Triangular Flows for Generative Modeling: Statistical Consistency, Smoothness Classes, and Fast Rates

Nicholas Irons
University of Washington

Meyer Scetbon
CREST, ENSAE

Soumik Pal
University of Washington

Zaid Harchaoui
University of Washington

Abstract

Triangular flows, also known as Knöthe-Rosenblatt measure couplings, comprise an important building block of normalizing flow models for generative modeling and density estimation, including popular autoregressive flows such as real-valued non-volume preserving transformation models (Real NVP). We present statistical guarantees and sample complexity bounds for triangular flow statistical models. In particular, we establish the statistical consistency and the finite sample convergence rates of the minimum Kullback-Leibler divergence statistical estimator of the Knöthe-Rosenblatt measure coupling using tools from empirical process theory. Our results highlight the anisotropic geometry of function classes at play in triangular flows, shed light on optimal coordinate ordering, and lead to statistical guarantees for Jacobian flows. We conduct numerical experiments to illustrate the practical implications of our theoretical findings.

1 INTRODUCTION

Triangular flows are popular generative models that allow one to define complex multivariate distributions via push-forwards from simpler multivariate distributions (Kobyzev et al., 2020). Triangular flow models target the *Knöthe-Rosenblatt map* (Spantini et al., 2018), which originally appeared in two independent papers by Knöthe (1957) and Rosenblatt (1952). The Knöthe-Rosenblatt map is a multivariate function S^* from $\mathbb{R}^d$ onto itself pushing a Lebesgue probability density f onto another one g:

$$S^* \# f = g$$

The Knöthe-Rosenblatt map has the striking property of being *triangular* in that its Jacobian is an upper triangular matrix. This map and its properties have been studied in probability theory, nonparametric statistics, and optimal transport, under the name of Knöthe-Rosenblatt (KR) coupling or rearrangement.

KR can be used to synthesize a sampler of a probability distribution given data drawn from that probability distribution. Moreover, any probability distribution can be well approximated via a KR map. This key property has been an important motivation of a number of flow models (Dinh et al., 2015, 2017; El Moselhy and Marzouk, 2012; Kingma and Dhariwal, 2018; Kobyzev et al., 2020; Marzouk et al., 2016; Spantini et al., 2018; Huang et al., 2018; Papamakarios et al., 2017; Wehenkel and Louppe, 2019; Germain et al., 2015) in machine learning, statistical science, computational science, and AI domains. Flow models or normalizing flows have achieved great success in a number of settings, allowing one to generate images that look realistic as well as texts that look as if they were written by humans. A general theory of normalizing flows is yet a huge undertaking as many challenges arise at the same time: the recursive construction of a push forward, the learning objective to estimate the push forward, the neural network functional parameterization, and the statistical modeling of complex data.

We focus in this paper on the Knöthe-Rosenblatt coupling, and more generally triangular flows, as it can be seen as the statistical backbone of normalizing flows. Spantini et al. (2018) showed how to estimate KR from data by minimizing a Kullback-Leibler objective. The estimated KR can then be used to sample at will from the probability distribution at hand. The learning objective of Spantini et al. (2018) has the benefit of being statistically classical, hence amenable to detailed analysis compared to adversarial learning objectives which are still subject to active research. The sample complexity or rate of convergence of this KR estimator is however, to this day, unknown. On the other hand, KR and its multiple relatives are frequently motivated from a universal approximation perspective (Huang et al.,

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

2018; Kobyzev et al., 2020), which underscores their expressiveness yet also only considers the approximation theory viewpoint. Indeed, while universal approximation holds for many proposed models, slow rates of statistical convergence can occur.

Contributions We present a theoretical analysis of the Knöthe-Rosenblatt coupling, from its statistical framing to convergence rates. We put forth a simple example of slow rates showing the limitations of a viewpoint based on universal approximation only. This leads us to identify the function classes that the KR maps belong to and bring to light their anisotropic geometry. We then establish finite sample rates of convergence using tools from empirical process theory. Our analysis delineates different regimes of statistical convergence, depending on the dimension and the sample. Our theoretical results hold under general conditions. Assuming that the source density is log-concave, we establish fast rates of Sobolev-type convergence in the smooth regime. We outline direct implications of our results on Jacobian flows. We provide numerical illustrations on synthetic data to highlight potential implications of our theoretical results. Additional details can be found in the longer version (Irons et al., 2021).

2 TRIANGULAR FLOWS

Knöthe-Rosenblatt Rearrangement KR originated from independent works of M. Rosenblatt and H. Knöthe and has spawned fruitful applications in diverse areas. Rosenblatt (1952) studied the KR map for statistical purposes, specifically multivariate goodness-of-fit testing. Knöthe (1957), on the other hand, elegantly utilized the KR map to extend the Brunn-Minkowski inequality, which can be used to prove the celebrated isoperimetric inequality. More recently, triangular flows have been proposed as simple and expressive building blocks of generative models, for the problems of sampling and density estimation, among others (Kobyzev et al., 2020; Spantini et al., 2018; Marzouk et al., 2016; El Moselhy and Marzouk, 2012). A triangular flow can be used to approximate the KR map between a source density and a target density from their respective samples.

Consider two Lebesgue probability densities f and g on $\mathbb{R}^d$. The Knöthe-Rosenblatt map $S^* : \mathbb{R}^d \to \mathbb{R}^d$ between f and g is the unique monotone non-decreasing upper triangular measurable map pushing f forward to g, written $S^* \# f = g$ (Spantini et al., 2018). The KR map is upper triangular in the sense that its Jacobian is an upper triangular matrix, since S^* takes the form

$$S^*(x) = \begin{bmatrix} S_1^*(x_1, \ldots, x_d) \\ S_2^*(x_2, \ldots, x_d) \\ \vdots \\ S_{d-1}^*(x_{d-1}, x_d) \\ S_d^*(x_d) \end{bmatrix}.$$

Furthermore, S^* is monotone non-decreasing in the sense that the univariate map $x_k \mapsto S_k^*(x_k, \ldots, x_d)$ is monotone non-decreasing for any $(x_{k+1}, \ldots, x_d) \in \mathbb{R}^{d-k}$ for each $k \in \{1, \ldots, d\} =: [d]$. The components of the KR map can be defined recursively via the monotone transport between the univariate conditional densities of f and g. Let $F_k(x_k|x_{(k+1):d})$ denote the cdf of the conditional density $f_k(x_k|x_{(k+1):d})$. Similarly, let $G_k(y_k|y_{(k+1):d})$ denote the conditional cdf of g. In particular, when $k = d$ we obtain $F_d(x_d)$, the cdf of the d-th marginal density $f_d(x_d) = \int f(x_1, \ldots, x_d) dx_1 \cdots dx_{d-1}$, and similarly for g.

Assuming $g(y) > 0$ everywhere, the maps $y_k \mapsto G_k(y_k|y_{(k+1):d})$ are strictly increasing and therefore invertible. The dth component of S^* is defined as the monotone transport between $f_d(x_d)$ and $g_d(y_d)$ that is $S_d^*(x_d) = G_d^{-1}(F_d(x_d))$ (Santambrogio, 2015).

From here the kth component of S^* is given by

$$\begin{aligned} &S_k^*(x_k, \ldots, x_d) \\ &\qquad = G_k^{-1}\left(F_k(x_k|x_{(k+1):d}) \middle| S^*_{(k+1):d}(x_{(k+1):d})\right) \end{aligned}$$

for $k \in [d-1]$, where $x_{(k+1):d} = (x_{k+1}, \ldots, x_d)$ and

$$S^*_{(k+1):d}(x_{(k+1):d}) = (S^*_{k+1}(x_{(k+1):d}), \ldots, S_d^*(x_d)).$$

That S^* is upper triangular and monotonically non-decreasing is clear from the construction. Under tame assumptions on f and g discussed below, S^* is invertible. We denote $T^* = (S^*)^{-1}$, which is the KR map from g to f. In this paper, we shall be interested in the asymptotic and non-asymptotic convergence of statistical estimators towards S^* and T^*, respectively.

From The Uniform Distribution To Any Distribution In his seminal paper, Rosenblatt (1952) considered the special case in which g is the uniform density on the hypercube $[0,1]^d$. This implies that the conditional cdfs are identity maps, $G_k(y_k|y_{k+1}, \ldots, y_d) = y_k$, and so the KR map from f to g consists simply of the conditional cdfs, with $S_d^*(x_d) = P(X_d \le x_d)$, $S_{d-1}^*(x_{d-1}, x_d) = P(X_{d-1} \le x_{d-1}|X_d = x_d)$, and $S_1^*(x_1, \ldots, x_d) = P(X_1 \le x_1|X_2 = x_2, \ldots, X_d = x_d)$. Figure 1 exhibits heat maps of the target density $f(x_1, x_2)$ and the first component of the KR map $F_1(x_1|x_2)$ for the choice $\mu_1 = \mu_2 = 0, \sigma_1 = \sigma_2 = 1$, with

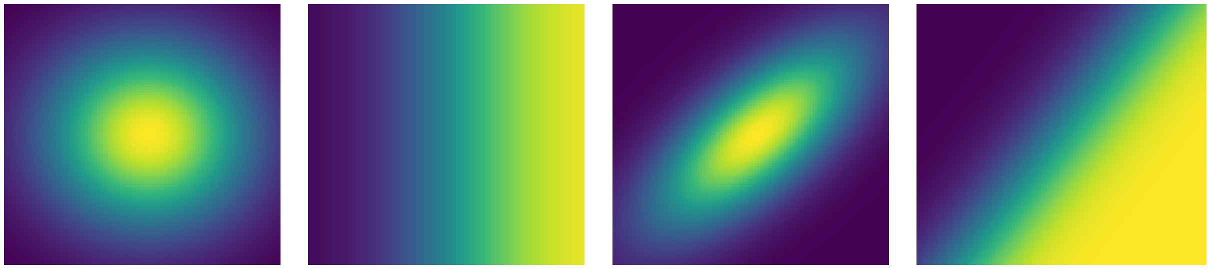

Figure 1: **Left:** Heat map of standard bivariate Gaussian density $f(x_1, x_2)$ with correlation $\rho = 0$ (leftmost) and the first component of the KR map $F_1(x_1|x_2)$ from f to g, the uniform density on $[0,1]^2$. **Right:** Bivariate Gaussian density with $\rho = 0.7$ and first component of the KR map $F_1(x_1|x_2)$ (rightmost).

$\rho = 0$ (left panels) and $\rho = 0.7$ (right panels), where f is a bivariate Gaussian $N\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}\right)$.

Kullback-Leibler Objective Spantini et al. (2018) proposed to estimate the KR map using a minimum distance approach based on the Kullback-Leibler divergence. Since S^* is the unique monotone upper triangular map satisfying $S\#f = g$, it is the unique such map satisfying $\mathrm{KL}(S\#f|g) = 0$, where KL denotes the Kullback-Leibler divergence,

$$\mathrm{KL}(p|q) = \int p(x)\log\frac{p(x)}{q(x)}dx$$

for Lebesgue densities p, q on $\mathbb{R}^d$. As such, S^* uniquely solves the variational problem $\min_{S\in\mathcal{T}} \mathrm{KL}(S\#f|g)$, where $\mathcal{T}$ denotes the convex cone of monotone non-decreasing upper triangular maps $S : \mathbb{R}^d \to \mathbb{R}^d$. The change of variables formula for densities states that

$$(S^{-1}\#g)(x) = g(S(x))|\det(JS(x))|, \tag{1}$$

where $JS(x)$ denotes the Jacobian matrix of S evaluated at x. Applying this formula to $\mathrm{KL}(S\#f|g) = \mathrm{KL}(f|S^{-1}\#g)$, we rewrite $\min_{S\in\mathcal{T}} \mathrm{KL}(S\#f|g)$ as

$$\min_{S\in\mathcal{T}} \mathbb{E}\left\{\log\left[\frac{f(X)}{g(S(X))}\right] - \sum_{k=1}^{d}\log D_k S_k(X)\right\}, \tag{2}$$

where X is a random variable on $\mathbb{R}^d$ with density f and D_k is shorthand for differentiation with respect to the kth component $\frac{\partial}{\partial x_k}$. By monotonicity, $D_k S_k(x)$ is defined Lebesgue almost everywhere for every $S \in \mathcal{T}$. The relation (2) is proved in the Supplement.

2.1 Statistical Estimator Of The KR Map

We shall study an estimator S^n of S^* derived from the sample average approximation to (2), which yields the minimization problem (Spantini et al., 2018)

$$\min_{S\in\mathcal{S}} \frac{1}{n}\sum_{i=1}^{n}\left\{\log\left[\frac{f(X^i)}{g(S(X^i))}\right] - \sum_{k=1}^{d}\log D_k S_k(X^i)\right\}, \tag{3}$$

where $X^1, \ldots, X^n$ is an i.i.d random sample from f and $\mathcal{S}$ is a hypothesis function class. In generative modeling, we have a finite sample from f, perhaps an image dataset, that we use to train a map that can generate more samples from f. In this case, f is the target density and g, the source density, is a degree of freedom in the problem. In practice, g should be chosen so that it is easy to sample from, e.g., a multivariate normal density. The target density f could be also unknown in practice, and if necessary we can omit the terms involving f from the objective function in (3), since they do not depend on the argument S.

With an estimator S^n in hand, which approximately solves the sample average problem (3), we can generate approximate samples from f by pulling back samples from g under S^n, or equivalently by pushing forward samples from g under $T^n = (S^n)^{-1}$. As S^n is defined via KL projection, it can also be viewed as a non-parametric maximum likelihood estimator (MLE). Universal approximation is insufficient to reason about nonparametric estimators, since slow rates can happen, as we shall show next. In practice, S^n can be estimated by parameterizing the space of triangular maps via neural networks or some basis expansion.

Figure 2 illustrates the problem at hand in the case where the target density f is an unbalanced mixture of three bivariate normal distributions centered on the vertices of an equilateral triangle with spherical covariance. Panel (a) of Figure 2 displays level curves of f. The source density g is a standard bivariate normal density. We solved for S^n by parametrizing its components via a Hermite polynomial basis, as described by Marzouk et al. (2016). Since g is log-concave, the optimization problem (3) is convex and can be minimized efficiently using standard convex solvers. By the change of variables formula, $f(x) = g(S^*(x))|\det(JS^*(x)|$, where $|\det(JS^*(x))|$ denotes the Jacobian determinant of the KR map S^* from f to g. Hence we take $f_n(x) = g(S^n(x))|\det(JS^n(x))|$ as an estimate of the target density. Panels (b)-(d) of Figure 2 display the level curves of f_n as the sample size n increases from 1500 to 5000. The improving accuracy of the density estimates f_n as the sample size grows is consistent with our convergence results in Section 3.

Slow Rates Without combining both a tail condition (e.g., common compact support) and a smooth regularity condition (e.g., uniformly bounded derivatives) on the function class $\mathcal{F}$ of the target density f, we show that convergence of any estimator T^n of the direct map T from g to f can occur at an arbitrarily slowly rate.

Theorem 2.1. *Let $\mathcal{F}$ denote the class of infinitely continuously differentiable Lebesgue densities supported on the d-dimensional hypercube $[0,1]^d$ and uniformly bounded by 2, i.e., $\sup_{f\in\mathcal{F}} \|f\|_\infty \leq 2$. Let g be any Lebesgue density on $\mathbb{R}^d$.*

For any $n \in \mathbb{N}$, the minimax risk in terms of KL divergence is bounded below as

$$\inf_{T^n} \sup_{f\in\mathcal{F}} \mathbb{E}_f[KL(f|f_n))] \geq 1/2,$$

where $T^n : \mathbb{R}^d \to [0,1]^d$ is any estimate of the KR map from g to f based on an iid sample of size n from f, and $f_n = T^n\#g$ is the density estimate of f.

Theorem 2.1 underscores the importance of going beyond universal approximation results to study the sample complexity and statistical performance of KR estimation. The proof of this "no free lunch" theorem follows an idea of Birgé (1986); see also Birgé (1983); Bousquet et al. (2004); Devroye (1983, 1995); Devroye et al. (1996); Györfi et al. (2002). We construct a family of densities in $\mathcal{F}$ built from rapidly oscillating perturbations of the uniform distribution on $[0,1]^d$. Such densities are intuitively difficult to estimate. As is evident from the construction, however, a suitable uniform bound on the derivatives of the functions in $\mathcal{F}$ would preclude the existence of such pathological examples. As such, in what proceeds we aim to derive convergence rate bounds under the assumption that the target and source densities f and g, respectively, are compactly supported and sufficiently regular, in the sense that they lie in a Sobolev space of functions with continuous and uniformly bounded partial derivatives up to order s for some $s \geq 1$. For simplicity, our theoretical treatment assumes that f and g are fixed, but our convergence rate bounds as stated also bound the worst-case KL risk over any f and g lying in the L^∞ Sobolev ball $\mathcal{F} = \left\{h : \sum_{|\alpha|\leq s} \|D^\alpha h\|_\infty \leq B\right\}$ for any fixed $B > 0$.

Theorem 2.1 provides a lower bound on the minimax KL risk of KR estimation over the hypothesis class of target densities $\mathcal{F}$. The following stronger result, based on work of Devroye (1983, 1995); Devroye et al. (1996), demonstrates that convergence can still occur arbitrarily slowly for the task of estimating a single target density.

Theorem 2.2. *Let g be any Lebesgue density on $\mathbb{R}^d$ and $\{a_n\}_{n=1}^\infty$ any sequence converging to zero with $1/512 \geq a_1 \geq a_2 \geq \cdots \geq 0$. For every sequence of KR map estimates $T^n : \mathbb{R}^d \to \mathbb{R}^d$ based on a random sample of size n, there exists a target distribution f on $\mathbb{R}^d$ such that*

$$\mathbb{E}_f[KL(f|f_n)] \geq a_n,$$

where $f_n = T^n\#g$ is the density estimate of f.

3 STATISTICAL CONSISTENCY

Setting the stage for the theoretical results, let us introduce the main assumptions.

Assumption 3.1. *The Lebesgue densities f and g have convex compact supports $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$, respectively.*

Assumption 3.2. *The densities f, g are bounded away from 0, i.e., $\inf_{x\in\mathcal{X}, y\in\mathcal{Y}}\{f(x), g(y)\} > 0$*

Assumption 3.3. *Let $s \geq 1$ be a positive integer. The densities f and g are s-smooth on their supports, in the sense that*

$$D^\alpha f(x) := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} f(x)$$

is continuous for every multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{Z}_+^d$ satisfying $|\alpha| := \sum_{k=1}^d \alpha_k \leq s$ and similarly for g.

It is well known that the KR map S^* from f to g is as smooth as the densities f and g, but not smoother (Santambrogio, 2015). As such, under Assumptions 3.1-3.3 we can restrict our attention from $\mathcal{T}$, the set of monotone non-decreasing upper triangular maps, to the smaller function class of monotone upper triangular maps that are s-smooth, of which the KR map S^* is an element. That is, we can limit our search for an estimator S^n solving (3) to a space of functions with more structure. This restriction is crucial to establishing a rate of convergence of the estimator S^n, as we can quantitatively bound the complexity of spaces of smooth maps. We discuss these developments in further detail below. Proofs of all results are included in the Supplement.

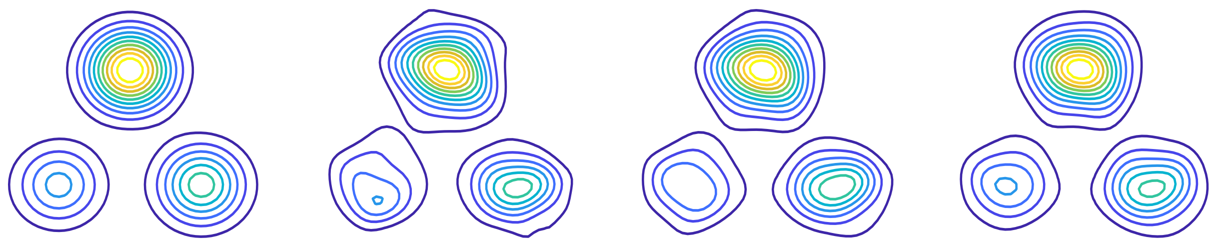

Figure 2: (a) Level curves of the normal mixture density $f(x) = g(S^*(x))|\det(JS^*(x))|$; (b)-(d) Level curves of the density estimates $f_n(x) = g(S^n(x))|\det(JS^n(x))|$ trained on $n = 1500, 3000, 5000$ samples, respectively.

3.1 Upper Bounds On Metric Entropy

We first derive useful estimates of the metric entropy of function classes previously introduced. Assumptions 3.1-3.3 allow us focus on smooth subsets of the class of monotone upper triangular maps $\mathcal{T}$.

Definition 3.1. *Let $M > 0$. For $\mathfrak{s} \in \mathbb{Z}_+^d$ and $k \in [d]$, let $\tilde{\mathfrak{s}}_k = (s_k + 1, s_{k+1}, \ldots, s_d)$. Define $\mathcal{T}(\mathfrak{s}, d, M) \subset \mathcal{T}$ as the convex subset of strictly increasing upper triangular maps $S : \mathcal{X} \to \mathcal{Y}$ satisfying:*

1. $\inf_{k\in[d],x\in\mathcal{X}} D_k S_k(x) \geq 1/M$,
2. $\|D^\alpha S_k\|_\infty \leq M$ *for all* $k \in [d]$ *and* $\alpha_{k:d} \preceq \tilde{\mathfrak{s}}_k$.

For $s \in \mathbb{N}$, we also define the homogeneous smoothness class $\mathcal{T}(s, d, M) = \mathcal{T}((s, \ldots, s), d, M)$.

Condition 1 guarantees that the Jacobian term in (3) is bounded, and condition 2 guarantees smoothness of the objective. In Section 4 below we consider densities with anisotropic smoothness, in which case the number of continuous derivatives s_k varies with the coordinate x_k. For simplicity and clarity of exposition, we first focus on the case when f and g are smooth in a homogeneous sense, as in Assumption 3.3, and we work in the space $\mathcal{T}(s, d, M)$. As remarked above, the KR map S^* from f to g lies in $\mathcal{T}(s, d, M^*)$ under Assumptions 3.1-3.3 when M^* is sufficiently large. The same is true of the direct map T^* from g to f. In fact, all of the results stated here for the sampling map S^* also hold for the direct map T^*, possibly with minor changes (although the proofs are generally more involved). For brevity, we mainly discuss S^* and direct the interested reader to the Supplement.

Henceforth, we consider estimators S^n lying in $\mathcal{T}(s, d, M^*)$ that minimize the objective in (3). We leave the issue of model selection, i.e., determining a sufficiently large M^* such that $\mathcal{T}(s, d, M^*)$ contains the true KR map S^*, for future work. In the Supplement, we calculate explicit quantitative bounds on the complexity of this space as measured by the metric entropy in the d-dimensional L^∞ norm $\|S\|_{\infty,d} := \max_{k\in[d]} \|S_k\|_\infty$. The compactness of $(\overline{\mathcal{T}(s, d, M^*)}, \|\cdot\|_{\infty,d})$ derived as a corollary of this result is required to establish the convergence of a sequence of estimators S^n to S^*, and the entropy bound on the corresponding class of Kullback-Leibler loss functions over $\mathcal{T}(s, d, M^*)$ in Proposition 3.3 below allows us to go further by deriving bounds on the rate of convergence in KL divergence. This result builds off known entropy estimates for function spaces of Besov type (Birgé, 1983, 1986; Nickl and Pötscher, 2007).

Definition 3.2. *For a map $S \in \mathcal{T}$, define the loss function $\psi_S : \mathcal{X} \to \mathbb{R}$ by*

$$\psi_S(x) = \log f(x) - \log g(S(x)) - \sum_{k=1}^d \log D_k S_k(x).$$

We also define the class $\Psi(s, d, M^*)$ of loss functions over $\mathcal{T}(s, d, M^*)$ as

$$\Psi(s, d, M^*) := \{\psi_S : S \in \mathcal{T}(s, d, M^*)\}.$$

By (2) we have $\mathbb{E}[\psi_S(X)] = \mathrm{KL}(S\#f|g)$, where $X \sim f$. Similarly, the sample average of ψ_S is the objective in (3). Hence, to derive finite sample bounds on the expected KL loss, we must study the sample complexity of the class $\Psi(s, d, M^*)$. Define $N(\epsilon, \Psi(s, d, M^*), \|\cdot\|_\infty)$ as the ϵ-covering number of $\Psi(s, d, M^*)$ with respect to the uniform norm $\|\cdot\|_\infty$, and the metric entropy

$$H(\epsilon, \Psi(s, d, M^*), \|\cdot\|_\infty) = \log N(\epsilon, \Psi(s, d, M^*), \|\cdot\|_\infty).$$

Proposition 3.3. *Under Assumptions 3.1-3.3, the metric entropy of $\Psi(s, d, M^*)$ in $L^\infty(\mathcal{X})$ is bounded as*

$$H(\epsilon, \Psi(s, d, M^*), \|\cdot\|_\infty) \lesssim \epsilon^{-d/s}.$$

Consequently, $\Psi(s, d, M^*)$ is totally bounded and therefore precompact in $L^\infty(\mathcal{X})$.

Here, for functions $a(\epsilon), b(\epsilon)$ (or sequences a_n, b_n) we write $a(\epsilon) \lesssim b(\epsilon)$ (resp. $a_n \lesssim b_n$) if $a(\epsilon) \leq Cb(\epsilon)$ (resp. $a_n \leq Cb_n$) for all ϵ (resp. n) for some constant $C > 0$. For brevity, in this result and those that follow, we suppress scalar prefactors that do not depend on the sample size n. As our calculations in the Supplement demonstrate, the constant prefactors in this and subsequent bounds are polynomial in the $\|\cdot\|_\infty$ radius M^* and exponential in the dimension d. This dependence resembles other results on sample complexity of transport map estimators (Hütter and Rigollet, 2021).

3.2 Statistical Consistency

For the sake of concision, we introduce empirical process notation (Dudley, 1967, 1968; van der Vaart and Wellner, 1996; Wainwright, 2019). Let $\mathcal{H}$ be a collection of functions from $\mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$ measurable and square integrable with respect to P, a Borel probability measure on $\mathbb{R}^d$. Let P_n denote the empirical distribution of an iid random sample $X_1, \ldots, X_n$ drawn from P. For a function $h \in \mathcal{H}$ we write $Ph := \mathbb{E}[h(X)]$, $P_n h := \frac{1}{n}\sum_{i=1}^n h(X_i)$, and $\|P_n - P\|_{\mathcal{H}} := \sup_{h\in\mathcal{H}} |(P_n - P)h|$.

Let P denote the probability measure with density f. With these new definitions, the sample average minimization objective in (3) can be expressed $P_n\psi_S$, while the population counterpart in (2) reads as $P\psi_S = \mathrm{KL}(S\#f|g)$. Suppose the estimator S^n of the sampling map S^* is a random element in $\mathcal{T}(s, d, M^*)$ obtained as a near-minimizer of $P_n\psi_S$. Let

$$R_n = P_n\psi_{S^n} - \inf_{S\in\mathcal{T}(s,d,M)} P_n\psi_S \geq 0$$

denote the approximation error of our optimization algorithm. Our goal is to bound the loss $P\psi_{S^n}$. Fix $\epsilon > 0$ and let $\tilde{S}$ be any deterministic element of $\mathcal{T}(s, d, M^*)$ that nearly minimizes $P\psi_S$, i.e., suppose

$$P\psi_{\tilde{S}} \leq \inf_{S\in\mathcal{T}(s,d,M^*)} P\psi_S + \epsilon.$$

It follows that

$$P\psi_{S^n} - \inf_S P\psi_S \leq 2\|P_n - P\|_{\Psi(s,d,M^*)} + R_n + \epsilon.$$

As $\epsilon > 0$ was arbitrary, we conclude that

$$P\psi_{S^n} - \inf_S P\psi_S \leq 2\|P_n - P\|_{\Psi(s,d,M^*)} + R_n. \quad (4)$$

Controlling the deviations of the empirical process $\|P_n - P\|_{\Psi(s,d,M^*)}$ as in Lemma 3.4 allows us to bound the loss of the estimator S^n and establish consistency and a rate of convergence in KL divergence.

Lemma 3.4. *Under Assumptions 3.1-3.3, we have*

$$\mathbb{E}\|P_n - P\|_{\Psi(s,d,M^*)} \lesssim \begin{cases} n^{-1/2}, & d < 2s, \\ n^{-1/2}\log n, & d = 2s, \\ n^{-s/d}, & d > 2s. \end{cases}$$

Remark 3.5. Consider the case where $2s > d$, for example when both f, g are the densities of the standard normal distribution, then we have by the central limit theorem that for any $S \in \mathcal{T}(s, d, M^*)$, $\sqrt{n}\,[P_n\psi_S - P\psi_S]$ converge in law towards a centered Gaussian distribution. Therefore, for any $S \in \mathcal{T}(s, d, M^*)$, $\mathbb{E}\|P_n - P\|_{\Psi(s,d,M^*)}$ is at least as large as

$$\sup_{S\in\mathcal{T}(s,d,M^*)} \mathbb{E}\left[|P_n\psi_S - P\psi_S|\right] \gtrsim n^{-1/2}$$

which shows that our results are tight at least in the smooth regime.

The proof of Lemma 3.4 relies on metric entropy integral bounds established by Dudley (1967) and van der Vaart and Wellner (1996). Although we have phrased the sample complexity bounds in Lemma 3.4 in terms of the expectation of the empirical process $\|P_n - P\|_{\Psi(s,d,M)}$, high probability bounds can be obtained similarly (Wainwright, 2019).

Hence, the following KL consistency theorem is obtained as a direct result of Lemma 3.4 and the risk decomposition (4).

Theorem 3.6. *Suppose Assumptions 3.1-3.3 hold. Let S^n be a near-optimizer of the functional $S \mapsto P_n\psi_S$ on $\mathcal{T}(s, d, M^*)$ with remainder R_n given by*

$$R_n = P_n\psi_{S^n} - \inf_{S\in\mathcal{T}(s,d,M^*)} P_n\psi_S = o_P(1).$$

Then $P\psi_{S^n} \xrightarrow{p} P\psi_{S^*} = 0$, i.e., S^n is a consistent estimator of S^* with respect to KL divergence.

Moreover, if R_n is bounded in expectation as

$$\mathbb{E}[R_n] \lesssim \mathbb{E}\|P_n - P\|_{\Psi(s,d,M^*)},$$

then the expected KL divergence of S^n is bounded as

$$\mathbb{E}[P\psi_{S^n}] \lesssim \begin{cases} n^{-1/2}, & d < 2s, \\ n^{-1/2}\log n, & d = 2s, \\ n^{-s/d}, & d > 2s. \end{cases}$$

Remark 3.7. Note that, as we work on a compact set, we also obtain the rates of convergence of S^n with respect to the Wasserstein metric thanks to Pinsker's inequality:

$$\mathbb{E}[W(S^n\#f, g)] \lesssim \begin{cases} n^{-1/4}, & d < 2s, \\ n^{-1/4}\log n, & d = 2s, \\ n^{-s/2d}, & d > 2s. \end{cases}$$

where for any probability measures μ and ν on $\mathbb{R}^d$ with finite first moments, we denote $W(\mu, \nu)$ the Wasserstein-1 distance between μ and ν.

Uniform Convergence Although Theorem 3.6 only establishes a weak form of consistency in terms of the KL divergence, we leverage this result to prove strong consistency, in the sense of uniform convergence in probability, in Theorem 3.8. The proof requires understanding the regularity of the KL divergence with respect to the topology induced by the $\|\cdot\|_{\infty,d}$ norm. In the Supplement, we establish that KL is lower semicontinuous in $\|\cdot\|_{\infty,d}$ utilizing the weak lower semicontinuity of KL proved by Donsker and Varadhan (1975).

Theorem 3.8. *Suppose Assumptions 3.1-3.3 hold. Let S^n be any near-optimizer of the functional $S \mapsto P_n\psi_S$ on $\mathcal{T}(s,d,M^*)$, i.e., suppose*

$$P_n\psi_{S^n} = \inf_{S\in\mathcal{T}(s,d,M^*)} P_n\psi_S + o_P(1).$$

Then $\|S^n - S^*\|_{\infty,d} \xrightarrow{p} 0$, i.e, S^n is a consistent estimator of S^* with respect to the uniform norm $\|\cdot\|_{\infty,d}$.

Inverse Consistency We have proved consistency of the estimator S^n of the sampling map S^*, pushing forward the target f to the source g. We can also get the consistency and an identical rate of convergence of $T^n = (S^n)^{-1}$ estimating $T^* = (S^*)^{-1}$, although the proof of the analog to Theorem 3.8 establishing uniform consistency of T^n is much more involved. We defer to the Supplement for details.

4 LOG-CONCAVITY, DIMENSION ORDERING, JACOBIAN FLOWS

Sobolev-type Rates Under Log-concavity Suppose the source density g is log-concave. Then $\min_{S\in\mathcal{T}(s,d,M)} P_n\psi_S$ is a convex problem; moreover if g is strongly log-concave, strong convexity follows. The user can choose a convenient g, such as a multivariate Gaussian with support truncated to a compact convex set. In this case, we can establish a bound on the rate of convergence of S^n to S^* in the L^2 Sobolev-type norm

$$\|S\|^2_{H_f^{1,2}(\mathcal{X})} := \sum_{k=1}^{d}\left\{\|S_k\|^2_{L_f^2(\mathcal{X})} + \|D_kS_k\|^2_{L_f^2(\mathcal{X})}\right\}.$$

Here $\|\cdot\|_{L_f^2(\mathcal{X})}$ denotes the usual L^2 norm integrating against the target density f.

Theorem 4.1. *Suppose Assumptions 3.1-3.3 hold. Assume further that g is m-strongly log-concave. Let S^n be a near-optimizer of the functional $S \mapsto P_n\psi_S$ on $\mathcal{T}(s,d,M^*)$ with remainder R_n satisfying*

$$\mathbb{E}[R_n] \lesssim \mathbb{E}\|P_n - P\|_{\Psi(s,d,M^*)}.$$

Then S^n converges to the true sampling map S^* with respect to the norm $H_f^{1,2}(\mathcal{X})$ norm with rate

$$\mathbb{E}\|S^n - S^*\|^2_{H_f^{1,2}(\mathcal{X})} \lesssim \begin{cases} n^{-1/2}, & d < 2s,\\ n^{-1/2}\log n, & d = 2s,\\ n^{-s/d}, & d > 2s,\end{cases}$$

With more work, we can establish Sobolev convergence rates of the same order in n for $T^n = (S^n)^{-1}$ to $T^* = (S^*)^{-1}$, yet now in the appropriate norm $\|\cdot\|_{H_g^{1,2}(\mathcal{Y})}$; see details in the Supplement.

Remark 4.2. Note that here we obtain the rates with respect to the $H_f^{1,2}(\mathcal{X})$ norm from which we deduce immediately similar rates for the $H_f^{1,p}(\mathcal{X})$ norm with $p \geq 2$. In addition, for fixed p, under Assumptions 3.1-3.3, the $H_f^{1,p}(\mathcal{X})$ norm is equivalent to $H^{1,p}(\mathcal{X})$ norm with the Lebesgue measure as the reference measure. Finally, here we show the rates of the strong consistency of our estimator with respect $H_f^{1,\cdot}(\mathcal{X})$ norm as we do not require any additional assumptions on the higher order derivatives of the maps living in $\mathcal{T}(s,d,M^*)$. Additional assumptions on these derivatives may lead to the convergence rates of higher order derivatives of our estimate with respect to the $H_f^{k,p}(\mathcal{X}), k > 1$ norm which is beyond the scope of the present paper.

Dimension Ordering Suppose now that the smoothness of the target density f is anisotropic. That is, assume $f(x_1,\ldots,x_d)$ is s_k-smooth in x_k for each $k \in [d]$. As there are $d!$ possible ways to order the coordinates, the question arises: how should we arrange $(x_1,\ldots,x_d)$ such that the estimator S^n converges to the true KR map S^* at the fastest possible rate? Papamakarios et al. (2017) provide a discussion of this issue in the context of autoregressive flows in Section 2.1 therein; they construct a simple 2D example in which the model fails to learn the target density if the wrong order of the variables is chosen.

This relates to choices made in neural architectures for normalizing flows on images and texts. Our results suggest here that one would rather start with the coordinates (i.e. data parts) that are the least smooth and make their way through the triangular construction to the most smooth ones. We formalize the anisotropic smoothness of the KR map as follows.

Assumption 4.1. *Let $\mathfrak{s} = (s_1,\ldots,s_d) \in \mathbb{Z}_+^d$ be a multi-index with $s_k \geq 1$ for all $k \in [d]$. The density f is $\mathfrak{s}$-smooth on its support, in the sense that*

$$D^\alpha f(x) := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}} f(x)$$

exists and is continuous for every multi-index $\alpha = (\alpha_1,\ldots,\alpha_d) \in \mathbb{Z}_+^d$ satisfying $\alpha \preceq \mathfrak{s}$, i.e., $\alpha_k \leq s_k$

for every $k \in [d]$. Furthermore, we assume that $g(y)$ is $(\|\mathfrak{s}\|_\infty, \ldots, \|\mathfrak{s}\|_\infty)$-smooth with respect to $y = (y_1, \ldots, y_d)$ on $\mathcal{Y}$.

As the source density g is a degree of freedom in the problem, we are free to impose this assumption on g. Note that $(\|\mathfrak{s}\|_\infty, \ldots, \|\mathfrak{s}\|_\infty)$-smoothness of g is equivalent to $\|\mathfrak{s}\|_\infty$-smoothness of g as defined in Assumption 3.3. The results that follow are slight variations on those in Sections 3.1 and 3.2 adapted to the anisotropic smoothness of the densities posited in Assumption 4.1.

Under Assumptions 3.1, 3.2, and 4.1, there exists some $M^* > 0$ such that the KR map S^* from f to g lies in $\mathcal{T}(\mathfrak{s}, d, M^*)$; see the Supplement for a proof. We also define the class of loss functions

$$\Psi(\mathfrak{s}, d, M^*) = \{\psi_S : S \in \mathcal{T}(\mathfrak{s}, d, M^*)\},$$

which appear in the objective $P_n \psi_S$. Hence, we can proceed as above to bound the metric entropy and obtain uniform convergence and Sobolev-type rates for estimators in the function class $\mathcal{T}(\mathfrak{s}, d, M^*)$. Appealing to metric entropy bounds for anisotropic smoothness classes (Birgé, 1986, Proposition 2.2), we have the following analog of Lemma 3.4.

Lemma 4.3. *For $k \in [d]$, let $d_k = d - k + 1$ and $\sigma_k = d_k \left(\sum_{j=k}^d s_j^{-1}\right)^{-1}$. Under Assumptions 3.1, 3.2, and 4.1, we have*

$$\mathbb{E}\|P_n - P\|_{\Psi(\mathfrak{s},d,M^*)} \lesssim \sum_{k=1}^d c_{n,k}.$$

where we define

$$c_{n,k} = \begin{cases} n^{-1/2}, & d_k < 2\sigma_k, \\ n^{-1/2}\log n, & d_k = 2\sigma_k, \\ n^{-\sigma_k/d_k}, & d_k > 2\sigma_k. \end{cases}$$

Lemma 4.3 is proved via a chain rule decomposition of relative entropy which relies upon the triangularity of the hypothesis maps $\mathcal{T}(\mathfrak{s}, d, M^*)$. From here we can repeat the analysis in Section 3.2 to obtain consistency and bounds on the rate of convergence of the estimators S^n and $T^n = (S^n)^{-1}$ of the sampling map S^* and the direct map $T^* = (S^*)^{-1}$, respectively, in the anisotropic smoothness setting of Assumption 4.1. All the results in these Sections are true with $\sum_k c_{n,k}$ replacing the rate under isotropic smoothness.

In order to minimize this bound to obtain an optimal rate of convergence, we should order the coordinates $(x_1, \ldots, x_d)$ such that σ_k is as large as possible for each $k \in [d]$. Inspecting the definition of σ_k in Lemma 4.3, we see that this occurs when $s_1 \leq \cdots \leq s_d$.

Theorem 4.4. *The bound on the rate of convergence $\sum_k c_{n,k}$ is minimized when $s_1 \leq \cdots \leq s_d$, i.e., when the smoothness of the target density f in the direction x_k increases with $1 \leq k \leq d$.*

Our result on the optimal ordering of coordinates complements the following theorem of Carlier et al. (2010).

Theorem 4.5. *Let f and g be compactly supported Lebesgue densities on $\mathbb{R}^d$. Let $\epsilon > 0$ and let γ^ϵ be an optimal transport plan between f and g for the cost*

$$c_\epsilon(x, y) = \sum_{k=1}^d \lambda_k(\epsilon)(x_k - y_k)^2,$$

for some weights $\lambda_k(\epsilon) > 0$. Suppose that for all $k \in \{1, \ldots, d-1\}$, $\lambda_k(\epsilon)/\lambda_{k+1}(\epsilon) \to 0$ as $\epsilon \to 0$. Let S^* be the Knöthe-Rosenblatt map between f and g and $\gamma^* = (id \times S^*)\#f$ the associated transport plan. Then $\gamma^\epsilon \rightsquigarrow \gamma^*$ as $\epsilon \to 0$. Moreover, should the plans γ^ϵ be induced by transport maps S^ϵ, then these maps would converge to S^* in $L^2(f)$ as $\epsilon \to 0$.

With this theorem in mind, the KR map S^* can be viewed as a limit of optimal transport maps S^ϵ for which transport in the dth direction is more costly than in the $(d-1)$th, and so on. The anisotropic cost function $c_\epsilon(x, y)$ inherently promotes increasing regularity of S^ϵ in x_k for larger $k \in [d]$. Theorem 4.4 establishes the same heuristic for learning triangular flows based on KR maps to build generative models.

In particular, our result suggests that we should order the coordinates such that f is smoothest in the x_d direction. Intuitively, this is because the component maps $S_k^*, k \in [d]$ all depend on the dth coordinate of the data. As such, we should leverage smoothness in x_d to stabilize all of our estimates as much as possible. If not, we risk propagating error throughout. In comparison, only the estimate of the first component S_1^* depends on x_1. Since our estimator depends comparatively little on x_1, we should order the coordinates so that f is least smooth in the x_1 direction.

Jacobian Flows Suppose now we solve

$$\inf_S \mathrm{KL}(S\#f|g) = \inf_S P\psi_S$$

where the candidate map S is a composition of smooth monotone increasing upper triangular maps U^j and orthogonal linear transformations Σ^j for $j \in [m]$, i.e.,

$$S(x) = U^m \circ \Sigma^m \circ \cdots \circ U^1 \circ \Sigma^1(x). \tag{5}$$

We call S a *Jacobian flow* of order m. This model captures many popular autoregressive flows (Kobyzev et al., 2020), such as Real NVP (Dinh et al., 2017), in which the Σ^j are "masking" permutation matrices.

For simplicity, in this section we assume that f and g are supported on the unit ball $B_d(0,1) \subset \mathbb{R}^d$ centered at the origin. Since Σ^j are orthogonal, we have $\Sigma^j(B_d(0,1)) = B_d(0,1)$. Hence, to accommodate the setup of the preceding sections, we can guarantee that S maps from $\mathcal{X} = B_d(0,1)$ to $\mathcal{Y} = B_d(0,1)$ by requiring $\sup_{x\in B_d(0,1)} \|U^j(x)\|_2 \le 1$ for $j \in [m]$.

Definition 4.6. *Define the class $\mathcal{J}_m(\Sigma, s, M)$ of s-smooth Jacobian flows of order m to consist of those maps S of the form (5) such that*

1. $\Sigma = (\Sigma^1, \ldots, \Sigma^m)$ *are fixed orthogonal matrices,*
2. $\sup_{x\in B_d(0,1)} \|U^j(x)\|_2 \le 1$ *for* $j \in [m]$,
3. $U^j \in \mathcal{T}(s, d, M)$ *for* $j \in [m]$.

We also define $\Psi_m(\Sigma, s, M) = \{\psi_S : S \in \mathcal{J}_m(\Sigma, s, M)\}$.

By expanding our search to $\mathcal{J}_m(\Sigma, s, M)$, we are not targeting the KR map S^* and we are no longer guaranteed uniqueness of a KL minimizer. Nevertheless, we can study the performance of estimators $S^n \in \overline{\mathcal{J}_m(\Sigma, s, M)}$ as compared to minimizers of KL in $\overline{\mathcal{J}_m(\Sigma, s, M)}$, which are guaranteed to exist by compactness of $\overline{\mathcal{J}_m(\Sigma, s, M)}$ and lower semicontinuity of KL in $\|\cdot\|_{\infty,d}$.

Since the Σ^j are orthogonal, we have $|\det(J\Sigma^j(x))| = |\det(\Sigma^j)| = 1$, and therefore

$$\psi_S(x) = \log[f(x)/g(S(x))] - \sum_{j=1}^{m}\sum_{k=1}^{d} \log D_k U_k^j(x^j),$$

where we define

$$x^j = \Sigma^j \circ U^{j-1} \circ \Sigma^{j-1} \circ \cdots \circ U^1 \circ \Sigma^1(x), \quad j \in [m].$$

Hence, with a loss decomposition mirroring that in Definition 3.2, we can apply the methods of the preceding sections to establish quantitative limits on the loss incurred by the estimates S^n in finite samples.

Theorem 4.7. *Suppose f, g are s-smooth and supported on $B_d(0,1)$. Let S^n be a near-optimizer of the functional $S \mapsto P_n\psi_S$ on $\mathcal{J}_m(\Sigma, s, M)$ with*

$$R_n = P_n\psi_{S^n} - \inf_{S\in\mathcal{J}_m(\Sigma,s,M)} P_n\psi_S = o_P(1).$$

Further, let S^0 be any minimizer of $S \mapsto P\psi_S$ on $\overline{\mathcal{J}_m(\Sigma, s, M)}$. It follows that

$$P\psi_{S^n} \xrightarrow{p} P\psi_{S^0}.$$

Moreover, if R_n is bounded in expectation as

$$\mathbb{E}[R_n] \lesssim \mathbb{E}\|P_n - P\|_{\Psi_m(\Sigma,s,M)},$$

then the expected KL divergence of S^n is bounded as

$$\mathbb{E}[P\psi_{S^n}] - P\psi_{S^0} \lesssim \begin{cases} n^{-1/2}, & d < 2s, \\ n^{-1/2}\log n, & d = 2s, \\ n^{-s/d}, & d > 2s. \end{cases}$$

5 DISCUSSION

Related Work Previous work on KR couplings has focused on existence and approximation questions in relation to universal approximation (Bogachev et al., 2005; Alexandrova, 2006; Huang et al., 2018). This universal approximation property of KR maps has frequently motivated normalizing flows from a theoretical viewpoint; see Sec. 3 in Kobyzev et al. (2020). KR maps have been used for various learning and inference problems (Kobyzev et al., 2020; Spantini et al., 2018; Marzouk et al., 2016; El Moselhy and Marzouk, 2012).

Kong and Chaudhuri (2020) study the expressiveness of basic types of normalizing flows, such as planar flows, Sylvester flows, and Householder flows. For $d = 1$, they show that such flows are universal approximators. However, when the distributions lives in a d-dimensional space with $d \ge 2$, the authors provide a partially negative answer to the universal approximation power of these flows. For example, they exhibit cases where Sylvester flows cannot recover the target distributions. Their results can be seen as complementary to ours as we give examples of arbitrary slow statistical rates and we develop the consistency theory of KR-type flows.

Jaini et al. (2020) investigate the properties of the increasing triangular map required to push a tractable source density with known tails onto a desired target density. Then they consider the general d-dimensional case and show similarly that imposing smoothness condition on the increasing triangular map will result in a target density with the same tail properties as the source. Such results suggest that without any assumption on the target distribution, the transport map might be too irregular to be estimated. These results echo our assumptions on the target to obtain fast rates and complement ours by focusing on the tail behavior while we focus on the consistency and the rates.

Conclusion We have established the uniform consistency and convergence rates of statistical estimators of the Knöthe-Rosenblatt rearrangement, highlighting the anisotropic geometry of function classes at play in triangular flows. Our results also lead to statistical guarantees for Jacobian flows. Identifying other function classes of source densities that lead to faster rates is an interesting venue for future work.

Acknowledgements The authors would like to thank Z. Kong and K. Chaudhuri for sharing their code to reproduce their results (Kong and Chaudhuri, 2020). N. Irons is supported by a Shanahan Endowment Fellowship and a Eunice Kennedy Shriver National Institute of Child Health and Human Development training grant, T32 HD101442-01, to the Center for Studies in Demography & Ecology at the University of Washington, and NSF CCF-2019844. M. Scetbon is supported by "Chaire d'excellence de l'IDEX Paris Saclay". S. Pal is supported by NSF DMS-2052239 and a PIMS CRG (PIHOT). Z. Harchaoui is supported by NSF CCF-2019844, NSF DMS-2134012, NSF DMS-2023166, CIFAR-LMB, and faculty research awards. Part of this work was done while Z. Harchaoui was visiting the Simons Institute for the Theory of Computing. The authors would like to thank the Kantorovich Initiative of the Pacific Institute for the Mathematical Sciences (PIMS) for supporting this collaboration. This work was first presented at the Joint Statistical Meetings in August 2021.

References

Alexandrova, D. E. (2006). Convergence of triangular transformations of measures. *Theory of Probability & Its Applications*, 50(1):113–118.

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrscheinlichkeitstheor. Verw. Geb.*, 65:181–237.

Birgé, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Th. Rel. Fields*, 71:271–291.

Bogachev, V. I., Kolesnikov, A. V., and Medvedev, K. V. (2005). Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309.

Bousquet, O., Boucheron, S., and Lugosi, G. (2004). Introduction to statistical learning theory. In Bousquet, O., von Luxburg, U., and Rätsch, G., editors, *Advanced Lectures on Machine Learning*, pages 169–207. Springer, Berlin, Heidelberg.

Carlier, G., Galichon, A., and Santambrogio, F. (2010). From Knöthe's transport to Brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41(6):2554–2576.

Devroye, L. (1983). On arbitrarily slow rates of global convergence in density estimation. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 62:475–483.

Devroye, L. (1995). Another proof of a slow convergence result of Birgé. *Statistics and Probability Letters*, 23:63–67.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.

Dinh, L., Krueger, D., and Bengio, Y. (2015). NICE: Non-linear Independent Components Estimation. In *ICLR Workshop*.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2017). Density estimation using Real NVP. In *ICLR*.

Donsker, M. and Varadhan, S. (1975). Asymptotic evaluation of certain Markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, 28(1):1–47.

Dudley, R. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1:290–330.

Dudley, R. (1968). The speed of mean Glivenko-Cantelli convergence. *Ann. Math. Statist.*, 40:40–50.

El Moselhy, T. A. and Marzouk, Y. M. (2012). Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850.

Germain, M., Gregor, K., Murray, I., and Larochelle, H. (2015). MADE: Masked Autoencoder for Distribution Estimation. In *ICML*.

Grathwohl, W., Chen, R., Bettencourt, J., Sutskever, I., and Duvenaud, D. (2019). FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *ICLR*.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*, volume 1. Springer.

Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. (2018). Neural autoregressive flows. In *ICML*.

Hütter, J.-C. and Rigollet, P. (2021). Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194.

Irons, N. J., Scetbon, M., Pal, S., and Harchaoui, Z. (2021). Triangular flows for generative modeling: Statistical consistency, smoothness classes, and fast rates. *CoRR*, abs/2112.15595.

Jaini, P., Kobyzev, I., Yu, Y., and Brubaker, M. (2020). Tails of Lipschitz triangular flows. In *International Conference on Machine Learning*, pages 4673–4681. PMLR.

Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR*.

Kingma, D. and Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215—10224.

Knöthe, H. (1957). Contributions to the theory of convex bodies. *The Michigan Mathematical Journal*, 4(1):39–52.

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*.

Kong, Z. and Chaudhuri, K. (2020). The expressive power of a class of normalizing flow models. In *International Conference on Artificial Intelligence and Statistics*, pages 3599–3609. PMLR.

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

Nickl, R. and Pötscher, B. (2007). Bracketing metric entropy rates and empirical central limit theorems for function classes of Besov- and Sobolev-Type. *J Theor Probab*, 20:177–199.

Papamakarios, G., Pavlakou, T., and Murray, I. (2017). Masked autoregressive flow for density estimation. In *NeurIPS*.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472.

Rudin, W. (1976). *Principles of Mathematical Analysis*. McGraw-Hill, 3rd edition.

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians*. Birkhäuser.

Spantini, A., Bigoni, D., and Marzouk, Y. (2018). Inference via low-dimensional couplings. *The Journal of Machine Learning Research*, 19(1):2639–2709.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

Wainwright, M. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.

Wehenkel, A. and Louppe, G. (2019). Unconstrained monotonic neural networks. In *Advances in Neural Information Processing Systems*, pages 1543–1553.

This Supplement collects the numerical illustrations to illustrate the theoretical results in the main text, as well as the detailed proofs of the theoretical results stated in the main text. Sec. B.1 details the derivations of the Kullback-Leibler objective. Sec. B.2 details the proof of the slow rates in Sec. 2 of main text. Sec. B.3 provides estimates of metric entropy from Sec. 3 of main text. Sec. B.4 provides the proofs of the statistical consistency from Sec. 3 of main text. Sec. B.5 provides the proofs of the Sobolev rates under log-concavity of g from Sec. 4 of main text. Sec. B.6 expands on the dimension ordering from Sec. 4 of main text. Sec. B.7 expands on the extension to Jacobian flow from Sec. 4 of main text.

A Numerical illustrations

We conducted numerical experiments to illustrate our theoretical results. Code to reproduce our experiments is available at https://github.com/njirons/krc-stat.

To estimate the KR map, we used Unconstrained Monotonic Neural Networks Masked Autoregressive Flows (UMNN-MAF), a particular triangular flow introduced in Wehenkel and Louppe (2019), with code to implement the model provided therein. UMNN-MAF learns an invertible monotone triangular map targeting the KR rearrangement via KL minimization

$$S(x;\theta) = \begin{bmatrix} S_1(x_1,\dots,x_d;\theta) \\ S_1(x_2,\dots,x_d;\theta) \\ \vdots \\ S_{d-1}(x_{d-1},x_d;\theta) \\ S_d(x_d;\theta) \end{bmatrix},$$

where each component $S_k(x_{k:d};\theta)$ is parametrized as the output of a neural network architecture that can learn arbitrary monotonic functions. Specifically, we have

$$S_k(x_{k:d},\theta) = \int_0^{x_k} f_k(t, h_k(x_{(k+1):d};\phi_k);\psi_k)dt + \beta_k(h_k(x_{(k+1):d};\phi_k);\psi_k), \tag{6}$$

where $h_k(\cdot;\phi_k) : \mathbb{R}^{d-k-1} \to \mathbb{R}^q$ is a q-dimensional neural embedding of $x_{(k+1):d}$ and $\beta_k(\cdot;\psi_k) : \mathbb{R}^q \to \mathbb{R}$ is parametrized by a neural network. Each f_k is a strictly positive function parametrized by a neural network, which guarantees that $\int_0^{x_k} f_k$ is increasing in x_k. Here the total parameter θ is defined as $\theta = \{(\psi_k,\phi_k)\}_{k=1}^d$. Further details of the model are provided in Wehenkel and Louppe (2019). We note that UMNN-MAF is captured by the model setup in our theoretical treatment of KR map estimation.

The model was trained via log-likelihood maximization using minibatch gradient descent with the Adam optimizer (Kingma and Ba, 2015) with minibatch size 64, learning rate 10^{-4}, and weight decay 10^{-5}. The integrand network architecture defined in equation (6) consisted of 4 hidden layers of width 100. Following Wehenkel and Louppe (2019), the architecture of the embedding networks is the best performing MADE network (Germain et al., 2015) used in NAF (Huang et al., 2018). We used 20 integration steps to numerically approximate the integral in equation (6). The source density g is a bivariate standard normal distribution. The population negative log-likelihood loss, which differs from the KL objective by a constant factor (namely, the negative entropy of the target density f), was approximated by the empirical negative log-likelihood on a large independently generated test set of size $N = 10^5$.

Figure 3 exhibits our results for UMNN-MAF trained on 8 two-dimensional datasets considered in Wehenkel and Louppe (2019) and Grathwohl et al. (2019). Heatmaps of the target densities are displayed in the top rows, while the bottom rows show the log-likelihood convergence rates as the sample size increases from $n = 10^3$ to $n = 10^4$ on a log-log scale. For each training sample size, we repeated the experiment 100 times. We report the mean of the loss over the 100 replicates with 95% error bars.

These experiments highlight the impact of the ordering of coordinates on the convergence rate, as predicted by Theorem 4.4. The blue curves correspond to first estimating the KR map along the horizontal x_1 axis, then the vertical axis conditional on the horizontal, $x_2|x_1$. The orange curves show the reverse order, namely estimating the KR map along the vertical x_2 axis first. The 5 densities in the top rows (2 spirals, pinwheel, moons, and banana) and the bottom right (swiss roll) are asymmetric in (x_1,x_2) (i.e., $f(x_1,x_2)$ is not exchangeable in (x_1,x_2)). These densities exhibit different convergence rates depending on the choice of order. The remaining 3 densities in the

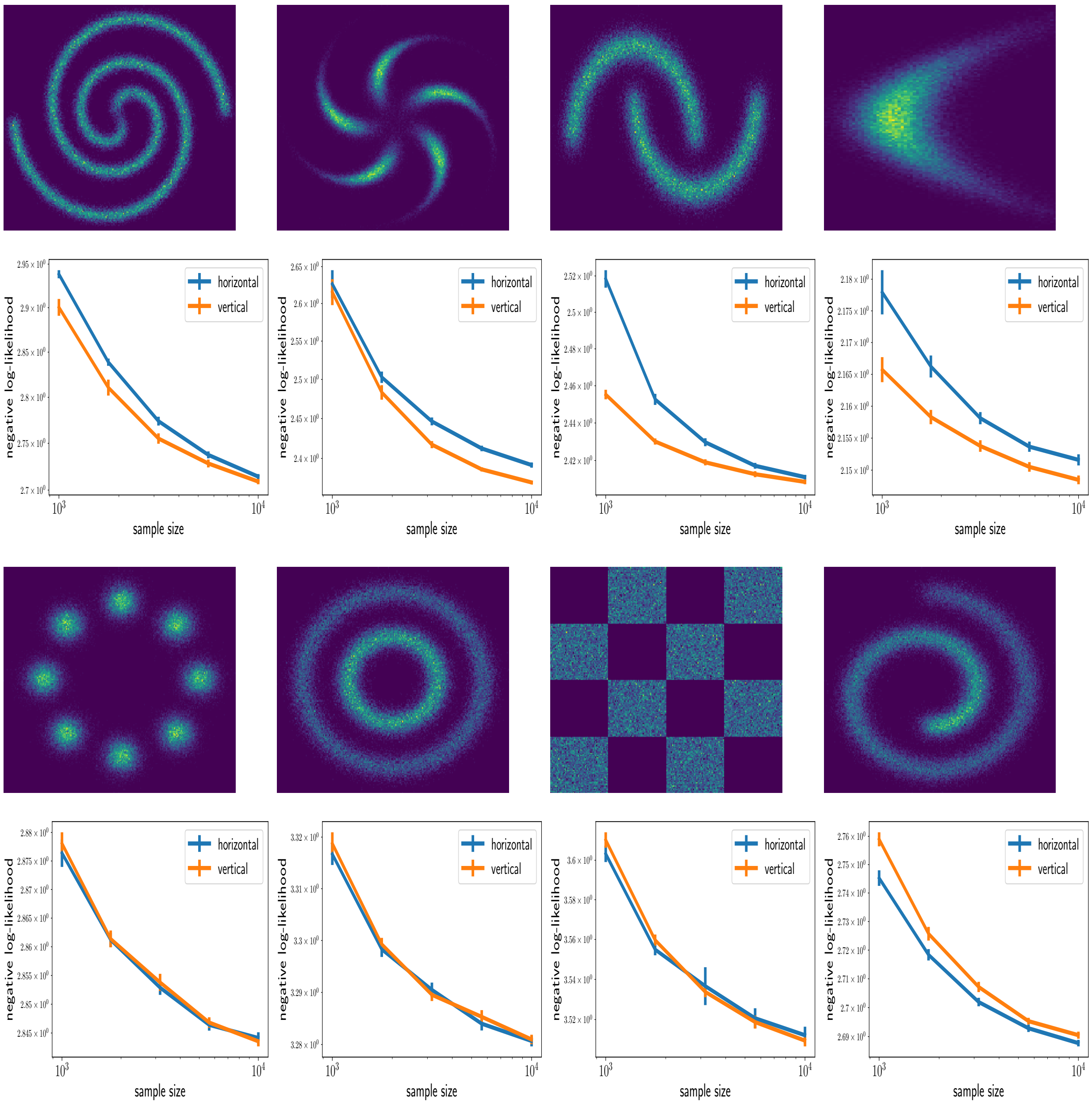

Figure 3: Experimental comparison of convergence rates on 2D examples (top: density; bottom: rates) for the 8 datasets considered in Wehenkel and Louppe (2019); Grathwohl et al. (2019).

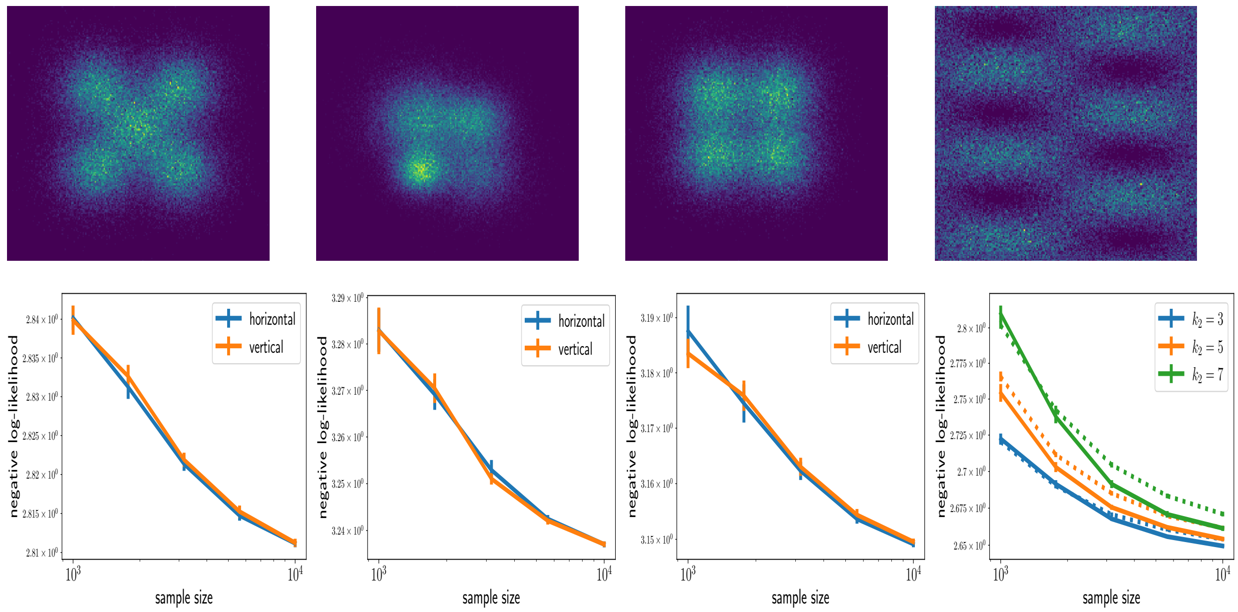

Figure 4: Experimental comparison of convergence rates on 2D examples (top: density; bottom: rates) for the 3 datasets considered in Kong and Chaudhuri (2020) and the sine density with varying frequency $k_2 \in \{3, 5, 7\}$.

bottom rows (8 gaussians, 2 circles, checkerboard) are symmetric in (x_1, x_2) and do not exhibit this behavior, which is to be expected. We also note that a linear trend (on the log-log scale) is visible in the plots for smaller n, which aligns with the convergence rates established in Theorem 3.6. For larger n, however, approximation error dominates and the loss plateaus.

As an illustrative example, we focus in on the top right panel of Figure 3, which plots the *banana* density $f(x_1, x_2)$ corresponding to the random variables

$$
\begin{aligned}
X_2 &\sim N(0, 1) \\
X_1 | X_2 &\sim N(X_2^2/2, 1/2).
\end{aligned}
$$

It follows that f is given by

$$
\begin{aligned}
f(x_1, x_2) &= f(x_1|x_2) f(x_2) \\
&\propto \exp\left\{-\left(x_1 - x_2^2/2\right)^2\right\} \cdot \exp\left\{-x_2^2/2\right\}.
\end{aligned}
$$

Intuitively, estimating the normal conditional $f(x_1|x_2) = N(x_2^2/2, 1/2)$ and the standard normal marginal $f(x_2)$ should be easier than estimating $f(x_2|x_1)$ and $f(x_1)$. Indeed, as x_1 increases, we see that $f(x_2|x_1)$ transitions from a unimodal to a bimodal distribution. As such, we expect that estimating the KR map $S^{(21)}$ from $f(x_2, x_1)$ to the source density g should be more difficult than estimating the KR map $S^{(12)}$ from $f(x_1, x_2)$ to g. This is because the first component of $S^{(21)}$ targets the conditional distribution $f(x_2|x_1)$ and the second component targets $f(x_1)$, while the first component of $S^{(12)}$ targets $f(x_1|x_2)$ and the second component targets $f(x_2)$. Indeed, this is what we see in the top right panels of Figure 3, which shows the results of fitting UMNN-MAF to estimate $S^{(12)}$ (orange) and $S^{(21)}$ (blue). As expected, we see that estimates of $S^{(21)}$ converge more slowly than those of $S^{(12)}$. These results are consistent with the findings of Papamakarios et al. (2017), who observed this behavior in estimating the banana density with MADE (Germain et al., 2015).

In the first 3 panels of Figure 4, we repeat the experimental setup on the 3 normal mixture densities considered by Kong and Chaudhuri (2020). The conclusions drawn from Figure 3 are echoed here. We observe no dependence in the convergence rates on the choice of variable ordering, since the target densities $f(x_1, x_2)$ are exchangeable in (x_1, x_2). Furthermore, we observe a linear convergence rate as predicted by Theorem 3.6.

Inspired by the pathological densities constructed in the proof of the "no free lunch" Theorem 2.1, which are rapidly oscillating perturbations of the uniform density on the hypercube, we now consider the sine density on the hypercube, defined as

$$f(x_1, \dots, x_d) = 1 + \prod_{j=1}^{d} \sin(2\pi k_j x_j)$$

where $k_j \in \mathbb{Z}$ for $j \in [d]$, and $x = (x_1, \dots, x_d) \in [0,1]^d$. The smoothness of f, as measured by any L^p norm of its derivative(s), decreases as the frequency $|k_j|$ increases. As such, f parametrizes a natural family of functions to test our theoretical results concerning the statistical performance of KR map estimation as a function of the sample size n, the smoothness of the underlying target density, and the order of coordinates. The rightmost panels of Figure 4 plot the sine density with $k_1 = 1, k_2 = 3$ (top row) and convergence rates for the choices $k_1 = 1$ and $k_2 \in \{3, 5, 7\}$ (bottom row). The dashed lines correspond to estimating the marginal x_1 first, followed by $x_2|x_1$; the solid lines indicate the reverse order. We again see an effect of coordinate ordering on convergence rates. It is also apparent that convergence slows down as k_2 increases and f becomes less smooth.

B Detailed proofs

B.1 Kullback-Leibler objective

Derivation of (2). By the change of variables formula (1), the density $S\#f$ is given by

$$(S\#f)(y) = f(S^{-1}(y))|\det(J(S^{-1})(y))|.$$

Consequently, $\mathrm{KL}(S\#f|g)$ rewrites as

$$\begin{aligned}
\mathrm{KL}(S\#f|g) &= \int_Y (S\#f)(y) \log\left(\frac{(S\#f)(y)}{g(y)}\right) dy \\
&= \int_{\mathcal{Y}} f(S^{-1}(y))|\det(J(S^{-1})(y))| \log\left(\frac{f(S^{-1}(y))|\det(J(S^{-1})(y))|}{g(y)}\right) dy \\
&= \int_{\mathcal{Y}} f(S^{-1}(y))|\det(JS(S^{-1}(y)))|^{-1} \log\left(\frac{f(S^{-1}(y))}{g(y)|\det(JS(S^{-1}(y)))|}\right) dy \quad \text{(inverse function theorem)} \\
&= \int_{\mathcal{X}} f(x)|\det(JS(x))|^{-1} \log\left(\frac{f(x)}{g(S(x))|\det(JS(x))|}\right) \cdot |\det(JS(x))|\, dx \qquad (x := S^{-1}(y)) \\
&= \int_{\mathcal{X}} f(x) \log\left(\frac{f(x)}{g(S(x))|\det(JS(x))|}\right) dx \\
&= \mathbb{E}_{X\sim f}\left\{\log f(X) - \log g(S(X)) - \log|\det(JS(x))|\right\} \\
&= \mathbb{E}_{X\sim f}\left\{\log f(X) - \log g(S(X)) - \sum_{k=1}^{d} \log D_k S_k(X)\right\}.
\end{aligned}$$

The last line follows because S is assumed to be upper triangular and monotone non-decreasing, and therefore

$$|\det(JS(x))| = \prod_{k=1}^{d} D_k S_k(x).$$

Note that in the above calculation, we have also established that

$$\mathrm{KL}(S\#f|g) = \mathrm{KL}(f|S^{-1}\#g),$$

for any diffeomorphism $S : \mathbb{R}^d \to \mathbb{R}^d$, since

$$\begin{aligned}
\mathrm{KL}(S\#f|g) &= \int_{\mathcal{X}} f(x) \log\left(\frac{f(x)}{g(S(x))|\det(JS(x))|}\right) dx \\
&= \int_{\mathcal{X}} f(x) \log\left(\frac{f(x)}{(S^{-1}\#g)(x)}\right) dx \qquad \text{(change of variables)} \\
&= \mathrm{KL}(f|S^{-1}\#g).
\end{aligned}$$

This completes the proof. □

B.2 Slow rates

B.2.1 Proof of Theorem 2.1

Our proof follows the argument in Section V of Birgé (1986).

Proof of Theorem 2.1. For fixed $\epsilon \in (0,1)$, let $\tilde{h}(x;\epsilon)$ be a C^∞ bump function on $\mathbb{R}$ satisfying

1. $0 \le \tilde{h}(x;\epsilon) \le 1 \qquad \forall x \in \mathbb{R}$,
2. $\tilde{h}(x;\epsilon) = 1$ on the interval $[\epsilon/4, 1/2 - \epsilon/4]$,
3. $\tilde{h}(x;\epsilon) = 0$ outside of the interval $[0, 1/2]$.

Now for $r \in \mathbb{N}$, define the function $h_{\epsilon,r} : [0,1]^d \to [-1,1]$ by

$$h_{\epsilon,r}(x) = \tilde{h}(x_1 r;\epsilon) - \tilde{h}(x_1 r - 1/2;\epsilon).$$

It is clear that $h_{\epsilon,r}$ is smooth, $\sup_{x\in[0,1]^d} |h_{\epsilon,r}(x)| \le 1$, and $\int h_{\epsilon,r}(x)dx = 0$. Therefore, $1 + h_{\epsilon,r} \in \mathcal{F}$ is a smooth Lebesgue density on $[0,1]^d$ uniformly bounded by 2. Also note that $|h_{\epsilon,r}(x)| = 1$ whenever $[\epsilon/4 \le x_1 r \le 1/2 - \epsilon/4]$ or $[1/2 + \epsilon/4 \le x_1 r \le 1 - \epsilon/4]$. Furthermore, the support of $h_{\epsilon,r}$ is contained in the set $[0, 1/r] \times [0,1]^{d-1}$. It follows that

$$\begin{aligned}
\mathrm{TV}(1 + h_{\epsilon,r}, 1 - h_{\epsilon,r}) &= \frac{1}{2}\int |(1 + h_{\epsilon,r}(x)) - (1 - h_{\epsilon,r}(x))| dx \\
&= \int |h_{\epsilon,r}(x)| dx \\
&\ge \int_{[\epsilon/4 \le x_1 r \le 1/2 - \epsilon/4] \cup [1/2 + \epsilon/4 \le x_1 r \le 1 - \epsilon/4]} 1 \; dx_1 \\
&= \frac{1 - \epsilon}{r},
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{TV}(1 + h_{\epsilon,r}, 1 - h_{\epsilon,r}) &= \int |h_{\epsilon,r}(x)| dx \\
&\le \int_{[0,1/r]} 1 \; dx_1 \\
&= 1/r.
\end{aligned}$$

As we will see, these bounds on the total variation imply that the perturbations $1 + h_{\epsilon,r}, 1 - h_{\epsilon,r}$ are sufficiently similar to make identification a challenging task, but sufficiently different to incur significant loss when mistaken for each other.

Now define the translates $h_i(x;\epsilon,r) = h_{\epsilon,r}\left(x_1 - \frac{i-1}{r}, x_2, \ldots, x_d\right)$, which are disjointedly supported with support contained in $H_i = [(i-1)/r, i/r] \times [0,1]^{d-1}$ for $i = 1, \ldots, r$. Hereafter, we suppress dependence of h_i on ϵ, r for notational convenience. Consider the family of densities

$$\mathcal{F}(\epsilon, r) = \left\{ 1 + \sum_{i=1}^{r} \delta_i h_i : \delta_i = \pm 1 \right\} \subset \mathcal{F}$$

with cardinality 2^r. For $\delta \in \{\pm 1\}^r$ we write

$$f_\delta = 1 + \sum_{i=1}^{r} \delta_i h_i.$$

The worst-case KL risk on $\mathcal{F}$ of any density estimate $f_n = T^n \# g$ derived from a KR map estimate T^n can be bounded below as

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \mathbb{E}_f[\mathrm{KL}(f|f_n)] &\geq \sup_{f\in\mathcal{F}(\epsilon,r)} \mathbb{E}_f[\mathrm{KL}(f|f_n)] \\
&\geq \sup_{f\in\mathcal{F}(\epsilon,r)} \mathbb{E}_f[2\mathrm{TV}(f,f_n)^2] && \text{(Pinsker's inequality)} \\
&\geq \sup_{f\in\mathcal{F}(\epsilon,r)} 2\mathbb{E}_f[\mathrm{TV}(f,f_n)]^2. && \text{(Jensen's inequality)}
\end{aligned}$$

We aim to lower bound the total variation risk on $\mathcal{F}(\epsilon, r)$. We have

$$\begin{aligned}
\sup_{f\in\mathcal{F}(\epsilon,r)} \mathbb{E}_f[\mathrm{TV}(f,f_n)] &\geq 2^{-r} \sum_{f\in\mathcal{F}(\epsilon,r)} \mathbb{E}_f[\mathrm{TV}(f,f_n)] \\
&= 2^{-r} \sum_{\delta\in\{\pm1\}^r} \mathbb{E}_{f_\delta}\left[\mathrm{TV}\left(f_\delta, f_n\right)\right],
\end{aligned}$$

i.e, the worst-case risk is larger than the Bayes risk associated to the uniform prior on $\mathcal{F}(\epsilon, r)$. Now note that

$$\begin{aligned}
\mathrm{TV}\left(f_\delta, f_n\right) &= \frac{1}{2}\int |f_n(x) - f_\delta(x)| dx \\
&= \frac{1}{2}\int \left| f_n(x) - \left(1 + \sum_{i=1}^r \delta_i h_i(x)\right)\right| dx \\
&= \frac{1}{2}\sum_{i=1}^r \int_{H_i} |f_n(x) - (1+\delta_i h_i(x))| dx. && (\{H_i\} \text{ are disjoint})
\end{aligned}$$

Define

$$\ell_i(f_n) = \frac{1}{2}\int_{H_i} |f_n(x) - (1+h_i(x))| dx, \qquad \ell_i'(f_n) = \frac{1}{2}\int_{H_i} |f_n(x) - (1-h_i(x))| dx$$

and note that, by the triangle inequality,

$$\ell_i(f_n) + \ell_i'(f_n) \geq \frac{1}{2}\int_{H_i} |(1+h_i(x)) - (1-h_i(x))| dx = \int |h_i(x)| dx \geq \frac{1-\epsilon}{r}.$$

Writing F_δ^n to denote the cdf of an iid sample of size n from f_δ, we then have

$$\begin{aligned}
2^{-r}\sum_\delta \mathbb{E}_{f_\delta}\left[\mathrm{TV}\left(f_\delta, f_n\right)\right] &= 2^{-r}\sum_{i=1}^r \left\{ \sum_{\delta_i=1} \mathbb{E}_{f_\delta}[\ell_i(f_n)] + \sum_{\delta_i=-1} \mathbb{E}_{f_\delta}[\ell_i'(f_n)] \right\} \\
&= \frac{1}{2}\sum_{i=1}^r \left\{ \int \ell_i(f_n) d\left[2^{1-r}\sum_{\delta_i=1} F_\delta^n\right] + \int \ell_i'(f_n) d\left[2^{1-r}\sum_{\delta_i=-1} F_\delta^n\right]\right\} \\
&\geq \frac{1}{2}\sum_{i=1}^r \left\{ \int [\ell_i(f_n) + \ell_i'(f_n)] d\left(2^{1-r}\sum_{\delta_i=1} F_\delta^n \wedge 2^{1-r}\sum_{\delta_i=-1} F_\delta^n\right)\right\} \\
&\geq \frac{1-\epsilon}{2r}\sum_{i=1}^r \int d\left(2^{1-r}\sum_{\delta_i=1} F_\delta^n \wedge 2^{1-r}\sum_{\delta_i=-1} F_\delta^n\right) \\
&:= \frac{1-\epsilon}{2r}\sum_{i=1}^r \pi\left(2^{1-r}\sum_{\delta_i=1} F_\delta^n, 2^{1-r}\sum_{\delta_i=-1} F_\delta^n\right),
\end{aligned}$$

where $x \wedge y = \min(x, y)$. In the last line we defined the testing affinity π between two distribution functions F_p, F_q with Lebesgue densities p, q,

$$\pi(F_p, F_q) = \int d(F_p \wedge F_q) = \int (p \wedge q) dx,$$

which satisfies the well known identity

$$\pi(F_p, F_q) = 1 - \mathrm{TV}(p, q).$$

Since min is concave, Jensen's inequality implies that

$$\pi\left(2^{1-r}\sum_{\delta_i=1} F_\delta^n, 2^{1-r}\sum_{\delta_i=-1} F_\delta^n\right) \geq 2^{1-r}\sum_{(\delta,\delta')\in\Delta_i} \pi(F_\delta^n, F_{\delta'}^n),$$

where $\Delta_i = \{(\delta, \delta') : \delta_i = 1, \delta'_i = -1, \delta_j = \delta'_j \quad \forall j \neq i\}$. For any $(\delta, \delta') \in \Delta_i$, we have

$$\begin{aligned}
\pi(F_\delta^n, F_{\delta'}^n) &= 1 - \mathrm{TV}(f_\delta, f'_\delta) \\
&= 1 - \frac{1}{2}\int |f_\delta(x) - f_{\delta'}(x)|dx \\
&= 1 - \int |h_i(x)|dx \\
&\geq 1 - \frac{1}{r}.
\end{aligned}$$

Hence, we conclude that

$$\pi\left(2^{1-r}\sum_{\delta_i=1} F_\delta^n, 2^{1-r}\sum_{\delta_i=-1} F_\delta^n\right) \geq 2^{1-r}\sum_{\Delta_i}\left(1 - \frac{1}{r}\right) = 1 - \frac{1}{r}.$$

Thus, putting this all together, we have shown that

$$\begin{aligned}
\sup_{f\in\mathcal{F}(\epsilon,r)} \mathbb{E}_f[\mathrm{TV}(f, f_n)] &\geq \frac{1-\epsilon}{2r}\sum_{i=1}^{r}\left(1 - \frac{1}{r}\right) \\
&= \frac{1-\epsilon}{2}\left(1 - \frac{1}{r}\right).
\end{aligned}$$

Sending $\epsilon \to 0$ and $r \to \infty$, it follows that

$$\sup_{f\in\mathcal{F}(\epsilon,r)} \mathbb{E}_f[\mathrm{TV}(f, f_n)] \geq 1/2,$$

and hence

$$\sup_{f\in\mathcal{F}} \mathbb{E}_f[\mathrm{KL}(f|f_n)] \geq \sup_{f\in\mathcal{F}(\epsilon,r)} 2\mathbb{E}_f[\mathrm{TV}(f, f_n)]^2 \geq 1/2.$$

This completes the proof. Note that, with a little extra work, the lower bound can be improved to $d/2$ using the chain rule of entropy, since we have considered here perturbations $h_{\epsilon,r}$ varying only in the x_1 dimension. □

B.2.2 Proof of Theorem 2.2

Proof. Let h_n denote the first marginal of the density estimate $f_n = T^n\#g$. Note that h_n is a density on $\mathbb{R}$. Defining $\pi_1 : \mathbb{R}^d \to \mathbb{R}$ to be the projection along the first factor, we have

$$h_n = \pi_1\#f_n = (\pi_1 \circ T^n)\#g.$$

By Problem 7.5 in Devroye et al. (1996), for any positive sequence $1/16 \geq b_1 \geq b_2 \cdots$ converging to zero and any density estimate h_n there exists a density h on $\mathbb{R}$ such that

$$\mathbb{E}\left\{\int |h(x) - h_n(x)|dx\right\} \geq b_n.$$

Letting TV denote the total variation distance, this inequality can be rewritten as

$$2\mathbb{E}[\mathrm{TV}(h, h_n)] \geq b_n.$$

Setting $a_n = b_n^2/2$, which satisfies $a_1 \leq 1/512$, we find that

$$\begin{aligned}
\mathbb{E}[\mathrm{KL}(h|h_n))] &\geq \mathbb{E}[2 \cdot \mathrm{TV}(h, h_n)^2] && \text{(Pinsker's inequality)} \\
&\geq 2\mathbb{E}[\mathrm{TV}(h, h_n)]^2 && \text{(Jensen's inequality)} \\
&\geq b_n^2/2 \\
&= a_n.
\end{aligned}$$

Finally, let $f = h^{\otimes d}$ be the density on $\mathbb{R}^d$ defined as a d-fold product of h. By the chain rule of relative entropy, it follows that

$$\mathbb{E}[\mathrm{KL}(f|f_n)] \geq \mathbb{E}[\mathrm{KL}(h|h_n)] \geq a_n.$$

This completes the proof. Note that, with a little extra work, the lower bound can be improved to $d \cdot a_n$ by looking at the univariate conditional densities of f_n and using the chain rule of entropy, since we have only considered the first marginal of f_n here. $\square$

B.3 Upper bounds on metric entropy

We begin by defining relevant Sobolev function spaces, for which metric entropy bounds are known.

Definition B.1. *For $\mathcal{X} \subseteq \mathbb{R}^d$, define the function space*

$$D_s(\mathcal{X}) = \{\phi : D^\alpha \phi \text{ are uniformly continuous for all } |\alpha| \leq s\}.$$

and its subset

$$C_s(\mathcal{X}) = \left\{ \phi : \mathcal{X} \to \mathbb{R} : \sum_{0 \leq |\alpha| \leq s} \|D^\alpha \phi\|_\infty < \infty \right\} \cap D_s(\mathcal{X}).$$

endowed with the Sobolev norm

$$\|\phi\|_{H^{s,\infty}(\mathcal{X})} = \sum_{|\alpha| \leq s} \|D^\alpha \phi\|_\infty.$$

Proposition B.2 (Corollary 3, Nickl and Pötscher (2007))**.** *Assume $\mathcal{X} \subset \mathbb{R}^d$ is compact and let $\mathcal{F}$ be a bounded subset of $C_s(\mathcal{X})$ with respect to $\|\cdot\|_{H^{s,\infty}(\mathcal{X})}$ for some $s > 0$. The metric entropy of $\mathcal{F}$ in the L^∞ norm is bounded as*

$$H(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \lesssim \epsilon^{-d/s}.$$

With this result in hand, we can proceed to the proof of Proposition 3.3.

Proof of Proposition 3.3. This is a direct consequence of Proposition B.2. Indeed, under Assumptions 3.1-3.3 and Definition 3.1, for every $S \in \mathcal{T}(s, d, M)$, every term in $\exp(\psi_S(x))$ is bounded away from 0 and s-smooth with uniformly bounded derivatives. Since log is smooth away from 0, it follows that that every $\psi_S \in \Psi(s, d, M)$ is s-smooth with uniformly bounded derivatives. Consequently, $\Psi(s, d, M)$ is a bounded subset of $C_s(\mathcal{X})$ for every $M > 0$. $\square$

We also provide a metric entropy bound for the space $\mathcal{T}(\mathfrak{s}, d, M)$, which in particular establishes compactness of $\mathcal{T}(\mathfrak{s}, d, M)$ with respect to $\|\cdot\|_{\infty,d}$ (although this can also be proved with the Arzelà-Ascoli theorem).

Proposition B.3. *Let $\mathfrak{s} = (s_1, \ldots, s_d) \in \mathbb{Z}_+^d$, $d_k = d - k + 1$, and $\tilde{\sigma}_k = d_k \left((s_k + 1)^{-1} + \sum_{j=k+1}^d s_j^{-1} \right)^{-1}$ for $k \in [d]$. Under Assumptions 3.1, 3.2, and 4.1 the space $\mathcal{T}(\mathfrak{s}, d, M)$ is totally bounded (and therefore precompact) with respect to the uniform norm $\|\cdot\|_{\infty,d}$ with metric entropy satisfying*

$$H(\epsilon, \mathcal{T}(\mathfrak{s}, d, M), \|\cdot\|_{\infty,d}) \leq \sum_{k=1}^d c_k (\epsilon/2M)^{-d_k/\tilde{\sigma}_k}$$

for some positive constants $c_k, k \in [d]$ independent of ϵ and M.

This result relies on known metric entropy bounds for anisotropic smoothness classes.

Proposition B.4 (Prop. 2.2, Birgé (1986)). *Let* $\mathfrak{s} = (s_1, \dots, s_d) \in \mathbb{Z}_+^d$ *and* $\sigma = d\left(\sum_{j=1}^d s_j^{-1}\right)^{-1}$. *Assume that* Φ *is a family of functions* $\mathbb{R}^d \to \mathbb{R}$ *with common compact convex support of dimension* d *and satisfying*

$$\sup_{\phi \in \Phi, \alpha \preceq \mathfrak{s}} \|D^\alpha \phi\|_\infty < \infty.$$

The metric entropy of Φ *in the* L^∞ *norm is bounded as*

$$H(\epsilon, \Phi, \|\cdot\|_\infty) \lesssim \epsilon^{-d/\sigma}.$$

We now proceed to the proof of Proposition B.3.

Proof of Proposition B.3. For every $k \in [d]$, define the set of functions $\mathcal{X}_{k:d} \to \mathbb{R}$ given by

$$\mathcal{T}_k = \{S_k : S \in \mathcal{T}(\mathfrak{s}, d, M)\}.$$

By Definition 3.1, for each $k \in [d]$ we have that $\mathcal{T}_k/M$ satisfies the assumptions of Proposition B.4, and hence

$$n_k := H(\epsilon, \mathcal{T}_k, L^\infty) \le c_k(\epsilon/M)^{-d_k/\tilde{\sigma}_k},$$

for some $c_k > 0$ independent of ϵ and M.

Now note that $\mathcal{T}(\mathfrak{s}, d, M) \subseteq \prod_{k=1}^d \mathcal{T}_k$. For each $k \in [d]$, let $\{g_{k,1}, \dots, g_{k,n_k}\}$ be a minimal ϵ-cover of $\mathcal{T}_k$ with respect to the L^∞ norm, and define the subset

$$\mathcal{E} = \{f_{i_1,\dots,i_d} = (g_{1,i_i}, \dots, g_{d,i_d}) : 1 \le i_k \le n_k\} \subset \prod_{k=1}^d \mathcal{T}_k$$

which has cardinality $\prod_{k=1}^d n_k$. Now fix an arbitrary $f \in \prod_{k=1}^d \mathcal{T}_k$. For each $k \in [d]$, we can find some g_{k,i_k} such that $\|f_k - g_{k,i_k}\|_\infty \le \epsilon$. It follows that

$$\|f - f_{i_1,\dots,i_d}\|_{\infty,d} = \max_{k \in [d]} \|f_k - g_{k,i_k}\|_\infty \le \epsilon.$$

Hence, $\mathcal{E}$ is an ϵ-cover of $\prod_{k=1}^d \mathcal{T}_k$ and so

$$H\left(\epsilon, \prod_{k=1}^d \mathcal{T}_k, \|\cdot\|_{\infty,d}\right) \le \log\left(\prod_{k=1}^d n_k\right) = \sum_{k=1}^d n_k.$$

To conclude the proof, we claim that $\mathcal{T}(\mathfrak{s}, d, M) \subseteq \prod_{k=1}^d \mathcal{T}_k$ implies that

$$H\left(\epsilon, \mathcal{T}(\mathfrak{s}, d, M), \|\cdot\|_{\infty,d}\right) \le H\left(\epsilon/2, \prod_{k=1}^d \mathcal{T}_k, \|\cdot\|_{\infty,d}\right).$$

Indeed, suppose $\{f_1, \dots, f_m\} \subseteq \prod_{k=1}^d \mathcal{T}_k$ is a finite $(\epsilon/2)$-cover of $\prod_{k=1}^d \mathcal{T}_k$. For each $j = 1, \dots, m$, if $\mathcal{T}(\mathfrak{s}, d, M) \cap B_\infty(f_j, \epsilon/2)$ is non-empty, we define g_j to be an element of this intersection. Here $B_\infty(f_j, \epsilon/2)$ denotes the ball of radius $\epsilon/2$ in $\prod_{k=1}^d \mathcal{T}_k$ centered at f_j with respect to the norm $\|\cdot\|_{\infty,d}$. Now let $g \in \mathcal{T}(\mathfrak{s}, d, M)$ be arbitrary. Since $g \in \prod_{k=1}^d \mathcal{T}_k$, there is some $j \in \{1, \dots, m\}$ such that $\|g - f_j\|_{\infty,d} \le \epsilon/2$. This implies that g_j is defined and hence

$$\|g - g_j\|_{\infty,d} \le \|g - f_j\|_{\infty,d} + \|f_j - g_j\|_{\infty,d} \le \epsilon/2 + \epsilon/2 = \epsilon.$$

It follows that $\{g_j\}$ is a finite ϵ-cover of $\mathcal{T}(\mathfrak{s}, d, M)$ with respect to $\|\cdot\|_{\infty,d}$, which establishes the claim and completes the proof.

∎

B.4 Statistical consistency

B.4.1 Proof of Lemma 3.4

Proof of Lemma 3.4. Assume $d < 2s$. By Theorem 2.14.2 in van der Vaart and Wellner (1996), since functions in $\Psi(s,d,M)$ are by definition uniformly bounded there exist positive constants C_1, C_2 such that

$$\begin{aligned}\mathbb{E}\|P_n - P\|_{\Psi(s,d,M)} &\leq \frac{C_1 C_2}{\sqrt{n}} \int_0^1 \sqrt{1 + H(C_2\epsilon, \Psi(s,d,M), \|\cdot\|_\infty)} d\epsilon \\ &\lesssim \frac{1}{\sqrt{n}} \int_0^1 \sqrt{1 + \epsilon^{-d/s}} d\epsilon && \text{(Proposition 3.3)} \\ &\lesssim n^{-1/2}. && (d < 2s)\end{aligned}$$

The last line follows since $d < 2s$ implies that the integral on the right side is finite.

When $d \geq 2s$, the metric entropy integral above is no longer finite. In this case, we appeal to Dudley's metric entropy integral bound (Dudley, 1967) (see also Theorem 5.22 in Wainwright (2019)), which states that there exists positive constants $C_3, D > 0$ for which

$$\begin{aligned}\mathbb{E}\|P_n - P\|_{\Psi(s,d,M)} &\leq \min_{\delta \in [0,D]} \left\{ \delta + \frac{C_3}{\sqrt{n}} \int_\delta^D \sqrt{H(\epsilon, \Psi(s,d,M), \|\cdot\|_\infty)} d\epsilon \right\} \\ &\leq \min_{\delta \in [0,D]} \left\{ \delta + \frac{C_3\sqrt{c}}{\sqrt{n}} \int_\delta^D \epsilon^{-d/2s} d\epsilon \right\}. && \text{(Proposition 3.3)}\end{aligned}$$

First assume $d = 2s$. Evaluating the integral in Dudley's bound, we obtain

$$\mathbb{E}\|P_n - P\|_{\Psi(s,d,M)} \leq \min_{\delta \in [0,D]} \left\{ \delta + \frac{C_3\sqrt{c}}{\sqrt{n}} [\log D - \log \delta] \right\}.$$

To minimize the expression on the right side in δ, we differentiate with respect to δ and find where the derivative vanishes. The bound is optimized by choosing δ proportional to $n^{-1/2}$, which implies that $\mathbb{E}\|P_n - P\|_{\Psi(s,d,M)} \lesssim n^{-1/2} \log n$.

Now assume $d > 2s$. Evaluating the integral in Dudley's bound, we obtain

$$\begin{aligned}\mathbb{E}\|P_n - P\|_{\Psi(s,d,M)} &\leq \min_{\delta \in [0,D]} \left\{ \delta + \frac{C_3\sqrt{c}}{\sqrt{n}} \int_\delta^\infty \epsilon^{-d/2s} d\epsilon \right\} \\ &= \min_{\delta \in [0,D]} \left\{ \delta + \frac{C_4}{\sqrt{n}} \delta^{1-d/2s} \right\}.\end{aligned}$$

We optimize this bound by choosing δ proportional to $n^{-s/d}$, which implies that $\mathbb{E}\|P_n - P\|_{\Psi(s,d,M)} \lesssim n^{-s/d}$. $\square$

B.4.2 Proof of Theorem 3.6

Before continuing on to the proof of Theorem 3.6, we first prove that indeed $S^* \in \mathcal{T}(s,d,M^*)$ for $M^* > 0$ sufficiently large, and similarly for T^*.

Lemma B.5. *Let $\mathfrak{s} = (s_1, \dots, s_d) \in \mathbb{Z}_+^d$. Under Assumptions 3.1, 3.2, and 4.1, there exists some $M^* > 0$ such that the KR map S^* from f to g lies in $\mathcal{T}(\mathfrak{s}, d, M)$ for all $M \geq M^*$.*

By symmetry, if we switch the roles of f and g in Assumption 4.1, we see that the same is true for the KR map T^* from g to f. In particular, if $\mathfrak{s} = (s, \dots, s)$ for some $s \in \mathbb{N}$, then we are in the case of Assumption 3.3. It then follows from Lemma B.5 that $S^*, T^* \in \mathcal{T}(s,d,M^*)$ for $M^* > 0$ sufficiently large.

The proof of Lemma B.5 requires an auxiliary result establishing the regularity of marginal and conditional pdfs and quantile functions associated to a smooth density.

Lemma B.6. *Let h be a density on $\mathbb{R}^d$ with compact support $\mathcal{Z}$. Assume further that h is strictly positive on $\mathcal{Z}$ and s-smooth in the sense of Assumption 4.1, where $s = (s_1, \dots, s_d) \in \mathbb{Z}_+^d$. The following hold.*

1. *For any index set $J \subseteq [d]$, the marginal density $h(x_J) = \int_{\mathbb{R}^{d-|J|}} h(x_J, x_{-J}) dx_{-J}$ is s_J-smooth. Similarly, the conditional density $h(x_J|x_{-J}) = \frac{h(x_J, x_{-J})}{h(x_{-J})}$ is s-smooth as a function of $x = (x_J, x_{-J})$.*
2. *For any $J \subseteq [d], k \in J^c$, the univariate conditional cdf $H(x_k|x_J) = \int_{-\infty}^{x_k} h(y_k|x_J) dy_k$ is $(s_k + 1, s_J)$-smooth as a function of (x_k, x_J).*
3. *Assume further that $\mathcal{Z}$ is convex. Then the conditional quantile function $H^{-1}(p_k|x_J)$ is $(s_k + 1, s_J)$-smooth as a function of (p_k, x_J) for any $J \subseteq [d], k \in J^c, p_k \in [0, 1]$.*

Proof. For (1), the regularity assumptions on h imply that we can differentiate under the integral. For any multi-index $\alpha \preceq s$ satisfying $\alpha_{-J} = 0$ we have

$$D^\alpha h(x_J) = \int D^\alpha h(x_J, x_{-J}) dx_{-J}.$$

Since $D^\alpha h(x_J, x_{-J})$ is continuous and compactly supported by hypothesis, the dominated convergence theorem implies that for any sequence $x_{J,n} \to x_J$ we have

$$D^\alpha h(x_{J,n}) = \int D^\alpha h(x_{J,n}, x_{-J}) dx_{-J} \to \int D^\alpha h(x_J, x_{-J}) dx_{-J} = D^\alpha h(x_J).$$

This proves that $h(x_J)$ is s_J-smooth. Since $h > 0$ on its support and $x \mapsto 1/x$ is a smooth function for $x > 0$, this implies that $1/h(x_{-J})$ is s_{-J}-smooth. As products of differentiable functions are differentiable, we conclude that $h(x_J|x_{-J}) = h(x_J, x_{-J})/h(x_{-J})$ is s-smooth.

For (2), the result from (1) implies that the univariate conditional density $h(x_k|x_J)$ is (s_k, s_J)-smooth. For any multi-index $\alpha \preceq s$ satisfying $\alpha_k = 0$, we can repeat the differentiation-under-the-integral argument above, combined with the dominated convergence theorem, to see that $D^\alpha H(x_k|x_J)$ is continuous. If $\alpha_k \neq 0$, we apply the fundamental theorem of calculus to take care of one derivative in x_k and then recall that $h(x_k|x_J)$ is (s_k, s_J)-smooth to complete the proof, appealing to Clairaut's theorem to exchange the order of differentiation. Note that this shows that $H(x_k|x_J)$ is $(s_k + 1, s_J)$-smooth as a function of (x_k, x_J).

For (3), the assumption that $\mathcal{Z}$ is convex, combined with the fact that $h > 0$ on its support, implies that $H(x_k|x_J)$ is strictly increasing in x_k. As such, $H^{-1}(p_k|x_J)$ is defined, strictly increasing, and continuous in p_k. Since $H(x_k|x_J)$ is differentiable in x_k, we have

$$\frac{\partial}{\partial p_k} H^{-1}(p_k|x_J) = \frac{1}{\frac{\partial}{\partial x_k} H(H^{-1}(p_k|x_J)|x_J)} = \frac{1}{h(H^{-1}(p_k|x_J)|x_J)},$$

which is continuous in p_k as a composition of continuous functions. Continuing in this way, we see that $H^{-1}(p_k|x_J)$ is $(s_k + 1)$-smooth in p_k.

For the other variables, note that the following relation holds generally:

$$H(H^{-1}(p_k|x_J)|x_J) = p_k.$$

Defining the function

$$u(p_k, x_J, x_k) = H(x_k|x_J) - p_k,$$

we see that u is $(\infty, s_J, s_k + 1)$-smooth in its arguments (p_k, x_J, x_k) and satisfies

$$u(p_k, x_J, H^{-1}(p_k|x_J)) = 0.$$

Furthermore $\frac{\partial}{\partial x_k} u > 0$ by assumption, since $h > 0$ on its support. Hence we can appeal to the implicit function theorem (stated precisely below) to conclude that $H^{-1}(p_k|x_J)$ is $(s_k + 1, s_J)$-smooth in (p_k, x_J). For example, for any $j \in J$ we can evaluate the partial derivative in x_j by implicitly differentiating the above relation. Letting $x_k(p_k, x_J) = H^{-1}(p_k|x_J)$ we obtain

$$\begin{aligned}
\frac{\partial}{\partial x_j} H(x_k(p_k, x_J)|x_J) &= \frac{\partial H}{\partial x_k}(x_k(p_k, x_J)|x_J) \cdot \frac{\partial x_k}{\partial x_j}(p_k, x_J) + \frac{\partial H}{\partial x_j}(x_k|x_J) \\
&= \frac{\partial p_k}{\partial x_j} \\
&= 0.
\end{aligned}$$

Rearranging terms and noting that $\frac{\partial x_k}{\partial x_j}(p_k, x_J) = \frac{\partial}{\partial x_j} H^{-1}(p_k|x_J)$, we find that

$$\frac{\partial}{\partial x_j} H^{-1}(p_k|x_J) = -\frac{\frac{\partial H}{\partial x_j}(H^{-1}(p_k|x_J)|x_J)}{h(H^{-1}(p_k|x_J)|x_J)}.$$

□

Theorem B.7 (Implicit function theorem (Rudin, 1976)). *Let $u : \mathbb{R}^{n+m} \to \mathbb{R}^m$ be continuously differentiable and suppose $u(a, b) = 0$ for some $(a, b) \in \mathbb{R}^n \times \mathbb{R}^m$. For $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$ define the partial Jacobians*

$$J_n u(x, y) = \left[\frac{\partial}{\partial x_j} u_i(a, b)\right]_{(i,j)\in[m]\times[n]} \qquad J_m u(x, y) = \left[\frac{\partial}{\partial y_j} u_i(a, b)\right]_{(i,j)\in[m]\times[m]},$$

and suppose that $J_m u(a, b)$ is invertible. Then there exists an open neighborhood $V \subseteq \mathbb{R}^n$ of a such that there exists a unique continuously differentiable function $v : V \to \mathbb{R}^m$ satisfying $v(a) = b$, $u(x, v(x)) = 0$ for all $x \in V$, and

$$Jv(x) = -J_m u(x, v(x))^{-1} J_n u(x, v(x)).$$

We now proceed to the proof of Lemma B.5.

Proof of Lemma B.5. By definition of the KR map, for each $k \in [d]$ we have

$$S_k^*(x_k|x_{(k+1):d}) = G_k^{-1}(F_k(x_k|x_{(k+1):d})|S^*_{(k+1):d}(x)).$$

The requisite smoothness in Definition 3.1 of $S^*_{k:d}$ then follows from the chain rule of differentiation, appealing to Assumption 4.1, and an application of Lemma B.6. Note also that $x_k \mapsto G_k^{-1}(F_k(x_k|x_{(k+1):d})|S^*_{(k+1):d}(x))$ is strictly increasing as a composition of strictly increasing functions, since f, g are bounded away from 0 on their supports by assumption. Defining $\mathcal{T}(\mathfrak{s}, d) \subset \mathcal{T}$ as the subset of strictly increasing triangular maps that are $\mathfrak{s}$-smooth (although not necessarily with uniformly bounded derivatives), we see that $S^* \in \mathcal{T}(\mathfrak{s}, d)$. Since $\mathcal{T}(\mathfrak{s}, d) = \cup_{M>0}\mathcal{T}(\mathfrak{s}, d, M)$, there exists some $M^* > 0$ for which $S^* \in \mathcal{T}(\mathfrak{s}, d, M^*)$. Since $\mathcal{T}(\mathfrak{s}, d, M_1) \subseteq \mathcal{T}(\mathfrak{s}, d, M_2)$ for all $M_1 \leq M_2$ by definition, we conclude that $S^* \in \mathcal{T}(\mathfrak{s}, d, M)$ for all $M \geq M^*$. □

With Lemma B.5 in hand, we can now prove Theorem 3.6.

Proof of Theorem 3.6. The theorem is a direct result of inequality (4) and Lemma 3.4. Indeed, since $S^* \in \mathcal{T}(s, d, M^*)$ by Lemma B.5, we have $\inf_{S\in\mathcal{T}(s,d,M^*)} P\psi_S = P\psi_{S^*} = 0$, and therefore

$$\begin{aligned}
\mathbb{E}[P\psi_{S^n}] &= \mathbb{E}\left\{P\psi_{S^n} - \inf_{S\in\mathcal{T}(s,d,M^*)} P\psi_S\right\} \\
&\leq \mathbb{E}\left\{2\|P_n - P\|_{\Psi(s,d,M^*)} + R_n\right\} && \text{(inequality (4))} \\
&\lesssim \mathbb{E}\|P_n - P\|_{\Psi(s,d,M^*)}.
\end{aligned}$$

Appealing to Lemma 3.4 establishes the stated bound on $\mathbb{E}[P\psi_{S^n}]$. To conclude that $P\psi_{S^n} \xrightarrow{p} 0$, we apply Markov's inequality:

$$P(P\psi_{S^n} \geq \epsilon) \leq \epsilon^{-1}\mathbb{E}[P\psi_{S^n}] \to 0.$$

As $\epsilon > 0$ was arbitrary, this completes the proof. □

B.4.3 Proof of KL lower semicontinuity

Although Theorem 3.6 only establishes a weak form of consistency in terms of the KL divergence, we leverage this result to prove strong consistency, in the sense of uniform convergence of S^n to S^* in probability, in Theorem 3.8. The proof requires understanding the regularity of the KL divergence with respect to the topology induced by the $\|\cdot\|_{\infty,d}$ norm. Lemma B.8 establishes that KL is lower semicontinuous with respect to this topology. It relies on the weak lower semicontinuity of KL proved by Donsker and Varadhan using their dual representation in Lemma 2.1 of Donsker and Varadhan (1975).

Lemma B.8. *Under Assumptions 3.1-3.3, the functional $S \mapsto P\psi_S$ on the domain $\mathcal{T}$ is lower semicontinuous with respect to the uniform norm $\|\cdot\|_{\infty,d}$.*

Proof. Assume $\|S^n - S\|_{\infty,d} \to 0$ for fixed maps $S^n, S \in \mathcal{T}$. We first claim that $\nu_n := S^n \# \mu \rightsquigarrow S\#\mu =: \nu$ for any probability measure μ. We will show that $\int h\ d\nu_n \to \int h\ d\nu$ for all bounded continuous functions $h \in C_b(\mathbb{R}^d)$. Note that $\int h\ d\nu_n = \int h \circ S^n\ d\mu$. By assumption, $h \circ S^n$ is bounded and measurable with $\|h \circ S^n\|_\infty \leq \|h\|_\infty$ for all n. Furthermore, $h \circ S^n \to h \circ S$ pointwise, since h is continuous. Then by the dominated convergence theorem, $\int |h \circ S^n - h \circ S|\ d\mu \to 0$. In particular, this implies that $\int h\ d\nu_n = \int h \circ S^n\ d\mu \to \int h \circ S\ d\mu = \int h\ d\nu$. This proves the claim.

Combining this fact with the weak lower semicontinuity of KL divergence (Donsker and Varadhan, 1975) proves the lemma. Indeed, since $\|S^n - S\|_{\infty,d} \to 0$ implies that $S^n\#\mu \rightsquigarrow S\#\mu$ for any μ, it follows that

$$\begin{aligned} P\psi_S &= \mathrm{KL}(S\#f|g) \\ &\leq \liminf_{n\to\infty} \mathrm{KL}(S^n\#f|g) && \text{(KL is weakly lower semicontinuous)} \\ &= \liminf_{n\to\infty} P\psi_{S^n}. \end{aligned}$$

Thus, $S \mapsto P\psi_S$ on $\mathcal{T}$ is lower semicontinuous with respect to $\|\cdot\|_{\infty,d}$. ☐

As a corollary, we also obtain an existence result for minimizers of $S \mapsto P\psi_S$ on $\overline{\mathcal{T}(s,d,M)}$.

Corollary B.9. *Under Assumptions 3.1-3.3, for any $M > 0$ the minimum $\inf_{S\in\overline{\mathcal{T}(s,d,M)}} P\psi_S$ is attained.*

Proof. This follows from the direct method of calculus of variations, since $\overline{\mathcal{T}(s,d,M)}$ is compact with respect to $\|\cdot\|_{\infty,d}$ (Proposition B.3) and $S \mapsto P\psi_S$ is bounded below and lower semicontinuous with respect to $\|\cdot\|_{\infty,d}$. ☐

B.4.4 Proof of Theorem 3.8

Proof. Recall that $\overline{\mathcal{T}(s,d,M^*)}$ is compact with respect to $\|\cdot\|_{\infty,d}$. Together, lower semicontinuity of KL in $\|\cdot\|_{\infty,d}$ and compactness guarantee that the KR map S^*, which is the unique minimizer of $S \mapsto P\psi_S$ over $\overline{\mathcal{T}(s,d,M^*)}$, is well-separated in $\mathcal{T}(s,d,M^*)$. In other words, for any $\epsilon > 0$,

$$P\psi_{S^*} < \inf_{S\in\mathcal{T}(s,d,M^*):\|S-S^*\|_{\infty,d}\geq\epsilon} P\psi_S.$$

Indeed, suppose to the contrary that we can find a deterministic sequence $\tilde{S}^n \in \mathcal{T}(s,d,M^*)$ satisfying $\|\tilde{S}^n - S^*\|_{\infty,d} \geq \epsilon$ such that $P\psi_{\tilde{S}^n} \to P\psi_{S^*}$. Since $\mathcal{T}(s,d,M^*)$ is precompact with respect to $\|\cdot\|_{\infty,d}$, we can extract a subsequence $\tilde{S}^{n_k}$ converging to some $\tilde{S}^* \in \overline{\mathcal{T}(s,d,M^*)}$, which necessarily satisfies $\|\tilde{S}^* - S^*\|_{\infty,d} \geq \epsilon$. By lower semicontinuity of $S \mapsto P\psi_S$ with respect to $\|\cdot\|_{\infty,d}$, it follows that

$$P\psi_{\tilde{S}^*} \leq \liminf_{k\to\infty} P\psi_{\tilde{S}^{n_k}} = P\psi_{S^*}.$$

We have a contradiction, since the KR map S^* is the unique minimizer of $S \mapsto P\psi_S$. This proves the claim that S^* is a well-separated minimizer.

Now fix $\epsilon > 0$ and define

$$\delta = \inf_{S\in\mathcal{T}(s,d,M^*):\|S-S^*\|_{\infty,d}\geq\epsilon} P\psi_S - P\psi_{S^*} > 0.$$

It follows that

$$\{\|S^n - S^*\|_{\infty,d} \geq \epsilon\} \subseteq \{P\psi_{S^n} - P\psi_{S^*} \geq \delta\}.$$

We have shown $P\psi_{S^n} \xrightarrow{p} P\psi_{S^*}$ in Theorem 3.6. As a consequence,

$$P(\|S^n - S^*\|_{\infty,d} \geq \epsilon) \leq P(P\psi_{S^n} - P\psi_{S^*} \geq \delta) \to 0.$$

As $\epsilon > 0$ was arbitrary, we have $\|S^n - S^*\|_{\infty,d} \xrightarrow{p} 0$. ☐

B.4.5 Uniform consistency of the inverse map

We have proved consistency of the estimator S^n of the sampling map S^*, which pushes forward the target density f to the source density g. However, we require knowledge of the direct map $T^* = (S^*)^{-1}$ to generate new samples from f by pushing forward samples from g under T^*. In this section we prove consistency and a rate of convergence of the estimator $T^n = (S^n)^{-1}$ of the direct map T^*.

First note that $\mathrm{KL}(S\#f|g) = \mathrm{KL}(f|S^{-1}\#g)$ for any diffeomorphism $S : \mathbb{R}^d \to \mathbb{R}^d$, as we proved in Section B.1. As such, the consistency and rate of convergence of $\mathrm{KL}(S^n\#f|g)$ obtained in Theorem 3.6 yield the same results for the estimator T^n in terms of $\mathrm{KL}(f|T^n\#g)$ under identical assumptions. We can also establish uniform consistency of T^n as we did for S^n in Theorem 3.8, although the proof of this fact requires a bit more work.

Theorem B.10. *Suppose Assumptions 3.1-3.3 hold. Let S^n be any near-optimizer of the functional $S \mapsto P_n\psi_S$ on $\mathcal{T}(s, d, M^*)$, i.e., suppose*

$$P_n\psi_{S^n} = \inf_{S\in\mathcal{T}(s,d,M^*)} P_n\psi_S + o_P(1).$$

Let $T^n = (S^n)^{-1}$. Then $\|T^n - T^*\|_{\infty,d} \xrightarrow{p} 0$, i.e, T^n is a uniformly consistent estimator of T^*.

The proof of Theorem B.10 relies heavily on the bounds on S^n and its derivatives posited in Definition 3.1, which we utilize in conjunction with the inverse function theorem to uniformly bound the Jacobian $JT^n(y) = (JS^n(T^n(y)))^{-1}$ over $y \in \mathcal{Y}$ and $n \in \mathbb{N}$, thereby establishing uniform equicontinuity of the family of estimators $\{T^n\}_{n=1}^\infty$. We combine this uniform equicontinuity with the uniform consistency of S^n from Theorem 3.8 to complete the proof.

First we establish a lemma that allows us to bound the derivatives of the inverse map estimates T^n.

Lemma B.11. *Suppose $A \in \mathbb{R}^{d\times d}$ is an invertible upper triangular matrix satisfying*

$$\max_{i,j\in[d],i<j} |A_{ij}| \le L \qquad \textit{and} \qquad \min_{j\in[d]} |A_{jj}| \ge 1/M$$

for some positive $L, M > 0$. Then A^{-1} is upper triangular and the diagonal entries are bounded as

$$\max_{j\in[d]} |A^{-1}_{jj}| \le M.$$

Furthermore, the superdiagonal terms $i, j \in [d]$ with $i < j$ are bounded as

$$|A^{-1}_{ij}| \le M^2L(ML + 1)^{j-i-1}.$$

Proof. Let $D = \mathrm{diag}(A)$ denote the matrix with diagonal entries $D_{jj} = A_{jj}$ for $j \in [d]$ and zeros elsewhere. Note that D is invertible since A is, and D^{-1} is diagonal with entries $D^{-1}_{jj} = 1/A_{jj}$ for $j \in [d]$. Let $U = A - D$ denote the strictly upper triangular part of A. Now note that

$$A^{-1} = (D + U)^{-1} = [D(I + D^{-1}U)]^{-1} = (I + D^{-1}U)^{-1}D^{-1}.$$

To calculate $(I + D^{-1}U)^{-1}$ we make use of the matrix identity

$$(I - X)\left[\sum_{k=0}^{d-1} X^k\right] = I - X^d.$$

We plug in $X = -D^{-1}U$ and note that $D^{-1}U$ is strictly upper triangular, which implies that it is nilpotent of degree at most d, i.e., $(D^{-1}U)^d = 0$. Hence, we obtain

$$(I + D^{-1}U)\left[\sum_{k=0}^{d-1}(-D^{-1}U)^k\right] = I - (-D^{-1}U)^d = I,$$

which implies that $(I + D^{-1}U)^{-1} = \sum_{k=0}^{d-1}(-D^{-1}U)^k$. It follows that

$$A^{-1} = \sum_{k=0}^{d-1}(-D^{-1}U)^kD^{-1}.$$

This shows that A^{-1} is upper triangular as a sum of products of upper triangular matrices. Note also that the terms in the sum with $k > 0$ are strictly upper triangular. Hence, we see that $\mathrm{diag}(A^{-1}) = D^{-1}$ and therefore the diagonal entries of A^{-1} satisfy $|A^{-1}_{jj}| = \frac{1}{|A_{jj}|} \leq M$ by hypothesis.

Now we bound the superdiagonal entries. By repeated application of the triangle inequality and appealing to the bounds on A_{ij}, we can bound A^{-1}_{ij} for $i < j$ as

$$\begin{aligned}
|A^{-1}_{ij}| &= \left|\sum_{k=0}^{d-1}[(D^{-1}U)^k D^{-1}]_{ij}\right| \\
&= \left|\sum_{k=1}^{d-1}[(D^{-1}U)^k D^{-1}]_{ij}\right| && (i < j) \\
&\leq \sum_{k=1}^{d-1} |[(D^{-1}U)^k D^{-1}]_{ij}| \\
&\leq \sum_{k=1}^{d-1} |[(MU)^k M]_{ij}| && (|D^{-1}_{jj}| = |A^{-1}_{jj}| \leq M) \\
&= \sum_{k=1}^{d-1} M^{k+1} |(U^k)_{ij}|.
\end{aligned}$$

Now let V denote the $d \times d$ matrix with ones strictly above the diagonal and zeros elsewhere, i.e.,

$$V_{ij} = \begin{cases} 1, & i < j, \\ 0, & i \geq j. \end{cases}$$

We now claim that for $k \geq 1$,

$$(V^k)_{ij} = \begin{cases} \binom{j-i-1}{k-1}, & k \leq j - i \\ 0, & \text{else.} \end{cases}$$

Our proof of the claim proceeds by induction. The statement is clearly true for the base case $k = 1$ by definition of V. Suppose the claim holds up to k. It follows that

$$\begin{aligned}
(V^{k+1})_{ij} &= (V^k \cdot V)_{ij} \\
&= \sum_{\ell=1}^{d} (V^k)_{i\ell} V_{\ell j} \\
&= \sum_{\ell=1}^{d} \binom{\ell-i-1}{k-1} 1_{[k \le \ell - i]} \cdot 1_{[\ell < j]} && \text{(inductive step)} \\
&= 1_{[k+1 \le j-i]} \sum_{\ell=i+k}^{j-1} \binom{\ell-i-1}{k-1} \\
&= 1_{[k+1 \le j-i]} \sum_{p=k-1}^{j-i-2} \binom{p}{k-1} && (p := \ell - i - 1) \\
&= 1_{[k+1 \le j-i]} \left\{ 1 + \sum_{p=k}^{j-i-2} \binom{p}{k-1} \right\} \\
&= 1_{[k+1 \le j-i]} \left\{ 1 + \sum_{p=k}^{j-i-2} \left[\binom{p+1}{k} - \binom{p}{k} \right] \right\} && \text{(Pascal's formula)} \\
&= 1_{[k+1 \le j-i]} \left\{ 1 + \binom{j-i-1}{k} - \binom{k}{k} \right\} && \text{(telescoping sum)} \\
&= 1_{k+1 \le j-i} \binom{j-i-1}{k}.
\end{aligned}$$

This proves the claim for $k+1$. The result then follows by induction.

Note now that $|U_{ij}| \le LV_{ij}$ for all $i, j \in [d]$ by hypothesis. Again applying the triangle inequality and the assumption $\max_{i<j} |A_{ij}| \le L$, it follows that

$$\begin{aligned}
|A_{ij}^{-1}| &\le \sum_{k=1}^{d-1} M^{k+1} |(U^k)_{ij}| \\
&\le \sum_{k=1}^{d-1} M^{k+1} |((LV)^k)_{ij}| \\
&= \sum_{k=1}^{d-1} M^{k+1} L^k \binom{j-i-1}{k-1} 1_{[k \le j-i]} \\
&= \sum_{k=1}^{j-i} M^{k+1} L^k \binom{j-i-1}{k-1} \\
&= M^2 L \sum_{\ell=0}^{j-i-1} (ML)^{\ell} 1^{j-i-1-\ell} \binom{j-i-1}{\ell} && (\ell := k-1) \\
&= M^2 L (ML+1)^{j-i-1}. && \text{(binomial formula)}
\end{aligned}$$

This completes the proof. □

We now use this lemma to establish strong consistency of the direct map estimator.

Proof of Theorem B.10. First note that convexity of $\mathcal{X}$ along with positivity $f > 0$ on $\mathcal{X}$ implies that the inverse KR map $T^* = (S^*)^{-1}$ is well-defined and continuous, as noted in Lemma B.5. Furthermore, by definition of $\mathcal{T}(s, d, M^*)$, the inverse maps $T^n = (S^n)^{-1}$ exist and are continuous also.

We first prove that the sequence $\{T^n\}_{n=1}^{\infty}$ is uniformly equicontinuous with respect to the ℓ_∞ norm $\|\cdot\|_\infty$ on $\mathbb{R}^d$. For a function $S:\mathcal{X}\to\mathcal{Y}$ let $JS(x)$ denote the Jacobian matrix at $x\in\mathcal{X}$. For a matrix $A:\mathbb{R}^d\to\mathbb{R}^d$ let $\|A\|_p$ denote the operator norm induced by the ℓ_p norm on $\mathbb{R}^d$, i.e.,

$$\|A\|_p = \sup\left\{\|Ax\|_p : x\in\mathbb{R}^d, \|x\|_p = 1\right\}.$$

When $p=\infty$, this norm is simply the maximum absolute row sum of the matrix:

$$\|A\|_\infty = \max_{1\le i\le d}\sum_{j=1}^{d}|A_{ij}|.$$

We aim to bound $\|JT^n(y)\|_\infty$ uniformly in $y\in\mathcal{Y}, n\in\mathbb{N}$. Since $\{S^n\}\subset\mathcal{T}(s,d,M^*)$, the Jacobian matrix $JS^n(x)$ is upper triangular for every $x\in\mathcal{X}$ and

$$\max_{i<j}\sup_{x\in\mathcal{X}}|[JS^n(x)]_{ij}| = \max_{i<j}\|D_jS_i^n\|_\infty \le M^*.$$

Similarly, we also have that

$$\min_{j\in[d]}\inf_{x\in\mathcal{X}}|[JS^n(x)]_{jj}| = \min_{j\in[d]}\inf_{x\in\mathcal{X}}|D_jS_j^n(x)| \ge 1/M^*.$$

It follows that for each $x\in\mathcal{X}$, the Jacobian $JS^n(x)$ satisfies the hypotheses of Lemma B.11 with $L=M^*$. Applying the inverse function theorem, we conclude from the lemma that the entries of $JT^n(y) = (JS^n(T^n(y)))^{-1}$ are bounded as

$$\begin{aligned}
&\max_{i<j}\sup_{y\in\mathcal{Y}}|[JT^n(y)]_{ij}| \le (M^*)^3((M^*)^2+1)^{j-i-1},\\
&\max_{j\in[d]}\sup_{y\in\mathcal{Y}}|[JT^n(y)]_{jj}| \le M^*.
\end{aligned}$$

It follows that the ℓ_∞ operator norm is bounded as

$$\begin{aligned}
\sup_{y\in\mathcal{Y}}\|JT^n(y)\|_\infty &\le M^* + \sum_{j=2}^{d}(M^*)^3((M^*)^2+1)^{j-2}\\
&= \begin{cases} M^* + (M^*)^3\cdot\frac{1-((M^*)^2+1)^{d-2}}{1-((M^*)^2+1)}, & d\ge 2,\\ M^*, & d=1.\end{cases} \qquad \text{(partial geometric series)}\\
&= \begin{cases} M^*((M^*)^2+1)^{d-2}, & d\ge 2,\\ M^*, & d=1.\end{cases}\\
&= M^*\cdot\max\{1,((M^*)^2+1)^{d-2}\}\\
&:= C(d,M^*).
\end{aligned}$$

Here we have used the convention that $\sum_{j=2}^{1}a_j = 0$ for any sequence $\{a_j\}$. Now we apply the mean value inequality for vector-valued functions to deduce that the $\{T^n\}$ are uniformly equicontinuous. Indeed, for any $y_1,y_2\in\mathcal{Y}$, we have

$$\begin{aligned}
\|T^n(y_1)-T^n(y_2)\|_\infty &\le \sup_{y\in\mathcal{Y}}\|JT^n(y)\|_\infty\|y_1-y_2\|_\infty\\
&\le C(d,M^*)\|y_1-y_2\|_\infty.
\end{aligned}$$

Now let $y\in\mathcal{Y}$ and note that $y=S^*(x)$ for some $x\in\mathcal{X}$. We then have

$$\begin{aligned}
\|T^n(y)-T^*(y)\|_\infty &= \|T^n(y)-T^*(S^*(x))\|_\infty\\
&= \|T^n(y)-x\|_\infty\\
&= \|T^n(S^*(x))-T^n(S^n(x))\|_\infty\\
&\le C(d,M^*)\|S^*(x)-S^n(x)\|_\infty\\
&\le C(d,M^*)\|S^*-S^n\|_{\infty,d}.
\end{aligned}$$

Since $y \in \mathcal{Y}$ was arbitrary, we can take the supremum over y on the left side to obtain the desired result:

$$\|T^n - T^*\|_{\infty,d} \leq C(d, M^*)\|S^* - S^n\|_{\infty,d} \xrightarrow{p} 0. \qquad \text{(Theorem 3.8)}$$

$\square$

B.5 Sobolev rates under log-concavity

B.5.1 Proof of Theorem 4.1

Suppose the source density g is log-concave, which implies that $S \mapsto P\psi_S$ and $S \mapsto P_n\psi_S$ are strictly convex functionals. Since $\mathcal{T}(s, d, M)$ is convex,

$$\min_{S \in \mathcal{T}(s,d,M)} P\psi_S, \qquad \min_{S \in \mathcal{T}(s,d,M)} P_n\psi_S$$

are convex optimization problems. If in addition g is strongly log-concave, we obtain strong convexity of the objective, as we show in Lemma B.12.

Lemma B.12. *Suppose Assumptions 3.1-3.3 hold. Assume further that the source density g is m-strongly log-concave for some $m > 0$:*

$$[\nabla \log g(y_1) - \nabla \log g(y_2)]^T (y_1 - y_2) \leq m\|y_1 - y_2\|_2^2 \qquad \forall y_1, y_2 \in \mathcal{Y}.$$

Then the map $S \mapsto P\psi_S$ on $\mathcal{T}(s, d, M)$ is $\min\{m, M^{-2}\}$-strongly convex with respect to the L^2 Sobolev-type norm

$$\|S\|^2_{H_f^{1,2}(\mathcal{X})} := \|S\|^2_{L_f^2(\mathcal{X})} + \sum_{k=1}^{d} \|D_k S_k\|^2_{L_f^2(\mathcal{X})}.$$

Proof. We first calculate the Gâteaux derivative of $S \mapsto P\psi_S$ in the direction $A \in \mathcal{T}(s, d, M)$.

$$\begin{aligned}
\nabla P\psi_S(A) &= \lim_{t\to 0} \frac{P\psi_{S+tA} - P\psi_S}{t} \\
&= \lim_{t\to 0} -t^{-1}\mathbb{E}\Bigg\{ [\log g((S + tA)(X)) - \log g(S(X))] \\
&\qquad\qquad + \sum_{k=1}^{d} [\log D_k(S_k + tA_k)(X) - \log D_k S_k(X)] \Bigg\} \\
&= -\mathbb{E}\left\{ \nabla \log g(S(X))^T A(X) + \sum_k \frac{D_k A_k(X)}{D_k S_k(X)} \right\}.
\end{aligned}$$

We can differentiate under the integral by the dominated convergence theorem, since the integrand is smooth and compactly supported by Assumptions 3.1-3.3.

Now note that $\nabla P\psi_S(A)$ is a bounded linear operator in A. Furthermore, since the KR map S^* is the global minimizer of $S \mapsto P\psi_S$, we have $\nabla P\psi_{S^*}(A) = 0$ for all $A \in \mathcal{T}(s, d, M)$ satisfying $S^* + tA \in \mathcal{T}(s, d, M)$ for all t sufficiently small. We now check the strong convexity condition. Assume $A, B \in \mathcal{T}(s, d, M)$ for some $M > 0$. We have

$$\begin{aligned}
(\nabla P\psi_A - \nabla P\psi_B)(A-B) &= \nabla P\psi_A(A-B) - \nabla P\psi_B(A-B) \\
&= \mathbb{E}\{[\nabla \log g(B(X)) - \nabla \log g(A(X))]^T(A(X)-B(X)) \\
&\quad + \sum_k \left[\frac{D_k(A_k-B_k)(X)}{D_kB_k(X)} - \frac{D_k(A_k-B_k)(X)}{D_kA_k(X)}\right]\} \\
&\geq \mathbb{E}\left\{m\|A(X)-B(X)\|_2^2 + \sum_k \left[\frac{D_k(A_k-B_k)(X)}{D_kB_k(X)} - \frac{D_k(A_k-B_k)(X)}{D_kA_k(X)}\right]\right\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(strong log-concavity of } g\text{)} \\
&= \mathbb{E}\left\{m\|A(X)-B(X)\|_2^2 + \sum_k \frac{[D_k(A_k-B_k)(X)]^2}{D_kA_k(X)D_kB_k(X)}\right\} \\
&\geq \mathbb{E}\left\{m\|A(X)-B(X)\|_2^2 + \frac{1}{M^2}\sum_k [D_k(A_k-B_k)(X)]^2\right\} \qquad (A,B\in\mathcal{T}(s,d,M)) \\
&= m\|A-B\|_{L^2(f)}^2 + \frac{1}{M^2}\sum_k \|D_k(A_k-B_k)\|_{L^2(f)}^2 \\
&\geq \min\{m, M^{-2}\}\|A-B\|_{H^1(f)}^2.
\end{aligned}$$

Hence, $S\mapsto P\psi_S$ satisfies the first-order strong convexity condition. □

We now proceed to prove Theorem 4.1.

Proof of Theorem 4.1. By Lemma B.12, strong convexity of $S\mapsto P\psi_S$ with respect to $\|\cdot\|_{H_f^{1,2}(\mathcal{X})}$ implies that

$$\begin{aligned}
P\psi_{S^n} &= P\psi_{S^n} - P\psi_{S^*} \\
&\geq \nabla P\psi_{S^*}(S^n - S^*) + \frac{\min\{m,(M^*)^{-2}\}}{2}\|S^n - S^*\|_{H_f^{1,2}(\mathcal{X})}^2 \\
&= \frac{\min\{m,(M^*)^{-2}\}}{2}\|S^n - S^*\|_{H_f^{1,2}(\mathcal{X})}^2,
\end{aligned}$$

since $\nabla P\psi_{S^*}(S^n - S^*) = 0$, as S^* minimizes $P\psi_S$. We complete the proof by appealing to the bound on $\mathbb{E}[P\psi_{S^n}]$ established in Theorem 3.6. □

B.5.2 Sobolev rates for the inverse map

Now we prove a rate of convergence of the inverse map estimator in the L^2 Sobolev norm $\|\cdot\|_{H_g^{1,2}(\mathcal{Y})}$ assuming strong log-concavity, as in Theorem 4.1.

Theorem B.13. *Suppose Assumptions 3.1-3.3 hold. Assume further that g is m-strongly log-concave. Let S^n be a near-optimizer of the functional $S\mapsto P_n\psi_S$ on $\mathcal{T}(s,d,M^*)$ with remainder R_n satisfying*

$$\mathbb{E}[R_n] = \mathbb{E}\left\{P_n\psi_{S^n} - \inf_{S\in\mathcal{T}(s,d,M^*)} P_n\psi_S\right\} \lesssim \mathbb{E}\|P_n - P\|_{\Psi(s,d,M^*)}.$$

Then $T^n = (S^n)^{-1}$ converges to $T^* = (S^*)^{-1}$ with respect to the norm $\|\cdot\|_{H_g^{1,2}(\mathcal{Y})}$ with rate

$$\frac{\min\{m,(M^*)^{-2}\}}{2}\mathbb{E}\|T^n - T^*\|_{H^1(g)}^2 \lesssim \begin{cases} n^{-1/2}, & d<2s, \\ n^{-1/2}\log n, & d=2s, \\ n^{-s/d}, & d>2s. \end{cases}$$

Proof. We aim to bound $\|T^n - T^*\|_{H_g^{1,2}(\mathcal{Y})}$ by a multiple of $\|S^n - S^*\|_{H_f^{1,2}(\mathcal{X})}$, which will establish the same rate of convergence for the inverse map (up to constant factors) as derived for S^n in Theorem 4.1.

In the proof of Theorem B.10 we showed that we can bound the ℓ_∞ matrix norm of the Jacobian $JT^n(y)$ uniformly as

$$\sup_{y\in\mathcal{Y}} \|JT^n(y)\|_\infty \leq C(d, M^*) := M^* \cdot \max\{1, ((M^*)^2 + 1)^{d-2}\}.$$

Since $JT^n(y)$ is upper triangular, we can use the exact same argument to arrive at the same bound on the ℓ_1 matrix norm, which equals the maximum absolute column sum of the matrix,

$$\|A\|_1 = \max_{1\leq j\leq d} \sum_{i=1}^d |A_{ij}|.$$

Hence we conclude that $\sup_{y\in\mathcal{Y}} \|JT^n(y)\|_1 \leq C(d, M^*)$. Now we apply Hölder's inequality for matrix norms to conclude that

$$\sup_{y\in\mathcal{Y}} \|JT^n(y)\|_2 \leq \sup_{y\in\mathcal{Y}} \sqrt{\|JT^n(y)\|_1 \|JT^n(y)\|_\infty} \leq C(d, M^*).$$

Consequently, we can bound the first term in $\|\cdot\|_{H_g^{1,2}(\mathcal{Y})}$ as

$$\begin{aligned}
\|T^n - T^*\|^2_{L_g^2(\mathcal{Y})} &= \int \|T^n(y) - T^*(y)\|_2^2 \; g(y)dy \\
&= \int \|T^n(S^*(x)) - T^*(S^*(x))\|_2^2 \; |\det JS^*(x)| g(S^*(x))dx && \text{(Change of variables)} \\
&= \int \|T^n(S^*(x)) - T^*(S^*(x))\|_2^2 \; f(x)dx && (g = T_\# f) \\
&= \int \|T^n(S^*(x)) - T^n(S^n(x))\|_2^2 \; f(x)dx \\
&\leq \int \sup_{y\in\mathcal{Y}} \|JT^n(y)\|_2^2 \|S^*(x) - S^n(x)\|_2^2 \; f(x)dx \\
&\leq C(d, M^*)^2 \int \|S^*(x) - S^n(x)\|_2^2 \; f(x)dx \\
&= C(d, M^*)^2 \|S^* - S^n\|^2_{L_f^2(\mathcal{X})}.
\end{aligned}$$

To bound the deviations of the first derivatives, note that

$$\begin{aligned}
|D_k T_k^n(y) - D_k T_k^*(y)| &= \left| \frac{1}{D_k S_k^n(T^n(y))} - \frac{1}{D_k S_k^*(T^*(y))} \right| \\
&= \left| \frac{D_k S_k^n(T^n(y)) - D_k S_k^*(T^*(y))}{[D_k S_k^n(T^n(y))][D_k S_k^*(T^*(y))]} \right| \\
&\leq (M^*)^2 \left| D_k S_k^n(T^n(y)) - D_k S_k^*(T^*(y)) \right| && (S, S^n \in \mathcal{T}(s, d, M^*)) \\
&\leq (M^*)^2 \left| D_k S_k^n(T^n(y)) - D_k S_k^n(T^*(y)) \right| \\
&\quad + (M^*)^2 \left| D_k S_k^n(T^*(y)) - D_k S_k^*(T^*(y)) \right| \\
&\leq (M^*)^2 \sup_{x\in\mathcal{X}} \|\nabla(D_k S_k^n)(x)\|_2 \|T^n(y) - T^*(y)\|_2 \\
&\quad + (M^*)^2 \left| D_k S_k^n(T^*(y)) - D_k S_k^*(T^*(y)) \right|. && \text{(mean value inequality)}
\end{aligned}$$

Now note that S_k^n depends only on $(x_k, \ldots, x_d)$, and therefore, since $S^n \in \mathcal{T}(s, d, M^*)$ implies that

$\max_{j\in[d]} \|D_j D_k S_k^n\|_\infty \le M^*$, we have

$$\begin{aligned}
\sup_{x\in\mathcal{X}} \|\nabla(D_k S_k^n)(x)\|_2 &= \sup_{x\in\mathcal{X}} \sqrt{\sum_{j=1}^d |D_j D_k S_k^n(x)|^2} \\
&= \sup_{x\in\mathcal{X}} \sqrt{\sum_{j=k}^d |D_j D_k S_k^n(x)|^2} \\
&\le \sqrt{\sum_{j=k}^d (M^*)^2} \\
&= M^*\sqrt{d-k+1}.
\end{aligned}$$

Hence, we conclude that

$$\begin{aligned}
|D_k T_k^n(y) - D_k T_k^*(y)|^2 &\le \{(M^*)^3\sqrt{d-k+1}\|T^n(y) - T^*(y)\|_2 \\
&\quad + (M^*)^2 \left|D_k S_k^n(T^*(y)) - D_k S_k^*(T^*(y))\right|\}^2 \\
&\le 2(M^*)^6(d-k+1)\|T^n(y) - T^*(y)\|_2^2 \\
&\quad + 2(M^*)^4 \left|D_k S_k^n(T^*(y)) - D_k S_k^*(T^*(y))\right|^2. \qquad ((a+b)^2 \le 2(a^2+b^2))
\end{aligned}$$

Summing over k and integrating against the density g, we obtain

$$\begin{aligned}
\sum_{k=1}^d \|D_k T_k^n - D_k T_k^*\|_{L_g^2(\mathcal{Y})}^2 &= \sum_{k=1}^d \int |D_k T_k^n(y) - D_k T_k^*(y)|^2 g(y)dy \\
&\le \sum_{k=1}^d \int 2(M^*)^6(d-k+1)\|T^n(y) - T^*(y)\|_2^2 \, g(y)dy \\
&\quad + \sum_{k=1}^d \int 2(M^*)^4 \left|D_k S_k^n(T^*(y)) - D_k S_k^*(T^*(y))\right|^2 g(y)dy \\
&= (M^*)^6 d(d+1)\|T^n - T^*\|_{L_g^2(\mathcal{Y})}^2 \\
&\quad + 2(M^*)^4 \sum_{k=1}^d \int \left|D_k S_k^n(x) - D_k S_k^*(x)\right|^2 g(S^*(x))|\det JS^*(x)|dx \\
&= (M^*)^6 d(d+1)\|T^n - T^*\|_{L_g^2(\mathcal{Y})}^2 \\
&\quad + 2(M^*)^4 \sum_{k=1}^d \int \left|D_k S_k^n(x) - D_k S_k^*(x)\right|^2 f(x)dx \\
&= (M^*)^6 d(d+1)\|T^n - T^*\|_{L_g^2(\mathcal{Y})}^2 + 2(M^*)^4 \sum_{k=1}^d \|D_k(S_k^n - S_k^*)\|_{L_f^2(\mathcal{X})}^2 \\
&\le (M^*)^6 d(d+1)C(d,M^*)^2\|S^n - S^*\|_{L_f^2(\mathcal{X})}^2 + 2(M^*)^4 \sum_{k=1}^d \|D_k(S_k^n - S_k^*)\|_{L_f^2(\mathcal{X})}^2
\end{aligned}$$

Putting all of these calculations together, we have shown that

$$\begin{aligned}\|T^n - T^*\|^2_{H^{1,2}_g(\mathcal{Y})} &= \|T^n - T^*\|^2_{L^2_g(\mathcal{Y})} + \sum_{k=1}^d \|D_k(T^n_k - T^*_k)\|^2_{L^2_g(\mathcal{Y})} \\ &\leq C(d, M^*)^2\|S^n - S^*\|^2_{L^2_f(\mathcal{X})} + (M^*)^6 d(d+1)C(d, M^*)^2\|S^n - S^*\|^2_{L^2_f(\mathcal{X})} \\ &\quad + 2(M^*)^4 \sum_{k=1}^d \|D_k(S^n_k - S^*_k)\|^2_{L^2_f(\mathcal{X})} \\ &\leq \tilde{C}(d, M^*)^2\|S^n - S^*\|^2_{H^{1,2}_f(\mathcal{X})},\end{aligned}$$

where we define

$$\tilde{C}(d, M)^2 = \max\{C(d, M)^2[1 + M^6 d(d+1)], 2M^4\}.$$

Finally, we appeal to the bound on $\|S^n - S^*\|^2_{H^{1,2}_f(\mathcal{X})}$ derived in Theorem 4.1 to conclude the proof:

$$\begin{aligned}\frac{\min\{m, (M^*)^{-2}\}}{2}\mathbb{E}\|T^n - T^*\|^2_{H^{1,2}_g(\mathcal{Y})} &\leq \frac{\min\{m, (M^*)^{-2}\}}{2}\tilde{C}(d, M^*)^2\mathbb{E}\|S^n - S^*\|^2_{H^{1,2}_f(\mathcal{X})} \\ &\lesssim \begin{cases} n^{-1/2}, & d < 2s, \\ n^{-1/2}\log n, & d = 2s, \\ n^{-s/d}, & d > 2s. \end{cases}\end{aligned}$$

□

B.6 Dimension ordering

B.6.1 Proof of Lemma 4.3

Now we establish a rate of convergence in the anisotropic smoothness setting. Define

$$\psi^k_S(x) = \log f_k(x_k|x_{(k+1):d}) - \log g_k(S_k(x)|S_{(k+1):d}(x)) + \log D_k S_k(x),$$

which is $(s_k, \ldots, s_d)$-smooth in $(x_k, \ldots, x_d)$ whenever each S_k is $(s_k + 1, s_{k+1}, \ldots, s_d)$-smooth in $(x_k, x_{k+1}, \ldots, x_d)$ by Assumption 4.1 and Lemma B.6. Since

$$f(x) = \prod_{k=1}^d f_k(x_k|x_{(k+1):d})$$

by the chain rule of densities, and similarly for g, we have

$$(P_n - P)\psi_S = \sum_{k=1}^d (P_n - P)\psi^k_S.$$

Note that ψ^k_S is a function of $S_{k:d}$ and the d_k variables $(x_k, \ldots, x_d)$ only. Defining

$$\Psi_k(\mathfrak{s}, d, M) = \{\psi^k_S : S \in \mathcal{T}(\mathfrak{s}, d, M)\},$$

we have the following analog of Proposition 3.3.

Proposition B.14. *Under Assumptions 3.1, 3.2, and 4.1, the metric entropy of $\Psi_k(\mathfrak{s}, d, M)$ in the L^∞ norm is bounded as*

$$H(\epsilon, \Psi_k(\mathfrak{s}, d, M), \|\cdot\|_\infty) \lesssim \epsilon^{-d_k/\sigma_k}.$$

Consequently, $\Psi_k(\mathfrak{s}, d, M)$ is totally bounded and therefore precompact in $L^\infty(\mathcal{X}_{k:d})$.

Hence, we obtain in Lemma 4.3 a bound on the supremum process analogous to Lemma 3.4 now applied to the family $\Psi_k(\mathfrak{s}, d, M)$ of $(s_k, \ldots, s_d)$-smooth maps ψ^k_S.

Proof of Lemma 4.3. The metric entropy integral bounds utilized in the proof of Lemma 3.4, combined with the entropy bound on $\Psi_k(\mathfrak{s}, d, M)$ derived in Proposition B.14, yield the following rate

$$\mathbb{E}\|P_n - P\|_{\Psi_k(\mathfrak{s},d,M)} = \mathbb{E}\left\{\sup_{S\in\mathcal{T}(\mathfrak{s},d,M)} |(P_n - P)\psi_S^k|\right\} \lesssim c_{n,k}.$$

Applying the triangle inequality then yields

$$\begin{aligned}
\mathbb{E}\|P_n - P\|_{\Psi(\mathfrak{s},d,M)} &= \mathbb{E}\left\{\sup_{S\in\mathcal{T}(\mathfrak{s},d,M)} |(P_n - P)\psi_S|\right\} \\
&= \mathbb{E}\left\{\sup_{S\in\mathcal{T}(\mathfrak{s},d,M)} \left|\sum_{k=1}^{d}(P_n - P)\psi_S^k\right|\right\} \\
&\le \sum_{k=1}^{d}\mathbb{E}\left\{\sup_{S\in\mathcal{T}(\mathfrak{s},d,M)} \left|(P_n - P)\psi_S^k\right|\right\} \\
&= \sum_{k=1}^{d}\mathbb{E}\|P_n - P\|_{\Psi_k(\mathfrak{s},d,M)} \\
&\lesssim \sum_{k=1}^{d} c_{n,k}.
\end{aligned}$$

□

B.7 Jacobian flows

Proof of Theorem 4.7. The proof is practically identical to that of Theorem 3.6, since the functions in $\Psi_m(\Sigma, s, M)$ are s-smooth with uniformly bounded derivatives (analogous to $\Psi(s, d, M)$) by definition of $\mathcal{J}_m(\Sigma, s, M)$, the chain rule of differentiation, and the relation

$$\psi_S(x) = \log[f(x)/g(S(x))] - \sum_{j=1}^{m}\sum_{k=1}^{d}\log D_k U_k^j(x^j),$$

where we define

$$x^j = \Sigma^j \circ U^{j-1} \circ \Sigma^{j-1} \circ \cdots \circ U^1 \circ \Sigma^1(x), \quad j \in [m].$$

Hence, the entropy estimates for $\Psi(s, d, M)$ in Proposition 3.3 hold also for $\Psi_m(\Sigma, s, M)$. Thus, we obtain similar bounds on $\mathbb{E}\|P_n - P\|_{\Psi_m(\Sigma,s,M)}$ as in Lemma 3.4. Combining this with the risk decomposition (4) and the argument in Theorem 3.6 completes the proof. □

B.8 On separability

Suppose the source g is a product density, i.e., $g(y) = \prod_{k=1}^d g_k(y_k)$ for some smooth densities $g_k : \mathbb{R} \to \mathbb{R}$. As the source density g is a degree of freedom in our problem, we are free to choose g to factor as such. For example, g could be the standard normal density in d-dimensions or the uniform density on a box $\mathcal{Y} \subset \mathbb{R}^d$. We will show that the task of estimating the KR map S^* is amenable to distributed computation in this case.

Assumption B.1. *The source density g factors as a product: $g(y) = \prod_{k=1}^d g_k(y_k)$.*

Recall the minimization objective defining our estimator:

$$-\frac{1}{n}\sum_{i=1}^{n}\left[\log g(S(X^i)) + \sum_{k=1}^{d}\log D_k S_k(X^i)\right],$$

where $X^i = (X_1^i, \ldots, X_d^i), i = 1, \ldots, n$ is an iid random sample from f. Here we omit the entropy term involving f that does not depend on S without loss of generality. For a general density g, we can simplify the above expression by appealing to the chain rule for densities

$$g(y) = \prod_{k=1}^{d} g_k(y_k|y_{k+1}, \ldots, y_d),$$

where

$$g_k(y_k|y_{k+1}, \ldots, y_d) = \frac{\int g(y_1, \ldots, y_d)dy_1 \cdots dy_{k-1}}{\int g(y_1, \ldots, y_d)dy_1 \cdots dy_k}.$$

The objective then becomes

$$-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{d}\left[\log g_k(S_k(X^i)|S_{k+1}(X^i), \ldots, S_d(X^i)) + \log D_k S_k(X^i)\right].$$

When g is a product density we have $g_k(y_k|y_{k+1}, \ldots, y_d) = g_k(y_k)$ and we obtain

$$\sum_{k=1}^{d}\left\{-\frac{1}{n}\sum_{i=1}^{n}\left[\log g_k(S_k(X^i)) + \log D_k S_k(X^i)\right]\right\},$$

which is a separable objective over the component maps S_k. In this case, we can find our estimator $S^n = (S_1^n, \ldots, S_d^n)$ by solving for the components S_k^n in parallel.