability distributions, with applications in density estimation, Bayesian inference, generative modeling, and beyond. Monotone triangular transport maps—approximations of the Knothe–Rosenblatt (KR) rearrangement—are a canonical choice for these tasks. Yet the representation and parameterization of such maps have a significant impact on their generality and expressiveness, and on properties of the optimization problem that arises in learning a map from data (e.g., via maximum likelihood estimation). We present a general framework for representing monotone triangular maps via invertible transformations of smooth functions. We establish conditions on the transformation such that the associated infinite-dimensional minimization problem has no spurious local minima, i.e., all local minima are global minima; and we show for target distributions satisfying certain tail conditions that the unique global minimizer corresponds to the KR map. Given a sample from the target, we then propose an adaptive algorithm that estimates a sparse semi-parametric approximation of the underlying KR map. We demonstrate how this framework can be applied to joint and conditional density estimation, likelihood-free inference, and structure learning of directed graphical models, with stable generalization performance across a range of sample sizes.

Keywords and phrases: Transportation of measure, Knothe–Rosenblatt rearrangement, normalizing flows, monotone functions, infinite-dimensional optimization, adaptive approximation, multivariate polynomials, wavelets, density estimation.
MSC2020 subject classifications: 65C20, 49Q22, 62G07.

Contents

1. Introduction

Many sampling, estimation, and inference algorithms seek to characterize a somehow intractable or complex probability distribution $\boldsymbol{\mu}$ on $\mathbb{R}^d$. Transportation of measure provides a useful and versatile approach to this problem. The underlying idea is to construct a coupling of $\boldsymbol{\mu}$ with a tractable "reference" distribution $\boldsymbol{\nu}$ on $\mathbb{R}^d$—for instance, a standard normal. Formally, one jointly constructs a pair of random variables $(\boldsymbol{X}, \boldsymbol{Z})$ such that $\boldsymbol{X} \sim \boldsymbol{\mu}$ and $\boldsymbol{Z} \sim \boldsymbol{\nu}$. A special class of couplings is given by *deterministic* transformations $S\colon \mathbb{R}^d \to \mathbb{R}^d$ such that $S(\boldsymbol{X}) = \boldsymbol{Z}$ in distribution, and a transformation S that satisfies this property is called a *transport map* [67]. As a result, if the transport map is invertible, one can generate realizations of $\boldsymbol{X}$ by first simulating $\boldsymbol{Z}$. If $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ have densities π and η with respect to a common base measure, one can also explicitly represent the target density π as a transformation of the reference density η. The construction of such transport maps has found numerous applications: density estimation [60, 2, 16, 20]; variational Bayesian inference [19, 48, 9]; generative modeling of images, video, and other structured objects [40, 24]; likelihood-free inference [41, 45, 31]; and beyond.

In general, there exist infinitely many transport maps between two absolutely continuous distributions on $\mathbb{R}^d$. In this paper, we focus on a specific, canonical choice: *triangular* transport maps [7] of the form

$$S(\boldsymbol{x}) = \begin{bmatrix} S_1(x_1) \\ S_2(x_1, x_2) \\ \vdots \\ S_d(x_1, \dots, x_d) \end{bmatrix}, \tag{1}$$

where each component function S_k depends only on the k variables $\boldsymbol{x}_{\leq k} \coloneqq (x_1, \dots, x_k)$ and is monotone increasing with respect to the last input x_k. In particular, each function S_k encodes an increasing rearrangement [67] between ordered marginal conditionals of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, i.e., $\boldsymbol{\mu}(\mathrm{d}x_k|\boldsymbol{x}_{<k})$ and $\boldsymbol{\nu}(\mathrm{d}z_k|\boldsymbol{z}_{<k})$. Later we will discuss properties of such triangular maps—known also as Knothe–Rosenblatt (KR) rearrangements [67, 49, 51]—more precisely, but we comment here on their utility. First of all, triangular structure facilitates computational tractability: S is easy to invert and the determinant of its Jacobian (a lower triangular matrix) is easy to evaluate (see Section 2). Triangular maps have thus been used extensively in the density estimation, inference, and generative modeling applications noted above. For example, triangular maps are the building blocks of many *normalizing flows*, popularized by the machine learning community [25, 42]; specifically, autoregressive normalizing flows define triangular maps [22] via particular structural choices and parameterizations, and compose these maps to produce more expressive transformations. More fundamentally, because triangular maps expose certain conditionals of $\boldsymbol{\mu}$, they are particularly well suited to *conditional density estimation* and *conditional sampling*; we will describe this link in Section 2. Triangular maps also inherit *sparsity* from the conditional independence properties of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, as detailed in [59].

It is worth noting, of course, that other canonical choices of transport have different attractive features. For instance, optimal transport maps are invariant under relabeling of the coordinates or more general isometries on $\mathbb{R}^d$ (unlike triangular maps), and have deep links to partial differential equations [67]. But optimal transport maps are in general more challenging to represent, evaluate, and estimate from data in the continuous setting; moreover, they do not enjoy such a direct link to conditioning or to graphical models.

Many representations and finite-dimensional parameterizations of monotone triangular maps have been proposed in recent years. These include representations based on polynomials [33, 22], radial basis functions [60], neural networks of varying capacity [17, 24], and tensor decompositions [14, 15]. A core challenge in this setting is to satisfy the monotonicity constraint $\partial_{x_k} S_k > 0$. For instance,

one might enforce the monotonicity constraint at a finite collection of points in the support of π [44], but this approach cannot in general guarantee that S is monotone over the entire support of π. Other efforts have sought to enforce monotonicity by construction—via the parameterization of S itself. For example, [43] employs map components with affine dependence on the last variable, i.e., $S_k(\boldsymbol{x}_{\leq k}) = \alpha(\boldsymbol{x}_{<k}) + \exp(\beta(\boldsymbol{x}_{<k}))x_k$, where α and β are neural networks. While S is then guaranteed to be monotone, it can only represent a restricted class of distributions $\boldsymbol{\mu}$. (If $\boldsymbol{\nu}$ is Gaussian, then $\boldsymbol{\mu}$ can only be a product of Gaussian marginal conditionals.) Such representations therefore cannot consistently approximate the *true* KR rearrangement for general $\boldsymbol{\mu}$. Recent work [61] has shown that a *composition* of such affine maps, interleaved with rotations and permutations, can approximate a general class of distributions, though approximation rates remain unknown. The required rotations or permutations break the overall triangular structure of the transformation, however. Alternatively, to increase the "expressiveness" of a given triangular function, [69, 21, 18] have introduced more general parametric representations of the monotone component functions S_k. For distributions with analytic densities on *bounded domains*, a complete approximation theory for the KR map was recently developed in [71, 72], using polynomial or ReLU neural network representations of range-constrained monotone triangular functions.

Despite these myriad proposals, relatively little attention has been paid to the structure and tractability of the *optimization problem* involved in learning triangular maps (e.g., in estimating a map given an i.i.d. sample $\{\boldsymbol{X}^i\}$ from $\boldsymbol{\mu}$). Properties of this optimization problem are intimately tied to the representation and parameterization of the associated map. It is desirable to have a flexible and general representation—one that can consistently recover the KR rearrangement for a broad class of distributions $(\boldsymbol{\mu}, \boldsymbol{\nu})$—that at the same time makes optimization tractable. It is also desirable to have adaptivity: a parameterization whose size or complexity can be adapted to properties of the target distribution and the available sample size, for good empirical statistical performance.

This paper directly addresses these desiderata. We do so by developing and analyzing a *functional framework* for representing and learning triangular maps. Our main contributions are as follows. We propose a general representation of monotone triangular functions, based on a *rectification* operator $\mathcal{R}_k$ that transforms sufficiently smooth non-monotone functions $f_k : \mathbb{R}^k \to \mathbb{R}$ into monotone component functions S_k of a triangular map. This operator takes the form

$$\mathcal{R}_k(f_k)(\boldsymbol{x}_{\leq k}) = f_k(\boldsymbol{x}_{<k}, 0) + \int_0^{x_k} g\big(\partial_k f_k(\boldsymbol{x}_{<k}, t)\big)\mathrm{d}t, \tag{2}$$

where $g\colon \mathbb{R} \to \mathbb{R}_{>0}$ is a positive function. We then analyze the infinite-dimensional optimization problem associated with learning maps from data, recast as optimization over functions $\{f_k\}_{k=1}^d$. Specifically, we establish conditions on the rectification operator and on the target distribution such that the resulting optimization problem is well-posed, smooth, and has *no spurious local minima*. Under further conditions on the target distribution (essentially that it has Gaussian tails), we show that the optimization problem has a *unique global minimizer* corresponding to the KR map.

These theoretical results guarantee, in practice, fast and reliable learning of monotone triangular maps given an appropriate function space V_k for each f_k. The second main contribution of our paper is algorithmic: given a hierarchical basis (e.g., polynomials or wavelets) for each V_k, we propose a greedy adaptive procedure to learn parametric representations of f_k. The procedure naturally produces map representations that are *sparse* and interpretable—in particular, it exploits and implicitly discovers conditional independence. We use these learned maps for density estimation, given an i.i.d. sample $\{\boldsymbol{X}^i\}$ from π. Maintaining a strict triangular structure also exposes marginal conditionals of the target density, thus enabling *conditional density estimation*. Our numerical experiments show that the algorithm provides robust performance at small-to-moderate sample sizes, and constitutes a *semi-parametric* approach that naturally links map complexity to the size of the data.

The remainder of the paper is organized as follows. Section 2 recalls properties of triangular transport maps and introduces some estimation problems of interest. Our main theoretical contributions are in Section 3, which introduces a framework for representing monotone triangular maps and analyzes the resulting optimization problem. Section 4 then introduces our greedy adaptive procedure for learning maps, and Section 5 contains numerical experiments. Proofs of certain theoretical results are deferred to the appendix.

2. Triangular transport for density estimation and simulation

Consider the unsupervised learning problem of approximating a target probability density function π defined on $\mathbb{R}^d$, given an i.i.d. sample from π. Our goal is to construct a sufficiently smooth and invertible map $S\colon \mathbb{R}^d \to \mathbb{R}^d$ such that the pullback density

$$S^\sharp\eta(\boldsymbol{x}) = \eta\left(S(\boldsymbol{x})\right) |\det \nabla S(\boldsymbol{x})|, \tag{3}$$

is a good approximation to π, where η is a simple/tractable probability density function on $\mathbb{R}^d$. The choice of η is a degree of freedom of the method, and here we take $\eta(\boldsymbol{x}) \propto \exp(-\|\boldsymbol{x}\|^2/2)$; i.e., η is the density of the standard normal distribution on $\mathbb{R}^d$, where $\|\cdot\|$ is the canonical norm of $\mathbb{R}^d$.

To ensure invertibility of S, a common practice is to constrain S to be an increasing lower triangular map of the form (1), where each component $S_k\colon \mathbb{R}^k \to \mathbb{R}$ is such that $x_k \mapsto S_k(\boldsymbol{x}_{<k}, x_k)$ is increasing for all $\boldsymbol{x}_{<k} = (x_1, \dots, x_{k-1}) \in \mathbb{R}^{k-1}$. Such a map is easy to invert[1] and has a lower triangular Jacobian $\nabla S(\boldsymbol{x})$ so that $|\det \nabla S(\boldsymbol{x})| = \prod_{k=1}^d \partial_k S_k(\boldsymbol{x}_{\le k})$ is readily computable. This structure in fact corresponds to the *Knothe–Rosenblatt* (KR) rearrangement S_{KR}: the increasing lower triangular map satisfying

$$\pi(\boldsymbol{x}) = S_{\mathrm{KR}}^\sharp \eta(\boldsymbol{x}).$$

For a measure $\boldsymbol{\mu}$ on $\mathbb{R}^d$ that is absolutely continuous with respect to a Gaussian measure $\boldsymbol{\nu}$ (and hence has a density π with respect to the Lebesgue measure), the KR rearrangement S_{KR} exists and is the *unique* map of the form (1) that *pulls back* η to π (or equivalently *pushes forward* π to η), up to sets of measure zero [7].

A useful way to measure the difference between π and its approximation $S^\sharp\eta$ is the Kullback–Leibler (KL) divergence $\mathcal{D}_{\mathrm{KL}}(\pi||S^\sharp\eta) = \int \log(\pi/S^\sharp\eta)\mathrm{d}\pi$. As explained below, this choice has direct links to maximum likelihood estimation of S. Furthermore, the following inequality shows that convergence in the KL sense $\mathcal{D}_{\mathrm{KL}}(\pi||S^\sharp\eta) \to 0$ implies convergence of S towards S_{KR} in the L^2_π sense. This result is a direct consequence of Corollary 3.10 in [7]; see Appendix A.1 for a proof.

Proposition 1. *Let S_{KR} be the KR rearrangement pushing forward a distribution with density π on $\mathbb{R}^d$ to the standard normal distribution on $\mathbb{R}^d$, with density η. For any map $S : \mathbb{R}^d \to \mathbb{R}^d$ as in (1), we have*

$$\int \|S_{\mathrm{KR}}(\boldsymbol{x}) - S(\boldsymbol{x})\|^2 \mathrm{d}\pi(\boldsymbol{x}) \le 2\mathcal{D}_{KL}(\pi||S^\sharp\eta). \tag{4}$$

Since the standard normal density η is a product of its marginal densities, the KL divergence decomposes as

$$\mathcal{D}_{\mathrm{KL}}(\pi||S^\sharp\eta) = \sum_{k=1}^d \mathcal{J}_k(S_k) - \mathcal{J}_k(S_{\mathrm{KR},k}), \tag{5}$$

[1] For any $\boldsymbol{z} \in \mathbb{R}^d$, $\boldsymbol{x} = S^{-1}(\boldsymbol{z})$ can be computed recursively by $x_k = T^k(\boldsymbol{x}_{<k}, z_k)$ for $k = 1, \dots, d$, where the function $T^k(\boldsymbol{x}_{<k}, \cdot)$ is the inverse of $x_k \mapsto S_k(\boldsymbol{x}_{<k}, x_k)$.

where the functionals $\mathcal{J}_1, \ldots, \mathcal{J}_d$ are given by

$$\mathcal{J}_k(s) = \int \left(\frac{1}{2} s(\boldsymbol{x}_{\leq k})^2 - \log \left| \partial_k s(\boldsymbol{x}_{\leq k}) \right| \right) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{6}$$

Minimizing the KL divergence (5) over triangular maps S of the form (1) is therefore equivalent to *independently* minimizing each objective functional $\mathcal{J}_k$ to find the associated map component S_k [33]. Solution of these optimization problems is thus embarrassingly parallel. This parallel structure was also exploited for Cholesky factorization via KL minimization in [52]. In addition, minimizing each objective $s \mapsto \mathcal{J}_k(s)$ over functions $s \colon \mathbb{R}^k \to \mathbb{R}$ that are strictly increasing in the last variable is a *strictly convex* optimization problem; see Lemma 10.

Given an i.i.d. sample $\{\mathbf{X}^i\}_{i=1}^n$ from π, we can replace the expectation in (6) by the sample average, which yields the objective

$$\widehat{\mathcal{J}}_k(s) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{2} s(\mathbf{X}^i_{\leq k})^2 - \log \left| \partial_k s(\mathbf{X}^i_{\leq k}) \right| \right). \tag{7}$$

Minimizing (7) under the constraint $\partial_k s(\boldsymbol{x}_{\leq k}) > 0$ produces an estimator $\widehat{S}_k$ of $S_{\mathrm{KR},k}$. The collection of all such component functions defines an estimator $\widehat{S} = (\widehat{S}_1, \ldots, \widehat{S}_d)$ of the KR rearrangement, along with an estimate of the density π as $\widehat{\pi}(\boldsymbol{x}) \coloneqq \widehat{S}^\sharp \eta(\boldsymbol{x})$. This $\widehat{S}$ is also the maximum likelihood estimator of S_{KR}, i.e.,

$$\widehat{S} = \underset{\{S \in (1),\ \partial_k S_k > 0,\ k=1,\ldots,d\}}{\arg\max} \sum_{i=1}^{n} \log S^\sharp \eta(\boldsymbol{X}^i),$$

with the optimization being over the space of monotone increasing and triangular maps of the form (1). This connection between maximum likelihood estimation and minimization of an empirical forward KL divergence is standard. Furthermore, convexity of the optimization objective is preserved when replacing the expectation in (6) by the sample average in (7).

A core question when estimating maps and densities by minimizing (7) is how to parameterize sufficiently expressive monotone map components S_k—i.e., maps capable of representing a wide class of distributions π—while ensuring that the optimization problem can be solved efficiently. As explained in the introduction, this question is intimately tied to the monotonicity constraint $\partial_k S_k(\boldsymbol{x}_{\leq k}) > 0$. For example, choosing a linear parameterization for S_k that admits only affine dependence on x_k (to easily enforce monotonicity) allows the map component be identified efficiently through the solution of a least-squares problem (see [58, Appendix A]), but such maps can only capture distributions that factor into a sequence of Gaussian marginal conditionals. On the other hand, a more complex ansatz for S_k will often yield a much more difficult optimization problem. Note that with any nonlinear parameterization of S_k, the convexity of (6) and (7) with respect to S_k does not in general yield convexity in the parameters. As we shall demonstrate later, many nonlinear parameterizations that enforce monotonicity yield optimization problems that may not even be smooth, and that have many local minima. We will address these issues in Section 3.

Conditional density estimation and sampling Another important feature of the triangular structure (1) is that each component of the map represents one marginal conditional density of π. More precisely, in the present setting where η is a product density, $S_{\mathrm{KR},k}$ pushes forward the marginal conditional $\pi_k(x_k|\boldsymbol{x}_{<k})$ to the k-th marginal of the reference $\eta_k(z_k)$ [51, Section 2.3]. Now partition $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{w})$, where $\boldsymbol{y} \in \mathbb{R}^m$ and $\boldsymbol{w} \in \mathbb{R}^p$, with $d = m + p$. This property of the KR map lets us estimate the conditional probability density function $\pi(\boldsymbol{w}|\boldsymbol{y})$, for any value of $\boldsymbol{y}$, given a sample

$\{(\boldsymbol{Y}^i, \boldsymbol{W}^i)\}_{i=1}^n$ from the joint density $\pi(\boldsymbol{y}, \boldsymbol{w})$. Observe that the KR map immediately has the block structure:

$$S(\boldsymbol{y}, \boldsymbol{w}) = \begin{bmatrix} S^{\mathcal{Y}}(\boldsymbol{y}) \\ S^{\mathcal{W}}(\boldsymbol{y}, \boldsymbol{w}) \end{bmatrix}, \tag{8}$$

where $\boldsymbol{y} \mapsto S^{\mathcal{Y}}(\boldsymbol{y})\colon \mathbb{R}^m \to \mathbb{R}^m$ and $\boldsymbol{w} \mapsto S^{\mathcal{W}}(\boldsymbol{y}^*, \boldsymbol{w})\colon \mathbb{R}^p \to \mathbb{R}^p$ are increasing lower triangular maps, the latter for any $\boldsymbol{y}^* \in \mathbb{R}^m$. Recall that the reference density η is the standard normal on $\mathbb{R}^d$, and thus is a product of two standard normals, $\eta_1(\boldsymbol{z}')\eta_2(\boldsymbol{z})$, where $\boldsymbol{z}' \in \mathbb{R}^m$ and $\boldsymbol{z} \in \mathbb{R}^p$. Using a KR map of the form (8), we can write the marginal density of $\boldsymbol{Y}$ as $\pi(\boldsymbol{y}) = (S^{\mathcal{Y}}_{\mathrm{KR}})^\sharp \eta_1(\boldsymbol{y})$ and, more interestingly, the conditional density of $\boldsymbol{W}$ as $\pi(\boldsymbol{w}|\boldsymbol{y}) = S^{\mathcal{W}}_{\mathrm{KR}}(\boldsymbol{y}, \cdot)^\sharp \eta_2(\boldsymbol{w})$.

Each of the last p components of the KR rearrangement, $S^{\mathcal{W},k}_{\mathrm{KR}}(\boldsymbol{y}, \boldsymbol{w}_{\le k})$, $1 \le k \le p$, can be estimated from a sample $\{(\boldsymbol{Y}^i, \boldsymbol{W}^i)\}_{i=1}^n$ by minimizing

$$\widehat{\mathcal{J}}_k(s) = \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{2} s(\boldsymbol{Y}^i, \boldsymbol{W}^i_{\le k})^2 - \log \left| \partial_{m+k} s(\boldsymbol{Y}^i, \boldsymbol{W}^i_{\le k}) \right| \right), \tag{9}$$

under the constraints $\partial_{m+k} s(\boldsymbol{y}, \boldsymbol{w}_{\le k}) > 0$. This produces an estimator $\widehat{S}^{\mathcal{W}}$ of $S^{\mathcal{W}}_{\mathrm{KR}}$, which in turn yields an estimator of the conditional density $\pi(\boldsymbol{w}|\boldsymbol{y})$ as $\widehat{\pi}(\boldsymbol{w}|\boldsymbol{y}) \coloneqq \widehat{S}^{\mathcal{W}}(\boldsymbol{y}, \cdot)^\sharp \eta_2(\boldsymbol{w})$. This property has been used to perform conditional density estimation (CDE) in [33, 41, 27, 15].

One important application of CDE is likelihood-free inference, where $\boldsymbol{W}$ represents a parameter to be inferred and $\boldsymbol{Y}$ are data whose conditional density $\pi(\boldsymbol{y}|\boldsymbol{w})$ is computationally intractable or unavailable in closed form. This setting arises in many applications: inference in stochastic models, or inference in the presence of high-dimensional nuisance parameters or latent variables (including parameter inference for state-space models). Given a joint sample $\{(\boldsymbol{Y}^i, \boldsymbol{W}^i)\}_{i=1}^n$, we can estimate the map component $S^{\mathcal{W}}$ and use it to simulate from the estimated conditional density $\widehat{\pi}(\boldsymbol{w}|\boldsymbol{y}^*)$ given any realization of the data $\boldsymbol{y}^*$, simply by sampling $\boldsymbol{z}^i \sim \eta_2$ and solving the triangular system $\widehat{S}^{\mathcal{W}}(\boldsymbol{y}^*, \boldsymbol{w}^i) = \boldsymbol{z}^i$ for each $\boldsymbol{w}^i$. Note that we can simulate from or evaluate the estimated conditional densities for multiple realizations of the data $\boldsymbol{y}^*$, including values that are not present in the dataset $\{\boldsymbol{Y}^i\}_{i=1}^n$. Thus, learning a single map $S^{\mathcal{W}}$ parameterized by $\boldsymbol{y}$ is said to *amortize* the cost of conditional sampling over the data.

3. Representing and learning continuous monotone functions

In this section, we define a general representation for components of monotone triangular maps and present our main theoretical results. The essential idea, as mentioned in the introduction, is to express each component S_k as a nonlinear transformation $S_k = \mathcal{R}_k(f)$ of a smooth function f, where $\mathcal{R}_k$ (2) is an operator that enforces the monotonicity constraint by construction. We then identify the map component by solving the re-parameterized optimization problem

$$\min_{f \in V_k} \mathcal{L}_k(f) = \mathcal{J}_k(\mathcal{R}_k(f)), \tag{10}$$

where V_k is a linear space of functions in which we seek f. With this nonlinear transformation, we lose the convexity of the constrained problem $\min_{\{s\colon \partial_k s > 0\}} \mathcal{J}_k(s)$, but obtain an unconstrained minimization problem instead. It turns out that, with appropriate conditions on $\mathcal{R}_k$, the transformed optimization problem retains many desirable properties.

In Section 3.1, we discuss the construction of $\mathcal{R}_k$ and motivate certain critical choices therein. Then, in Section 3.2, we analyze the regularity of the *composed* objective functional $\mathcal{L}_k$ for a certain choice of the function space V_k. In Section 3.3 we discuss the existence and uniqueness of solutions to (10), and describe conditions under which solving (10) exactly recovers the KR rearrangement.

3.1. The rectification operator

Recalling (2), for any sufficiently smooth $f\colon \mathbb{R}^k \to \mathbb{R}$ we let $\mathcal{R}_k(f)\colon \mathbb{R}^k \to \mathbb{R}$ be the function defined by

$$\mathcal{R}_k(f)(\boldsymbol{x}_{\leq k}) = f(\boldsymbol{x}_{<k}, 0) + \int_0^{x_k} g\big(\partial_k f(\boldsymbol{x}_{<k}, t)\big)\mathrm{d}t,$$

where $g\colon \mathbb{R} \to \mathbb{R}_{>0}$ is a positive function. We call the operator $\mathcal{R}_k\colon f \mapsto \mathcal{R}_k(f)$ a *rectifier* because it transforms any f into a function which is increasing in the k-th variable, i.e., $\partial_k \mathcal{R}_k(f)(\boldsymbol{x}_{\leq k}) = g(\partial_k f(\boldsymbol{x}_{\leq k})) > 0$. As a simple example, functions of the form $f(\boldsymbol{x}_{\leq k}) = \alpha(\boldsymbol{x}_{<k}) + \beta(\boldsymbol{x}_{<k})x_k$ are transformed into $\mathcal{R}_k(f)(\boldsymbol{x}_{\leq k}) = \alpha(\boldsymbol{x}_{<k}) + g(\beta(\boldsymbol{x}_{<k}))x_k$. Figure 1 illustrates the application of the rectifier to a nonlinear univariate function f.

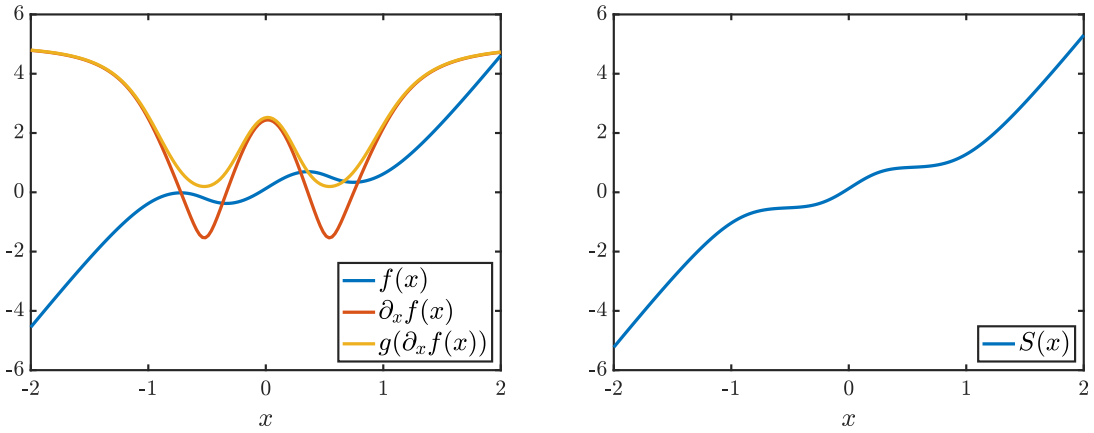


Fig 1: The rectifier (2) transforms the non-monotone function f into the monotone function $S = \mathcal{R}(f)$. Here we choose $g(\cdot) = \log(1 + \exp(\cdot))$. S is an increasing transport that pushes forward a one-dimensional mixture of Gaussians $\pi(x) = 0.5\mathcal{N}(x; -1, 1) + 0.5\mathcal{N}(x; 1, 1)$ to the standard Gaussian reference density η.

The choice of the function g in (2) has a crucial impact on properties of the optimization problem (10). One possible choice, proposed in [22, 9], is the square function $g(\xi) = \xi^2$. While this choice permits closed-form computation of the integral in (3.1) when f is polynomial, it yields an optimization problem (10) which possesses many spurious local minima and is in fact non-smooth; see Figure 2 for an illustration. This can be explained in part by the fact that this g is not bijective.

Instead, one can choose g to be a bijective function from $\mathbb{R}$ to $\mathbb{R}_{>0}$, such as the soft-plus function,

$$g(\xi) = \log(1 + \exp(\xi)), \tag{11}$$

whose inverse is $g^{-1}(\xi) = \log(\exp(\xi) - 1)$. Another example of a bijective function, considered in [69], is the shifted exponential linear unit (ELU),

$$g(\xi) = \begin{cases} \exp(\xi) & \xi < 0 \\ \xi + 1 & \xi \geq 0 \end{cases}, \tag{12}$$

whose inverse is $g^{-1}(\xi) = \xi - 1$ if $\xi \geq 1$ and $g^{-1}(\xi) = \log(\xi)$ otherwise.

As a consequence of g being bijective, the inverse of the rectifier $\mathcal{R}_k^{-1}(s)$ exists for any sufficiently smooth $s\colon \mathbb{R}^k \to \mathbb{R}$ with $\partial_k s(\boldsymbol{x}_{\leq k}) > 0$ and can be written as

$$\mathcal{R}_k^{-1}(s)(\boldsymbol{x}_{\leq k}) = s(\boldsymbol{x}_{<k}, 0) + \int_0^{x_k} g^{-1}\big(\partial_k s(\boldsymbol{x}_{<k}, t)\big)\mathrm{d}t. \tag{13}$$

More importantly, the fact that g is invertible yields an objective function $\mathcal{L}_k$ that is far better behaved than with $g(\xi) = \xi^2$; see the numerical illustration in Figure 2, with both soft-plus and shifted ELU g. With these choices of g, we observe that $\mathcal{L}_k = \mathcal{J}_k \circ \mathcal{R}_k$, though non-convex in general, has no local minima and no saddle points. In the next sections, we will analyze these properties of the optimization problem (10) (e.g., smoothness, existence and uniqueness of solutions) and elucidate their precise dependence on g, V_k, and π.

Remark 1. *In the finite-dimensional setting, it can easily be shown that composing a smooth and (strictly) convex objective function with a $\mathcal{C}^1$-diffeomorphic map preserves the (unique) global minima of the objective. Such diffeomorphic maps have been explored for accelerating optimization on finite-dimensional manifolds; see, e.g., [29]. Establishing that the* operator $\mathcal{R}_k$ *in* (3.1) *is $\mathcal{C}^1$-diffeomorphic is, however, non-trivial. We will show in Section 3.3 that, under additional assumptions, our reparameterization yields an infinite-dimensional optimization problem where local minima are global minima.*

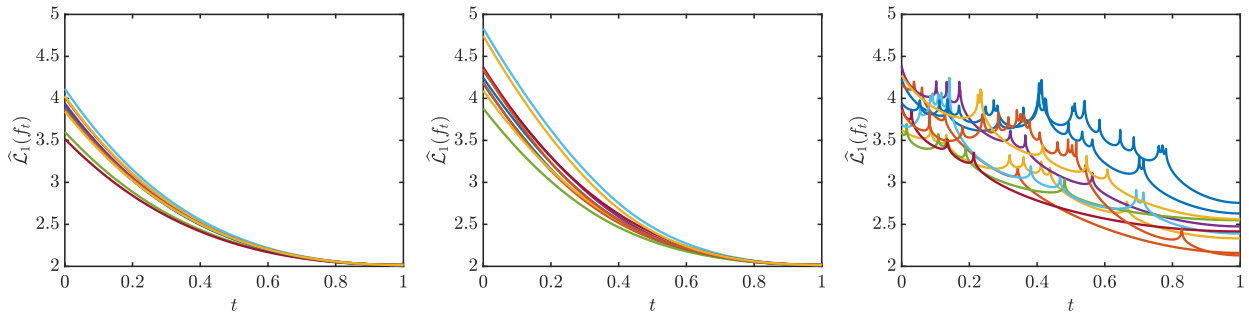


Fig 2: Objective function $\widehat{\mathcal{L}}_1 = \widehat{\mathcal{J}}_1 \circ \mathcal{R}_1$ where the rectifier $\mathcal{R}_1$ is defined using the soft-plus g (11) (*left*), the shifted ELU g (12) (*middle*), or the square function $g(\xi) = \xi^2$ (*right*). Here, $\pi(x) = 1/2\mathcal{N}(x; -2, 0.5) + 1/2\mathcal{N}(x; 2, 2)$ is a univariate Gaussian mixture, and we use $n = 50$ to define $\mathcal{J}_1$ with f_1 represented using a linear combination of Hermite functions up to degree 10. The objective is evaluated along line segments that interpolate between random initial maps ($t = 0$) and critical points resulting from a gradient-based optimization method ($t = 1$). Observe that with bijective g (*left* and *middle*) the algorithm always arrives at the same optimal value, whereas with the square function g (*right*) the algorithm gets stuck in local minima and rarely attains the optimal value.

3.2. Smoothness of the optimization problem

In this section we give sufficient conditions on the function g and the target density π to guarantee that the objective function $\mathcal{L}_k$ is smooth over an appropriate function space for f. We introduce this function space V_k and show continuity properties of the rectifier that hold for functions $f \in V_k$, before stating our main result in Proposition 4.

We begin by defining the weighted Sobolev space

$$V_k = \left\{ f \colon \mathbb{R}^k \to \mathbb{R} \text{ such that } \|f\|_{V_k}^2 := \int \Big(|f(\boldsymbol{x})|^2 + |\partial_k f(\boldsymbol{x})|^2 \Big) \eta_{\leq k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} < +\infty \right\}, \tag{14}$$

where $\eta_{\leq k}$ is the standard normal density on $\mathbb{R}^k$. By [28, Theorem 1.11], this space is complete and thus is a Hilbert space. The space V_k has sufficient regularity for $\partial_k f$ to exist, but also to

permit the pointwise evaluation $f(\boldsymbol{x}_{<k}, 0)$, as required in the definition (3.1) of the rectifier. This property is formalized by the following proposition, a trace theorem, which shows that for any $f \in V_k$, the function $\boldsymbol{x}_{<k} \mapsto f(\boldsymbol{x}_{<k}, 0)$ is a function in $L^2_{\eta_{<k}}$, the $\eta_{<k}$-weighted space of square integrable functions.

Proposition 2. *There exists a constant $C_T < \infty$ such that for any $f \in V_k$*

$$\int f(\boldsymbol{x}_{<k}, 0)^2 \, \eta_{<k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \leq C_T \|f\|^2_{V_k}. \tag{15}$$

Proof. See Appendix A.3. □

Remark 2. *Notice that $H^1_{\eta_{\leq k}} \subset V_k \subset L^2_{\eta_{\leq k}}$, where $L^2_{\eta_{\leq k}} = \{f \colon \mathbb{R}^k \to \mathbb{R} : \|f\|^2_{L^2_{\eta_{\leq k}}} := \int f^2 \mathrm{d}\eta_{\leq k} < \infty\}$ and $H^1_{\eta_{\leq k}} = \{f \colon \mathbb{R}^k \to \mathbb{R} : \|f\|^2_{H^1_{\eta_{\leq k}}} := \int f^2 + \|\nabla f\|^2 \mathrm{d}\eta_{\leq k} < \infty\}$ are the standard weighted Sobolev spaces. Given that the standard normal density factorizes as $\eta_{\leq k}(\boldsymbol{x}) = \prod_{i=1}^k \eta_i(x_i)$, the space V_k admits the following tensor product structure*

$$V_k = L^2_{\eta_1} \otimes \cdots \otimes L^2_{\eta_{k-1}} \otimes H^1_{\eta_k}, \tag{16}$$

and the norm $\|\cdot\|_{V_k}$ is a product norm.[2] This tensor product structure will be used later in Section 4 to construct a numerical scheme for approximating the map components S_k.

The following proposition shows that, under mild assumptions on g, the rectifier $\mathcal{R}_k$ is a Lipschitz continuous operator from V_k to itself. The proof relies on the Hardy inequality [37] and on Proposition 2.

Proposition 3. *Let $g \colon \mathbb{R} \to \mathbb{R}_{>0}$ be a Lipschitz function, i.e., there exists a constant $L < \infty$ so that*

$$|g(\xi) - g(\xi')| \leq L|\xi - \xi'|, \tag{17}$$

holds for any $\xi, \xi' \in \mathbb{R}$. Then $\mathcal{R}_k(f) \in V_k$ for any $f \in V_k$, where $\mathcal{R}_k(f)$ is defined in (3.1). Furthermore there exists a constant $C < \infty$ such that

$$\|\mathcal{R}_k(f_1) - \mathcal{R}_k(f_2)\|_{V_k} \leq C\|f_1 - f_2\|_{V_k}. \tag{18}$$

holds for any $f_1, f_2 \in V_k$.

Proof. See Appendix A.4. □

The next result relies on Proposition 3 to show that the objective functional $\mathcal{L}_k$ in (10), seen as a function from V_k to $\mathbb{R}$, is well-defined, continuous, and differentiable.

Theorem 4. *Let π be a probability density function on $\mathbb{R}^d$ satisfying*

$$\pi(\boldsymbol{x}) \leq C_\pi \eta(\boldsymbol{x}), \tag{19}$$

for all $\boldsymbol{x} \in \mathbb{R}^d$ and a constant $C_\pi < \infty$, and let $g \colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ be an increasing function where

$$|g(\xi) - g(\xi')| \leq L|\xi - \xi'|, \tag{20}$$

$$|\log \circ g(\xi) - \log \circ g(\xi')| \leq L|\xi - \xi'|, \tag{21}$$

[2] That is $\|v_1 \otimes \cdots \otimes v_k\|_{V_k} = \|v_1\|_{L^2_{\eta_1}} \|v_2\|_{L^2_{\eta_2}} \cdots \|v_{k-1}\|_{L^2_{\eta_{k-1}}} \|v_k\|_{H^1_{\eta_k}}$ for any $v_j \in L^2_{\eta_k}$ and $v_k \in H^1_{\eta_k}$.

holds for any $\xi, \xi' \in \mathbb{R}$ *and a constant* $L < \infty$. *Then*

$$\mathcal{L}_k(f) \coloneqq \mathcal{J}_k(\mathcal{R}_k(f)) < \infty,$$

for any $f \in V_k$, *where* $\mathcal{J}_k(s) = \int \left(\frac{1}{2}s(\boldsymbol{x}_{\leq k})^2 - \log|\partial_k s(\boldsymbol{x}_{\leq k})|\right)\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ *as in* (6) *and where* $\mathcal{R}_k(f)$ *is defined as in* (3.1). *Moreover, the function* $\mathcal{L}_k \colon V_k \to \mathbb{R}$ *is continuous and differentiable at any* $f \in V_k$ *and its gradient* $\nabla\mathcal{L}_k(f) \in V_k$ *is given by*

$$\begin{aligned}\langle \nabla\mathcal{L}_k(f), \varepsilon\rangle_{V_k} = &\int \mathcal{R}_k(f)(\boldsymbol{x})\left(\varepsilon(\boldsymbol{x}_{<k}, 0) + \int_0^{x_k} g'\big(\partial_k f(\boldsymbol{x}_{<k}, t)\big)\partial_k\varepsilon(\boldsymbol{x}_{<k}, t)\mathrm{d}t\right)\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &- \int (\log\circ g)'(\partial_k f(\boldsymbol{x}))\partial_k\varepsilon(\boldsymbol{x})\ \pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \end{aligned} \tag{22}$$

for any $\varepsilon \in V_k$, *where* $\langle\cdot,\cdot\rangle_{V_k}$ *is the scalar product in* V_k.

Proof. See Appendix A.5. □

It is useful to discuss some implications of Theorem 4. A smooth objective function permits us to use deterministic first or second-order optimization algorithms. Furthermore, in Section 4, we exploit the gradients of $\mathcal{J}_k$ in (22) to construct an adaptive polynomial (or wavelet) basis for V_k. We also note that many functions g satisfy the conditions (20) and (21) in the proposition. For example, these include the soft-plus (11) and shifted ELU (12) functions considered in Figure 2.

Remark 3 (Assumption (19) implies Gaussian tails)**.** *The condition* (19) *implies that* $\mathbb{P}_{\boldsymbol{X}\sim\pi}(\|\boldsymbol{X}\|_2 > t) \leq C_\pi\mathbb{P}_{\boldsymbol{Z}\sim\eta}(\|\boldsymbol{Z}\|_2 > t)$ *holds for any* $t \geq 0$, *and hence that* π *has sub-Gaussian tails [65]. The reverse, however, does not hold in general. For instance, a compactly supported random variable with density* $\pi(x) \propto 1/\sqrt{|x|}\mathbb{1}_{\{x\in[-1,1]\}}$ *is sub-Gaussian, but the density is unbounded at* $x = 0$ *and thus does not satisfy* (19). *We also note that* (19) *can be relaxed with the (less interpretable) assumption that there exist a constant* $C_\pi < \infty$ *and diagonal scalings* $d_k > 0$ *so that* $\pi(\boldsymbol{x}) \leq C_\pi\prod_{k=1}^d \eta_k(d_k x_k)$ *for all* $\boldsymbol{x} \in \mathbb{R}^d$. *In practice, we can always normalize the marginals of* π *to match the mean and variance of a standard Gaussian reference. Thus, without loss of generality, we will use the assumption above for the remainder of this article.*

Remark 4. *Under additional assumptions on* g, *the gradient* $f \mapsto \nabla\mathcal{L}_k(f)$ *is locally Lipschitz from* $\overline{V}_k$ *to* V_k, *where* $\overline{V}_k = \{f \in V_k, \partial_k f \in L^\infty\}$ *is the space endowed with the norm* $\|f\|_{\overline{V}_k} = \|f\|_{V_k} + \|\partial_k f\|_{L^\infty}$. *More specifically, if* g' *and* $(\log\circ g)'$ *are Lipschitz functions, then there exists a constant* $M < \infty$ *such that*

$$\|\nabla\mathcal{L}_k(f_1) - \nabla\mathcal{L}_k(f_2)\|_{V_k} \leq M(1 + \|f_2\|_{V_k})\|f_1 - f_2\|_{\overline{V}_k}, \tag{23}$$

holds for any $f_1, f_2 \in \overline{V}_k$. *See the derivation of this result in Appendix A.6. Such local Lipschitz regularity is useful to analyze the convergence of backtracking gradient descent procedures, i.e., gradient descent with an inexact line search such as Armijo's rule, to a stationary point; see Theorem 2.1 in [63] and Proposition 2.1.1 in [4]. We leave the extension of such optimization guarantees for solving* $\min_{f\in V_k}\mathcal{L}_k(f)$ *to future work.*

3.3. Existence and uniqueness of solutions

In this section, we show that the optimization problem (10) does not admit any spurious local minima, meaning that local minimizers are in fact global minimizers. We also show that problem (10) admits a unique global minimizer which permits us to recover the KR rearrangement. To prove these results, we will need the following proposition, which provides conditions ensuring that the *inverse* rectifier $\mathcal{R}_k^{-1}$ is continuous.

Proposition 5. *Let g be a bijective function from $\mathbb{R}$ to $\mathbb{R}_{>0}$ such that for any $c > 0$ there exists a constant $L_c < \infty$ so that*

$$|g^{-1}(\xi) - g^{-1}(\xi')| \leq L_c|\xi - \xi'|, \tag{24}$$

holds for any $\xi, \xi' \geq c$. Then, for any $s \in V_k$ such that $\operatorname{ess\,inf} \partial_k s > 0$, we have $\mathcal{R}_k^{-1}(s) \in V_k$ and $\operatorname{ess\,inf} \partial_k \mathcal{R}_k^{-1}(s) > -\infty$. Furthermore for any $c > 0$, there exists a constant $C_c < \infty$ such that

$$\|\mathcal{R}_k^{-1}(s_1) - \mathcal{R}_k^{-1}(s_2)\|_{V_k} \leq C_c\|s_1 - s_2\|_{V_k}, \tag{25}$$

holds for any $s_1, s_2 \in V_k$ such that $\operatorname{ess\,inf} \partial_k s_i \geq c$.

Proof. See Appendix A.8. □

Note that the softplus function (11) and the shifted exponential linear unit (12) satisfy (24) with $L_c = (1 - e^{-c})^{-1}$ and $L_c = \max\{1/c, 1\}$, respectively.

3.3.1. Local minima are global minima

The following proposition shows that, under certain conditions on the function g, the image set $\mathcal{R}_k(V_k) = \{\mathcal{R}_k(f) : f \in V_k\}$ is convex.

Proposition 6. *Let $g\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ be an increasing function such that $\xi \mapsto (g^{-1}(\xi))^2$ is a convex function. Let $\mathcal{R}_k(f)$ be defined as in (3.1) and V_k as in (14). Then $\mathcal{R}_k(V_k) = \{\mathcal{R}_k(f) : f \in V_k\}$ is convex.*

Proof. See Appendix A.7. □

When g is the softplus function (11) or the exponential linear unit (12), the mapping $\xi \mapsto (g^{-1}(\xi))^2$ is indeed convex. Interestingly, however, the exponential function $g(\xi) = \exp(\xi)$—which has been used to parameterize monotone maps in [33, 59] and for monotone regression in [46, 55]—does not satisfy this property, and there is no guarantee for $\mathcal{R}_k(V_k)$ to be convex in this case.

An important consequence of Proposition 6 is that the constrained problem $\min_{s \in \mathcal{R}_k(V_k)} \mathcal{L}_k(s)$ remains convex. Hence, this optimization problem has a unique global minimizer by the strict convexity of $\mathcal{L}_k$ (see Appendix A.2). The next result shows that the local minima of the unconstrained problem (10) are in fact global minima.

Theorem 7. *Under the assumptions of Propositions 5 and 6, let $f^* \in V_k$ be a local minimizer of $f \mapsto \mathcal{J}_k(\mathcal{R}_k(f))$, meaning that there exists $\rho > 0$ such that*

$$\mathcal{J}_k(\mathcal{R}_k(f^*)) \leq \mathcal{J}_k(\mathcal{R}_k(f)), \quad \forall f \in V_k \text{ such that } \|f - f^*\|_{V_k} \leq \rho. \tag{26}$$

If $\operatorname{ess\,inf} \partial_k f^* > -\infty$, then f^* is a global minimizer, meaning that

$$\mathcal{J}_k(\mathcal{R}_k(f^*)) \leq \mathcal{J}_k(\mathcal{R}_k(f)), \quad \forall f \in V_k.$$

Proof. Let $f \in V_k$. For any $0 < t < 1$ we let $s_t = t\mathcal{R}_k(f) + (1 - t)\mathcal{R}_k(f^*)$. By Proposition 6, $\mathcal{R}_k(V_k)$ is convex so that $s_t \in \mathcal{R}_k(V_k)$. By convexity of $\mathcal{J}_k$ (*c.f.* Appendix A.2), we can write $\mathcal{J}_k(s_t) \leq t\mathcal{J}_k(\mathcal{R}_k(f)) + (1 - t)\mathcal{J}_k(\mathcal{R}_k(f^*))$, or equivalently

$$\mathcal{J}_k(s_t) - \mathcal{J}_k(\mathcal{R}_k(f^*)) \leq t\Big(\mathcal{J}_k(\mathcal{R}_k(f)) - \mathcal{J}_k(\mathcal{R}_k(f^*))\Big). \tag{27}$$

Next we show that there exists a sufficiently small $t > 0$ such that the above left-hand side is positive. As a consequence, the right hand side of (27) will be positive, which will conclude the proof.

Let $b = \operatorname{ess\,inf} \partial_k f^*$ and $c = g(b)/2 > 0$. For any $t \leq 1/2$, as have $\partial_k s_t = tg(\partial_k f) + (1-t)g(\partial_k f^*) \geq 1/2 g(\partial_k f^*)$ so that $\operatorname{ess\,inf} \partial_k s_t \geq c$. In addition, we have $\operatorname{ess\,inf} \partial_k \mathcal{R}_k(f^*) = \operatorname{ess\,inf} g(\partial_k f^*) \geq c$. Thus, by Proposition 5, there exists a constant $C_c < \infty$ such that

$$\begin{aligned}\|\mathcal{R}_k^{-1}(s_t) - f^*\|_{V_k} &= \|\mathcal{R}_k^{-1}(s_t) - \mathcal{R}_k^{-1}(\mathcal{R}_k(f^*))\|_{V_k} \\ &\leq C_c \|s_t - \mathcal{R}_k(f^*)\|_{V_k} \\ &= tC_c \|\mathcal{R}_k(f) - \mathcal{R}_k(f^*)\|_{V_k}.\end{aligned}$$

By letting $t = \rho/(C_c\|\mathcal{R}_k(f) - \mathcal{R}_k(f^*)\|_{V_k})$, we have $\|\mathcal{R}_k^{-1}(s_t) - f^*\|_{V_k} \leq \rho$. Therefore, setting $f = \mathcal{R}_k^{-1}(s_t)$ in (26) ensures that $0 \leq \mathcal{J}_k(s_t) - \mathcal{J}_k(\mathcal{R}_k(f^*))$. □

Let us remark that Theorem 7 holds for general target densities π without the bound assumed in Theorem 4. Applying this result, however, depends on the existence of a local minimizer in the function space V_k, which must be established for a given target density. The following section studies a relatively broad class of distributions where the k-th component of the KR rearrangement is in $\mathcal{R}_k(V_k)$, thereby providing a candidate solution to apply Theorem 7.

3.3.2. Uniqueness of the global minimizer and recovery of the KR rearrangement

As discussed in Section 2, the Knothe–Rosenblatt rearrangement S_{KR} is the unique lower triangular and monotone map such that $\mathcal{D}_{\mathrm{KL}}(\pi||S_{\mathrm{KR}}^\sharp \eta) = 0$; see [7]. The decomposition (5) of the KL divergence $\mathcal{D}_{\mathrm{KL}}(\pi||S^\sharp\eta)$ thus permits us to write

$$\mathcal{J}_k(S_{\mathrm{KR},k}) \leq \mathcal{J}_k(\mathcal{R}_k(f)), \tag{28}$$

for any $f \in V_k$ and for any $1 \leq k \leq d$. Indeed, by letting S be the map such that $S_k = \mathcal{R}_k(f)$ and $S_i = S_{\mathrm{KR},i}$ for $i \neq k$, (5) yields (28). Thus, if there exists a function $f_{\mathrm{KR},k} \in V_k$ such that $S_{\mathrm{KR},k} = \mathcal{R}_k(f_{\mathrm{KR},k})$, then $f_{\mathrm{KR},k}$ is a global minimizer of $f \mapsto \mathcal{J}_k(\mathcal{R}_k(f))$ over $f \in V_k$. To show this, we first need the following intermediate result.

Proposition 8. *Let π be a probability density function on $\mathbb{R}^d$ such that*

$$c\eta(\boldsymbol{x}) \leq \pi(\boldsymbol{x}) \leq C\eta(\boldsymbol{x}), \tag{29}$$

for all $\boldsymbol{x} \in \mathbb{R}^d$ with some constants $0 < c \leq C < \infty$. Then, for all $\boldsymbol{x}_{<k} \in \mathbb{R}^{k-1}$ and $k = 1, \ldots, d$, $S_{KR,k}(\boldsymbol{x}_{<k}, x_k) = \mathcal{O}(x_k)$ and $\partial_k S_{KR,k}(\boldsymbol{x}_{<k}, x_k) = \mathcal{O}(1)$ as $|x_k| \to \infty$. Furthermore, we have $S_{KR,k} \in V_k$ and $\operatorname{ess\,inf} \partial_k S_{KR,k} > 0$ for all $k = 1, \ldots, d$.

Proof. See Appendix A.9. □

Let us comment on the assumption (29). While the results in Section 3.2 assume only an upper bound on the joint density π, here we have both an upper and a lower bound. These bounds together imply that the k-th marginal conditional density, for $1 \leq k \leq d$, satisfies $(c/C)\eta_k(x_k) \leq \pi(x_k|\boldsymbol{x}_{<k}) \leq (C/c)\eta_k(x_k)$ for all $\boldsymbol{x}_{\leq k} \in \mathbb{R}^k$. This means that the marginal conditionals of the target density have Gaussian tails, rather than potentially being lighter than Gaussian (as in assumption (19)). This condition guarantees that the asymptotic behavior of each map component $S_{\mathrm{KR},k}$ in its last variable (as $|x_k| \to \infty$) is affine.

We now combine our previous results to show that under the assumption (29) on the target density π, we have $f_{\mathrm{KR},k} \in V_k$ and the existence of a unique solution to $\min_{f \in V_k} \mathcal{J}_k(\mathcal{R}_k(f))$.

Corollary 9. *Under the assumptions on π in Proposition 8, let g be a Lipschitz bijection from $\mathbb{R}$ to $\mathbb{R}_{\geq 0}$ such that $\xi \mapsto (g^{-1}(\xi))^2$ is convex and such that, for any $c > 0$ there exists a constant $L_c < \infty$ such that $|g^{-1}(\xi) - g^{-1}(\xi')| \leq L_c|\xi - \xi'|$ for any $\xi, \xi' \geq c$. Then, for any $1 \leq k \leq d$, there exists a unique function $f_{KR,k} \in V_k$ that satisfies $\operatorname{ess\,inf} \partial_k f_{KR,k} > -\infty$ such that $\mathcal{R}_k(f_{KR,k})$ is the k-th component of the KR rearrangement S_{KR}. As a consequence, $\min_{f \in V_k} \mathcal{J}_k(\mathcal{R}_k(f))$ admits a unique solution and*

$$\mathcal{J}_k(\mathcal{R}_k(f_{KR,k})) = \min_{f \in V_k} \mathcal{J}_k(\mathcal{R}_k(f)).$$

Proof. Combining Proposition 8 and Proposition 5, there exists a function $f_{\mathrm{KR},k} \in V_k$ that satisfies $\operatorname{ess\,inf} \partial_k f_{\mathrm{KR},k} > -\infty$ such that $\mathcal{R}_k(f_{\mathrm{KR},k}) = S_{\mathrm{KR},k}$. Then from Theorem 7 and inequality (28), we have that $f_{\mathrm{KR},k}$ is a global minimizer of $f \mapsto \mathcal{J}_k(\mathcal{R}_k(f))$ over $f \in V_k$. From the uniqueness of the KR rearrangement $S_{\mathrm{KR},k}$ among monotone triangular maps and the injectivity of $\mathcal{R}_k^{-1}$, $f_{\mathrm{KR},k} = \mathcal{R}_k^{-1}(S_{\mathrm{KR},k})$ is unique. □

4. Adaptive parameterization of transport maps

Given an i.i.d. sample $\{\mathbf{X}^i\}_{i=1}^n \sim \pi$, we now propose an adaptive algorithm to build $f \in V_k$ that minimizes the empirical objective $\widehat{\mathcal{L}}_k := \widehat{\mathcal{J}}_k \circ \mathcal{R}_k(f)$, where $\widehat{\mathcal{J}}_k$ is as in (7). Doing so for each $k = 1, \ldots, d$ yields a monotone triangular map, as described in Section 2. For each k, we represent $f \in V_k$ with an m-term expansion

$$f(\boldsymbol{x}_{\leq k}) = \sum_{\boldsymbol{\alpha} \in \Lambda} c_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{x}_{\leq k}), \tag{30}$$

where Λ is a set of multi-indices $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k) \in \mathbb{N}_0^k$ with $\#\Lambda = m$, $c_{\boldsymbol{\alpha}} \in \mathbb{R}$ are coefficients, and $\psi_{\boldsymbol{\alpha}} \colon \mathbb{R}^k \to \mathbb{R}$ are basis functions for V_k, constructed as products of univariate functions,

$$\psi_{\boldsymbol{\alpha}}(\boldsymbol{x}_{\leq k}) = \prod_{j=1}^{k} \psi_{\alpha_j}^j(x_j).$$

Here, $\{\psi_\alpha^j\}_{\alpha \in \mathbb{N}_0}$ is chosen to be a basis of $L^2_{\eta_j}$ if $j < k$ and of $H^1_{\eta_k}$ if $j = k$. Because V_k possesses a tensor product structure (16) and because its norm $\|\cdot\|_{V_k}$ is a product norm, the basis $\{\psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathbb{N}_0^k}$ is orthonormal if $\{\psi_\alpha^j\}_{\alpha \in \mathbb{N}_0}$ is an orthonormal basis for all $j \leq k$. Since f depends linearly on the coefficients $c_{\boldsymbol{\alpha}}$, the smoothness properties of the objective $\mathcal{L}_k$ transfer to the objective function parameterized by the coefficients $c_{\boldsymbol{\alpha}}$, $\boldsymbol{\alpha} \in \Lambda$.

Sections 4.1 and 4.2 describe two different choices of basis: polynomials and wavelets, respectively. Then, in Section 4.3, we present a greedy algorithm for building the multi-index set Λ, applicable to any such hierarchical basis. In addition, we propose a cross validation procedure to determine when to stop enriching Λ in order to avoid over-fitting to the finite sample.

4.1. Polynomial basis

Probabilists' Hermite polynomials $\{\varphi_\alpha\}_{\alpha \in \mathbb{N}}$ form an orthogonal basis for L^2_η where η is the univariate standard Gaussian density. φ_α is a polynomial of degree $\alpha \geq 0$ defined as

$$\varphi_\alpha(x) = (-1)^\alpha \exp(x^2/2) \frac{\mathrm{d}^\alpha}{\mathrm{d}x^\alpha} \exp(-x^2/2).$$

Furthermore, we have $\langle \varphi_\alpha, \varphi_\beta \rangle_{L^2_\eta} = \alpha!\, \delta_{\alpha,\beta}$ so that $\{\varphi_\alpha / \sqrt{\alpha!}\}_{\alpha \in \mathbb{N}_0}$ forms an orthonormal basis for L^2_η. Similarly, $\langle \varphi_\alpha, \varphi_\beta \rangle_{H^1_\eta} = (\alpha+1)!\, \delta_{\alpha,\beta}$ so that $\{\varphi_\alpha / \sqrt{(\alpha+1)!}\}_{\alpha \in \mathbb{N}_0}$ forms a orthonormal basis for H^1_η; see [53, Proposition 1.3]. We can thus tensorize these basis functions to obtain an orthonormal basis for the anisotropic space V_k (16).

In practice, we modify the univariate Hermite polynomials to be linear outside of an arbitrary compact domain. Following [44], we let

$$\psi_{\alpha_j}(x_j) = \frac{1}{\sqrt{Z_{\alpha_j}}} \begin{cases} \varphi_{\alpha_j}(x_j^a) + \varphi'_{\alpha_j}(x_j^a)(x_j - x_j^a) & \text{if } x_j < x_j^a \\ \varphi_{\alpha_j}(x_j) & \text{if } x_j^a \le x \le x_j^b \\ \varphi_{\alpha_j}(x_j^b) + \varphi'_{\alpha_j}(x_j^b)(x_j - x_j^b) & \text{if } x_j > x_j^b \end{cases} \tag{31}$$

where $Z_{\alpha_j} = \alpha_j!$ for $j = 1, \dots, k-1$ and $Z_{\alpha_j} = (\alpha_j + 1)!$ for $j = k$. In our numerical experiments, we set x_j^a and x_j^b to be the 0.01 and 0.99 (empirical) quantiles, respectively, of the marginal sample $\{X_j^i\}_{i=1}^n$. The basis $\{\psi_{\alpha_j}\}_{\alpha_j \in \mathbb{N}_0}$, albeit not exactly orthonormal, is close to being orthonormal for sufficiently small x_j^a and large x_j^b. This is important for the numerical stability of the algorithm presented in the following section.

The basis $\{\psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}}$ generated by the piecewise polynomials (31) has numerical and structural advantages over the Hermite polynomials. First, these functions avoid the fast growth rate of standard polynomials as $\|\boldsymbol{x}\| \to \infty$, providing a more stable extrapolation of the map to the tails of π, where there are typically no samples available to otherwise constrain its growth. Second, a map that reverts to linearity far from the origin yields densities in the class considered in Subsection 3.3.2; see condition (29). Indeed, each map component $S_k = \mathcal{R}_k(f)$ with $f \in \operatorname{span}\{\psi_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \Lambda\}$ behaves linearly as $|x_k| \to \infty$, so that the pullback density $S^\sharp \eta$ has Gaussian tails like the reference η.

4.2. Wavelet basis

Wavelets are a popular approach for approximating functions that are not uniformly smooth [12, 66]. In particular, wavelet techniques define a multi-resolution approximation to a function that can better capture local features than, for instance, global polynomials. Given a compactly supported function $\psi \colon \mathbb{R} \to \mathbb{R}$ called the mother wavelet, and indices $(l, q) \in \mathbb{Z}^2$, the function

$$\psi_{(l,q)} : x \mapsto 2^{l/2} \psi(2^l x - q),$$

is the qth wavelet at the level l. Common choices for the mother wavelet include the Haar, the Daubechies, and the Meyer wavelets.

In our numerical experiments, we use the continuous Mexican hat mother wavelet

$$\psi(x) = \frac{2}{\sqrt{3\sigma}\pi^{1/4}} (1 - (x/\sigma)^2) \exp(-x^2/(2\sigma^2)), \tag{32}$$

with scale parameter $\sigma = 1$ [32], which is obtained from the second derivative of the univariate Gaussian density. The Mexican hat function is also commonly referred to as the Ricker wavelet in the geophysics community. We treat this function as having essentially compact support on $[-6, 6]$, up to numerical precision.

Tensorizing a univariate wavelet in the kth coordinate direction with modified Hermite polynomials in the first $k-1$ directions yields an expansion of the form in (30), where the indices in $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k)$ are tuples of the form

$$\alpha_1, \dots, \alpha_{k-1} \in \mathbb{N}_0, \quad \alpha_k = (l_k, q_k) \in \mathbb{Z}^2. \tag{33}$$

In practice, we consider a truncated set of wavelets by first rescaling the mother wavelet ψ to have the same support as the interval between the 0.01 and 0.99 empirical quantiles of the marginal sample $\{X_j^i\}_{i=1}^n$. Then, we constrain $l_j \geq 0$ and consider $l_j = 0$ to be the coarsest level of the approximation. Next, for any l_j, the translation parameter q_j is bounded as $0 \leq q_j \leq 2^{l_j} - 1$. Thus, we can consider

$$\alpha_1, \ldots, \alpha_{k-1} \in \mathbb{N}_0, \quad \alpha_k = (l_k, q_k) \in \mathbb{N}_0^2.$$

rather than (33).

Lastly, we note that any function f of the form (30) expanded with compactly supported wavelets will decay to 0 as $|x_k| \to \infty$. Thus, the map $S_k = \mathcal{R}_k(f)$ will have asymptotic linear growth as $|x_k| \to \infty$, similarly to the modified Hermite polynomial expansions.

4.3. Adaptive transport map algorithm

Now we propose a greedy method for constructing the multi-index set $\Lambda = \Lambda_t$ in (30). For simplicity, we consider only the case of single indices $\alpha_j \in \mathbb{N}_0$. The extension to the case $\alpha_j \in \mathbb{N}_0^2$ (as in the wavelet construction above) is described in Appendix B. At each greedy iteration t, we add one multi-index $\boldsymbol{\alpha}_t^*$ to Λ_t. Starting with $\Lambda_0 = \emptyset$ we let

$$\Lambda_{t+1} = \Lambda_t \cup \{\boldsymbol{\alpha}_t^*\}, \tag{34}$$

where $\boldsymbol{\alpha}_t^* \notin \Lambda_t$ is a multi-index that yields the *best* improvement of $\widehat{\mathcal{L}}_k$ in a sense to be defined below. Borrowing ideas from [34, 35], we constrain the sets $\{\Lambda_t\}_{t\geq 0}$ to be *downward closed* [11, 13], meaning that they satisfy the property

$$\boldsymbol{\alpha} \in \Lambda_t \text{ and } \boldsymbol{\alpha}' \leq \boldsymbol{\alpha} \;\Rightarrow\; \boldsymbol{\alpha}' \in \Lambda_t, \tag{35}$$

where $\boldsymbol{\alpha}' \leq \boldsymbol{\alpha}$ denotes $\alpha_i' \leq \alpha_i$ for all $1 \leq i \leq k$. Intuitively, (35) means that Λ_t has a pyramidal shape containing no holes. Downward-closed sets are known to preserve good approximation properties and permit a tractable construction of Λ_t; see [13]. Indeed, Λ_{t+1} remains downward-closed if and only if the multi-index $\boldsymbol{\alpha}_t^*$ is selected from the so-called *reduced margin* of Λ_t, defined by

$$\Lambda_t^{\mathrm{RM}} = \{\boldsymbol{\alpha} \notin \Lambda_t \,:\, \boldsymbol{\alpha} - \mathbf{e}_i \in \Lambda_t \text{ for all } 1 \leq i \leq k \text{ with } \alpha_i \neq 0\},$$

where $\mathbf{e}_i$ denotes the i-th canonical vector of $\mathbb{N}^k$. The reduced margin is a subset of the margin set Λ_t^{M} (i.e., multi-indices $\boldsymbol{\alpha} \notin \Lambda_t$ such that $\exists\, i > 0$ where $\boldsymbol{\alpha} - \mathbf{e}_i \in \Lambda_t$); see Figure 3. The size of the reduced margin typically grows more slowly with respect to the dimension k than the margin itself. For instance, if Λ_t contains all multi-indices in a hypercube of width p in dimension k, the margin has cardinality $(p+1)^k - p^k$, while the reduced margin has cardinality k.

Denoting by f_t the minimizer of $f \mapsto \widehat{\mathcal{L}}_k(f)$ over $f \in \mathrm{span}\{\psi_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \Lambda_t\}$, we select $\boldsymbol{\alpha}_t^*$ in the reduced margin Λ_t^{RM} with the following heuristic

$$\boldsymbol{\alpha}_t^* \in \operatorname*{arg\,max}_{\boldsymbol{\alpha} \in \Lambda_t^{\mathrm{RM}}} |\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}_k(f_t)|. \tag{36}$$

Here, the notation $\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}_k(f_t)$ denotes the derivative of $c_{\boldsymbol{\alpha}} \mapsto \widehat{\mathcal{L}}_k(f_t + c_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}})$ evaluated at $c_{\boldsymbol{\alpha}} = 0$. In other words, we select the multi-index $\boldsymbol{\alpha}_t^*$ by choosing the largest functional derivative of $\widehat{\mathcal{L}}_k$ at f_t with respect to the basis functions not contained in the current expansion. This greedy procedure for learning each map component is presented in detail in Algorithm 1.

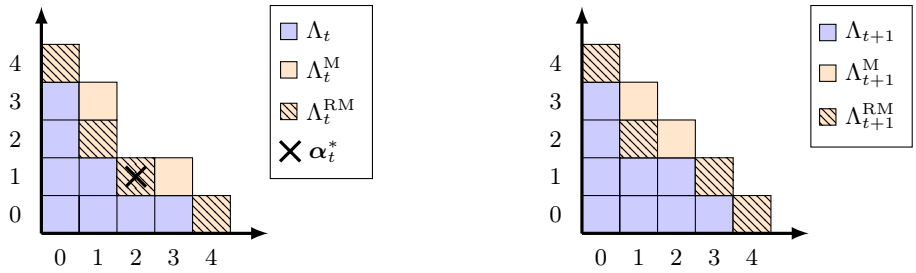


Fig 3: A $k = 2$ dimensional downward-closed active set of multi-indices Λ_t with its margin Λ_t^{M} and reduced margin Λ_t^{RM}. The margin and reduced margins are plotted before (*left*) and after (*right*) adding to $\boldsymbol{\alpha}_t^* = (2, 1)$, that is denoted with a cross, to Λ_t.

Algorithm 1 Estimate map component S_k

Input: training sample $\{\mathbf{X}_{1:k}^i\}_{i=1}^n$, maximum cardinality m for Λ_t
Initialize $\Lambda_0 = \emptyset$, $f_0 = 0$
for $t = 0, \dots, m-1$ **do**
  Construct the reduced margin: Λ_t^{RM}
  Select the new multi-index: $\boldsymbol{\alpha}_t^* \in \arg\max_{\boldsymbol{\alpha} \in \Lambda_t^{\mathrm{RM}}} |\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}_k(f_t)|$
  Update the active set: $\Lambda_{t+1} = \Lambda_t \cup \{\boldsymbol{\alpha}_t^*\}$
  Update the approximation: $f_{t+1} = \arg\min_{f \in \mathrm{span}\{\psi_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \Lambda_{t+1}\}} \widehat{\mathcal{L}}_k(f)$
end for
Output: $\widehat{S}_k = \mathcal{R}_k(f_m)$

Remark 5. *If the objective function $\widehat{\mathcal{L}}_k$ is the linear least-squares loss, then Algorithm 1 corresponds to orthogonal matching pursuit for sparse linear regression with normalized basis functions. An alternative heuristic for selecting $\boldsymbol{\alpha}_t^*$ that requires second-order information from the objective is $\boldsymbol{\alpha}_t^* \in \arg\max_{\boldsymbol{\alpha} \in \Lambda_t^{RM}} |\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}(f_t)|^2 / |\nabla_{\boldsymbol{\alpha}}^2 \widehat{\mathcal{L}}(f_t)|$. We found this criterion to perform similarly to (36) for the polynomial basis, but it led to faster convergence of the objective when working with the wavelet basis.*

Remark 6. *A potential drawback of Algorithm 1 is that the greedy enrichment procedure cannot "see" behind the reduced margin. For instance, if a relevant multi-index is located far beyond the reduced margin and the gradient $\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}_k(f_t)$ vanishes for all $\boldsymbol{\alpha} \in \Lambda_t^{RM}$, then the algorithm will be stuck. [35] suggested a safeguard mechanism to avoid this behaviour: arbitrarily activate the most ancient index from the reduced margin every $t_{sg} \in \mathbb{N}$ iterations. This modification, however, was not needed in our numerical experiments.*

Remark 7. *As pointed out in [13, 34, 35], adding multiple multi-indices at each greedy iteration could also yield better performance compared to adding only one multi-index at a time. The so-called "bulk-chasing" procedure identifies a subset $\lambda_t^* \subset \Lambda_t^{RM}$ of multi-indices that capture a fraction of the L_2-norm of the gradient $\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}_k(f_t)$ along the reduced margin.*

To determine the maximal cardinality of Λ_t (i.e., when to stop adaptation) we use ν-fold cross-validation as in [68, 5]. For each fold, we run Algorithm 1 with $\nu - 1$ folds of the training data for $m = n\frac{\nu-1}{\nu}$ iterations, and evaluate the objective function in (7) on the held-out test set. We then select the number of terms $m^* \leq n$ in the expansion (30) that minimizes the test error averaged over the ν folds. The full sample is then used to run Algorithm 1 for m^* iterations. The complete procedure for learning each map component S_k, with cross-validation, is called the Adaptive Transport Map (`ATM`) algorithm. This procedure is detailed in Algorithm 2. In practice, we also stop the training on

each fold early if the log-likelihood of the test set does not continue to increase for more than 20 iterations.

The cross-validation procedure produces an approximation for f with an expansion whose cardinality is *adapted* to the size of the training sample n. With a larger sample, we observe that the cross-validation procedure reliably adds more parameters, when needed, to reduce the bias in the approximation while controlling the variance of the estimated map parameters. Thus, we consider `ATM` a *semi-parametric* approach for approximating monotone triangular transport maps. Adaptation of the map complexity to the sample size n will be demonstrated in the numerical results to follow.

Algorithm 2 `ATM` algorithm for learning map component S_k

Input: Training sample $\boldsymbol{\chi} = \{\mathbf{X}_{1:k}^i\}_{i=1}^n$, number of folds ν
Partition data into ν folds of equal size
for $j = 1, \dots, \nu$ **do**
  Partition $\boldsymbol{\chi}$: $\boldsymbol{\chi}_j^{\text{test}}$ is the jth subset of $\boldsymbol{\chi}$ and $\boldsymbol{\chi}_j^{\text{train}} = \boldsymbol{\chi} \setminus \boldsymbol{\chi}_j^{\text{test}}$
  Estimate f_t using Algorithm 1 for $m = |\boldsymbol{\chi}_j^{\text{train}}|$ iterations with sample $\boldsymbol{\chi}_j^{\text{train}}$
  Store iterates $\widehat{S}_k^{j,1} \coloneqq \mathcal{R}_k(f_1), \dots, \widehat{S}_k^{j,m} \coloneqq \mathcal{R}_k(f_m)$
  Evaluate log-likelihood $\log(\widehat{S}_k^{j,t})^\sharp \eta$ for iterations $t = 1, \dots, m$ with sample $\boldsymbol{\chi}_j^{\text{test}}$
end for
Define $m^* \in \{1, \dots, m\}$ as the minimizer of negative log-likelihood $t \mapsto \sum_{j=1}^{\nu} -\log(\widehat{S}_k^{j,t})^\sharp \eta(\boldsymbol{\chi}_j^{\text{test}})$
Estimate map f^* using Algorithm 1 for m^* iterations with sample $\boldsymbol{\chi}$
Output: $\widehat{S}_k = \mathcal{R}_k(f^*)$

5. Numerical experiments

In this section, we evaluate the performance of the `ATM` algorithm for learning monotone triangular transport maps associated with a variety of target distributions. Sections 5.1 and 5.2 evaluate and visualize the approximation power of the proposed method for strongly non-Gaussian targets. Sections 5.3 illustrates the benefits of adaptivity over a range of sample sizes, and Section 5.4 demonstrates how the estimated maps reveal and exploit conditional independence structure in the target distribution. Section 5.5 presents joint and conditional density estimation results for a suite of UCI datasets. Code for reproducing all of these numerical results is available online.[3]

In all of the numerical experiments, we pre-process the data by standardizing each marginal, i.e., subtracting the empirical mean and dividing each component of $\mathbf{X}$ by its empirical standard deviation. Within the rectifier $\mathcal{R}_k$, we employ the modified soft-plus function $g(\xi) = \log(1 + 2^\xi)/\log(2)$, which satisfies the conditions of Propositions 4, 5, and 6. This function also satisfies $g(0) = 1$, so that $f(\boldsymbol{x}_{1:k}) = 0$ is transformed into $\mathcal{R}_k(f)(\boldsymbol{x}_{1:k}) = x_k$. We evaluate the integral in (3.1) numerically using an adaptive quadrature method based on Clenshaw-Curtis points with a relative error of 10^{-3}. At each iteration of the `ATM` algorithm, we optimize over the coefficients $c_{\boldsymbol{\alpha}}$ for $\boldsymbol{\alpha} \in \Lambda_t$ using a BFGS (quasi-Newton) method [38].

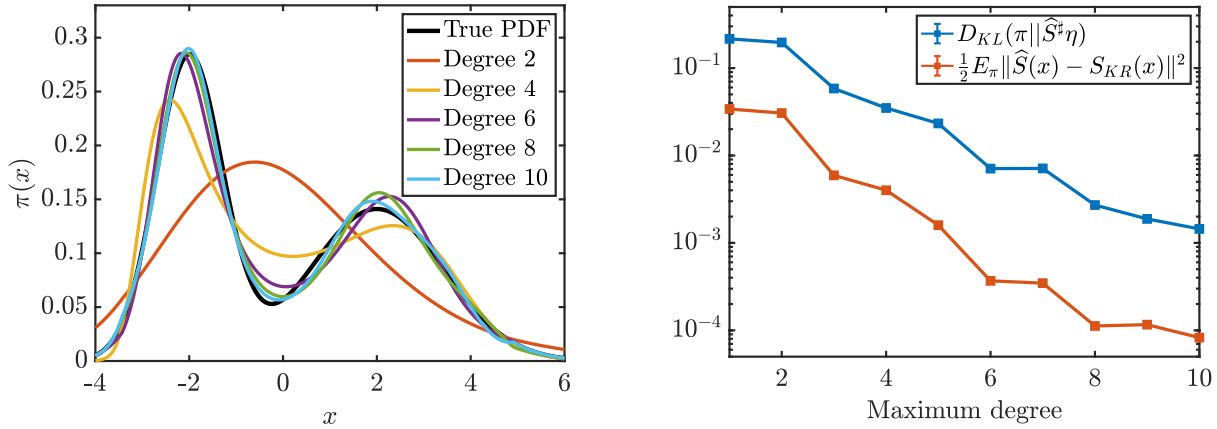
5.1. One-dimensional examples

We first consider a one-dimensional mixture of Gaussians with density $\pi(x) = 0.5\mathcal{N}(x; -2, 0.5) + 0.5\mathcal{N}(x; 2, 2)$. To estimate π, we generate $n = 10^4$ training samples $\{X^i\}_{i=1}^n$ and use them to estimate the map S_{KR} that pushes forward π to η. In one dimension, the KR rearrangement (which coincides with the optimal transport map [51]) is simply the increasing function $S_{\text{KR}} = F_\eta^{-1} \circ F_\pi$, where

[3] https://github.com/baptistar/ATM

F_η and F_π denote the cumulative distribution functions of the reference and target distributions, respectively. We can thus compare our estimated maps to this exact expression.

We approximate S as the transformation $\mathcal{R}(f)$ of a smooth non-monotone function f, seeking f in the finite-dimensional space $V_1^p \subset V_1$ spanned by polynomials of degree at most p. Figure 4a plots the estimated pullback densities $\widehat{\pi} = \widehat{S}^\sharp \eta$ for increasing polynomial degree p, or equivalently an increasing number of terms $\#\Lambda$ in f (30). Figure 4b shows convergence of this approximation in KL divergence, again for increasing p. We also observe convergence of the estimated map $\widehat{S}$ towards the true KR rearrangement in the L^2_π norm, as expected from the upper bound in (4). Figure 5 illustrates the approximation to the map S and the associated non-monotone function f for different choices of basis $\{\psi_\alpha\}_\alpha$. Here we observe that approximating f using the modified Hermite polynomials (31) very closely tracks the KR rearrangement S_{KR} and the function $f_{\mathrm{KR}} = \mathcal{R}^{-1}(S_{\mathrm{KR}})$ as $x \to \pm\infty$, as compared to standard Hermite polynomials or Hermite functions [8]. Enforcing linear asymptotic behavior of the transport map ensures that the approximate distribution has Gaussian tails in regions where there are few samples.



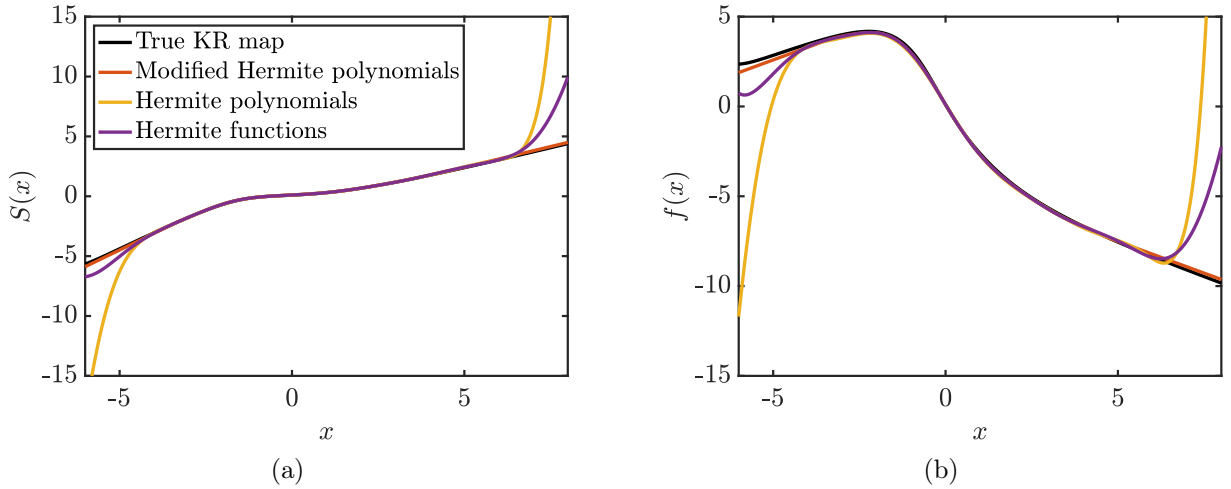
(a) (b)

Fig 4: (a) The pullback density $\widehat{S}^\sharp\eta$ approaches the target Gaussian mixture density π when increasing the maximum polynomial degree p of the space $V_1^p \subset V_1$. (b) With increasing p, the pullback density converges to π in KL divergence, and the estimated map converges to S_{KR} in L^2_π.

Next we consider a Gaussian mixture with density $\pi(x) = 0.5\mathcal{N}(x; 0, 1) + 0.5\mathcal{N}(x; 0, 0.025)$.This mixture is a common test case for sampling and density estimation methods, as it evaluates their ability to capture densities with multiple scales [57]. Given $n = 10^4$ samples, we compute the approximate map using either modified Hermite polynomials (Section 4.1) or Mexican hat wavelets (Section 4.2). Figure 6(a) plots approximations to the target density using an $m = 15$ term expansion (30) for f, while Figure 6(b) shows convergence in KL divergence for both the polynomial and wavelet bases. We observe that the polynomial approximation suffers from oscillations, while the wavelets better capture localized features. This results in a much faster convergence of the KL divergence for wavelets than for polynomials.

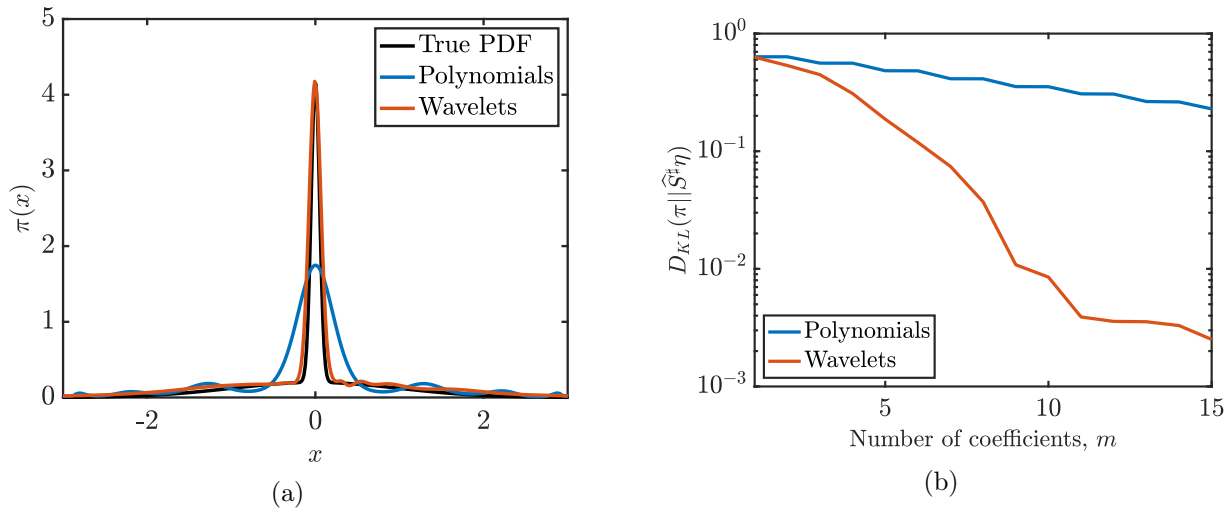
5.2. Two-dimensional datasets

To demonstrate the expressiveness of the modified Hermite polynomial basis (31), we use the `ATM` algorithm to approximate the Knothe–Rosenblatt rearrangement for several two-dimensional distributions with strongly non-Gaussian geometries: the "banana," "funnel," "cosine," "mixture



(a) (b)

Fig 5: (a) The approximate transport maps $\widehat{S}$ compared to S_{KR} (black), and (b) the corresponding non-monotone functions f compared to $f_{\mathrm{KR}} \coloneqq \mathcal{R}^{-1}(S_{\mathrm{KR}})$. Both subfigures illustrate different choices of basis $\{\psi_\alpha\}_\alpha$, for the Gaussian mixture target of Figure 4. The modified Hermite polynomial basis provides the closest approximation to S_{KR} and f_{KR}.



(a) (b)

Fig 6: Approximation of a scale mixture of Gaussians, using Hermite polynomials or Ricker wavelet expansions. (a) A wavelet expansion with $m = 15$ coefficients approximates the target density better than a polynomial expansion of the same size, and (b) yields significantly lower KL divergence for this example.

of Gaussians (MoG)," and "ring" distributions π considered in [70, 22]. For each distribution, we generate an i.i.d. sample of size $n = 10^4$ from π and apply a random rotation to the data using an angle uniformly distributed in $[0, \pi/2]$. Figure 7 plots the true densities π and the approximate densities $\widehat{\pi} \coloneqq \widehat{S}^\sharp\eta$ found using `ATM`. We use 5-fold cross-validation to identify the optimal number of elements in the multi-index set for each map component: S_1 and S_2. Figure 7 indicates the total number of coefficients $\#\Lambda_{m^*}$ across both map components, underscoring the parsimony of the approximation.

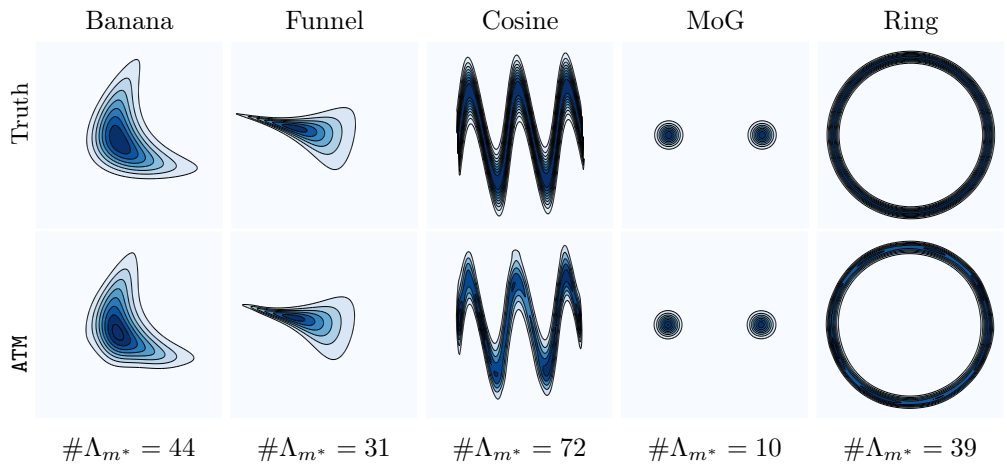


Fig 7: Five different densities π (top row) and their approximations $\widehat{\pi}$ (bottom row) built using the `ATM` algorithm with a Hermite polynomial basis for V_k, given a sample of size $n = 10^4$ from π. The total number of coefficients $\#\Lambda_{m^*}$ in both map components is indicated below each density.

5.3. Random mixture of Gaussians

In this example, we evaluate the stability and accuracy of the `ATM` algorithm for learning transport maps in the small sample regime. We consider a three-dimensional distribution π defined as a mixture of Gaussians centered at the eight vertices of the hypercube $[-4, 4]^3$, each with covariance matrix I_d. The weights of the mixture components are randomly sampled from a uniform distribution $\mathcal{U}([0, 1])$ and then normalized so that $\int_{\mathbb{R}^3} \mathrm{d}\pi = 1$. To estimate π, we generate a training sample $\{\boldsymbol{X}^i\}_{i=1}^n \sim \pi$ and use the `ATM` algorithm to estimate the KR rearrangement that pushes forward π to η using 5-fold cross validation.

Figure 8a plots the KL divergence of $\widehat{\pi}$ from π averaged over 10 experiments. Here, the training sets used to build $\widehat{\pi} = \widehat{S}^\sharp \eta$ are of varying size $n \in [10^1, 10^4]$ and the reported KL divergence is computed on a common test set of 10^4 samples. Figure 8b plots the total number of coefficients $\#\Lambda_{m^*}$ identified by the `ATM` algorithm. The performance of `ATM` is compared to a non-adaptive method where $\Lambda = \Lambda(p) \coloneqq \{\boldsymbol{\alpha} \in \mathbb{N}_0^k, \|\boldsymbol{\alpha}\|_1 \leq p\}$ is arbitrarily fixed, for $p = 1, 3, 5$. Note that $\Lambda(p)$ corresponds to polynomial f with total degree p. For each sample size n, `ATM` consistently finds a better estimator of π (in the sense of KL divergence) than a non-adaptive method with a fixed degree p, by adaptively identifying the basis $\{\psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \Lambda_t}$ to represent the map components. In addition, the `ATM` estimator achieves smaller KL divergence with fewer total map coefficients, as seen in Figure 8b.

5.4. Stochastic volatility

Next we consider data from a Markov process that describes the volatility of the return on a financial asset over time. The model has two hyperparameters μ and ϕ, with a state $(Z_k)_k$ that represents the log-volatility at times $k = 1, \ldots, T$. The two hyperparameters are drawn from the distributions $\mu \sim \mathcal{N}(0, 1)$ and $\phi = 2\exp(\phi^*)/(1+\exp(\phi^*))$ for $\phi^* \sim \mathcal{N}(3, 1)$. The log-volatility obeys the order-one autoregressive process $Z_{k+1} = \mu + \phi(Z_k - \mu) + \epsilon_k$ for all $k > 1$, where $\epsilon_k \sim \mathcal{N}(0, 1)$ is independent of all other variables and $Z_0|\mu, \phi \sim \mathcal{N}(\mu, \frac{1}{1-\phi^2})$. While the states are Gaussian conditioned on the hyperparameters, the joint distribution of

$$\boldsymbol{X} = (\mu, \phi, Z_1, \ldots, Z_T),$$

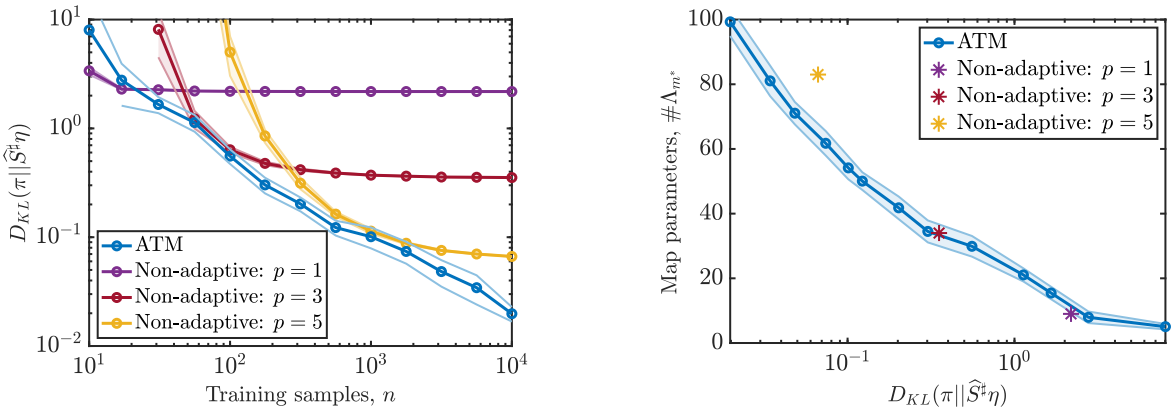


Fig 8: (a) KL divergence over 10 sets of training samples is lower for `ATM` than for non-adapted expansions. (b) Trade-off between the approximation quality and number of coefficients using the adaptive algorithm for different n. For any given number of map coefficients, the `ATM` algorithm finds a representation with similar or lower KL divergence than the lowest achievable error of each non-adaptive approximation (indicated with stars in plot (b)).

is non-Gaussian. In this example, the dimension of $\boldsymbol{X}$ is made arbitrarily large by increasing T.

A key property of triangular transport maps is that they inherit sparsity from the conditional independence structure of π. [59] showed that the Markov structure of π yields a lower bound on the sparsity of the map S (i.e., the functional dependence of each component S^k on the input variables $\boldsymbol{x}_{1:k-1}$). This sparsity was exploited to learn the structure of undirected probabilistic graph models from data in [36, 3]. From the conditional independence properties of the stochastic volatility model described above, we know that the KR rearrangement S_{KR} between the joint distribution of $\boldsymbol{X}$ and a standard Gaussian reference η is sparse. Moreover, the exact sparsity of S_{KR} can be derived from Theorem 5.1 in [59]. We now show that the `ATM` algorithm is capable of detecting and exploiting this structure—without knowing in advance that it exists.

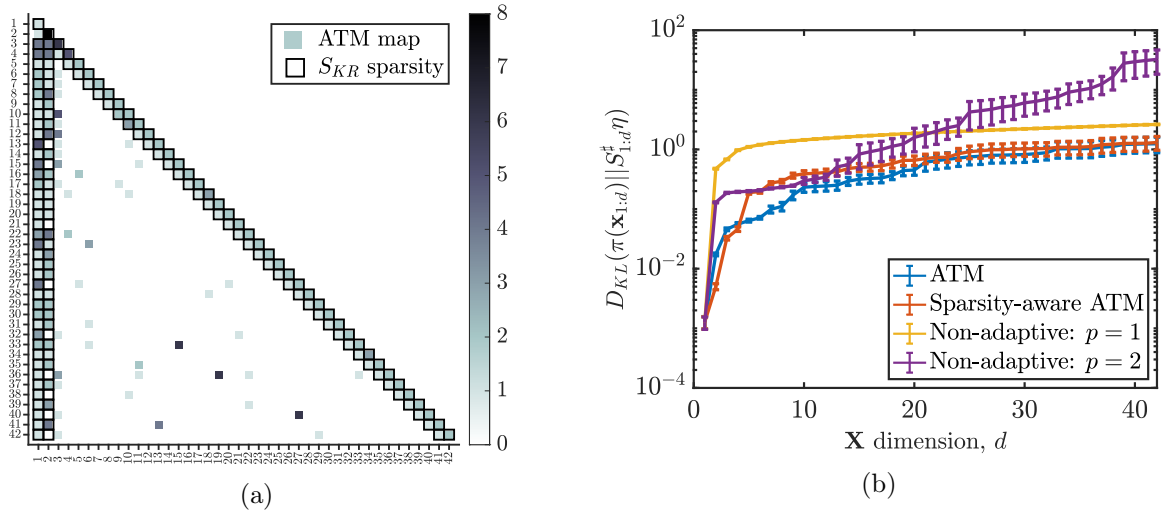
Figure 9a compares the variable dependence of the true KR rearrangement S_{KR} and the map $\widehat{S}$ learned by the `ATM` algorithm for a distribution with $T = 40$ using a sample of size $n = 10^3$ from π; a non-filled entry (j, k) entry of the plot indicates that k-th map component does not depend on variable x_j. The dependence of component S_k on (x_k, x_{k-1}) shows that each state Z_k strongly depends on the previous state in time. Most of the map components also show dependence on the hyperparameters (μ, ϕ). The estimated sparsity closely matches the exact sparsity of the KR rearrangement. Furthermore, the sparse variable dependence of the kth map component $\widehat{S}_k$ on parent nodes $\text{Pa}(k) \subseteq \{1, \ldots, k-1\}$ produces an approximation to the kth marginal conditional density given by

$$\widehat{\pi}(x_k|\boldsymbol{x}_{<k}) = \widehat{\pi}(x_k|\boldsymbol{x}_{\text{Pa}(k)}).$$

By identifying the parent nodes $\text{Pa}(k)$ of each variable k, we also learn a sparse Bayesian network or directed acyclic graphical (DAG) model representing the target distribution [26]. As a result, we can see the `ATM` algorithm as a technique for learning DAGs from samples, given a prescribed variable ordering.

Next, we consider the approximation of S_{KR} for Markov models of *increasing* state dimension T and hence increasing map dimension $d = T + 2$; these experiments use a fixed sample size $n = 10^3$ as before. Figure 9b plots the KL divergence from the `ATM` approximation to the joint density of

$\boldsymbol{X} \in \mathbb{R}^d$, for increasing dimensions d. We compare `ATM` with a variant of `ATM` where the exact sparsity pattern of the the KR rearrangement is provided to the algorithm in advance. This variant, labeled "sparsity-aware `ATM`" in Figure 9b, differs from the `ATM` algorithm in that it only activates multi-indices $\boldsymbol{\alpha}_t^*$ which match the sparsity of S_{KR} (meaning of the form $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, 0, \ldots, 0, \alpha_{k-1}, \alpha_k)$). We also compare `ATM` with non-adaptive maps of degree $p = 1$ and $p = 2$; these maps do not exploit the conditional independence structure of π, and hence depend on all input variables. For low dimensions, we see that degree $p = 2$ maps can better capture the non-Gaussian target distributions than $p = 1$ linear maps. As the dimension d increases, however, the growing number of coefficients in the $p = 2$ maps results in higher-variance estimators, outweighing their smaller bias, and hence we see a rapidly increasing KL divergence as the dimension d increases (with crossover around $d = 21$). In contrast, `ATM` outperforms the non-adaptive maps for all d and, more importantly, achieves a KL divergence that grows linearly with d; indeed, it performs similarly to the best case "oracle sparsity" of sparsity-aware `ATM`.



(a) (b)

Fig 9: (a) The sparsity of an `ATM` map $\widehat{S}$ and of the KR rearrangement S_{KR} for the stochastic volatility model with $T = 40$ time steps. (b) KL divergence using `ATM` with increasing dimension d, compared to adaptive map estimators with known sparsity and non-adaptive maps estimators.

5.5. Tabular datasets

Lastly, we evaluate `ATM`'s performance for density estimation on a suite of UCI datasets [30] that were also considered in [64]. These datasets have dimensionalities between $d = 10$ and 15 and sample sizes between $n = 506$ and 5875. We pre-process each dataset to eliminate discrete-valued variables and one variable in every pair that has Pearson correlation coefficient greater than 0.98, following the procedure in [64]. We consider 10 splits of the data. For each split, we use one fold (i.e., 10% of the data) as a test set and the remaining 9 folds (i.e., 90% of the data) as the training set to build $\widehat{S}$. To assess the quality of our estimated map $\widehat{S}$, we evaluate the negative log-likelihood of the pullback density $\widehat{\pi} = \widehat{S}^\sharp \eta$ on the test set. The negative log-likelihood is an empirical estimator of $-\mathbb{E}_\pi[\log \widehat{S}^\sharp \eta] = \mathcal{D}_{\mathrm{KL}}(\pi || \widehat{S}^\sharp \eta) - \int \log \pi \mathrm{d}\pi$ and is, up to the unknown constant $-\int \log \pi \mathrm{d}\pi$, the KL divergence from $\widehat{S}^\sharp \eta$ to π. Table 1 presents the mean of the negative log-likelihoods over the 10 splits and a 95% confidence interval for the mean. For all datasets, we observe an improvement using `ATM`

over non-adaptive $p = 2$ maps and multivariate Gaussian approximations (i.e., $p = 1$ maps), with a lower number of total coefficients.

TABLE 1
Mean negative log-likelihood for UCI datasets over 10 sets of training samples. The estimator with best performance (lowest negative log-likelihood) is highlighted in bold.

Dataset	(d, n)	Gaussian	`ATM`	# coefficients	$p = 2$ maps	# coefficients
Boston	$(10, 506)$	11.3 ± 0.5	$\mathbf{3.1 \pm 0.6}$	228 ± 7	6.5 ± 0.4	285
Red Wine	$(11, 1599)$	13.2 ± 0.3	$\mathbf{9.8 \pm 0.4}$	289 ± 9	10.5 ± 0.2	363
White Wine	$(11, 4898)$	13.2 ± 0.5	$\mathbf{11.0 \pm 0.2}$	342 ± 26	12.0 ± 1.0	363
Parkinsons	$(15, 5875)$	10.8 ± 0.4	$\mathbf{2.8 \pm 0.4}$	783 ± 17	5.1 ± 0.4	815

Lastly, we evaluate `ATM`'s performance for *conditional* density estimation on a different suite of UCI datasets. We follow a similar procedure as above to pre-process each dataset. Each dataset has a one-dimensional predictor variable X_d and covariates $\boldsymbol{X}_{<d}$ of varying dimension. To approximate the conditional density $\pi_d(x_d|\boldsymbol{x}_{<d})$, we estimate one map component S_d as a function of the predictor and the covariates using joint samples $\{(\boldsymbol{X}^i_{<k}, X^i_k)\}_{i=1}^n$. Table 2 presents the mean negative conditional log-likelihoods over 10 splits of the training data. The negative conditional log-likelihood is an empirical estimator of $-\mathbb{E}_\pi[\log \widehat{S}_d^\sharp \eta] = \mathbb{E}_{\pi(\boldsymbol{x}_{<d})}[\mathcal{D}_{\mathrm{KL}}(\pi_d(x_d|\boldsymbol{x}_{<d})||\widehat{S}_d^\sharp \eta)] - \int \log \pi_d(x_d|\boldsymbol{x}_{<d}) \mathrm{d}\pi_d$ and is, up to the unknown constant $-\int \log \pi_d(x_d|\boldsymbol{x}_{<d}) \mathrm{d}\pi_d$, the expected KL divergence from $\widehat{S}_d^\sharp \eta$ to the marginal conditional $\pi_d(x_d|\boldsymbol{x}_{<d})$.

In Table 2, we compare `ATM` to conditional kernel density estimation (`CKDE`) [56], ϵ-neighborhood kernel density estimation (`NKDE`), and kernel mixture networks (`KMN`) [1] using the implementation provided by [50]. We also include results for high-capacity parametric conditional density estimation methods that rely on neural networks: mixture density networks (`MDN`) [6] and conditional normalizing flows (`NF`) [62]; implementation details of these methods are provided in Appendix C. For all datasets, we observe that `ATM` has performance comparable to the neural-network based methods and improved performance over the nonparametric approaches (e.g., on the Concrete and Yacht datasets).

In Table 3, we compare the number of coefficients in the `ATM` and `MDN` approximations and the computational time required to learn each of these conditional density representations. The runtime of the greedy procedure in Algorithm 1 (sequentially, for all d map components) is generally less than that of `MDN`, achieving similar approximation performance with about two orders of magnitude fewer coefficients. It is more realistic, however, to consider `ATM` with the cross-validation procedure of Algorithm 2 and to compare this with the hyperparameter tuning required to achieve the reported performance of `MDN`. `ATM` only requires performing cross-validation for a single hyper-parameter, i.e., the total number of coefficients m^* (see Algorithm 2). In comparison, tuning several hyperparameters in the neural-network based methods (e.g., the learning rate, and the structure of the networks) by performing cross-validation over a tensor product grid increases runtimes by more than an order of magnitude relative to `ATM`, as presented below.

TABLE 2
Mean negative conditional log-likelihood for UCI datasets over 10 sets of training samples. The method with best performance from both categories is highlighted in bold.

Dataset	(d, n)	`ATM`	`CKDE`	`NKDE`	`MDN`	`KMN`	`NF`
Boston	(12,506)	$\mathbf{2.6 \pm 0.2}$	$\mathbf{2.6 \pm 0.2}$	3.1 ± 0.2	$\mathbf{2.4 \pm 0.2}$	2.7 ± 0.2	$\mathbf{2.4 \pm 0.1}$
Concrete	(9,1030)	$\mathbf{3.1 \pm 0.1}$	3.2 ± 0.1	3.9 ± 0.1	$\mathbf{2.9 \pm 0.1}$	3.5 ± 0.1	3.2 ± 0.2
Energy	(10,768)	1.5 ± 0.1	$\mathbf{1.0 \pm 0.1}$	2.1 ± 0.2	$\mathbf{1.2 \pm 0.1}$	1.7 ± 0.1	1.7 ± 0.3
Yacht	(7,308)	$\mathbf{0.5 \pm 0.2}$	1.1 ± 0.3	3.8 ± 0.2	$\mathbf{0.7 \pm 0.2}$	1.8 ± 0.2	1.3 ± 0.5

TABLE 3
The number of coefficients in the `ATM` *and* `MDN` *approximations, as well as the runtimes (in seconds) to identify coefficients via optimization and to perform cross-validation of hyperparameters.*

	`ATM`			`MDN`		
Dataset	# coefficients	runtime (s)	runtime w/CV	# coefficients	runtime (s)	runtime w/CV
Boston	31 ± 3	6 ± 1	88 ± 14	$(5 \pm 2) \times 10^3$	18 ± 8	1584 ± 13
Concrete	42 ± 5	14 ± 3	230 ± 18	$(2.6 \pm 0.3) \times 10^3$	11 ± 1	1846 ± 10
Energy	41 ± 7	12 ± 5	193 ± 29	$(8 \pm 3) \times 10^3$	32 ± 12	1758 ± 17
Yacht	28 ± 4	7 ± 2	99 ± 20	$(6 \pm 2) \times 10^3$	23 ± 6	1453 ± 15

6. Conclusions

This paper has presented and analyzed a functional framework for learning monotone triangular transport maps. Our approach represents monotone component functions of a triangular map through the action of an invertible *rectification* operator $\mathcal{R}_k$ on smooth, generally non-monotone, functions. Imposing appropriate structure on this operator and the function space V_k on which it acts, along with conditions on the target density π, yields an unconstrained optimization problem for learning the map components that has many desirable and useful properties. First, under certain assumptions on $\mathcal{R}_k$ and π, we show that the optimization objective is bounded, continuous, and differentiable for all $f \in V_k$. This permits the use of deterministic gradient-based optimization methods to find the minimizers. Next, under certain conditions on $\mathcal{R}_k$, we show that the optimization problem has *no spurious local minima* on V_k. In practice, this yields important robustness to the initial guess and other parameters of the optimization algorithms. Finally, under the same (cumulative) conditions on $\mathcal{R}_k$ and some additional assumptions on the target density, we show that the optimization problem has a unique global minimizer on V_k that corresponds to the canonical KR rearrangement.

Our functional framework also enables the construction of novel transport map estimators, based on maximum likelihood, given a finite sample from the target. The procedure we develop here is semi-parametric, making use of different hierarchical bases for V_k. In particular, we propose a greedy, adaptive algorithm that identifies a sparse set of basis functions, automatically tailored to the target density and to the sample size. Sparsity in the basis selection reflects conditional independence structure in the target. More generally, the maps we build can capture the structure of complex probability distributions with sufficient data, but also are robust in settings with few observations. This is crucial to deploying these algorithms within large-scale applications, such as data assimilation, where triangular transport maps must be learned online [58]. We demonstrate good computational performance of our algorithm—and the parsimonious representations it produces—in a variety of joint and conditional density estimation problems. We outline some interesting future research directions below.

KR recovery for broader classes of target distributions While our result guaranteeing no spurious local minima (Theorem 7) makes no explicit assumptions on π, our subsequent results on recovery of the KR rearrangement (see Corollary 9) make particular use of Gaussian tail assumptions, relating the target π to the Gaussian reference η. In principle, however, the space V_k contains functions $\mathcal{R}_k^{-1}(S_{\mathrm{KR},k})$ corresponding to a broader class of target densities; V_k includes functions that grow up to sub-exponentially fast, as a result of its Gaussian weighting. Hence, it includes the inverse images of KR rearrangements for lighter-tailed π, whose component functions grow faster than linear as $x_k \to \pm\infty$. It may be possible to obtain unique recovery results analogous to Corollary 9 for such densities, although some modifications to our analysis would be needed.

Approximation theory and statistical guarantees An open research topic, to the best of our knowledge, is to analyze the finite-dimensional approximation of triangular transport maps on unbounded domains. (See [71, 72] for an approximation theory of triangular transports between analytic densities π, η on bounded domains.) These results would show how approximate maps converge to the KR rearrangement with different parameterizations (e.g., polynomial spaces or neural networks of increasing size), and we expect that our rectification framework could be a useful ingredient of such analyses. A parallel line of work is to develop non-asymptotic statistical convergence results for triangular transports (e.g., in the context of density estimation), to understand how the quality of density estimates and of map estimates depends on the sample size n. Combining these lines of work could provide lower bounds on the number of samples required to learn maps of a given complexity, and also some analytical guidance for how to navigate the bias-variance tradeoff of a map estimator given finite samples.

Other sources of low-dimensional structure The `ATM` algorithm in this paper identifies maps with sparse variable dependence by exploiting conditional independence structure in the target distribution [59]. In future work it will be interesting to investigate other notions of low-dimensional structure and how they could facilitate the learning of maps from small samples. One promising notion is when the target density π departs from the reference η only along a low-dimensional subspace. In this case, the triangular map S pushing forward π to η can be written as a low-dimensional perturbation of the identity map, after a variable rotation. See [9] for a recent contribution in this direction applied to variational inference.

Map ordering One disadvantage of triangular maps is that the approximation depends on the choice of variable ordering. In particular, each variable ordering yields a different factorization of the target density and a different Knothe–Rosenblatt rearrangement S_{KR}. Thus, it is of interest to develop variable ordering algorithms that minimize the finite-sample error of the estimated transport map $\widehat{S}$, the estimated pullback density $\widehat{\pi} = \widehat{S}^\sharp \eta$, or other goal-oriented metrics. For target distributions that induce sparse variable dependence in the map, one approach is to find the permutation that maximizes the sparsity of $\widehat{S}$. This is equivalent to finding the the sparsest Bayesian network or directed acyclic graphical (DAG) model for a distribution from samples. While this problem is in general NP-complete, effective algorithms [47] have been proposed. This algorithm reduces to finding a sparse maximum likelihood estimator for the Cholesky factor of the inverse covariance matrix of π in the Gaussian setting. Since a linear transport map is precisely this Cholesky factor for Gaussian π and standard normal η (see [3, Section 3]), we expect that such sparse permutation algorithms could be generalized to the nonlinear transport map setting.

Nonparametric methods Instead of finding a particular finite-dimensional basis for the map components $S_k = \mathcal{R}_k(f)$, nonparametric methods do not limit the functional form of f (e.g., as linear combinations of multivariate polynomials). A broadly useful nonparametric approach involves seeking f in a reproducing kernel Hilbert space (RKHS). Thanks to the representer theorem [54], the optimal f in the RKHS can still be identified by solving a finite-dimensional optimization problem. Choosing the kernel so that the RKHS is a suitable weighted Sobolev space (see, e.g., [39]) will yield map representations that fall within the framework of Section 3.2. A related approach uses Gaussian process representations for the map components [23]. It will be interesting to compare the finite-sample performance of such methods to the semi-parametric procedure proposed in this work.

Acknowledgments

RB, YM, and OZ gratefully acknowledge support from the INRIA associate team Unquestionable. RB and YM are also grateful for support from the AFOSR Computational Mathematics program (MURI award FA9550-15-1-0038) and the US Department of Energy AEOLUS center. RB acknowledges support from an NSERC PGSD-D fellowship. OZ also acknowledges support from the ANR JCJC project MODENA (ANR-21-CE46-0006-01).

Appendix A: Proofs and theoretical details

A.1. Proof of Proposition 1

Proof. Recall that the KR rearrangement S_{KR} is a transport map that satisfies $S_{\mathrm{KR}}^\sharp \eta = \pi$, where η is the density of the standard Gaussian measure on $\mathbb{R}^d$ and π is the target density. Corollary 3.10 in [7] states that for any PDF ϱ on $\mathbb{R}^d$ of the form $\varrho(\boldsymbol{x}) := f(\boldsymbol{x})\eta(\boldsymbol{x})$ with $f \log f \in L^1_\eta$, the inequality

$$\int \|\boldsymbol{x} - T(\boldsymbol{x})\|^2 \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \leq 2 \int f(\boldsymbol{x}) \log f(\boldsymbol{x}) \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{37}$$

holds, where T is the KR rearrangement such that $T_\sharp \eta = \varrho$. Let S be an increasing lower triangular map as in (1) and let $\varrho = S_\sharp \pi$. Thus we have $T = S \circ S_{\mathrm{KR}}^{-1}$ and so the left-hand side of (37) becomes

$$\int \|\boldsymbol{x} - T(\boldsymbol{x})\|^2 \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int \|\boldsymbol{x} - S \circ S_{\mathrm{KR}}^{-1}(\boldsymbol{x})\|^2 \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int \|S_{\mathrm{KR}}(\boldsymbol{x}) - S(\boldsymbol{x})\|^2 \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$

and the right-hand side becomes

$$2 \int f(\boldsymbol{x}) \log f(\boldsymbol{x}) \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 2\mathcal{D}_{\mathrm{KL}}(\varrho || \eta) = 2\mathcal{D}_{\mathrm{KL}}(\pi || S^\sharp \eta),$$

which yields (4). □

A.2. Convexity of $s \mapsto \mathcal{J}_k(s)$

Lemma 10. *The optimization problem* $\min_{\{s:\, \partial_k s > 0\}} \mathcal{J}_k(s)$ *is strictly convex.*

Proof. Let $s_1, s_2 \colon \mathbb{R}^k \to \mathbb{R}$ be two strictly increasing functions with respect to x_k, i.e., $\partial_k s_1(\boldsymbol{x}_{\leq k}) > 0$ and $\partial_k s_2(\boldsymbol{x}_{\leq k}) > 0$. For any $0 < t < 1$, the function $s_t = t s_1 + (1-t) s_2$ is also strictly increasing with respect to x_k. Finally, because both $\xi \mapsto \frac{1}{2}\xi^2$ and $\xi \mapsto -\log(\xi)$ are strictly convex functions, we have

$$\begin{aligned}
\mathcal{J}_k(s_t) &\overset{(6)}{=} \int \left(\frac{1}{2} s_t(\boldsymbol{x}_{\leq k})^2 - \log \partial_k s_t(\boldsymbol{x}_{\leq k}) \right) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&< \int \left(t \frac{1}{2} s_1(\boldsymbol{x}_{\leq k})^2 + (1-t) \frac{1}{2} s_2(\boldsymbol{x}_{\leq k})^2 \right) - \left(t \log \partial_k s_1(\boldsymbol{x}_{\leq k}) + (1-t) \log \partial_k s_2(\boldsymbol{x}_{\leq k}) \right) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&= t\mathcal{J}_k(s_1) + (1-t)\mathcal{J}_k(s_2),
\end{aligned}$$

which shows that $\mathcal{J}_k$ is strictly convex. □

A.3. Proof of Proposition 2

To prove Proposition 2 we need the following lemma.

Lemma 11. *Let*

$$H^1([0,1]) = \left\{ f\colon [0,1] \to \mathbb{R} \text{ such that } \|f\|^2_{H^1([0,1])} \coloneqq \int_0^1 f(t)^2 + f'(t)^2 \,\mathrm{d}t \right\}.$$

Then

$$|f(0)| \leq \sqrt{2}\|f\|_{H^1([0,1])}, \tag{38}$$

holds for any $f \in H^1([0,1])$.

Proof. Because $\mathcal{C}^\infty([0,1])$ is dense in $H^1([0,1])$, it suffices to show (38) for any $f \in \mathcal{C}^\infty([0,1])$. By the mean value theorem, there exists $0 \leq z \leq 1$ such that

$$f(z) = \frac{1}{1-0}\int_0^1 f(t)\mathrm{d}t.$$

Thus we can write

$$\begin{aligned}
|f(0)|^2 &\leq 2|f(z) - f(0)|^2 + 2|f(z)|^2 \\
&= 2\left|\int_0^z f'(t)\mathrm{d}t\right|^2 + 2\left|\int_0^1 f(t)\mathrm{d}t\right|^2 \\
&\leq 2\int_0^1 \left|f'(t)\right|^2 \mathrm{d}t + 2\int_0^1 |f(t)|^2 \,\mathrm{d}t.
\end{aligned}$$

This concludes the proof. □

We now prove Proposition 2.

Proof. For any $f \in V_k$, Lemma 11 permits us to write

$$\begin{aligned}
\int |f(\boldsymbol{x}_{<k},0)|^2 \eta_{<k}(\boldsymbol{x}_{<k})\mathrm{d}\boldsymbol{x}_{<k} &\overset{(38)}{\leq} \int \left(2\int_0^1 |f(\boldsymbol{x}_{<k},t)|^2 + |\partial_k f(\boldsymbol{x}_{<k},t)|^2 \mathrm{d}t\right)\eta_{<k}(\boldsymbol{x}_{<k})\mathrm{d}\boldsymbol{x}_{<k} \\
&\leq C_T \int\int_0^1 \Big(|f(\boldsymbol{x}_{<k},t)|^2 + |\partial_k f(\boldsymbol{x}_{<k},t)|^2\Big)\eta_{<k}(\boldsymbol{x}_{<k})\eta_1(t)\mathrm{d}\boldsymbol{x}_{<k}\mathrm{d}t \\
&\leq C_T \int\int_{-\infty}^{+\infty} \Big(|f(\boldsymbol{x}_{<k},t)|^2 + |\partial_k f(\boldsymbol{x}_{<k},t)|^2\Big)\eta_{\leq k}(\boldsymbol{x}_{<k},t)\mathrm{d}\boldsymbol{x}_{<k}\mathrm{d}t \\
&= C_T\|f\|^2_{V_k},
\end{aligned}$$

where $C_T = 2\sup_{0\leq t\leq 1}\eta_1(t)^{-1}$. □

A.4. Proof of Proposition 3

The proof relies on Proposition 2 and on the following generalized integral Hardy inequality, see [37].

Lemma 12. *Let* $\eta_{\leq k}$ *be the standard Gaussian density on* $\mathbb{R}^k$. *Then there exists a constant* C_H *such that for any* $v \in L^2_\eta(\mathbb{R}^k)$,

$$\int \left(\int_0^{x_k} v(\boldsymbol{x}_{<k},t)\mathrm{d}t\right)^2 \eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \leq C_H \int v(\boldsymbol{x})^2\eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{39}$$

Proof of Lemma 12. Let us recall the integral Hardy inequality [37].

Theorem 13 (from [37]). *For weight $\rho\colon \mathbb{R}_+ \to \mathbb{R}_+$ and $u \in L^2_\rho(\mathbb{R})$, there exists a constant $C_H < \infty$ such that*

$$\int_0^{+\infty} \left(\int_0^x u(t)\mathrm{d}t\right)^2 \rho(x)\mathrm{d}x \le C_H \int_0^{+\infty} u(x)^2\rho(x)\mathrm{d}x \tag{40}$$

if and only if

$$\sup_{x>0} \left(\int_x^{+\infty} \rho(t)\mathrm{d}t\right)^{1/2} \left(\int_0^x \rho(t)^{-1}\mathrm{d}t\right)^{1/2} < +\infty. \tag{41}$$

We apply Theorem 13 with the one-dimensional standard Gaussian density $\rho = \eta$ for $x > 0$. In order to check condition (41), we need to show that

$$D(x) \coloneqq \left(\int_x^{+\infty} \rho(t)\mathrm{d}t\right)^{1/2} \left(\int_0^x \rho(t)^{-1}\mathrm{d}t\right)^{1/2} = \left(\int_x^{+\infty} e^{-t^2/2}\mathrm{d}t\right)^{1/2} \left(\int_0^x e^{t^2/2}\mathrm{d}t\right)^{1/2},$$

is bounded. Since $x \mapsto D(x)$ is a continuous function with a finite limit as $x \to 0$, it is sufficient to show that $D(x)$ has a finite limit when $x \to \infty$. For $x > 1$, $\int_x^{+\infty} e^{-t^2/2}\mathrm{d}t \le e^{-x^2/2}$ and $D(x)^2 \le e^{-x^2/2}\int_0^x e^{t^2/2}\mathrm{d}t$. Furthermore, using integration-by-parts we have $\int_0^x e^{t^2/2}\mathrm{d}t = \int_0^1 e^{t^2/2}\mathrm{d}t + e^{x^2/2}/x - \sqrt{e} + \int_1^x e^{t^2/2}/t^2\mathrm{d}t$. As $x \to \infty$ the dominating term in the sum is $e^{x^2/2}/x$. Thus, $e^{-x^2/2}\int_0^x e^{t^2/2}\mathrm{d}t$ behaves asymptotically as $\mathcal{O}(\frac{1}{x})$, so that $D(x) \to 0$ when $x \to \infty$. Thus, condition (41) is satisfied.

Then, by the Hardy inequality in (40) for $u \in L^2_\eta(\mathbb{R})$ we have

$$\int_0^{+\infty} \left(\int_0^{x_k} u(t)\mathrm{d}t\right)^2 \eta(x_k)\mathrm{d}x_k \le C_H \int_0^{+\infty} u(x_k)^2\eta(x_k)\mathrm{d}x_k. \tag{42}$$

For the symmetric density $\eta(x_k) = \eta(-x_k)$ we also have

$$\int_{-\infty}^0 \left(\int_0^{x_k} u(t)\mathrm{d}t\right)^2 \eta(x_k)\mathrm{d}x_k \le C_H \int_{-\infty}^0 u(x_k)^2\eta(x_k)\mathrm{d}x_k. \tag{43}$$

Combining the results in (42) and (43) we have

$$\int_{-\infty}^{+\infty} \left(\int_0^{x_k} u(t)\mathrm{d}t\right)^2 \eta(x_k)\mathrm{d}x_k \le C_H \int_{-\infty}^{+\infty} u(x_k)^2\eta(x_k)\mathrm{d}x_k.$$

Setting $u(t) = v(\boldsymbol{x}_{<k}, t)$ and integrating both sides over $\boldsymbol{x}_{<k} \in \mathbb{R}^{k-1}$ with the standard Gaussian weight function $\eta(\boldsymbol{x}_{<k})$ gives the result. □

We now prove Proposition 3.

Proof. By Proposition 2, Lemma 12 and by the Lipschitz property of g, we can write

$$\begin{aligned}
\|\mathcal{R}_k(f_1) - \mathcal{R}_k(f_2)\|^2_{L^2_{\eta_{\le k}}} &\le 2\int \Big(f_1(\boldsymbol{x}_{<k}, 0) - f_2(\boldsymbol{x}_{<k}, 0)\Big)^2 \eta_{\le k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\quad + 2\int \Big(\int_0^{x_k} g\big(\partial_k f_1(\boldsymbol{x}_{<k}, t)\big) - g\big(\partial_k f_2(\boldsymbol{x}_{<k}, t)\big)\mathrm{d}t\Big)^2 \eta_{\le k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\le 2C_T\|f_1 - f_2\|^2_{V_k} + 2C_H\|g(\partial_k f_1) - g(\partial_k f_2)\|^2_{L^2_{\eta_{\le k}}} \\
&\le 2C_T\|f_1 - f_2\|^2_{V_k} + 2C_H L^2\|\partial_k f_1 - \partial_k f_2\|^2_{L^2_{\eta_{\le k}}} \\
&\le 2(C_T + C_H L^2)\|f_1 - f_2\|^2_{V_k},
\end{aligned} \tag{44}$$

for any $f_1, f_2 \in V_k$. Furthermore, using the Lipschitz property of g we have

$$\begin{aligned}\|\partial_k \mathcal{R}_k(f_1) - \partial_k \mathcal{R}_k(f_2)\|^2_{L^2_{\eta_{\le k}}} &= \int \Big(g(\partial_k f_1(\boldsymbol{x}_{<k}, t)) - g(\partial_k f_2(\boldsymbol{x}_{<k}, t))\Big)^2 \eta_{\le k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &\le L^2 \int \Big(\partial_k f_1(\boldsymbol{x}_{<k}, t) - \partial_k f_2(\boldsymbol{x}_{<k}, t)\Big)^2 \eta_{\le k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &\le L^2 \|f_1 - f_2\|^2_{V_k}. \end{aligned} \tag{45}$$

Combining (44) with (45), we obtain (18) with $C = \sqrt{2(C_T + C_H L^2) + L^2}$.

It remains to show that $\|\mathcal{R}_k(f)\|_{V_k} < \infty$ for any $f \in V_k$. Letting with $f_1 = f$ and $f_2 = 0$ in (18), the triangle inequality yields

$$\|\mathcal{R}_k(f)\|_{V_k} \le \|\mathcal{R}_k(0)\|_{V_k} + C\|f\|_{V_k}$$

Because $\mathcal{R}_k(0)$ is the affine function $\boldsymbol{x} \mapsto g(0)x_k$, we have that $\|\mathcal{R}_k(0)\|^2_{L^2_{\eta_{\le k}}} = g(0)^2 \int x_k^2 \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ and $\|\partial_k \mathcal{R}_k(0)\|^2_{L^2_{\eta_{\le k}}} = g(0)^2$ are finite and so is $\|\mathcal{R}_k(0)\|_{V_k}$. Thus, $\mathcal{R}_k(f) \in V_k$ for all $f \in V_k$. □

A.5. Proof of Proposition 4

Proof. For any $f \in V_k$ we have

$$\begin{aligned}|\mathcal{L}_k(f)| &= \left| \int \left(\frac{1}{2} \mathcal{R}_k(f)^2 - \log(\partial_k \mathcal{R}_k(f)) \right) \mathrm{d}\pi \right| \\ &\overset{(19)}{\le} \frac{C_\pi}{2} \|\mathcal{R}_k(f)\|^2_{L^2_{\eta_{\le k}}} + C_\pi \int |\log(g(\partial_k f))| \, \mathrm{d}\eta_{\le k} \\ &\le \frac{C_\pi}{2} \|\mathcal{R}_k(f)\|^2_{L^2_{\eta_{\le k}}} + C_\pi \int |\log(g(0))| + |\log(g(\partial_k f)) - \log(g(0))| \, \mathrm{d}\eta_{\le k} \\ &\overset{(21)}{\le} \frac{C_\pi}{2} \|\mathcal{R}_k(f)\|^2_{L^2_{\eta_{\le k}}} + C_\pi |\log(g(0))| + C_\pi L \int |\partial_k f - 0| \, \mathrm{d}\eta_{\le k} \\ &\le \frac{C_\pi}{2} \|\mathcal{R}_k(f)\|^2_{L^2_{\eta_{\le k}}} + C_\pi |\log(g(0))| + C_\pi L \|f\|^2_{V_k}. \end{aligned}$$

Because Proposition 3 ensures $\mathcal{R}_k(f) \in V_k \subset L^2_{\eta_{\le k}}$, we have that $\mathcal{L}_k(f)$ is finite for any $f \in V_k$. Now, for any $f_1, f_2 \in V_k$, we can write

$$\begin{aligned}|\mathcal{L}_k(f_1) - \mathcal{L}_k(f_2)| &= \left| \int \left(\frac{1}{2} \mathcal{R}_k(f_1)^2 - \frac{1}{2} \mathcal{R}_k(f_2)^2 - \log(\partial_k \mathcal{R}_k(f_1)) + \log(\partial_k \mathcal{R}_k(f_2)) \right) \mathrm{d}\pi \right| \\ &\overset{(19)}{\le} C_\pi \int \frac{1}{2} \Big| \mathcal{R}_k(f_1)^2 - \mathcal{R}_k(f_2)^2 \Big| + \Big| \log(g(\partial_k f_1)) - \log(g(\partial_k f_2)) \Big| \mathrm{d}\eta \\ &\overset{(21)}{\le} \frac{C_\pi}{2} \|\mathcal{R}_k(f_1) + \mathcal{R}_k(f_2)\|_{L^2_{\eta_{\le k}}} \|\mathcal{R}_k(f_1) - \mathcal{R}_k(f_2)\|_{L^2_{\eta_{\le k}}} + C_\pi L \|\partial_k f_1 - \partial_k f_2\|_{L^2_{\eta_{\le k}}} \\ &\overset{(18)}{\le} C_\pi \frac{\|\mathcal{R}_k(f_1)\|_{L^2_{\eta_{\le k}}} + \|\mathcal{R}_k(f_2)\|_{L^2_{\eta_{\le k}}}}{2} C \|f_1 - f_2\|_{V_k} + C_\pi L \|f_1 - f_2\|_{V_k}. \end{aligned}$$

This shows that $\mathcal{L}_k\colon V_k \to \mathbb{R}$ is continuous. To show that $\mathcal{L}_k$ is differentiable, we let $f, \varepsilon \in V_k$ so that

$$\begin{aligned}
\mathcal{L}_k(f+\varepsilon) &= \int \left(\frac{1}{2}\mathcal{R}_k(f+\varepsilon)^2 - \log(\partial_k \mathcal{R}_k(f+\varepsilon))\right) \mathrm{d}\pi \\
&= \int \left(\frac{1}{2}\left(f(\boldsymbol{x}_{<k},0) + \varepsilon(\boldsymbol{x}_{<k},0) + \int_0^{x_k} g\big(\partial_k(f+\varepsilon)(\boldsymbol{x}_{<k},t)\big)\mathrm{d}t\right)^2 - \log\circ g(\partial_k(f+\varepsilon)(\boldsymbol{x}))\right)\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \int \left(\frac{1}{2}\mathcal{R}_k(f)(\boldsymbol{x})^2 + \mathcal{R}_k(f)(\boldsymbol{x})\left(\varepsilon(\boldsymbol{x}_{<k},0) + \int_0^{x_k} g'\big(\partial_k f(\boldsymbol{x}_{<k},t)\big)\partial_k\varepsilon(\boldsymbol{x}_{<k},t)\mathrm{d}t\right)\right)\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\quad - \int \log\circ g(\partial_k f) + (\log\circ g)'(\partial_k f)\partial_k\varepsilon \mathrm{d}\pi + \mathcal{O}(\|\varepsilon\|_{V_k}^2) \\
&= \mathcal{L}_k(f) + \ell(\varepsilon) + \mathcal{O}(\|\varepsilon\|_{V_k}^2)
\end{aligned}$$

where $\ell\colon V_k \to \mathbb{R}$ is the linear form defined by

$$\ell(\varepsilon) = \int \mathcal{R}_k(f)(\boldsymbol{x})\left(\varepsilon(\boldsymbol{x}_{<k},0) + \int_0^{x_k} g'\big(\partial_k f(\boldsymbol{x}_{<k},t)\big)\partial_k\varepsilon(\boldsymbol{x}_{<k},t)\mathrm{d}t\right) - (\log\circ g)'(\partial_k f(\boldsymbol{x}))\partial_k\varepsilon(\boldsymbol{x})\,\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

If ℓ is continuous, meaning if there exists a constant C_ℓ such that $|\ell(\varepsilon)| \le C_\ell\|\varepsilon\|_{V_k}$ for any $\varepsilon \in V_k$, then the Riesz representation theorem states that there exists a vector $\nabla\mathcal{L}_k(f) \in V_k$ such that $\ell(\varepsilon) = \langle \nabla\mathcal{L}_k(f), \varepsilon\rangle_{V_k}$. This proves $\mathcal{L}_k$ is differentiable everywhere.

To show that ℓ is continuous, we write

$$\begin{aligned}
|\ell(\varepsilon)| &\overset{(19)}{\le} C_\pi \int \left|\mathcal{R}_k(f)(\boldsymbol{x})\left(\varepsilon(\boldsymbol{x}_{<k},0) + \int_0^{x_k} g'\big(\partial_k f(\boldsymbol{x}_{<k},t)\big)\partial_k\varepsilon(\boldsymbol{x}_{<k},t)\mathrm{d}t\right)\right| \eta_{\le k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\quad + C_\pi \int \left|(\log\circ g)'(\partial_k f(\boldsymbol{x}))\partial_k\varepsilon(\boldsymbol{x})\right|\eta_{\le k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\overset{(21)}{\le} C_\pi\|\mathcal{R}_k(f)\|_{L^2_{\eta_{\le k}}}\sqrt{\int\left|\varepsilon(\boldsymbol{x}_{<k},0) + \int_0^{x_k} g'\big(\partial_k f(\boldsymbol{x}_{<k},t)\big)\partial_k\varepsilon(\boldsymbol{x}_{<k},t)\mathrm{d}t\right|^2\eta_{\le k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}} + C_\pi L\|\partial_k\varepsilon\|_{L^2_{\eta_{\le k}}} \\
&\overset{(20)}{\le} C_\pi\|\mathcal{R}_k(f)\|_{L^2_{\eta_{\le k}}}\sqrt{2C_T\|\varepsilon\|_{V_k}^2 + 2C_H L^2\int\left|\partial_k\varepsilon(\boldsymbol{x}))\right|^2\eta_{\le k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}} + C_\pi L\|\partial_k\varepsilon\|_{L^2_{\eta_{\le k}}} \\
&\le C_\pi\Big(\|\mathcal{R}_k(f)\|_{L^2_{\eta_{\le k}}}\sqrt{2C_T + 2C_H L^2} + L\Big)\|\varepsilon\|_{V_k},
\end{aligned}$$

where the second last inequality also uses Proposition 2 and Lemma 12. This concludes the proof. □

A.6. Proof of the local Lipschitz regularity (23)

Proposition 14. *In addition to the assumptions of Theorem 4, we further assume there exists a constant $L < \infty$ such that for all $\xi, \xi' \in \mathbb{R}$ we have*

$$|g'(\xi) - g'(\xi')| \le L|\xi - \xi'| \tag{46}$$
$$|(\log\circ g)'(\xi) - (\log\circ g)'(\xi')| \le L|\xi - \xi'|. \tag{47}$$

Then there exists $M < \infty$ such that

$$\|\nabla\mathcal{L}_k(f_1) - \nabla\mathcal{L}_k(f_2)\|_{V_k} \le M(1 + \|\mathcal{R}_k(f_2)\|_{V_k})\|f_1 - f_2\|_{\overline{V}_k},$$

for any $f_1, f_2 \in \overline{V}_k$, where $\overline{V}_k = \{f \in V_k, \partial_k f \in L^\infty\}$ is the space endowed with the norm $\|f\|_{\overline{V}_k} = \|f\|_{V_k} + \|\partial_k f\|_{L^\infty}$.

Proof. Recall the definition (22) of $\nabla\mathcal{L}_k(f)$

$$\begin{aligned}\langle\nabla\mathcal{L}_k(f),\varepsilon\rangle_{V_k} = &\int \mathcal{R}_k(f)(\boldsymbol{x})\left(\varepsilon(\boldsymbol{x}_{<k},0)+\int_0^{x_k} g'\big(\partial_k f(\boldsymbol{x}_{<k},t)\big)\partial_k\varepsilon(\boldsymbol{x}_{<k},t)\mathrm{d}t\right)\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &-\int(\log\circ g)'(\partial_k f(\boldsymbol{x}))\partial_k\varepsilon(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned}$$

Then for any $f_1, f_2 \in \overline{V}_k$. we can write

$$\langle\nabla\mathcal{L}_k(f_1)-\nabla\mathcal{L}_k(f_2),\varepsilon\rangle_{V_k} = A+B+C+D,$$

where

$$\begin{aligned}A &= \int\Big(\mathcal{R}_k(f_1)(\boldsymbol{x})-\mathcal{R}_k(f_2)(\boldsymbol{x})\Big)\varepsilon(\boldsymbol{x}_{<k},0)\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ B &= \int\Big(\mathcal{R}_k(f_1)(\boldsymbol{x})-\mathcal{R}_k(f_2)(\boldsymbol{x})\Big)\left(\int_0^{x_k} g'(\partial_k f_1(\boldsymbol{x}_{<k},t))\partial_k\varepsilon(\boldsymbol{x}_{<k},t)\mathrm{d}t\right)\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ C &= \int\mathcal{R}_k(f_2)(\boldsymbol{x})\left(\int_0^{x_k}\Big(g'(\partial_k f_1(\boldsymbol{x}_{<k},t))-g'(\partial_k f_2(\boldsymbol{x}_{<k},t))\Big)\partial_k\varepsilon(\boldsymbol{x}_{<k},t)\mathrm{d}t\right)\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ D &= \int\Big((\log\circ g)'(\partial_k f_1(\boldsymbol{x}))-(\log\circ g)'(\partial_k f_2(\boldsymbol{x}))\Big)\partial_k\varepsilon(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned}$$

For the first term A we write

$$\begin{aligned}|A| &\overset{(19)}{\leq} C_\pi\int\Big|\mathcal{R}_k(f_1)(\boldsymbol{x})-\mathcal{R}_k(f_2)(\boldsymbol{x})\Big||\varepsilon(\boldsymbol{x}_{<k},0)|\eta(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &\leq C_\pi\|\mathcal{R}_k(f_1)-\mathcal{R}_k(f_2)\|_{V_k}\left(\int|\varepsilon(\boldsymbol{x}_{<k},0)|^2\eta_{<k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)^{1/2}\\ &\overset{(15)}{\leq} C_\pi\sqrt{C_T}\|\mathcal{R}_k(f_1)-\mathcal{R}_k(f_2)\|_{V_k}\|\varepsilon\|_{V_k}\\ &\overset{(18)}{\leq} C_\pi\sqrt{C_T}C\|f_1-f_2\|_{V_k}\|\varepsilon\|_{V_k}.\end{aligned}$$

For the second term B we write

$$\begin{aligned}|B| &\overset{(19)}{\leq} C_\pi\|\mathcal{R}_k(f_1)-\mathcal{R}_k(f_2)\|_{V_k}\left(\int\left(\int_0^{x_k} g'(\partial_k f_1(\boldsymbol{x}_{<k},t))\partial_k\varepsilon(\boldsymbol{x}_{<k},t)\mathrm{d}t\right)^2\eta(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)^{1/2}\\ &\overset{(39)}{\overset{(18)}{\leq}} C_\pi\sqrt{C_H}C\|f_1-f_2\|_{V_k}\Big(\int\Big(g'(\partial_k f_1(\boldsymbol{x}_{\leq k}))\partial_k\varepsilon(\boldsymbol{x}_{\leq k})\Big)^2\eta(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\Big)^{1/2}\\ &\overset{(20)}{\leq} C_\pi\sqrt{C_H}CL\|f_1-f_2\|_{V_k}\Big(\int\Big(\partial_k\varepsilon(\boldsymbol{x}_{\leq k})\Big)^2\eta(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\Big)^{1/2}\\ &\leq C_\pi\sqrt{C_H}CL\|f_1-f_2\|_{V_k}\|\varepsilon\|_{V_k}.\end{aligned}$$

For the third term C we write

$$
\begin{aligned}
|C| &\overset{(19)}{\leq} C_\pi \|\mathcal{R}_k(f_2)\|_{V_k} \left(\int \left(\int_0^{x_k} \Big(g'(\partial_k f_1(\boldsymbol{x}_{<k}, t)) - g'(\partial_k f_2(\boldsymbol{x}_{<k}, t)) \Big) \partial_k \varepsilon(\boldsymbol{x}_{<k}, t) \mathrm{d}t \right)^2 \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right)^{1/2} \\
&\overset{(39)}{\leq} C_\pi \sqrt{C_H} \|\mathcal{R}_k(f_2)\|_{V_k} \left(\int \left(\Big(g'(\partial_k f_1(\boldsymbol{x}_{\leq k})) - g'(\partial_k f_2(\boldsymbol{x}_{\leq k})) \Big) \partial_k \varepsilon(\boldsymbol{x}_{\leq k}) \right)^2 \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right)^{1/2} \\
&\leq C_\pi \sqrt{C_H} \|\mathcal{R}_k(f_2)\|_{V_k} \left(\operatorname{ess\,sup} \left| g' \circ \partial_k f_1 - g' \circ \partial_k f_2 \right| \right) \left(\int \left(\partial_k \varepsilon(\boldsymbol{x}_{\leq k}) \right)^2 \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right)^{1/2} \\
&\overset{(46)}{\leq} C_\pi \sqrt{C_H} L \|\mathcal{R}_k(f_2)\|_{V_k} \left(\operatorname{ess\,sup} \left| \partial_k f_1 - \partial_k f_2 \right| \right) \|\varepsilon\|_{V_k} \\
&\leq C_\pi \sqrt{C_H} L \|\mathcal{R}_k(f_2)\|_{V_k} \|f_1 - f_2\|_{\overline{V}_k} \|\varepsilon\|_{V_k}.
\end{aligned}
$$

For the last term D we write

$$
\begin{aligned}
|D| &\overset{(19)}{\leq} C_\pi \left(\int \Big((\log \circ g)'(\partial_k f_1(\boldsymbol{x})) - (\log \circ g)'(\partial_k f_2(\boldsymbol{x})) \Big)^2 \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right)^{1/2} \|\varepsilon\|_{V_k} \\
&\overset{(47)}{\leq} C_\pi L \left(\int \Big(\partial_k f_1(\boldsymbol{x}) - \partial_k f_2(\boldsymbol{x}) \Big)^2 \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right)^{1/2} \|\varepsilon\|_{V_k} \\
&\leq C_\pi L \|f_1 - f_2\|_{V_k} \|\varepsilon\|_{V_k}.
\end{aligned}
$$

Thus, because $\|f_1 - f_2\|_{V_k} \leq \|f_1 - f_2\|_{\overline{V}_k}$ we obtain

$$
\begin{aligned}
\frac{|\langle \nabla \mathcal{L}_k(f_1) - \nabla \mathcal{L}_k(f_2), \varepsilon \rangle_{V_k}|}{\|\varepsilon\|_{V_k}} &\leq C_\pi \Big(\sqrt{C_T} C + \sqrt{C_H} C L + \sqrt{C_H} L \|\mathcal{R}_k(f_2)\|_{V_k} + L \Big) \|f_1 - f_2\|_{\overline{V}_k} \\
&\leq M (1 + \|\mathcal{R}_k(f_2)\|_{V_k}) \|f_1 - f_2\|_{\overline{V}_k},
\end{aligned}
$$

where

$$
M = C_\pi \max\{ \sqrt{C_T} C + \sqrt{C_H} C L + L; \sqrt{C_H} L \}.
$$

This concludes the proof. □

A.7. Proof of Proposition 6

Proof. To show that $\mathcal{R}_k(V_k) = \{\mathcal{R}(f) : f \in V_k\}$ is convex, let $f_1, f_2 \in V_k$ and $0 \leq \alpha \leq 1$. We need to show that there exists $f_\alpha \in V_k$ such that $\mathcal{R}(f_\alpha) = S_\alpha$ where

$$
S_\alpha \coloneqq \alpha \mathcal{R}_k(f_1) + (1 - \alpha) \mathcal{R}_k(f_2).
$$

Let

$$
f_\alpha(\boldsymbol{x}_{\leq k}) \coloneqq \mathcal{R}^{-1}(S_\alpha)(\boldsymbol{x}_{\leq k}) = S_\alpha(\boldsymbol{x}_{<k}, 0) + \int_0^{x_k} g^{-1}(\partial_k S_\alpha(\boldsymbol{x}_{<k}, t)) \mathrm{d}t.
$$

It remains to show that $f_\alpha \in V_k$, meaning that $f_\alpha \in L^2_{\eta_{\leq k}}$ and $\partial_k f_\alpha \in L^2_{\eta_{\leq k}}$. By convexity of $\xi \mapsto g^{-1}(\xi)^2$, we have

$$\begin{aligned}
\|\partial_k f_\alpha\|^2_{L^2_{\eta_{\leq k}}} &= \int g^{-1}\big(\alpha \partial_k \mathcal{R}_k(f_1) + (1-\alpha)\partial_k \mathcal{R}_k(f_2)\big)^2 \mathrm{d}\eta_{\leq k} \\
&= \int g^{-1}\big(\alpha g(\partial_k f_1) + (1-\alpha) g(\partial_k f_2)\big)^2 \mathrm{d}\eta_{\leq k} \\
&\leq \int \alpha g^{-1}(g(\partial_k f_1))^2 + (1-\alpha) g^{-1}(g(\partial_k f_2))^2 \mathrm{d}\eta_{\leq k} \\
&= \alpha \|\partial_k f_1\|^2_{L^2_{\eta_{\leq k}}} + (1-\alpha)\|\partial_k f_2\|^2_{L^2_{\eta_{\leq k}}}. \qquad (48)
\end{aligned}$$

Thus $\partial_k f_\alpha \in L^2_{\eta_{\leq k}}$. Furthermore we have

$$\begin{aligned}
\|f_\alpha\|^2_{L^2_{\eta_{\leq k}}} &= \int \left(S_\alpha(\boldsymbol{x}_{<k}, 0) + \int_0^{x_k} g^{-1}(\partial_k S_\alpha(\boldsymbol{x}_{<k}, t))\mathrm{d}t \right)^2 \eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\leq 2\int S_\alpha(\boldsymbol{x}_{<k}, 0)^2 \eta_{<k}(\boldsymbol{x}_{<k})\mathrm{d}\boldsymbol{x} + 2\int \left(\int_0^{x_k} g^{-1}(\partial_k S_\alpha(\boldsymbol{x}_{<k}, t))\mathrm{d}t \right)^2 \eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.
\end{aligned}$$

To show that the above quantity is finite, Proposition 2 permits us to write

$$\begin{aligned}
\int S_\alpha(\boldsymbol{x}_{<k}, 0)^2 \eta_{<k}(\boldsymbol{x}_{<k})\mathrm{d}\boldsymbol{x} &= \int \Big(\alpha f_1(\boldsymbol{x}_{<k}, 0) + (1-\alpha) f_2(\boldsymbol{x}_{<k}, 0)\Big)^2 \eta_{<k}(\boldsymbol{x}_{<k})\mathrm{d}\boldsymbol{x} \\
&\leq C_T \|\alpha f_1 + (1-\alpha) f_2\|^2_{V_k},
\end{aligned}$$

which is finite. Finally, because $g^{-1}(\partial_k S_\alpha) = \partial_k f_\alpha \in L^2_{\eta_{\leq k}}$ by (48), Lemma 12 yields

$$\begin{aligned}
\int \left(\int_0^{x_k} g^{-1}(\partial_k S_\alpha(\boldsymbol{x}_{<k}, t))\mathrm{d}t \right)^2 \eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &\leq C_H \int g^{-1}(\partial_k S_\alpha(\boldsymbol{x}_{\leq k}))^2 \eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= C_H \|\partial_k f_\alpha\|^2_{L^2_{\eta_{\leq k}}},
\end{aligned}$$

which is finite. We deduce that $f_\alpha \in L^2_{\eta_{\leq k}}$ and therefore that $f_\alpha \in V_k$. □

A.8. Proof of Proposition 5

Proof. Let $s_1, s_2 \in V_k$ be strictly increasing functions with respect to x_k that satisfy $\partial_k s_i(\boldsymbol{x}_{\leq k}) \geq c$ for $i = 1, 2$ and all $\boldsymbol{x}_{\leq k} \in \mathbb{R}^k$. By the Lipschitz property of g^{-1} on the domain $[c, \infty)$ with constant L_c, we can write

$$\begin{aligned}
\|\partial_k \mathcal{R}_k^{-1}(s_1) - \partial_k \mathcal{R}_k^{-1}(s_2)\|^2_{L^2_{\eta_{\leq k}}} &= \int (g^{-1}(\partial_k s_1(\boldsymbol{x}_{\leq k})) - g^{-1}(\partial_k s_2(\boldsymbol{x}_{\leq k})))^2 \eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\leq L_c^2 \int (\partial_k s_1(\boldsymbol{x}_{\leq k}) - \partial_k s_2(\boldsymbol{x}_{\leq k}))^2 \eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\leq L_c^2 \|s_1 - s_2\|^2_{V_k}. \qquad (49)
\end{aligned}$$

Applying Proposition 2 to $s_1, s_2 \in V_k$ and Lemma 12 to $\partial_k \mathcal{R}_k^{-1}(s_i) = g^{-1}(\partial_k s_i) \in L^2_{\eta_{\leq k}}$ for $i = 1, 2$

we have

$$\begin{aligned}\|\mathcal{R}_k^{-1}(s_1) - \mathcal{R}_k^{-1}(s_2)\|^2_{L^2_{\eta_{\leq k}}} &\leq 2\int \left(s_1(\boldsymbol{x}_{<k},0) - s_2(\boldsymbol{x}_{<k},0)\right)^2 \eta_{\leq k}(\boldsymbol{x})d\boldsymbol{x} \\ &\quad + 2\int \left(\int_0^{x_k} g^{-1}(\partial_k s_1) - g^{-1}(\partial_k s_2)\mathrm{d}t\right)^2 \eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &\leq 2C_T\|s_1 - s_2\|^2_{V_k} \\ &\quad + 2C_H\int (g^{-1}(\partial_k s_1(\boldsymbol{x}_{\leq k})) - g^{-1}(\partial_k s_2(\boldsymbol{x}_{\leq k})))^2 \eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &\leq (2C_T + 2C_H L_c^2)\|s_1 - s_2\|^2_{V_k}, \end{aligned} \tag{50}$$

where the last inequality follows from (49).

It remains to show that $\|\mathcal{R}_k^{-1}(s)\|_{V_k} < \infty$ for any $s \in V_k$ such that $\operatorname{ess\,inf} \partial_k s > 0$. Letting $s_1 = s$ and $s_2 = g(0)x_k$, the triangle inequality combined with (25) yields

$$\|\mathcal{R}_k^{-1}(s)\|_{V_k} \leq \|\mathcal{R}_k^{-1}(g(0)x_k)\|_{V_k} + C_c\|s - g(0)x_k\|_{V_k}.$$

The function $\mathcal{R}_k^{-1}(g(0)x_k)$ is zero. Therefore, $\|\mathcal{R}_k^{-1}(s)\|_{V_k} \leq C_c\|s - g(0)x_k\|_{V_k} \leq C_c(\|s\|_{V_k} + \|g(0)x_k\|_{V_k})$. For a linear function, $\|g(0)x_k\|^2_{V_k} = \|g(0)x_k\|^2_{L^2_{\eta_{\leq k}}} + \|g(0)\|^2_{L^2_{\eta_{\leq k}}} = 2g(0)^2$ is finite, and so $\|\mathcal{R}_k^{-1}(s)\|_{V_k} < \infty$ for $s \in V_k$. Furthermore, if $\partial_k s \geq c > 0$, then $\partial_k \mathcal{R}_k^{-1}(s) = g^{-1}(\partial_k s) \geq g^{-1}(c) > -\infty$ and so $\operatorname{ess\,inf} \mathcal{R}_k^{-1}(s) > -\infty$. □

A.9. Proof for the KR rearrangement

Proof. Let $S_{\mathrm{KR},k}$ be the kth component of the KR rearrangement, given by composing the inverse CDF of the standard Gaussian marginal $F_{\eta,k}(x_k)$ with the CDF of the target's kth marginal conditional $F_{\pi_k}(x_k|\boldsymbol{x}_{<k})$. That is,

$$S_{\mathrm{KR},k}(\boldsymbol{x}_{\leq k}) = F_{\eta_k}^{-1} \circ F_{\pi_k}(x_k|\boldsymbol{x}_{<k}). \tag{51}$$

The goal is to show $S_{\mathrm{KR},k} \in V_k$, that is, $S_{\mathrm{KR},k} \in L^2_{\eta_{\leq k}}$ and $\partial_k S_{\mathrm{KR},k} \in L^2_{\eta_{\leq k}}$.

First we show $S_{\mathrm{KR},k} \in L^2_{\eta_{\leq k}}$. From condition (29), we have $F_{\eta_k}^{-1}(c_k F_{\eta_k}(x_k)) \leq S_{\mathrm{KR},k}(x_k|\boldsymbol{x}_{<k}) \leq F_{\eta_k}^{-1}(C_k F_{\eta_k}(x_k))$ so that

$$S_{\mathrm{KR},k}(x_k|\boldsymbol{x}_{<k})^2 \leq \max\{F_{\eta_k}^{-1}(c_k F_{\eta_k}(x_k))^2; F_{\eta_k}^{-1}(C_k F_{\eta_k}(x_k))^2\}, \tag{52}$$

for all $\boldsymbol{x}_{<k} \in \mathbb{R}^{k-1}$. To show that $S_{\mathrm{KR},k} \in L^2_{\eta_{\leq k}}$, it is sufficient to prove that any function of the form $x_k \mapsto F_{\eta_k}^{-1}(CF_{\eta_k}(x_k))$ is in $L^2_{\eta_{\leq k}}$ for any $C > 0$. From Theorems 1 and 2 in [10], there exists strictly positive constants $\alpha_i, \beta_i > 0$ for $i = 1, 2$ such that

$$\alpha_1 \exp(-\beta_1 x_k^2) \leq F_{\eta_k}(x_k) \leq \alpha_2 \exp(-\beta_2 x_k^2), \tag{53}$$

for $x_k < 0$. With a change of variable $u = F_{\eta_k}(x_k)$ we obtain $F_{\eta_k}^{-1}(u)^2 \leq 1/\beta_2 \log(\alpha_2/u)$ for all $u < F_{\eta_k}(0) = 1/2$. Letting $u = CF_{\eta_k}(x_k)$ yields

$$\begin{aligned} F_{\eta_k}^{-1}(CF_{\eta_k}(x_k))^2 &\leq \frac{1}{\beta_2}\log\left(\frac{\alpha_2}{CF_{\eta_k}(x_k)}\right), \\ &\overset{(53)}{\leq} \frac{1}{\beta_2}\log\left(\frac{\alpha_2}{C\alpha_1\exp(-\beta_1 x_k^2)}\right) \\ &= \frac{1}{\beta_2}\log\left(\frac{\alpha_2}{C\alpha_1}\right) + \frac{\beta_1}{\beta_2}x_k^2 \end{aligned}$$

for all $x_k < \min\{F_{\eta_k}^{-1}(1/(2C)), 0\}$. Using the same argument, we obtain a similar bound on $F_{\eta_k}^{-1}(CF_{\eta_k}(x_k))^2$ for all x_k larger than a certain value. Together with the continuity of $x_k \mapsto F_{\eta_k}^{-1}(CF_{\eta_k}(x_k))^2$ these bounds ensure that $x_k \mapsto F_{\eta_k}^{-1}(CF_{\eta_k}(x_k))^2$ is in $L^2_{\eta_{\leq k}}$ for any C. Then $S_{\mathrm{KR},k} \in L^2_{\eta_{\leq k}}$. Furthermore, we have $S_{\mathrm{KR},k}(\boldsymbol{x}_{\leq k}) = \mathcal{O}(x_k)$ as $|x_k| \to \infty$.

Now we show that $\partial_k S_{\mathrm{KR},k} \in L^2_{\eta_{\leq k}}$ by showing $\partial_k S_{\mathrm{KR},k}$ is a continuous and bounded function. From the absolute continuity of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, we have that

$$\partial_k S_{\mathrm{KR},k}(\boldsymbol{x}_{\leq k}) = \frac{\pi_k(x_k|\boldsymbol{x}_{<k})}{\eta_k(S_{\mathrm{KR},k}(\boldsymbol{x}_{\leq k}))} = \frac{\pi_k(F_{\pi_k}^{-1}(F_{\pi_k}(x_k|\boldsymbol{x}_{<k})|\boldsymbol{x}_{<k})|\boldsymbol{x}_{<k})}{\eta_k(F_{\eta_k}^{-1} \circ F_{\pi_k}(x_k|\boldsymbol{x}_{<k}))}, \tag{54}$$

is continuous, where $F_{\pi_k}^{-1}(\cdot|\boldsymbol{x}_{<k})$ denotes the inverse of the map $x_k \mapsto F_{\pi_k}(x_k|\boldsymbol{x}_{<k})$ for each $\boldsymbol{x}_{<k} \in \mathbb{R}^{k-1}$. Hence, it is sufficient to show that $\partial_k S_{\mathrm{KR},k}$ goes to a finite limit as $|x_k| \to \infty$. For the left-hand limit, we can write

$$\begin{aligned}
\lim_{x_k \to -\infty} \partial_k S_{\mathrm{KR},k}(\boldsymbol{x}_{\leq k}) &= \lim_{u \to 0^+} \frac{\pi_k(F_{\pi_k}^{-1}(u|\boldsymbol{x}_{<k})|\boldsymbol{x}_{<k})}{\eta_k(F_{\eta_k}^{-1}(u))} \\
&= \lim_{u \to 0^+} \frac{(F_{\eta,k}^{-1})'(u)}{(F_{\pi_k}^{-1})'(u|\boldsymbol{x}_{<k})} \\
&= \lim_{u \to 0^+} \frac{F_{\eta_k}^{-1}(u)}{F_{\pi_k}^{-1}(u|\boldsymbol{x}_{<k})},
\end{aligned} \tag{55}$$

where in the second equality we used the implicit function theorem and the third equality follows from l'Hôpital's rule. To analyze the ratio $F_{\eta_k}^{-1}(u)/F_{\pi_k}^{-1}(u|\boldsymbol{x}_{<k})$, we combine the lower bound in (29) and the bounds in (53) to get

$$\frac{F_{\eta_k}^{-1}(u)}{F_{\pi_k}^{-1}(u|\boldsymbol{x}_{<k})} \leq \frac{F_{\eta_k}^{-1}(u)}{F_{\eta_k}^{-1}(u/c_k)} \leq \sqrt{\frac{\beta_2(\log\alpha_1 - \log(u))}{\beta_1(\log\alpha_2 - \log(u/c_k))}}.$$

Similarly, from the upper bound in (29) and the bounds in (53), we have

$$\frac{F_{\eta_k}^{-1}(u)}{F_{\pi_k}^{-1}(u|\boldsymbol{x}_{<k})} \geq \frac{F_{\eta_k}^{-1}(u)}{F_{\eta_k}^{-1}(u/C_k)} \geq \sqrt{\frac{\beta_1(\log\alpha_2 - \log(u))}{\beta_2(\log\alpha_1 - \log(u/C_k))}}.$$

Thus, $\partial_k S_{\mathrm{KR},k}(\boldsymbol{x}_{\leq k}) = \mathcal{O}(1)$ as $x_k \to -\infty$, and we have $\partial_k S_{\mathrm{KR},k} \in L^2_{\eta_k}$.

Lastly, taking the limit in (55) we have $\lim_{x_k \to -\infty} \partial_k S_{\mathrm{KR},k}(\boldsymbol{x}_{\leq k}) \geq \sqrt{\beta_1/\beta_2}$. For a target distribution π with full support, all marginal conditional densities satisfy $\pi_k(x_k|\boldsymbol{x}_{<k}) > 0$ for each $\boldsymbol{x}_{\leq k} \in \mathbb{R}^k$. Given that the $\partial_k S_{\mathrm{KR},k}$ does not approach zero as $|x_k| \to \infty$, we can find a strictly positive constant $c_k > 0$ such that $\partial_k S_{\mathrm{KR},k}(\boldsymbol{x}_{\leq k}) \geq c_k$ for all $\boldsymbol{x}_{\leq k} \in \mathbb{R}^k$. This shows that $\operatorname{ess\,inf} \partial_k S_{\mathrm{KR},k} > 0$. □

Appendix B: Multi-index refinement for the wavelet basis

In this section we show how to greedily enrich the index set Λ_t for a one-dimensional wavelet basis parameterized by the tuple of indices (l, q) representing the level l and translation q of each wavelet $\psi_{(l,q)}$. To define the allowable indices, we construct a binary tree where each node is indexed by (l, q) and has two children with indices $(l+1, 2q)$ and $(l+1, 2q+1)$. The root of the tree has index $(0, 0)$ and corresponds to the mother wavelet ψ. Analogously to the downward closed property for

polynomial indices, we only add nodes to the tree (i.e., indices in Λ_t) if its parents have already been added. Given any set Λ_t, we define its reduced margin as

$$\Lambda_t^{\mathrm{RM}} = \left\{ \alpha = (l,q) \notin \Lambda_t \text{ such that } \begin{array}{ll} (l-1, q/2) \in \Lambda_t & \text{for odd } q \\ (l-1,(q-1)/2) \in \Lambda_t & \text{for even } q \end{array} \right\}.$$

Then, the `ATM` algorithm with a wavelet basis follows from Algorithm 1 with this construction for the reduced margin at each iteration.

Appendix C: Architecture details of alternative methods

In this section we present the details of the alternative methods to `ATM` that we consider in Section 5.

For each normalizing flow model, we use the recommended stochastic gradient descent optimizer with a learning rate of 10^{-3}. We partition 10% of the samples in each training set to be validation samples and use the remaining samples for training the model. We select the optimal hyper-parameters for each dataset by fitting the density with the training data and choosing the parameters that minimize the negative log-likelihood of the approximate density on the validation samples. We also use the validation samples to set the termination criteria during the optimization.

We follow the implementation of [50] to define the architectures of these models. The hyper-parameters we consider for the neural networks in the `MDN` and `NF` models are: 2 hidden layers, 32 hidden units in each layer, $\{5, 10, 20, 50, 100\}$ centers or flows, weight normalization, and a dropout probability of $\{0, 0.2\}$ for regularizing the neural networks during training. For `CKDE` and `NKDE` we select the bandwidth of the kernel estimators using 5-fold cross-validation.

References

[1] Ambrogioni, L., Güçlü, U., van Gerven, M. A. and Maris, E. (2017). The kernel mixture network: A nonparametric method for conditional density estimation of continuous random variables. *arXiv preprint arXiv:1705.07111*.

[2] Anderes, E. and Coram, M. (2012). A general spline representation for nonparametric and semiparametric density estimates using diffeomorphisms. *arXiv preprint arXiv:1205.5314*.

[3] Baptista, R., Marzouk, Y., Morrison, R. E. and Zahm, O. (2021). Learning non-Gaussian graphical models via Hessian scores and triangular transport. *arXiv preprint arXiv:2101.03093*.

[4] Bertsekas, D. P. (1997). Nonlinear programming. *Journal of the Operational Research Society* **48** 334–334.

[5] Bigoni, D., Marzouk, Y., Prieur, C. and Zahm, O. (2022). Nonlinear dimension reduction for surrogate modeling using gradient information. *Information and Inference: A Journal of the IMA*.

[6] Bishop, C. M. (1994). Mixture density networks Technical Report No. Neural Computing Research Group report: NCRG/94/004, Aston University.

[7] Bogachev, V. I., Kolesnikov, A. V. and Medvedev, K. V. (2005). Triangular transformations of measures. *Sbornik: Mathematics* **196** 309.

[8] Boyd, J. P. (1984). Asymptotic coefficients of Hermite function series. *Journal of Computational Physics* **54** 382–410.

[9] Brennan, M., Bigoni, D., Zahm, O., Spantini, A. and Marzouk, Y. (2020). Greedy inference with structure-exploiting lazy maps. *Advances in Neural Information Processing Systems* **33**.

[10] Chang, S.-H., Cosman, P. C. and Milstein, L. B. (2011). Chernoff-type bounds for the Gaussian error function. *IEEE Transactions on Communications* **59** 2939–2944.

[11] Chkifa, A., Cohen, A. and Schwab, C. (2015). Breaking the curse of dimensionality in sparse polynomial approximation of parametric PDEs. *Journal de Mathématiques Pures et Appliquées* **103** 400–428.

[12] Cohen, A. (2003). *Numerical analysis of wavelet methods*. Elsevier.

[13] Cohen, A. and Migliorati, G. (2018). Multivariate approximation in downward closed polynomial spaces. In *Contemporary Computational Mathematics-A celebration of the 80th birthday of Ian Sloan* 233–282. Springer.
[14] Cui, T. and Dolgov, S. (2021). Deep composition of Tensor-Trains using squared inverse Rosenblatt transports. *Foundations of Computational Mathematics* 1–60.
[15] Cui, T., Dolgov, S. and Zahm, O. (2021). Conditional Deep Inverse Rosenblatt Transports. *arXiv preprint arXiv:2106.04170*.
[16] Dinh, L., Sohl-Dickstein, J. and Bengio, S. (2017). Density estimation using Real NVP. In *International Conference on Learning Representations*.
[17] Dinh, L., Sohl-Dickstein, J. and Bengio, S. (2017). Density estimation using Real NVP. In *International Conference on Learning Representations*.
[18] Durkan, C., Bekasov, A., Murray, I. and Papamakarios, G. (2019). Neural spline flows. In *Advances in Neural Information Processing Systems* 7509–7520.
[19] El Moselhy, T. A. and Marzouk, Y. M. (2012). Bayesian inference with optimal maps. *Journal of Computational Physics* **231** 7815–7850.
[20] Huang, C.-W., Chen, R. T., Tsirigotis, C. and Courville, A. (2020). Convex Potential Flows: Universal Probability Distributions with Optimal Transport and Convex Optimization. In *International Conference on Learning Representations*.
[21] Huang, C.-W., Krueger, D., Lacoste, A. and Courville, A. (2018). Neural Autoregressive Flows. In *International Conference on Machine Learning* 2083–2092.
[22] Jaini, P., Selby, K. A. and Yu, Y. (2019). Sum-of-squares polynomial flow. In *International Conference on Machine Learning* 3009–3018.
[23] Katzfuss, M. and Schäfer, F. (2021). Scalable Bayesian transport maps for high-dimensional non-Gaussian spatial fields. *arXiv preprint arXiv:2108.04211*.
[24] Kingma, D. P. and Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems* 10215–10224.
[25] Kobyzev, I., Prince, S. and Brubaker, M. (2020). Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
[26] Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.
[27] Kovachki, N., Baptista, R., Hosseini, B. and Marzouk, Y. (2020). Conditional Sampling With Monotone GANs. *arXiv preprint arXiv:2006.06755*.
[28] Kufner, A. and Opic, B. (1984). How to define reasonably weighted Sobolev spaces. *Commentationes Mathematicae Universitatis Carolinae* **25** 537–554.
[29] Lezcano Casado, M. (2019). Trivializations for gradient-based optimization on manifolds. *Advances in Neural Information Processing Systems* **32** 9157–9168.
[30] Lichman, M. (2013). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml.
[31] Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P. and Macke, J. (2021). Benchmarking simulation-based inference. In *International Conference on Artificial Intelligence and Statistics* 343–351. PMLR.
[32] Mallat, S. (1999). *A wavelet tour of signal processing*. Elsevier.
[33] Marzouk, Y., Moselhy, T., Parno, M. and Spantini, A. (2016). *Sampling via Measure Transport: An Introduction* In *Handbook of Uncertainty Quantification* 1–41. Springer International Publishing.
[34] Migliorati, G. (2015). Adaptive polynomial approximation by means of random discrete least squares. In *Numerical Mathematics and Advanced Applications-ENUMATH 2013* 547–554. Springer.
[35] Migliorati, G. (2019). Adaptive approximation by optimal weighted least-squares methods. *SIAM Journal on Numerical Analysis* **57** 2217–2245.
[36] Morrison, R., Baptista, R. and Marzouk, Y. (2017). Beyond normality: Learning sparse probabilistic graphical models in the non-Gaussian setting. In *Advances in Neural Information Processing Systems* 2359–2369.
[37] Muckenhoupt, B. (1972). Hardy's inequality with weights. *Studia Mathematica* **44** 31–38.
[38] Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.
[39] Novak, E., Ullrich, M., Woźniakowski, H. and Zhang, S. (2018). Reproducing kernels of Sobolev spaces on $\mathbb{R}^d$ and applications to embedding constants and tractability. *Analysis and Applications* **16**

693–715.

[40] Oord, A. v. d., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G. v. d., Lockhart, E., Cobo, L. C., Stimberg, F. et al. (2017). Parallel WaveNet: Fast high-fidelity speech synthesis. *arXiv preprint arXiv:1711.10433*.

[41] Papamakarios, G. and Murray, I. (2016). Fast ε-free inference of simulation models with Bayesian conditional density estimation. In *Advances in Neural Information Processing Systems* 1028–1036.

[42] Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S. and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research* **22** 1–64.

[43] Papamakarios, G., Pavlakou, T. and Murray, I. (2017). Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems* 2338–2347.

[44] Parno, M. D. and Marzouk, Y. M. (2018). Transport map accelerated Markov chain Monte Carlo. *SIAM/ASA Journal on Uncertainty Quantification* **6** 645–682.

[45] Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L. and Köthe, U. (2020). BayesFlow: Learning complex stochastic models with invertible neural networks. *IEEE transactions on neural networks and learning systems*.

[46] Ramsay, J. O. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60** 365–375.

[47] Raskutti, G. and Uhler, C. (2018). Learning directed acyclic graph models based on sparsest permutations. *Stat* **7** e183.

[48] Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*.

[49] Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics* **23** 470–472.

[50] Rothfuss, J., Ferreira, F., Walther, S. and Ulrich, M. (2019). Conditional Density Estimation with Neural Networks: Best Practices and Benchmarks. *arXiv:1903.00954*.

[51] Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians*. Springer International Publishing.

[52] Schäfer, F., Katzfuss, M. and Owhadi, H. (2021). Sparse Cholesky Factorization by Kullback–Leibler Minimization. *SIAM Journal on Scientific Computing* **43** A2019–A2046.

[53] Schmuland, B. (1992). Dirichlet forms with polynomial domain. *Math. Japon* **37** 1015–1024.

[54] Schölkopf, B., Herbrich, R. and Smola, A. J. (2001). A generalized representer theorem. In *International conference on computational learning theory* 416–426. Springer.

[55] Shin, Y. E., Zhou, L. and Ding, Y. (2022). Joint estimation of monotone curves via functional principal component analysis. *Computational Statistics & Data Analysis* **166** 107343.

[56] Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics* 795–810.

[57] Sisson, S. A., Fan, Y. and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* **104** 1760–1765.

[58] Spantini, A., Baptista, R. and Marzouk, Y. (In press). Coupling techniques for nonlinear ensemble filtering. *SIAM Review*.

[59] Spantini, A., Bigoni, D. and Marzouk, Y. (2018). Inference via low-dimensional couplings. *The Journal of Machine Learning Research* **19** 2639–2709.

[60] Tabak, E. G. and Turner, C. V. (2013). A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics* **66** 145–164.

[61] Teshima, T., Ishikawa, I., Tojo, K., Oono, K., Ikeda, M. and Sugiyama, M. (2020). Coupling-based invertible neural networks are universal diffeomorphism approximators. In *Advances in Neural Information Processing Systems* **33** 3362–3373.

[62] Trippe, B. L. and Turner, R. E. (2018). Conditional density estimation with Bayesian normalising flows. In *Bayesian Deep Learning: NIPS 2017 Workshop*.

[63] Truong, T. T. and Nguyen, H.-T. (2021). Backtracking Gradient Descent Method and Some Applications in Large Scale Optimisation. Part 2: Algorithms and Experiments. *Applied Mathematics & Optimization* **84** 2557–2586.

[64] Uria, B., Murray, I. and Larochelle, H. (2013). RNADE: The real-valued neural autoregressive

density-estimator. *arXiv preprint arXiv:1306.0186*.
[65] Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science* **47**. Cambridge university press.
[66] Vidakovic, B. (2009). *Statistical modeling by wavelets* **503**. John Wiley & Sons.
[67] Villani, C. (2008). *Optimal transport: old and new* **338**. Springer Science & Business Media.
[68] Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.
[69] Wehenkel, A. and Louppe, G. (2019). Unconstrained monotonic neural networks. In *Advances in Neural Information Processing Systems* 1543–1553.
[70] Wenliang, L., Sutherland, D., Strathmann, H. and Gretton, A. (2019). Learning deep kernels for exponential family densities. In *International Conference on Machine Learning* 6737–6746.
[71] Zech, J. and Marzouk, Y. (2022). Sparse approximation of triangular transports. Part II: the infinite dimensional case. *Constructive Approximation* **55** 987–1036.
[72] Zech, J. and Marzouk, Y. (2022). Sparse Approximation of Triangular Transports. Part I: The Finite-Dimensional Case. *Constructive Approximation* **55** 919–986.