Journal of Machine Learning Research 25 (2024) 1-61 Submitted 10/23; Published 6/24

Distribution Learning via Neural Differential Equations: A Nonparametric Statistical Perspective

Youssef Marzouk YMARZ@MIT.EDU
Institute for Data, Systems, and Society
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Zhi Ren ZREN@MIT.EDU
Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Sven Wang SVENWANG@MIT.EDU
Institute for Data, Systems, and Society
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Jakob Zech JAKOB.ZECH@UNI-HEIDELBERG.DE
Interdisciplinary Center for Scientific Computing
Heidelberg University
Heidelberg 69120, Germany

Editor: Qiang Liu

Abstract

Ordinary differential equations (ODEs), via their induced flow maps, provide a powerful framework to parameterize invertible transformations for representing complex probability distributions. While such models have achieved enormous success in machine learning, little is known about their statistical properties. This work establishes the first general nonparametric statistical convergence analysis for distribution learning via ODE models trained through likelihood maximization. We first prove a convergence theorem applicable to *arbitrary* velocity field classes $\mathcal{F}$ satisfying certain simple boundary constraints. This general result captures the trade-off between the approximation error and complexity of the ODE model. We show that the latter can be quantified via the C^1-metric entropy of the class $\mathcal{F}$. We then apply this general framework to the setting of C^k-smooth target densities, and establish nearly minimax-optimal convergence rates for two relevant velocity field classes $\mathcal{F}$: C^k functions and neural networks. The latter is the practically important case of neural ODEs. Our results also provide insight on how the choice of velocity field class, and the dependence of this choice on sample size (e.g., the scaling of neural network classes), impact statistical performance.

Keywords: Neural differential equations, normalizing flows, density estimation, nonparametric statistics, M-estimation

1. Introduction

The interface of nonparametric statistics with complex models given by differential equations has been a major focus of contemporary statistics and applied mathematics. On the one hand, physically motivated differential equation models of data-generating processes are central to inverse problems and data assimilation. There has been considerable progress in understanding such models through a statistical lens, leading to structure-exploiting algorithms (Rudolf and Sprungk, 2018; Schillings and Schwab, 2016; Kim et al., 2023), new consistency and uncertainty quantification guarantees (Monard et al., 2019; Nickl et al., 2020; Nickl and Titi, 2023), and a growing understanding of computational complexity (Nickl and Wang, 2022; Nickl, 2023). On the other hand, differential equations underpin the construction of new expressive and flexible model classes for *representing* complex probability distributions, which have enjoyed enormous success in machine learning and data science. Examples of such models include neural ordinary differential equations (Chen et al., 2018), score-based diffusion models (Song et al., 2020b; Yang et al., 2022), and flow matching methods (Lipman et al., 2022; Liu et al., 2022; Albergo et al., 2023). In such models, a key aspect of the dynamics, e.g, the velocity field of an ordinary differential equation or the drift of a stochastic differential equation, is *learned from data* by minimizing a suitable objective. These approaches have achieved leading performance in diverse applications, ranging from generative modeling of images and video (Grathwohl et al., 2018; Song et al., 2020a; Ho et al., 2022) to density estimation in high-energy physics (Nachman and Shih, 2020) to conditional sampling and simulation-based Bayesian inference (Shi et al., 2022; Batzolis et al., 2021; Cranmer et al., 2020).

This paper develops statistical finite-sample guarantees for distribution learning with ordinary differential equation (ODE) models. These models are described via finite-time *flow maps* of ODEs (Arnold and Silverman, 1978) of the form

$$\begin{cases} \frac{d}{dt}X(x,t) &= f(X(x,t),t), \\ X(x,0) &= x, \end{cases} \qquad \text{for} \quad x \in D, \quad t \in [0,1]. \tag{1.1}$$

Here $D \subseteq \mathbb{R}^d$ ($d \geqslant 1$) is some domain and the velocity field $f \in \mathcal{F}$ belongs to some parametric function class $\mathcal{F}$. For each $f \in \mathcal{F}$, the collection of all trajectories of (1.1) is captured by the flow map $(x,t) \mapsto X^f(x,t)$, a continuous-time invertible transformation that can be used to push forward and pull back probability distributions. In the context of statistical learning, this framework is applied to infer complicated unknown distributions: a velocity field $\hat{f} \in \mathcal{F}$ is computed by minimizing a statistical objective, and the unknown distribution is then approximated as the *pullback* of a reference distribution (e.g., normal or uniform) under the terminal time (i.e., $t = 1$) flow map $x \mapsto X^{\hat{f}}(x,1)$. This approximation immediately provides a density estimate, and crucially further enables sampling (hence generative modeling) by evaluating the inverse of the flow map on reference samples. For details, see Section 2.2.

One way of interpreting such ODE-based models is to view them as a specific parameterizations of time-independent *transport maps* (Marzouk et al., 2016). However, a key practical advantage of the ODE formulation over models that directly represent transport maps is that for *any* sufficiently regular velocity field f, the ODE construction guarantees that the flow maps are invertible. Moreover, given an initial condition x at time $t = 0$ with known probability density, the density of the state at any intermediate time $t > 0$ can easily be evaluated via the so-called 'instantaneous change-of-variables' formula (Chen et al., 2018). Since these features hold for very generic choices of f, they permit using virtually any approximation class—such as polynomials, neural networks, or kernel representations (Owhadi and Yoo, 2019)—to describe the parameter space $\mathcal{F}$. By contrast, in models which directly parameterize transport maps, significant care is needed to ensure invertibility and tractable Jacobian determinants.

When the velocity field f is represented as a deep neural network, the system (1.1) is called a *neural ODE* (Chen et al., 2018); such models achieve state-of-the-art performance in density estimation (Grathwohl et al., 2018; Onken et al., 2021) and are competitive (by various sample quality metrics) for various generative modeling tasks. While this construction is powerful, most questions regarding theoretical performance guarantees for ODE-based methods remain unexplored; the notable exceptions (Ishikawa et al., 2022; Li et al., 2022; Ruiz-Balet and Zuazua, 2023b,a) will be discussed further below. To the best of our knowledge, current approximation results are limited to universal approximation (Teshima et al., 2020; Ishikawa et al., 2022; Li et al., 2022; Ruiz-Balet and Zuazua, 2023b), while quantitative approximation *rates* are yet unknown; our forthcoming companion paper (Marzouk et al., 2023) provides the first such approximation rate results. The present paper considers the yet more challenging task of giving *statistical* finite-sample convergence guarantees, which has thus far only been considered by Ruiz-Balet and Zuazua (2023a), whose proof approach and results are vastly different from ours; see below for further discussion. The observational setting we consider is that of *nonparametric density estimation* (e.g., Triebel (2008); Giné and Nickl (2016)), which also underlies generative modeling: a finite collection of independent and identically distributed (iid) samples is given,

$$Z_1, \dots, Z_n \overset{\text{iid}}{\sim} P_0,$$

and our goal is to characterize the unknown target distribution P_0. We consider general estimators $\hat{f}$ which arise as minimizers of a negative log-likelihood (or empirical Kullback–Leibler) objective over some class $\mathcal{F}$, a training strategy which is extremely common in practice (Grathwohl et al., 2018; Chen et al., 2018; Finlay et al., 2020).

1.1 Results and Contributions

To our knowledge, our paper provides the first rigorous statistical analysis of likelihood-based ODE density estimators, and specifically the first such statistical convergence results for *neural* ODEs. Our approach integrates tools from nonparametric M-estimation (van de Geer, 2000), recent advances in statistical and approximation theory for neural networks (e.g., Schmidt-Hieber (2020)), and ODE analytical theory (e.g., Hartman (2002)). Our results also create the first explicit framework for understanding the impact of choosing different velocity field classes on statistical performance.

In Section 2, we develop a statistical convergence result applicable to *general* ODE-parameterized maximum likelihood estimators (ODE-MLEs); see Theorem 2. Specifically, given any variational class $\mathcal{F}$ of velocity fields (satisfying mild boundedness assumptions), our result gives a bound on the rate of convergence as a sum of two terms reminiscent of the classical bias-variance tradeoff. The first term corresponds to the 'best approximation' of the ground truth measure P_0 by elements in the class $\mathcal{F}$, and the second term follows from the metric entropy of $\mathcal{F}$ in the C^1 norm. Our result thus identifies the latter as a natural *statistical complexity measure* that yields an upper bound on the stochastic fluctuations of any ODE-MLE; see Section 2 for details.

To obtain this result, we first derive natural boundary conditions on the variational class $\mathcal{F}$ to ensure that the statistical objective can be formulated over $\mathcal{F}$ in its standard form, by ensuring that all pullback distributions under the associated ODE flow maps possess the same support and are absolutely continuous. Then, to prove Theorem 2, we derive novel analytical Lipschitz estimates for ODEs—bounding the distance between terminal-time transport maps induced by ODE flows, and between their corresponding pullback distributions, in terms of the velocity fields that underlie them. Such Lipschitz properties hold true *locally* on sets of velocity fields which are uniformly bounded in a certain sense; see (2.8). These estimates, detailed in Section 2.4, are then combined with existing convergence theory for general sieved maximum likelihood estimators (van de Geer, 2000) in Hellinger loss. They crucially allow us to relate so-called bracketing entropy rates, which are commonly required in theory for M-estimation (van de Geer, 2000), to C^1-metric entropy rates of ODE-based estimators; see the proof of Theorem 2 for details.

Section 3 studies the case where P_0 possesses a C^k density and $\mathcal{F}$ likewise consists of C^k-smooth velocity fields. The main convergence theorem in this section is Theorem 10. A key intermediate result establishes the existence of a C^k velocity field coupling P_0 with the reference distribution *and* vanishing appropriately normal to the boundary. The existence of a C^k velocity field is established in our companion paper (Marzouk et al., 2023); it is constructed using a triangular Knothe–Rosenblatt (KR) map and straight-line trajectories. The required boundary behavior is proven here, using anisotropic regularity properties of KR maps shown in Wang and Marzouk (2022); see Theorem 9. Due to the additional dimension arising from the space-time structure of the ODEs, our rates of convergence are slightly suboptimal in a statistical minimax sense. Achieving minimax-optimality in this context will likely require a more refined choice of $\mathcal{F}$, e.g., as an anisotropic regularity class or via penalization; we leave this for future work. See also Remark 21.

Finally, Section 4 considers the case where $\mathcal{F}$ is given via neural network classes. Using scalings of ReLU network classes (i.e., width, depth, sparsity, and norm constraints scaling in the sample size n) derived in the seminal work of Schmidt-Hieber (2020), we prove the relevant metric entropy and approximation bounds needed to apply our general result from Theorem 2. In order to satisfy the regularity and boundary conditions required for our ODE setting, we make some modifications to the standard constructions of neural network classes: first, we need to work with the squared ReLU^2 activation functions to ensure C^1 regularity; and second, we multiply standard neural network classes with certain component-wise cutoff functions to create an ansatz space satisfying appropriate boundary conditions. See Section 4 for details. Our choice of a slightly more regular ReLU^2 activation, interestingly, may relate to the fact that smooth activation functions are often used in practical applications of continuous normalizing flows.

1.2 Related Work

The past decade has seen the emergence of increasingly expressive and powerful models for complex probability distributions that employ *transportation of measure*: The central idea is to express the "target" distribution of interest as the pullback or pushforward of a simple reference distribution (e.g., uniform or standard Gaussian) by a learned (measurable) map. Samples from the target distribution are then produced simply by evaluating this map on samples from the reference; this enables generative modeling (Kingma and Dhariwal, 2018). When the map is invertible and sufficiently smooth, the map and the reference density yield a closed-form expression for the target density, enabling density estimation (Tabak and Vanden-Eijnden, 2010; Anderes and Coram, 2011; Wang and Marzouk, 2022). Given a family of transport maps and a reference measure, variational inference can be cast as minimization of a suitable divergence over the resulting family of pushforward measures (Moselhy and Marzouk, 2012; Rezende and Mohamed, 2015).

A central question in designing these methods is how to represent or parameterize the map. Initial applications of transport in machine learning emphasized normalizing flows (Rezende and Mohamed, 2015; Papamakarios et al., 2021; Kobyzev et al., 2020), which are compositions of simple, parametric, invertible transformations whose Jacobian determinants are, by design, easy to evaluate. A considerable variety of such transformations have been proposed (Dinh et al., 2015; Kingma et al., 2016; Huang et al., 2018; Wehenkel and Louppe, 2019), sometimes under the broader label of "invertible neural networks." In other settings, triangular maps (Bogachev et al., 2005; Marzouk et al., 2016; Zech and Marzouk, 2022a,b; Baptista et al., 2022; Irons et al., 2022) and parametric approximations of optimal transport maps (Moselhy and Marzouk, 2012; Huang et al., 2020) have been popular. More recently, there has been considerable interest in "continuous-time" (i.e., differential) notions of normalizing flows. As explained earlier in this introduction, these models can be formalized as ODE systems (1.1) and are the central topic of this paper.

Questions of *function approximation* with neural ODEs have been studied in Li et al. (2022); Ishikawa et al. (2022). Ishikawa et al. (2022) show that neural ODEs are univeral approximators of smooth diffeomorphisms on $\mathbb{R}^d$ in appropriate Sobolev norms. Li et al. (2022) adapts ideas from dynamical systems to show that neural ODEs are universal approximators of continuous functions from $\mathbb{R}^d$ to $\mathbb{R}^m$ (hence, not only diffeomorphisms) in a L^2 sense, for $d \geqslant 2$. Both papers compose the flow map of the ODE with a terminal mapping, meant to represent a classification or regression layer. Yet these universal approximation results do not characterize approximation *rates*, e.g., relating bounds on an approximation error to the size of the network representing the velocity field. Function approximation is also different than our present focus of statistical recovery guarantees.

The results on diffeomorphism approximation in Ishikawa et al. (2022) do translate to universal approximation of certain classes of distributions, in a weak sense and in total variation. Ruiz-Balet and Zuazua (2023b) also proves universal approximation for certain target distributions, in Wasserstein-1 distance, using a rather different approach that is discrete and constructive. Our companion paper Marzouk et al. (2023), in contrast, establishes approximation *rates* for neural ODE representations of distributions with C^k-smooth densities, and shows that there exist neural network representations of the velocity field f, with size explicitly bounded in terms of the regularity of the densities, that achieve efficient approximation. There has also been relevant work on approximation theory for transport maps that are not constructed via ODEs. For example, Zech and Marzouk (2022a,b) investigate sparse polynomial and neural network approximations of triangular (Knothe–Rosenblatt) maps, formulating *a priori* descriptions of an ansatz space that achieves exponential convergence in the case of analytic densities. A broader framework for understanding the distributional errors of transport map approximations is proposed in Baptista et al. (2023).

From the statistical perspective, one must address the impact of using a finite number of samples n to estimate the ODE velocity field and the resulting pushforward or pullback densities. To our knowledge, there has been almost no statistical convergence analysis of neural ODEs. Perhaps the sole exception is Ruiz-Balet and Zuazua (2023a) (building on Ruiz-Balet and Zuazua (2023b)), which analyzes neural ODE-type models from a controllability perspective, explicitly constructing finite-difference approximations of the target density using a neural network velocity field with ReLU activations. Sample complexity results follow from assessing the convergence of the n-sample empirical measure to its finite-difference approximation. This construction is rather different from the maximum likelihood training typically used in neural ODEs, and similarly its analysis uses different tools than those we exploit here. Moreover, Ruiz-Balet and Zuazua (2023a,b) do not assume any smoothness in the reference and target densities.

For direct parameterization of transport maps, i.e., not using an ODE construction, Wang and Marzouk (2022) develop a general statistical convergence theory for transport-based estimation of Hölder-smooth densities, and we build on those results here. There is also a growing body of work on the statistical estimation of *optimal transport maps*; see, e.g., Manole et al. (2021); Divol et al. (2022). As a corollary of such results, for a fixed reference distribution, one can obtain rates of convergence for optimal transport-based density estimation in Wasserstein distances (Hütter and Rigollet, 2021, Remark 5). We emphasize, however, that these constructions are distinct from the ODE models of interest here.

Let us also comment briefly on sampling and generative modeling methods based *stochastic* differential equations (SDEs). As mentioned in the opening, such methods generally seek to learn the drift term of an SDE so that marginal distribution at a particular time (e.g., $t = 0$ or $t \to \infty$) is a good approximation of the target distribution. Score-based diffusion models (Song et al., 2020b, 2021; Yang et al., 2022) are a widely used approach of this type. Yet these models—and approaches for elucidating their approximation properties and statistical behavior—are rather different in character from deterministic ODEs, due to the presence of the diffusion term. Also, the estimation problem in score-based diffusions involves an objective that is quadratic in the desired score; this is much simpler than the log-likelihood we analyze here, which is highly nonlinear in the velocity f (see (2.5)). Very recent literature has established near-optimal minimax rates for the estimation of smooth densities (in, e.g., total variation distance) with score-based diffusion models (Oko et al., 2023); parallel efforts have analyzed the convergence of such models for target distributions supported on low-dimensional manifolds (Chen et al., 2023a). Yet it is worth noting that deterministic ODEs have a role in diffusion models as well. For instance, the deterministic "probability flow ODE" (Song et al., 2020b) (see also Song et al. (2020a)) is sometimes used instead of a time-reversed SDE for sampling in this context, as numerical integration of the

ODE can be more accurate and efficient than the comparable discretized SDE (Chen et al., 2023b).

2. General ODE-based Density Estimators

In this section we derive a key convergence result, Theorem 2, which characterizes a convergence rate for general ODE-based density estimators: specifically, we consider estimators obtained through a *velocity field* learned from the data, which in turn generates a pullback density estimate. In subsequent sections, we apply this result to two relevant classes of velocity fields: the class of k-times continuously differentiable velocity fields (Section 3) and neural network parameterizations of velocity fields (Section 4).

2.1 Notation

We introduce a number of notations and definitions which are needed throughout the paper.

Norms for vectors and matrices

For a vector, we denote by $\|\cdot\|_2$ its l^2-norm (the Euclidean norm), $\|\cdot\|_\infty$ its l^∞-norm, and $\|\cdot\|_0$ its l^0-norm (number of nonzero entries). For a matrix, we denote by $\|\cdot\|_2$ its operator norm induced by the l^2-norm on vectors, $\|\cdot\|_\infty$ its operator norm induced by the l^∞-norm on vectors, $\|\cdot\|_0$ its l^0-norm (number of nonzero entries), $\|\cdot\|_{\infty,\infty}$ its l^∞ norm (the maximum absolute value of its entries), and $\|\cdot\|_F$ its Frobenius norm.

Derivatives and function spaces

Let $d \geqslant 1$ and let $D \subseteq \mathbb{R}^d$ be an open domain. For $k \in \mathbb{N} := \{1, 2, \dots\}$, we denote by $C^k(D)$ the space of k-times continuously differentiable real-valued functions $f : D \to \mathbb{R}$ with uniformly continuous derivatives. Similarly, for $m \geqslant 1$ we shall write $C^k(D, \mathbb{R}^m)$ for the space of k-times differentiable vector-valued functions taking values in $\mathbb{R}^m$. When D is convex with Lipschitz boundary, then any function $f \in C^k(D)$ (and its derivatives) possesses a unique continuation onto the boundary (by uniform continuity). In this case, we shall also use the notation $f \in C^k(D)$ with closed, convex, sets D.

To denote partial derivatives of functions, we use standard multi-index notation. Given a multi-index $\boldsymbol{v} = (v_1, v_2, \dots, v_d) \in \mathbb{N}^d$, we will write $\partial^{\boldsymbol{v}} f(x) = \frac{\partial^{|\boldsymbol{v}|}}{\partial x_1^{v_1} \dots \partial x_d^{v_d}} f(x)$ for the $|\boldsymbol{v}|$-th order partial derivative of f, whenever it exists.

For $f \in C^1(D)$, we denote its gradient by $\nabla f : D \to \mathbb{R}^d$. Similarly, if $f \in C^1(D, \mathbb{R}^m)$ for $m \geqslant 1$, $\nabla f : \mathbb{R}^d \to \mathbb{R}^{m \times d}$ denotes the Jacobian (or gradient matrix) of f. If f depends on multiple variables—say a 'space variable' $x \in \mathbb{R}^d$ and a 'time variable' $t \in \mathbb{R}$—we will use the standard notation $\nabla_x f(x, t)$ for the x-gradient of f. Similarly, for a multi-index $\boldsymbol{v} \in \mathbb{N}^d$, $\partial_x^{\boldsymbol{v}} f(x, t)$ denotes the corresponding partial derivative with respect to x. For continuous $f : D \to \mathbb{R}$, we let $\|f\|_{C(D)} := \sup_{x \in D} |f(x)|$ and for a C^k function $f \in C^k(D)$ with $k > 1$, we let $\|f\|_{C^k(D)} = \sup_{|\boldsymbol{v}| \leqslant k} \|\partial^{\boldsymbol{v}} f\|_{C(D)}$. For a vector field $f \in C^k(D, \mathbb{R}^m)$, we define $\|f\|_{C^k(D, \mathbb{R}^m)} = \sup_{j \in \{1,2,\dots,m\}} \|f_j\|_{C^k(D)}$, and we may sometimes omit the $\mathbb{R}^m$ by abuse of notation. If $f : D \to \mathbb{R}^m$ is Lipschitz continuous, we write $|f|_{\mathrm{Lip}(D)}$ to denote its Lipschitz constant.

For $D \subseteq \mathbb{R}^d$ Borel measurable, a Borel measure μ on D, and $p \in [1, \infty]$, we write $L^p(D, \mu)$ to denote the usual space of p-integrable functions w.r.t. μ on D. If μ is the Lebesgue measure, we write $L^p(D)$ instead. In case there is no confusion about D, we also use the notation $L^p(\mu)$.

Transportation of measure

Let $d \geqslant 1$ and let D_1, $D_2 \subseteq \mathbb{R}^d$ be Borel measurable sets equipped with the Borel σ-algebra. Then, for any measurable function $T : D_1 \to D_2$ and probability distribution π on D_1, we denote the *pushforward distribution* of π under T by $T_\sharp\pi$, given by $T_\sharp\pi(A) = \pi\left(T^{-1}(A)\right)$ for any measurable subset $A \subseteq D_2$. Given another probability distribution ρ on D_2, we say that T pushes forward π to ρ if $T_\sharp\pi = \rho$. Since we will deal only with measures that possess densities with respect to Lebesgue measure, we will occasionally use the same symbol to represent a probability measure and its Lebesgue density, in a slight abuse of notation. If additionally T is bijective, differentiable, and invertible with a continuously differentiable inverse $T^{-1} : D_2 \to D_1$ (i.e., T is a diffeomorphism), then the pushforward density $T_\sharp\pi$ is given by $\rho(x) = \pi(T^{-1}(x))|\det \nabla T^{-1}(x)|$ (the change-of-variables formula). In this case, we also denote the *pullback* density of ρ under T by $T^\sharp\rho$, and it holds that

$$\pi(x) = [(T^{-1})_\sharp\rho](x) = [T^\sharp\rho](x) = \rho(T(x))|\det \nabla T(x)|.$$

2.2 Nonparametric Density Estimation via ODEs

For $d \geqslant 1$, we denote the d-dimensional unit cube by

$$D = [0,1]^d \subset \mathbb{R}^d$$

throughout.[1] We will be concerned with the problem of nonparametric density estimation on D, where the observations are given by independent and identically distributed (iid) samples

$$(Z_i : i = 1, \ldots, n), \ Z_i \overset{\text{iid}}{\sim} P_0, \tag{2.1}$$

for P_0 some unknown probability measure supported on D. Our goal is to infer P_0 from $(Z_i : i = 1, \ldots, n)$. We assume throughout that P_0 possesses a Lebesgue density which we denote by p_0. We denote the n-fold product measure of P_0 by P_0^n, and expectations with respect to P_0^n by $\mathbf{E}_{P_0}^n$.

Given any sufficiently regular 'velocity vector field' $f : D \times [0,1] \to \mathbb{R}^d$ and any initial condition $x \in D$, consider the following ordinary differential equation

$$\begin{cases} \frac{d}{dt}X^f(x,t) &= f(X^f(x,t),t), \qquad t \in [0,1], \\ X^f(x,0) &= x. \end{cases} \tag{2.2}$$

If f is Lipschitz continuous and if the 'flow lines' of f do not leave the domain D (a key technical condition to be discussed in more detail below), then, by the Picard–Lindelöf

1. Much of what follows could also be extended to more general (bounded and sufficiently regular) domains $D \subseteq \mathbb{R}^d$ at the expense of additional technicalities.

theorem, (2.2) is solvable and induces trajectories $t \mapsto X^f(x,t) : [0,1] \to D$ for each $x \in D$. They satisfy

$$X^f(x,t) = x + \int_0^t f(X^f(x,s),s)ds, \qquad t \in [0,1],\ x \in D. \tag{2.3}$$

We will refer to the mapping $x \mapsto X^f(x,t)$ as the time-t *flow map* of the ODE. The transport map obtained by evaluating this flow map at the terminal time $t = 1$ is denoted by $T^f := X^f(\cdot,1)$. Since the trajectories of the ODE (2.2) are unique and do not intersect, the inverse $(T^f)^{-1}$ is also well-defined as a map from $T^f(D)$ onto D; both maps T^f and $(T^f)^{-1}$ can then be used to transform probability measures.

Admissible velocity fields

Throughout, with $D = [0,1]^d$, we denote the cylindrical $d+1$-dimensional 'space-time' unit cube by

$$\Omega = D \times [0,1] \subset \mathbb{R}^{d+1}.$$

In the setting considered here, where the support of the unknown density is known, it is natural to consider only ODE flows which **(i)** do not leave the domain D, and **(ii)** for which flow maps $\{X^f(\cdot,t) : t \in [0,1]\}$ are diffeomorphisms $D \to D$. In order to ensure those properties, along with the existence and uniqueness of the solution to (2.2), we need to introduce boundary conditions on the class of velocity fields considered. Specifically, denoting by ν_x the outward pointing normal vector at any point $x \in \partial D$ where ν_x is well-defined, we let

$$\mathcal{V} = \Big\{ f \in C^1(\Omega, \mathbb{R}^d) \ : \ f(x,t)\cdot \nu_x \equiv 0 \text{ for all } (x,t) \in \partial D \times [0,1] \Big\}. \tag{2.4}$$

This condition ensures that there is no flow outside of D. In fact, it even implies the maps $x \mapsto X^f(x,t) : D \to D$ to be C^1-diffeomorphisms for every $t \in [0,1]$:

Lemma 1 *Suppose that $f \in \mathcal{V}$, for $\mathcal{V}$ given by (2.4). Then, for any $t \in [0,1]$, the ODE flow map $X^f(\cdot,t) : D \to D$ at time t is a diffeomorphism. In particular, the time-one map $T^f = X^f(\cdot,1) : D \to D$ is a diffeomorphism, and the pullback density $(T^f)^\sharp \rho$ of any density ρ supported on D is given by*

$$(T^f)^\sharp \rho(x) = \rho(T^f(x)) \det \nabla T^f(x), \qquad x \in D.$$

The proof of Lemma 1, which is based on tools from ODE theory (Hartman, 2002) as well as Grönwall's inequality, can be found in Appendix A. While the assumption that $f \in C^1$ ensures existence of a unique solution to the ODE (2.2), the additional requirement of the normal component $f \cdot \nu_x$ vanishing at the boundary ∂D guarantees that the trajectories remain inside the unit cube D at all times. In particular, if $\mathrm{supp}(\rho) = D$, then the 'interpolating' distributions $(X^f(\cdot,t))^\sharp \rho$ for $t \in [0,1]$, all possess common support D.

Maximum likelihood objective

Let $\mathcal{F} \subseteq \mathcal{V}$ be any class of admissible velocity fields, and let us fix some reference density ρ on $D = [0,1]^d$. Assume that ρ is strictly positive and upper bounded. By Lemma 1, each time-one flow map T^f is a diffeomorphism, so that we may form the collection of pullback densities $(T^f)^\sharp(x)\rho = \rho(T^f(x)) \det \nabla T^f(x)$ as an approximating class for the unknown ground truth distribution p_0. In this paper, we study estimators maximizing the likelihood: that is, with $Z_i \sim P_0$ i.i.d., we let

$$\hat{f} \in \arg\max_{f \in \mathcal{F}} \mathcal{J}(f), \qquad \mathcal{J}(f) := \left(\sum_{i=1}^{n} \log \rho(T^f(Z_i)) + \log \det \nabla T^f(Z_i) \right). \tag{2.5}$$

Any such $\hat{f}$ naturally gives rise to a plug-in estimator $(T^{\hat{f}})^\sharp \rho$ for the data-generating density p_0 via its pullback density

$$(T^{\hat{f}})^\sharp \rho = \rho(T^{\hat{f}}(x)) \det \nabla T^{\hat{f}}(x), \qquad x \in D. \tag{2.6}$$

The rate of convergence towards p_0 in terms of n and $\mathcal{F}$ will be the subject of our main results. Note that since all the $(T^f)^\sharp \rho$ have common support D, the likelihood objective is well-defined and finite for all $f \in \mathcal{F} \subseteq \mathcal{V}$. We will refer to estimators (2.6) as *ODE-MLE* estimators (over the class $\mathcal{F}$).

2.3 Main Convergence Result

We are now ready to formulate the first main result of this paper, Theorem 2, which provides a *general* convergence rate for ODE-MLE estimators. The result is stated in terms of two key characteristics of the class of velocity fields $\mathcal{F}$; the first of which is the 'best approximation error' $h\big((T^{f^*})^\# \rho, p_0\big)$ of p_0 by any pullback distribution $(T^f)^\sharp \rho$ over the class $f \in \mathcal{F}$. The second key quantity is the metric entropy of $\mathcal{F}$ in C^1, which is identified as a key *complexity measure* that gives an upper bound for the 'stochastic fluctuations' of ODE-MLE estimators over any class $\mathcal{F} \subseteq \mathcal{V}$ via the inequality (2.9); see Remark 3 for discussion.

We recall some standard definitions. Again let $D = [0,1]^d$ and $\Omega = D \times [0,1]$. The Hellinger distance $h(p_1, p_2)$ between any two probability densities $p_1,\ p_2 \in L^1(D)$ is

$$h(p_1, p_2) = \left(\int_D \left[\sqrt{p_1(x)} - \sqrt{p_2(x)} \right]^2 dx \right)^{1/2}.$$

For any normed space $(X, \|\cdot\|)$ and subset $A \subseteq X$, we denote the *metric entropy* of A by $H(A, X, \tau) = \log N(A, X, \tau)$ $(\tau > 0)$, where $N(A, X, \tau)$ is the *covering number* of A,

$$N(A, X, \tau) := \min \left\{ N \in \mathbb{N} \,\middle|\, \exists x_1, \ldots, x_N \in X \text{ such that } A \subseteq \bigcup_{j=1}^{N} B_\tau(x_j) \right\}.$$

Here, we used the notation $B_\varepsilon(x) := \{\tilde{x} \in X \mid \|x - \tilde{x}\| \leqslant \varepsilon\}$ for all $\varepsilon > 0$, $x \in X$.

Assumption 2.1 (Ground truth and reference density) *Let p_0, ρ be two probability densities such that ρ is Lipschitz continuous and for some $0 < \kappa \leqslant K < \infty$*

$$p_0(x) \leqslant K \qquad \textit{and} \qquad \kappa \leqslant \rho(x) \leqslant K \qquad \forall x \in D. \tag{2.7}$$

Now we define a useful subset of the admissible velocity fields $\mathcal{V}$ from (2.4).

Assumption 2.2 (Boundedness of $\mathcal{F}$) *Let $\mathcal{F} \subseteq \mathcal{V}$ be a class of admissible velocity fields such that for some $r > 0$,*

$$\sup_{f \in \mathcal{F}} \Big(\|f\|_{C^1(\Omega)} + \sup_{t \in [0,1]} |\nabla_x f(\cdot, t)|_{\mathrm{Lip(D)}} \Big) =: r < \infty. \tag{2.8}$$

Given a class $\mathcal{F} \subseteq \mathcal{V}$, we now define the crucial square root metric entropy integral of $\mathcal{F}$, which plays a key role in determining the convergence rate of the ODE-MLEs taken over $\mathcal{F}$. For any $R > 0$, we shall denote

$$I(\mathcal{F}, R) := R + \int_0^R H^{1/2}(\mathcal{F}, C^1(\Omega), \tau) d\tau \qquad \text{for all } R > 0.$$

For technical reasons, instead of working directly with $I(\mathcal{F}, R)$, we shall work with an *upper bound* for $I(\mathcal{F}, R)$. We fix any such upper bound Ψ, satisfying $\Psi(R) \geqslant I(\mathcal{F}, R)$ on $(0, \infty)$. The following assumption on the growth of Ψ is a standard technical requirement in the literature on nonparametric M-estimators; see, e.g., van de Geer (2000); Nickl et al. (2020) or Theorem 5 below. It is required in standard 'slicing' concentration arguments based on empirical processes (cf. the proofs of Theorems 7.4 and 10.13 in van de Geer (2000)), and is satisfied for all sufficiently smooth classes of functions; see for instance our examples in Sections 3 and 4 below.

Assumption 2.3 *Suppose that the upper bound $\Psi : (0, \infty) \to \mathbb{R}$ is such that $R \mapsto \Psi(R)/R^2$ is non-increasing on $(0, \infty)$.*

Theorem 2 (Convergence of general ODE-MLEs) *Suppose that p_0, ρ, and $\mathcal{F} \subseteq \mathcal{V}$ and Ψ are such that Assumptions 2.1, 2.2, and 2.3 are fulfilled with some constants $0 < \kappa < K$ and $r > 0$, and consider the i.i.d. sampling model (2.1) with p_0. Let $\hat{f} \in \arg\max_{f \in \mathcal{F}} \mathcal{J}(f)$ denote an ODE-MLE estimator as in (2.5). Then, there are constants $C, C' > 0$ only depending on d, κ, K, r, and $|\rho|_{\mathrm{Lip}(D)}$ such that for all $n \geqslant 1$ and $\delta_n > 0$ with*

$$\sqrt{n}\delta_n^2 \geqslant C\Psi(\delta_n), \tag{2.9}$$

all $f^* \in \mathcal{F}$ and all $\delta \geqslant \delta_n$, we have the concentration inequality

$$P_0^n\Big(h\big((T^{\hat{f}})^\sharp \rho, p_0\big) \geqslant C\big[h\big((T^{f^*})^\sharp \rho, p_0\big) + \delta\big]\Big) \leqslant C \exp\Big(-\frac{n\delta^2}{C}\Big), \tag{2.10}$$

and such that the mean squared error is bounded as follows:

$$\mathbf{E}_{P_0}^n [h^2((T^{\hat{f}})^\sharp \rho, p_0)] \leqslant C' \Big(h^2((T^{f^*})^\sharp \rho, p_0) + \delta_n^2 + \frac{1}{n}\Big). \tag{2.11}$$

The above theorem is *non-asymptotic* in that the constants involved are independent of n and of the variational class $\mathcal{F}$ of velocity fields. Thus, when applying the theorem, one may choose an 'approximating sequence' of classes $\mathcal{F} = \mathcal{F}_n$ as the number of statistical observations grows. We have omitted use of $\mathcal{F}_n$ merely for notational convenience. While some literature considers other loss functions (e.g., L^p-norms (Goldenshluger and Lepski, 2014)), the use of Hellinger loss is standard in our context, due to its natural relationship both to maximum likelihood theory and to density estimation; see van de Geer (2000) or Birgé (1986) for further discussion.

Remark 3 (C^1 metric entropy) *The C^1 metric entropy of $\mathcal{F}$ is a natural complexity measure in the context of ODE-MLEs, in light of the intuition that the pointwise distance between two pullback densities can be bounded by the C^1-norm of the corresponding inducing velocity fields. The latter is rigorously proven in Section 2.4 below, using analytical tools from ODE theory. The L^∞-norm complexity of the pullback densities induced by a class $\mathcal{F}$ in turn yields a bound for the bracketing metric entropy, which is well known to play a key role in quantifying convergence rates of maximum likelihood-type estimators; see, e.g., van de Geer (2000). Whether the C^1 norm is both necessary and sufficient for characterizing convergence rates, or whether a weaker norm than $\|\cdot\|_{C^1}$ would suffice (yielding smaller entropy integrals), is an interesting question for future research.*

2.4 Proof of Theorem 2

The proof of Theorem 2 relies on combining general convergence results for nonparametric M-estimation developed in (van de Geer, 2000, 2001) with key analytical Lipschitz estimates, which will allow us to derive metric entropy complexity bounds for the class of densities induced by any variational class $\mathcal{F}$ of velocity fields. A similar approach for obtaining convergence rates has been used before in the context of inverse problems, where Lipschitz properties for the 'forward map' permit to bound the metric entropy of the observed regression functions; see, e.g., Nickl et al. (2020); Giordano and Nickl (2020); Agapiou and Wang (2021).

A key statistical convergence rate result

To begin, we will derive a statistical convergence result in Theorem 5 which regards so-called *sieved* maximum likelihood estimators—that is, MLEs which are taken over growing approximating classes as the number of samples n increases. This convergence result follows (up to minor adaptations) from classical results in Chapter 10 of van de Geer (2000); we nevertheless include it here since it plays a key role in our derivations. Let again $Z_1, \dots, Z_n \in D$ be i.i.d. samples from some P_0 with Lebesgue density p_0. Suppose that $(\mathcal{P}_n)_{n\geqslant 1}$ is a sequence of approximating classes of densities on D. Then the sieved MLE is defined by

$$\hat{p}_n = \arg\max_{p\in\mathcal{P}_n} \sum_{i=1}^n \log p(Z_i).$$

In the following we will require the bracketing metric entropy for functions on $D = [0,1]^d$. This notion of entropy is different from the standard metric entropy due to its 'joint' L^2 and pointwise structure, but it can straightforwardly be compared to the L^∞ metric entropy; see Lemma 22 in the appendix. Let μ be a Borel measure on $D = [0,1]^d$ and recall the shorthand $L^2(\mu) := L^2(D, \mu)$.

Definition 4 (Bracketing metric entropy) *Let $\mathcal{G}$ be a class of real-valued functions on D. Then let $N_B(\mathcal{G}, L^2(\mu), \tau)$ be the smallest value of N such that there exist pairs of functions $\{g_j^L, g_j^U\}_{j=1}^N$ with $\|g_j^L - g_j^U\|_{L^2(\mu)} \leqslant \tau$ such that for every $g \in \mathcal{G}$, there exists j with*

$$g_j^L \leqslant g \leqslant g_j^U \quad \text{on } D.$$

The $L^2(\mu)$-bracketing metric entropy *of $\mathcal{G}$ is $H_B(\mathcal{G}, L^2(\mu), \tau) = \log N_B(\mathcal{G}, L^2(\mu), \tau)$.*

Given the approximating classes $\mathcal{P}_n$, it turns out that the key measure of statistical complexity featuring in our convergence rate result is the bracketing metric entropy of the *square root densities* induced by $\mathcal{P}_n$. Specifically, let us fix some element $p_n^* \in \mathcal{P}_n$ and denote

$$\mathcal{Q}_n^* := \Big\{ \sqrt{\frac{p + p_n^*}{2}} : p \in \mathcal{P}_n \Big\}.$$

While p_n^* can be chosen arbitrarily, typically one aims to choose it to be some 'best approximation' of p_0 within the class $\mathcal{P}_n$. We then define the *bracketing metric entropy integral*

$$I_B(\mathcal{P}_n, R, p_n^*) := R + \int_0^R H_B^{1/2}(\mathcal{Q}_n^*, L^2(p_n^*), \tau) d\tau \qquad \text{for all } R > 0. \tag{2.12}$$

As before, we will use the notation $\Psi : (0, \infty) \to \mathbb{R}$ for an upper bound satisfying $\Psi(R) \geqslant I_B(\mathcal{P}_n, R, p_n^*)$ for all $R > 0$.

Assumption 2.4 *Suppose that for some constants $0 < c < K$, we have that $p_0 \leqslant K$ and $p_n^* \geqslant c$ for all $n \geqslant 1$. Moreover, suppose that Ψ is such that $R \mapsto \Psi(R)/R^2$ is non-decreasing (for all $n \geqslant 1$).*

Theorem 5 (cf. Theorem 10.13 in van de Geer (2000)) *Suppose that p_0, p_n^*, and $\mathcal{P}_n$ satisfy Assumption 2.4 for some $0 < c < K$. There is a constant $C > 0$ depending only on c and K such that for any $n \geqslant 1$ and $\delta_n > 0$ satisfying*

$$\sqrt{n}\delta_n^2 \geqslant C\Psi(\delta_n) \tag{2.13}$$

and any $\delta \geqslant \delta_n$, we have the concentration inequality

$$P_0^n\Big(h(\hat{p}_n, p_0) \geqslant C\big[h(p_n^*, p_0) + \delta\big]\Big) \leqslant C \exp\Big(-\frac{n\delta^2}{C}\Big).$$

Theorem 5 is a variant of (van de Geer, 2000, Theorem 10.13). We provide the argument in Appendix A, indicating in particular the required modifications compared to (van de Geer, 2000, Theorem 10.13).

Analytical estimates for ODE-based measure transport

In order to utilize the convergence result from Theorem 5, we need a number of 'stability' properties which relate the distance between two ODE velocity fields to their corresponding transport maps and pullback distributions. Recall the notation $T^f = (X^f(\cdot, 1))$ for the time-one flow map. The following lemma shows that the map $f \mapsto T^f$ is locally Lipschitz continuous as a mapping from $C^1(\Omega)$ to $C^1(D)$, on sets of velocity fields which are uniformly bounded in an appropriate sense.

Lemma 6 *Fix $r > 0$. Then for all velocity fields f, $g \in \mathcal{V}$ (see (2.4)) satisfying*

$$\max\Big\{ \|f\|_{C^1(\Omega)}, \sup_{t \in [0,1]} |\nabla_x f(\cdot, t)|_{\mathrm{Lip(D)}} \Big\} \leqslant r, \qquad \|g\|_{C^1(\Omega)} \leqslant r, \tag{2.14}$$

it holds with $C := \max\{e^{dr}, \frac{re^{3dr} + 2de^{2dr}}{2\sqrt{dr}}\}$ that

$$\|T^f - T^g\|_{C^1(D)} \leqslant C\|f - g\|_{C^1(\Omega)}.$$

The proof relies on Grönwall-type estimates for ODEs, and can be found in Appendix A. The next lemma shows that the L^∞-norm between two pullback densities is bounded by the C^1-norm between their corresponding transport maps. Again, we defer the proof to Appendix A.

Theorem 7 *Let $\rho : D \to [0, \infty)$ be a Lipschitz probability density and $T, G : D \to D$ two diffeomorphisms. Let $\lambda_1(x) \geqslant \cdots \geqslant \lambda_d(x) > 0$ and $\eta_1(x) \geqslant \cdots \geqslant \eta_d(x) > 0$ be the singular values of $\nabla T(x)$ and $\nabla G(x)$ respectively. Then, it holds that*

$$\|T^\sharp \rho - G^\sharp \rho\|_{C(D)} \leqslant \|T - G\|_{C^1(D)} \Big(|\rho|_{\mathrm{Lip(D)}} \|T\|^d_{C^1(D)} + \tilde{C} d^2 \|\rho\|_{C(D)} \Big),$$

where

$$\tilde{C} := \sup_{x \in D} \frac{\exp\left(\sum_{i=1}^d \frac{|\lambda_i(x) - \eta_i(x)|}{\lambda_d(x)} \right) \prod_{i=1}^d \lambda_i(x)}{\min\{\lambda_d(x), \eta_d(x)\}}. \tag{2.15}$$

See Appendix A for the proof. Lemma 6 and Theorem 7 together yield that the map $f \mapsto (T^f)^\sharp \rho$ is locally Lipschitz continuous on classes of velocity fields for which the constants r and $\tilde{C}$ from (2.14) and (2.15) can be controlled uniformly. From (2.15), we can see that this requires uniform control over the largest and smallest singular values of the Jacobian matrix of ∇T^f. The next result states that for classes $\mathcal{F}$ which are bounded in $C^1(\Omega)$-norm, such uniform bounds hold true.

Theorem 8 *Let $\mathcal{F} \subseteq C^1(\Omega, \mathbb{R}^d)$ such that $\sup_{f \in \mathcal{F}} \|f\|_{C^1(\Omega)} =: M < \infty$. Then for all $f \in \mathcal{F}$*

$$\sup_{x \in D} \|\nabla(T^f)(x)\|_2 \leqslant 1 + dMe^{dM},$$

where $\|\cdot\|_2$ denotes the $\mathbb{R}^d \to \mathbb{R}^d$ operator norm. Consequently, the largest and smallest singular values $\lambda_1(x)$ and $\lambda_d(x)$ of $\nabla(T^f)(x)$, are respectively upper and lower bounded as

$$\sup_{f \in \mathcal{F}} \sup_{x \in D} \lambda_1^f(x) \leqslant 1 + dMe^{dM} \qquad \text{and} \qquad \inf_{f \in \mathcal{F}} \inf_{x \in D} \lambda_d^f(x) \geqslant \left(1 + dMe^{dM}\right)^{-1}.$$

Proof of Theorem 2

Given a class $\mathcal{F}_n \subseteq \mathcal{V}$ of velocity fields, define the set of corresponding pullback distributions as

$$\mathcal{P}_n := \{(T^f)^\sharp \rho : f \in \mathcal{F}_n\}.$$

Then, by definition, an ODE-MLE $\hat{f}$ as in (2.5) satisfies

$$(T^{\hat{f}})^\sharp \rho \in \arg\max_{p \in \mathcal{P}_n} \sum_{i=1}^n \log p(Z_i),$$

i.e., the pullback distribution $(T^{\hat{f}})^\sharp \rho$ constitutes an MLE over $\mathcal{P}_n$. Thus our strategy will be to verify that Theorem 5 can be suitably applied with approximating sieve classes $\mathcal{P}_n$.

Step 1: Uniform bounds on pullback densities. We prove that all densities in $\mathcal{P}_n$ are uniformly upper and lower bounded. First note that (2.7) and Theorem 8 imply the existence of constants $0 < C_1 < C_2 < \infty$ solely depending on r in (2.8) such that for all $f \in \mathcal{F}_n$ and $x \in D$, the spectrum $\sigma(\nabla T^f(x))$ of the Jacobian matrix $\nabla T^f(x) \in \mathbb{R}^{d\times d}$ satisfies

$$\sigma(\nabla T^f(x)) \in [C_1, C_2].$$

Using the change-of-variables formula

$$(T^f)^\sharp \rho(x) = \rho(T^f(x)) \det \nabla(T^f(x)),$$

and since $\kappa < \rho(x) < K$, we thus find that there exists $L = L(r,\kappa) > 0$ such that

$$\inf_{f\in\mathcal{F}_n} \inf_{x\in D} (T^f)^\sharp \rho(x) \geqslant L. \tag{2.16}$$

Similarly there exists $U = U(r, K)$ such that

$$\sup_{f\in\mathcal{F}_n} \sup_{x\in D} (T^f)^\sharp \rho(x) \leqslant U. \tag{2.17}$$

In particular, for any $f^* \in \mathcal{F}_n$, denoting $p^* = (T^{f^*})^\sharp\rho$, it holds that $p^* \geqslant L$ uniformly in D. Hence the assumption on p^* in Theorem 5 is fulfilled with $c = L$.

Step 2: Bounding the covering number via Lipschitz properties. Fix $f^* \in \mathcal{F}_n$ and denote again $p^* = (T^{f^*})^\sharp\rho$. Define

$$\mathcal{Q}_n^* := \left\{\sqrt{\frac{p+p^*}{2}} : p \in \mathcal{P}_n\right\} = \left\{\sqrt{\frac{(T^f)^\sharp\rho + p^*}{2}} : f \in \mathcal{F}_n\right\}.$$

Our goal is to bound the bracketing covering number $N_B(\mathcal{Q}_n^*, L^2(p^*), \tau)$; see Definition 4. To this end, we interpret $\mathcal{Q}_n^*$ as the image of $\mathcal{F}_n$ under two maps Φ_1, Φ_2 via

$$f \quad \overset{\Phi_1}{\mapsto} \quad (T^f)^\sharp\rho \quad \overset{\Phi_2}{\mapsto} \quad \sqrt{\frac{(T^f)^\sharp\rho + p^*}{2}},$$

and we now show that both maps are Lipschitz continuous.

We start with Φ_1. Recall that $\mathcal{F}_n$ is bounded in the sense (2.8), and $\rho : D \to \mathbb{R}$ is Lipschitz continuous. Thus Lemma 6, Theorem 7, as well as the bounds on the singular values of ∇T_f from Theorem 8 imply that there are constants $C_3 = C_3(r, d, |\rho|_{\mathrm{Lip}(D)}, K) > 0$ and $C_4 = C_4(r, d, K) > 0$ (cf. (2.7), (2.8), and (2.17)) such that for all $f, g \in \mathcal{F}_n$

$$\|(T^f)^\sharp\rho - (T^g)^\sharp\rho\|_{C(D)} \leqslant C_3\|T^f - T^g\|_{C^1(D)} \leqslant C_3C_4\|f-g\|_{C^1(\Omega)}.$$

That is,

$$\Phi_1 := \begin{cases} \mathcal{F}_n \subseteq C^1(\Omega) \to C(D) \\ f \mapsto (T^f)^\sharp\rho \end{cases} \qquad \text{has Lipschitz constant } C_3C_4. \tag{2.18}$$

To treat Φ_2, note that the uniform lower bound (2.16) also yields a lower bound for the corresponding square root densities:

$$\inf_{q\in\mathcal{Q}_n^*} \inf_{x\in D} q(x) = \inf_{f\in\mathcal{F}_n} \inf_{x\in D} \sqrt{\frac{(T^f)^\sharp\rho + (T^{f^*})^\sharp\rho}{2}(x)} \geqslant \sqrt{L}.$$

Since $\sqrt{\cdot}$ is Lipschitz continuous on the interval $[\sqrt{L},\infty)$, it follows that for all $f,g\in\mathcal{F}_n$ and some $C_5=C_5(L)=C_5(r,\kappa)$

$$\left\|\sqrt{\frac{(T^f)^\sharp\rho+(T^{f^*})^\sharp\rho}{2}}-\sqrt{\frac{(T^g)^\sharp\rho+(T^{f^*})^\sharp\rho}{2}}\right\|_{C(D)}\leqslant C_5\|(T^f)^\sharp\rho-(T^g)^\sharp\rho\|_{C(D)}.$$

That is,

$$\Phi_2:=\begin{cases}\mathcal{P}_n\subseteq C(D)\to C(D)\\ p\mapsto\sqrt{\frac{p+p^*}{2}}\end{cases}\qquad\text{has Lipschitz constant } C_5. \tag{2.19}$$

Applying first Lemma 22 (noting that $p^*(D)=1$) and then Lemma 23 with (2.19) and (2.18), we obtain for all $\tau>0$

$$\begin{aligned}N_B(\mathcal{Q}_n^*,L^2(p^*),\tau)&\leqslant N\Big(\mathcal{Q}_n^*,C(D),\frac{\tau}{2}\Big)\\ &\leqslant N\Big(\mathcal{P}_n,C(D),\frac{\tau}{4C_5}\Big)\leqslant N\Big(\mathcal{F}_n,C^1(\Omega),\frac{\tau}{8C_3C_4C_5}\Big).\end{aligned}$$

Step 3: Metric entropy integral bounds. In order to be able to apply Theorem 5, we need to verify that the metric entropy bound assumption (2.9) of Theorem 2 implies the corresponding condition (2.13) in Theorem 5.

Without loss of generality, we may assume that $C_3C_4C_5\geqslant 1$ (by choosing these constants larger than 1). Then we obtain the following estimate for the bracketing entropy integral:

$$R+\int_0^R H_B^{1/2}\big(\mathcal{Q}_n^*,L^2(p^*),\tau\big)d\tau\leqslant R+\int_0^R H\big(\mathcal{F}_n,C^1(\Omega),\frac{\tau}{2C_3C_4C_5}\big)d\tau \tag{2.20}$$

$$\leqslant R+2C_3C_4C_5\int_0^{R/C_3C_4C_5} H\big(\mathcal{F}_n,C^1(\Omega),\tau\big)d\tau \tag{2.21}$$

$$\leqslant 2C_3C_4C_5\Psi(R). \tag{2.22}$$

Now let C_6 be a constant with the same value as the constant C from Theorem 5, and let us define $\tilde{\Psi}(R):=2C_3C_4C_5\Psi(R)$. Clearly, $R\mapsto\tilde{\Psi}(R)/R^2$ is still a non-decreasing function. Moreover, any $n\geqslant 1,\delta_n>0$ satisfying

$$\sqrt{n}\delta_n^2\geqslant 2C_3C_4C_5C_6\Psi(\delta_n)$$

will also fulfill

$$\sqrt{n}\delta_n^2\geqslant C_6\tilde{\Psi}(\delta_n).$$

Finally, we may therefore apply Theorem 5 to those values of n,δ_n with $\tilde{\Psi}$ as an upper bound, and we obtain that for any $\delta\geqslant\delta_n$,

$$P_0^n\Big(h\big((T^{\hat f})^\sharp\rho,p_0\big)\geqslant C_6\big[h\big((T^{f^*})^\sharp\rho,p_0\big)+\delta\big]\Big)\leqslant C_6\exp\Big(-\frac{n\delta^2}{C_6}\Big).$$

This completes the proof of (2.10).

Finally, the bound (2.11) for the mean squared error follows from a standard integration argument (cf. the proof of Lemma 2.2 in van de Geer (2001)). Let us use the shorthand

$\hat{h} = h((T^{\hat{f}})^\sharp \rho, p_0)$ and $h = h((T^{f^*})^\sharp \rho, p_0)$. Then (2.10) implies that $P_0^n(\hat{h}^2 \geqslant 2C(h^2 + \delta^2)) \leqslant P_0^n(\hat{h} \geqslant C(h + \delta)) \leqslant C\exp(-\frac{n\delta^2}{C})$ for all $\delta \geqslant \delta_n$. Moreover, by assumption $\sqrt{n}\delta_n^2 \geqslant C\Psi(\delta_n) \geqslant C\delta_n$, such that $\delta_n \geqslant C/\sqrt{n}$. Thus, we obtain that

$$\mathbb{E}_{P_0}^n[\hat{h}^2] = \int_0^\infty P_0^n(\hat{h}^2 \geqslant t)dt \leqslant 2C^2(h^2 + \delta_n^2) + \int_{t > 2C^2(h^2+\delta_n^2)} P_0^n(\hat{h}^2 \geqslant t)dt.$$

The second term is further bounded by

$$\begin{aligned}\frac{1}{2C^2}\int_{\delta^2 > \delta_n^2} P_0^n(\hat{h}^2 \geqslant 2C^2(h^2 + \delta^2))d\delta^2 &\leqslant \frac{1}{2C^2}\int_{\delta^2 > \delta_n^2} C\exp\big(-\frac{n\delta^2}{C}\big)d\delta^2 \\ &\leqslant \frac{1}{2n}\exp\big(-\frac{n\delta_n^2}{C}\big) \leqslant \frac{1}{2ne}.\end{aligned}$$

■

3. Results for C^k Velocity Fields

We now apply the general theory from the preceding section to a canonical nonparametric density estimation setting, where the data-generating density p_0 is assumed to belong to a class of k-times differentiable functions. Again, let $D = [0, 1]^d$ denote the unit cube. Then, given some integer $k \geqslant 1$ and constants $0 < L_2 \leqslant L_1 < \infty$, let us introduce the following class of upper and lower bounded C^k probability densities on D:

$$\mathcal{M}(k, L_1, L_2) = \Big\{p \in C^k(D) : \inf_{x \in D} p(x) \geqslant L_2, \quad \|p\|_{C^k(D)} \leqslant L_1, \quad \int_D p(x)dx = 1\Big\}. \tag{3.1}$$

For Theorem 2 to yield ‘fast’ rates of convergence, it is essential to choose the variational class $\mathcal{F}$ of velocity fields appropriately. A canonical possibility is to choose the class ‘as small as possible’ such that there exists an element $f^* \in \mathcal{F}$ with $(T^{f^*})^\sharp \rho = p_0$. This leads to the following natural question: Given some density $p_0 \in \mathcal{M}(k, L_1, L_2)$, what is the regularity one can expect a velocity field coupling ρ with p_0 to have? Our first result of this section, Theorem 9, proves that there exists a velocity field which lies in $C^k \cap \mathcal{V}$ and exactly couples ρ with p_0. In other words, there exists a velocity field which is *at least as regular* as the densities which it couples.

Our result follows from proving the C^k-regularity of one specific velocity field, which is constructed using the Knothe–Rosenblatt (KR) transport (Santambrogio, 2015; Villani, 2009). Roughly speaking, the KR transport map is the triangular and monotone map $T : D \to D$ which couples ρ and p_0. By triangular, we mean that the l-th component function only depends on the first l variables $(x_1, \dots, x_l)$,

$$T(x) = \begin{bmatrix} T_1(x_1) \\ T_2(x_1, x_2) \\ \vdots \\ T_d(x_1, \dots, x_d) \end{bmatrix}, \qquad x \in D,$$

and by monotone we mean that each component function T_l is strictly increasing with respect to its last argument x_l. It is well known that the KR map is unique up to coordinate ordering, and that T actually possesses an explicit construction in terms of the CDFs of the marginal conditional densities of p_0 and ρ. We refer the reader to (Santambrogio, 2015, Chap. 2.3) or Zech and Marzouk (2022a) for this construction and for standard properties of KR maps.

C^k-regularity of a 'straight-line' velocity field

Given the KR map T, we now define our candidate velocity field which we will later prove to satisfy C^k-regularity. First, let $G : D\times[0,1] \to D\times[0,1]$ be the 'straight-line interpolation' (giving rise to an analogue of the displacement interpolation between ρ and p_0 (McCann, 1997)) between the identity map and T,

$$G_t(x) := tT(x) + (1-t)x. \tag{3.2}$$

In Marzouk et al. (2023), it is established that $G_t : D \to D$ is invertible for each $t \in [0,1]$. Then, let

$$F : D \times [0,1] \to D, \qquad F(x,t) = G_t^{-1}(x),$$

based on which we define the following velocity field

$$f^{\Delta}_{p_0}(y,s) = T(F(y,s)) - F(y,s), \qquad \forall (y,s) \in D \times [0,1]. \tag{3.3}$$

Then, the flow induced by $f^{\Delta}_{p_0} : D\times[0,1] \to D$ has the straight-line trajectories $X^{f^{\Delta}_{p_0}}(x,t) = tT(x) + (1-t)x$, and indeed pushes p_0 to ρ; see Marzouk et al. (2023) for details.

In order to state the next result, we require the following mild assumption on the reference density.

Assumption 3.1 *Let $\rho \in C^k(D)$ be uniformly lower bounded by $\kappa > 0$. Moreover, suppose that ρ factorizes into k-smooth marginal distributions; that is, there exist univariate densities $\rho_l \in C^k([0,1])$ such that $\rho(x) = \prod_{l=1}^{d} \rho_l(x_l)$.*

This assumption allows for many natural choices of reference distributions on the unit cube, such as the uniform distribution, or truncated Gaussian distributions with diagonal covariance matrix. We also note that the assumption of ρ being a product distribution is made for convenience, and can be relaxed at the expense of further technicalities; see Remark 12 for further details.

Theorem 9 *Let $k \geqslant 1$, and let ρ be some reference density satisfying Assumption 3.1. Moreover, suppose that $p_0 \in \mathcal{M}(k, L_1, L_2)$. Let $T : [0,1]^d \to [0,1]^d$ and $f^{\Delta}_{p_0}$ respectively denote the KR map and the straight-line velocity field between p_0 and ρ (constructed above). Then:*

1. *It holds that $f^{\Delta}_{p_0} \in C^k(\Omega)$ with $\|f^{\Delta}_{p_0}\|_{C^k(\Omega)} \leqslant C$, for some $C > 0$ that depends only on ρ, k, d, L_1, L_2.*
2. *For $g^{\Delta}_{p_0} : \Omega \to \mathbb{R}^d$ defined as*

$$[g^{\Delta}_{p_0}(x,s)]_j := \frac{(f^{\Delta}_{p_0}(x,s))_j}{x_j(1-x_j)}, \qquad j = 1,\ldots,d \tag{3.4}$$

it holds that $g_{p_0}^{\Delta} \in C^k(\Omega)$*, and there exists another constant* $\tilde{C} = \tilde{C}(d, L_1, L_2)$*, such that* $\|g_{p_0}^{\Delta}\|_{C^k(\Omega)} \leqslant \tilde{C}$*. In particular it holds that* $f_{p_0}^{\Delta} \in \mathcal{V}$ *(cf. (2.4)), i.e., the normal component* $f_{p_0}^{\Delta}(x,t) \cdot \nu_x$ *vanishes at every point* $(x,t) \in \partial D \times [0,1]$*.*

The above result shows that for C^k-regular target densities p_0, the velocity field $f_{p_0}^{\Delta}$ inherits C^k regularity. Crucially, Part 2 of the theorem also shows that $f_{p_0}^{\Delta}$ is an 'admissible' velocity field whose normal component vanishes on the 'tubular' boundary $\partial D \times [0,1]$. The proof uses certain anisotropic regularity results for KR maps developed in Wang and Marzouk (2022), along with technical results showing that this anisotropic regularity is preserved under composition and inversion of maps. In order to deduce the boundary properties in Part 2, we then use a so-called Hardy inequality. For the full proof, we refer to Appendix B.1.

Convergence theorem for estimators over C^k-classes

We are now ready to state the main theorem of this section, which gives a convergence rate for ODE-MLEs whenever $p_0 \in \mathcal{M}(k, L_1, L_2)$. For $r > 0$, define

$$\mathcal{F}(r) := \left\{ f \in C^k(\Omega, \mathbb{R}^d) : \|f\|_{C^k} \leqslant r \right\} \cap \mathcal{V}. \tag{3.5}$$

Theorem 10 *Let* $k > d/2 + 3/2$*,* $0 < \gamma < k - d/2 - 3/2$*,* $0 < L_2 \leqslant L_1 < \infty$*, and suppose* ρ *satisfies Assumption 3.1. Then, there exist constants* $r = r(k, L_1, L_2) > 0$ *and* $C = C(k, L_1, L_2) > 0$ *such that for any* $p_0 \in \mathcal{M}(k, L_1, L_2)$*, the velocity field* $\hat{f}$ *maximizing the objective (2.5) over* $\mathcal{F}(r)$ *satisfies*

$$\mathbf{E}_{P_0}^n\left[h^2((T^{\hat{f}})^{\sharp}\rho, p_0)\right] \leqslant Cn^{-\eta}, \qquad \text{with } \eta = \frac{2(k-1-\gamma)}{2(k-1-\gamma)+d+1} > 0.$$

The proof of Theorem 10 can be found in Appendix B.1. In essence, the result follows from an application of the general Theorem 2 with $\mathcal{F} = \mathcal{F}_n = \mathcal{F}(r)$ and with 'approximating' velocity field $f^* = f_n^* = f_{p_0}^{\Delta}$ given by Theorem 9, using also classical metric entropy estimates for C^k classes. Note that the approximation error $h((T^{f_n^*})^{\sharp}\rho, p_0)$ from (2.11) then vanishes, such that there is no need for $\mathcal{F}$ to depend on n.

Remark 11 (On the parameterization of $\mathcal{F}(r)$) The choice of $\mathcal{F} = \mathcal{F}(r)$ underlying Theorem 10 is informed by the regularity that we can expect a velocity between two C^k probability densities to have. In practical implementations, of course, one cannot employ the full class $\mathcal{F}(r)$ and must resort to a subclass of $\mathcal{F}(r)$ described by finitely many parameters, whose size would typically increase as n grows. One example are neural network-based parameterizations, which will be discussed in Section 4. Alternatively, one could use classical approximating classes such as polynomials, wavelets, or splines (Triebel, 2008; DeVore and Lorentz, 1993).

Typically, those approximating classes will *not* satisfy that the normal component of $f(x,t)$ vanishes at the boundary. In order to enforce this property, one can employ a boundary cut-off construction where one first chooses an approximating class (e.g., polynomials, wavelets, splines, neural networks) and then multiplies the field's j-th component by the 'cut-off' function $x_j(1-x_j)$ for all $j \in \{1, \ldots, d\}$. The fact that such a construction still yields a sufficiently rich class $\mathcal{F}(r)$ is implied by the regularity result in Theorem 9, part 2: Indeed, the theorem implies that the triangular velocity field $f^{\Delta}_{p_0}$ may be expressed as the product of some C^k-velocity field $\tilde{f} \in C^k(\Omega, \mathbb{R}^d)$ with the above component-wise cutoff:

$$[f^{\Delta}_{p_0}]_j = \tilde{f}_j \cdot x_j(1 - x_j), \qquad \forall j \in \{1, \dots, d\}.$$

This is precisely the construction that will be used to construct the neural network-based 'ansatz space' in Section 4 below.

Remark 12 (On Assumption 3.1) While our general Theorem 2 only required ρ to be Lipschitz continuous (and lower bounded), the present results hold under slightly more stringent requirements on ρ. The C^k regularity is crucial for guaranteeing the existence of a C^k transport map between ρ and p_0. In contrast, the assumption that ρ factorizes into its marginal distributions can be relaxed at the expense of further technicalities. An inspection of the proofs reveals that the factorization property is only needed in the proof of Theorem 9 because we cite a regularity result from Wang and Marzouk (2022) for Knothe–Rosenblatt maps which uses this assumption. The latter result, however, can be generalized to general C^k-smooth reference densities.

4. Neural ODEs: Neural Network Parameterization of Velocity Fields

In this section, we study the case where the underlying velocity field is parameterized by a neural network class, i.e., neural ODEs (Chen et al., 2018; Grathwohl et al., 2018). Like in Section 3, our strategy will be to apply Theorem 2, this time to classes of neural networks. To do so, we will separately study the metric entropy rates and the "best approximation" properties of the neural network classes defined below.

We now introduce our notation for neural network classes with ReLU^m activation function. Let $\eta_1(x) = \max\{x, 0\}$ being the ReLU activation function, and $\eta_m(x) = \max\{x, 0\}^m$ be the ReLU^m activation function.

Definition 13 *Let $m \geqslant 1$ and fix $d_1, d_2 \geqslant 1$. Then, the class of ReLU^m networks mapping from $[0,1]^{d_1}$ to $\mathbb{R}^{d_2}$, with height L, width W, sparsity constraint S, and norm constraint B, is defined by*

$$\begin{aligned}
&\Phi^{d_1,d_2}(L, W, S, B) = \Big\{ \big(W^{(L)}\eta_m(\cdot) + b^{(L)}\big) \circ \cdots \circ \big(W^{(1)}\eta_m(\cdot) + b^{(1)}\big) : \\
&W^{(L)} \in \mathbb{R}^{1\times W}, b^{(L)} \in \mathbb{R}^{d_2}, W^{(1)} \in \mathbb{R}^{W\times d_1}, b^{(1)} \in \mathbb{R}^{W}, W^{(l)} \in \mathbb{R}^{W\times W}, \\
&b^{(l)} \in \mathbb{R}^{W} (1 < l < L), \sum_{l=1}^{L} \big(\|W^{(l)}\|_0 + \|b^{(l)}\|_0\big) \leqslant S, \max_{1 \leqslant l \leqslant L} \big(\|W^{(l)}\|_{\infty,\infty} \vee \|b^{(l)}\|_\infty\big) \leqslant B \Big\}.
\end{aligned}$$

We refer to an element of $\Phi^{d_1,d_2}(L, W, S, B)$ as a ReLU^m network. *For any index $1 \leqslant l \leqslant L$, we write F_l for the network composed of the first l-layers, that is,*

$$F_l = \big(W_F^{(l)}\eta_m(\cdot) + b_F^{(l)}\big) \circ \cdots \circ \big(W_F^{(1)}\eta_m(\cdot) + b_F^{(1)}\big).$$

We refer to such networks as $l-$ReLU^m networks. We use $\Phi_l^{d_1,d_2}(L, W, S, B)$ to denote all such l-layer networks.

Since we will need the $C^1(\Omega)$ metric entropy of the above network classes, we shall also need the gradient space $\nabla\Phi^{d_1,d_2}(L,W,S,B)$. Note that for any $1 \leqslant l \leqslant L-1$, any l-ReLU^2 network $F_l \in \Phi_l^{d_1,d_2}(L,W,S,B)$ is a map from $\mathbb{R}^{d_1}$ to $\mathbb{R}^W$. For any $1 \leqslant j \leqslant W$, we use $F_{l,j}$ to denote the $j-$th component. Then, we may write $F_l(x)$ and its Jacobian $\nabla F_l(x)$ as follows:

$$F_l(x) = [F_{l,1}(x), F_{l,2}(x), \ldots, F_{l,W}(x)]^T,$$

$$\nabla F_l(x) = \begin{bmatrix} \frac{\partial}{\partial x_1}F_{l,1}(x) & \frac{\partial}{\partial x_2}F_{l,1}(x) & \cdots \frac{\partial}{\partial x_{d_1}}F_{l,1}(x) \\ \vdots & \ddots & \\ \frac{\partial}{\partial x_1}F_{l,W}(x) & \ddots & \frac{\partial}{\partial x_{d_1}}F_{l,W}(x) \end{bmatrix}$$

When $l = L$, F_L maps $\mathbb{R}^{d_1}$ to $\mathbb{R}^{d_2}$ and the Jacobian can be written as a $d_2 \times d_1$ matrix.

4.1 Metric Entropy Rates

In order to apply Theorem 2, we need to control the $C^1([0,1]^{d_1})$-metric entropy of these parametric classes. We now present our results on entropy rates of the NN class $\Phi^{d_1,d_2}(L,W,S,B)$ in $C^1([0,1]^{d_1})$ norm. Our results are similar to those in Lu et al. (2021) except that we use ReLU^2 networks, in place of the ReLU^3 networks considered there.

The following theorem gives an upper bound for the metric entropy rate of $\Phi^{d_1,1}(L,W,S,B)$, i.e., the case where $d_2 = 1$. The subsequent corollary will then deal with the case of multi-dimensional outputs.

Theorem 14 *Let $d_1 \in \mathbb{N}$. Consider the ReLU^2 network space $\Phi^{d_1,1}(L,W,S,B)$ with $L = \mathcal{O}(1), W = \mathcal{O}(N), S = \mathcal{O}(N)$ and $B = \mathcal{O}(N)$. Then*

$$H(\Phi^{d_1,1}(L,W,S,B), C^1([0,1]^{d_1}), \tau) = \mathcal{O}\big(N\log(\tau^{-1}) + N\log N\big).$$

Proof The proof of this theorem is based on translating covering numbers of the NN parameter space (in l^∞ norm) into covering numbers of the NN function space (in C^1 norm). For this purpose, we shall need Lipschitz-type estimates from the NN parameter space into the NN function space and its gradient space, which are respectively given by Lemma 30 and Lemma 32.

We first fix a sparsity pattern (i.e., the locations of the non-zero entries are fixed) and let $k = L$ in Lemma 30 and Lemma 32. Following the arguments in the proof of Lemma 3 in Suzuki (2019), we get the following upper bound for the covering number with respect to $C^1([0,1]^{d_1})$ norm:

$$\left(\frac{\tau}{\max\{N_L W^{2^{L-1}-1}(B \vee d_1)^{2^L+1}, A_L W^{2^{L-1}-1}(B \vee d_1)^{2^L}\}}\right)^{-S},$$

where A_L, N_L are the constants from Lemmata 30 and 32, which only depend on L. Note that the number of possible sparsity patterns is upper bounded by $\binom{(W+1)^L}{S} \leqslant (W+1)^{LS}$ (see Suzuki (2019); Schmidt-Hieber (2020); Lu et al. (2021)). Plugging in the magnitudes for the network parameters, we get the following metric entropy bound:

$$
\begin{aligned}
&H(\Phi^{d_1,1}(L,W,S,B), C^1([0,1]^{d_1}), \tau) = \log N(\Phi^{d_1,1}(L,W,S,B), C^1[0,1]^{d_1}, \tau) \\
&\leqslant \log\left[(W+1)^{LS}\left(\frac{\tau}{\max\{N_L W^{2^{L-1}-1}(B\vee d_1)^{2^L+1}, A_L W^{2^{L-1}-1}(B\vee d_1)^{2^L}\}}\right)^{-S}\right] \\
&\lesssim \max\Big\{S\log\Big[\tau^{-1}(W+1)^L N_L W^{2^{L-1}-1}(B\vee d_1)^{2^L+1}\Big], \\
&\qquad\qquad S\log\Big[\tau^{-1}(W+1)^L A_L W^{2^{L-1}-1}(B\vee d_1)^{2^L}\Big]\Big\} \\
&\lesssim S\Big[\log(\tau^{-1}) + 2^L\log(W(B\vee d_1))\Big] = \mathcal{O}\Big(N\log(\tau^{-1}) + N\log N\Big).
\end{aligned}
$$

■

For the purpose of modeling velocity fields as neural networks, we need to consider the above neural network classes with $d_1 = d+1$ and $d_2 = d$, i.e., as mappings from $\Omega = [0,1]^{d+1}$ to $\mathbb{R}^d$; this entropy rate is obtained by a tensorizing argument.

Corollary 15 *Let $d \geqslant 1$ be fixed and let $N \geqslant d$ be sufficiently large. Consider the ReLU^2 network class $\Phi^{d+1,d}(L,W,S,B)$ with $L = \mathcal{O}(1)$, $W = \mathcal{O}(N)$, $S = \mathcal{O}(N)$, and $B = \mathcal{O}(N)$. Then, the metric entropy satisfies*

$$H(\Phi^{d+1,d}(L,W,S,B), C^1(\Omega), \tau) = \mathcal{O}\Big(N\log(\tau^{-1}) + N\log N\Big).$$

Proof Let $\phi = [\phi_1,\ldots,\phi_d]^T \in \Phi^{d+1,d}(L,W,S,B)$. Then for each j it holds that $\phi_j \in \Phi^{d+1,1}(L,W,S,B)$ with $L = \mathcal{O}(1)$ and $W,S,B = \mathcal{O}(N)$.

For $j = 1,\ldots,d$, let $\{\psi_j^m\}_{m=1}^{M_j}$ be a τ-covering of the j-th coordinate. We now construct a covering set of $\Phi^{d+1,d}(L,W,S,B)$ by taking the product set $\Psi = \{\psi_1^m\}_{m=1}^{M_1} \times \cdots \times \{\psi_d^m\}_{m=1}^{M_d}$ To show that Cartesian product is indeed a covering set, note that for any member $\phi = [\phi_1,\ldots,\phi_d]^T \in \Phi^{d+1,d}(L,W,S,B)$, we can find $\psi = [\psi_1^{m_1},\ldots,\psi_d^{m_d}]^T$ such that $\|\phi_j - \psi_j^{m_j}\|_{C^1} \leqslant \tau$, where $1 \leqslant m_j \leqslant M_j$. It is then not hard to verify $\|\phi - \psi\|_{C^1} \leqslant \tau$.

Assume $M_j \leqslant \tilde{M}$ for $1 \leqslant j \leqslant d$, then the covering number satisfies $|\Psi| \leqslant \tilde{M}^d$ and the metric entropy is upper bounded by $d\log\tilde{M}$. From Theorem 14, $\tilde{M}$ is upper bounded as $\mathcal{O}(N\log(\tau^{-1}) + N\log N)$ and since we take d to be a fixed constant, the metric entropy for $\Phi^{d+1,d}(L,W,S,B)$ is the same asymptotically. ■

4.2 Approximation Theory

The goal of this section is to show that functions $f \in C^k(\Omega)$ can be efficiently approximated by neural networks of a certain architecture. Recall from the general Theorem 2 that we not only need our approximating NN class to be able to approximate the target function $f : \Omega \to \mathbb{R}^d$ in the $C^1(\Omega)$-norm, but also require its (spatial) gradient to be Lipschitz continuous.

Approximation results for C^k functions on compact domains with neural networks are by now standard, e.g., Pinkus (1999) or the more recent works Yarotsky (2017); Yarotsky and

Zhevnerchuk (2020). However, the specific statement we require, which involves the uniform approximation of some function and its (higher) derivatives, appears not to be readily available in the literature. We therefore provide a full proof in Appendix D. The argument leverages a widely recognized technique, first introduced in Mhaskar and Micchelli (1992); Mhaskar (1993), based on spline approximation. In principle, better convergence rates than those stated below are theoretically achievable with respect to the number of trainable parameters; see, for example, Pinkus (1999, Section 6) or Yarotsky and Zhevnerchuk (2020) regarding uniform approximation. However, such approximations typically result in neural networks that depend non-continuously on the function being approximated (DeVore et al., 1989), thus limiting their practical relevance.

The results that are directly related to our setting are the following theorem and corollary. Their proofs can be found in Appendix D.

Theorem 16 *Let k, d_1, $m \in \mathbb{N}$ and $k+1 \leqslant m$. Then there exists $C = C(d_1, k, m)$ such that for all $f \in C^k([0,1]^{d_1}, \mathbb{R})$ and all $N \in \mathbb{N}$ there exists a ReLU^{m-1} neural network $\tilde{f} \in \Phi^{d_1,1}(L, W, S, B)$ with*

$$L \leqslant C, \qquad W \leqslant N, \qquad S \leqslant N, \qquad B \leqslant C\|f\|_{C([0,1]^{d_1})} + N^{1/d_1} \tag{4.1}$$

such that $\tilde{f} \in C^{m-2}([0,1]^{d_1}, \mathbb{R})$ and

$$\|f - \tilde{f}\|_{W^{r,\infty}([0,1]^{d_1})} \leqslant CN^{-\frac{k-r}{d_1}} |f|_{C^k([0,1]^{d_1})} \qquad \forall r \in \{0, \ldots, k\}. \tag{4.2}$$

The next corollary shows that the assumption $m > k$ in Theorem 16 can be dropped. We emphasize, however, that a ReLU^{m-1} network always belongs to $W^{m-1,\infty}$ but it generally does not belong to $W^{m,\infty}$. Consequently, the network approximation $\tilde{f} \in W^{k,\infty}$, where $k \geqslant m$ is permitted, constructed in the following corollary is rather specific. Moreover, we state the result in the more general case of approximating a function $f = (f_j)_{j=1}^{d_2} : [0,1]^{d_1} \to \mathbb{R}^{d_2}$ for some $d_2 \in \mathbb{N}$, which is how we will use it in the following.

Corollary 17 *Let k, d_1, d_2, $m \in \mathbb{N}$, and $m \geqslant 3$. Then there exists $C = C(d_1, d_2, k, m)$ such that for all $f \in C^k([0,1]^{d_1}, \mathbb{R}^{d_2})$ and all $N \in \mathbb{N}$ there exists a ReLU^{m-1} neural network $\tilde{f} \in \Phi^{d_1,d_2}(L, W, S, B)$ with*

$$L \leqslant C, \qquad W \leqslant N, \qquad S \leqslant N, \qquad B \leqslant C\|f\|_{C([0,1]^{d_1}, \mathbb{R}^{d_2})} + N^{1/d_1} \tag{4.3}$$

such that $\tilde{f} \in C^{m-2}([0,1]^{d_1}, \mathbb{R}^{d_2})$ and for all $j \in \{1, \ldots, d_2\}$

$$\|f_j - \tilde{f}_j\|_{W^{r,\infty}([0,1]^{d_1})} \leqslant CN^{-\frac{k-r}{d_1}} |f_j|_{C^k([0,1]^{d_1})} \qquad \forall r \in \{0, \ldots, k\}. \tag{4.4}$$

4.3 Statistical Convergence Rates for Neural ODEs

Ansatz Space

As elaborated in Section 2, we need to ensure that the velocity fields in $\mathcal{F}$ satisfy certain boundary conditions in order for the pullback distributions $(T^f)^\sharp \rho$, $f \in \mathcal{F}$, to be supported on the same domain D. Lemma 1, Theorem 9 together with Remark 11 suggest that a suitable ansatz space can be formed by multiplying the preceding neural network classes by 'component-wise' cutoff functions.

Definition 18 *Let $\chi_d(x_1, \dots x_d) : D \to D$ be given by*

$$\chi_d(x_1, \dots x_d) = [x_1(1-x_1), \dots, x_d(1-x_d)]^T.$$

Let $\otimes$ be the coordinate-wise multiplication of two vectors (of the same dimension). Then for any velocity field $f : \Omega = [0,1]^d \times [0,1] \to \mathbb{R}^d$, $f \otimes \chi_d$ yields a vector field on D with vanishing normal components at the boundary. Similarly, we let $\oplus$ denote coordinate-wise division of two vectors.

Definition 19 *We let*

$$\Phi^{d+1,d}_{ansatz}(L, W, S, B) := \Big\{ f^{NN}(x_1, \dots, x_d, t) \otimes \chi_d(x_1, \dots, x_d),\ f^{NN} \in \Phi^{d+1,d}(L, W, S, B) \Big\},$$

where $\Phi^{d+1,d}(L, W, S, B)$ is the class of ReLU^2 networks defined in Definition 13 and L, W, S, B are the respective network parameters. For $r \geqslant 0$, we further define the following bounded sparse neural network classes

$$\mathcal{F}_{NN}(L, W, S, B, r) = \Phi^{d+1,d}_{ansatz}(L, W, S, B) \cap \{f \in W^{2,\infty}(\Omega) : \|f\|_{W^{2,\infty}(\Omega)} \leqslant r\}. \tag{4.5}$$

Main statistical convergence result

Finally, we obtain the following nonparametric convergence rate for neural ODEs, by combining the preceding results about approximation and statistical complexity.

Theorem 20 *Fix an integer $k \geqslant 1$ and constants $0 < L_2 \leqslant L_1 < \infty$, and suppose ρ is a reference density satisfying Assumption 3.1. Then there exist parameter choices $L = \mathcal{O}(1)$, $W = \mathcal{O}(n^{\frac{d+1}{d+1+2(k-1)}})$, $S = \mathcal{O}(n^{\frac{d+1}{d+1+2(k-1)}})$, $B = \mathcal{O}(n^{\frac{d+1}{d+1+2(k-1)}})$, and $r = \mathcal{O}(1)$ such that for all $p_0 \in \mathcal{M}(k, L_1, L_2)$, the neural ODE estimator $\hat{f}$ given by (2.5) over the class of velocity fields $\mathcal{F}_{NN}(L, W, S, B, r)$ satisfies the convergence rate*

$$\mathbf{E}^n_{P_0}\big[h^2((T^{\hat{f}})^\sharp \rho, p_0)\big] \lesssim n^{-\frac{2(k-1)}{d+1+2(k-1)}} \log n.$$

Proof Our proof strategy will be to apply our general Theorem 2 to the neural network classes of velocity fields $\Phi^{d+1,d}_{\text{ansatz}}(L, W, S, B)$. To this end, we bound the approximation error (Step 1) and the metric entropy rates (Step 2) separately.

Step 1: Approximation error. Suppose that $p_0 \in \mathcal{M}(k, L_1, L_2)$. By Theorem 9, there exists a velocity field $f^\Delta \in C^k(\Omega) \cap \mathcal{V}$ such that $(T^{f^\Delta})^\sharp \rho = p_0$ and such that for any $i \in [d]$, the i-th component f^Δ_i vanishes 'linearly' at the boundaries, i.e., $\frac{f^\Delta_i}{x_i(1-x_i)} \in C^k(\Omega)$.

Let us now define the velocity field

$$f^*(x_1, \dots, x_d) = f^\Delta \oplus \chi_d = \Big(\frac{f^\Delta_1(x_1)}{x_1(1-x_1)}, \dots, \frac{f^\Delta_d(x_1, \dots, x_d)}{x_d(1-x_d)} \Big)^T.$$

Theorem 9 moreover implies that for any k, L_1, L_2 there exists some constant $\tilde{C}$ such that

$$\sup_{p_0 \in \mathcal{M}(k, L_1, L_2)} \|f^*\|_{C^k(\Omega, \mathbb{R}^d)} \leqslant \tilde{C}.$$

Note that f^* does not necessarily satisfy the same boundary-vanishing properties as f^Δ. By Corollary 17 with $d_1 = d+1$ and $d_2 = d$, there exists a constant $C_{d,k}$ such that for all $N \geqslant 1$ and with

$$L \leqslant C_{d,k}, W \leqslant N, S \leqslant N, B \leqslant C_{d,k}\|f\|_{C^k(\Omega)} + N^{1/(d+1)}, \tag{4.6}$$

there is a ReLU^2 neural network $\tilde{f} \in \Phi^{d+1,d}(L, W, S, B)$ with $\tilde{f} \in C^1(\Omega)$, satisfying the approximation properties

$$\|\tilde{f} - f^*\|_{C^1(\Omega)} \leqslant C_{d,k} d^{\frac{k-1}{d+1}} N^{-\frac{k-1}{d+1}} \|f^*\|_{C^k(\Omega)}, \tag{4.7}$$

and

$$\|\tilde{f} - f^*\|_{W^{2,\infty}(\Omega)} \leqslant C_{d,k} d^{\frac{k-2}{d+1}} N^{-\frac{k-2}{d+1}} \|f^*\|_{C^k(\Omega)}. \tag{4.8}$$

Later in the proof, we will make a choice of N which balances the approximation error analysed here with the metric entropy term analysed in Step 2.

Defining $\hat{f}^\Delta = \tilde{f} \otimes \chi_d$, it then follows from standard multiplication inequalities that

$$\|\hat{f}^\Delta - f^\Delta\|_{C^1(\Omega)} = \|\big(\tilde{f} - f^*\big) \otimes \chi_d\|_{C^1(\Omega)} \lesssim \|\tilde{f} - f^*\|_{C^1(\Omega)} \|\chi_d\|_{C^1(\Omega)} \lesssim N^{-\frac{k-1}{d+1}}.$$

Similarly, we see that $\|f^\Delta - \hat{f}^\Delta\|_{W^{2,\infty}(\Omega)} = \mathcal{O}(N^{-\frac{k-2}{d+1}})$. Thus, using the triangle inequality and the fact that $f^\Delta \in C^k(\Omega)$, it follows that for some $r > 0$,

$$\sup_{p_0 \in \mathcal{M}(k, L_1, L_2)} \|\hat{f}^\Delta\|_{C^1(\Omega)} + \|\hat{f}^\Delta\|_{W^{2,\infty}(\Omega)} \leqslant r.$$

In summary, we have now proved the existence of an approximating element

$$\hat{f}^\Delta \in \mathcal{F}_{\mathrm{NN}}(L, W, S, B, r) = \Phi^{d+1,d}_{\mathrm{ansatz}}(L, W, S, B) \cap \{f \in W^{2,\infty}(\Omega) : \|f\|_{W^{2,\infty}(\Omega)} \leqslant r\}$$

which approximates f^Δ at rate $\|\hat{f}^\Delta - f^\Delta\|_{C^1(\Omega)} = \mathcal{O}(N^{-\frac{k-1}{d+1}})$. In particular, we may now also deduce an approximation for the corresponding pullback densities in Hellinger distance. Indeed, using the Lipschitz estimates from Lemma 6, Theorem 7, Theorem 8 and Lemma 24, we obtain that

$$h((T^{\hat{f}^\Delta})^\sharp \rho, (T^{f^\Delta})^\sharp \rho) = h((T^{\hat{f}^\Delta})^\sharp \rho, p_0) = \mathcal{O}(N^{-\frac{k-1}{d+1}}).$$

Step 2: Metric entropy bound. Given $N \geqslant 1$, we now derive the required upper bound for the square-root metric entropy for the neural network class $\mathcal{F}_{NN}(L, W, S, B, r)$, again with the choices from (4.6). Later on, we will choose N to be of the same order as B, so let us assume now that $B \leqslant N$. Note that for any $f, g \in \Phi^{d+1,d}(L, W, S, B)$ (such that $f \otimes \chi_d, g \otimes \chi_d \in \Phi^{d+1,d}_{\mathrm{ansatz}}(L, W, S, B)$) it holds that

$$\|f \otimes \chi_d - g \otimes \chi_d\|_{C^1(\Omega)} \lesssim \|f - g\|_{C^1(\Omega)} \|\chi_d\|_{C^1(\Omega)} \lesssim \|f - g\|_{C^1(\Omega)},$$

which implies that for some constant $c \geqslant 1$ and any $\tau > 0$,

$$N(\Phi^{d+1,d}_{\mathrm{ansatz}}(L, W, S, B), C^1(\Omega), \tau) \leqslant N(\Phi^{d+1,d}(L, W, S, B), C^1(\Omega), \tau/c).$$

Thus, using the upper bound from Corollary 15 regarding metric entropy of neural network classes, we obtain using (4.5) that

$$
\begin{aligned}
I(R) &= R + \int_0^R H^{1/2}(\mathcal{F}_{\mathrm{NN}}(L,W,S,B,r), C^1(\Omega), \tau) d\tau \\
&\lesssim R + \int_0^R H^{1/2}(\Phi^{d+1,d}_{\mathrm{ansatz}}(L,W,S,B), C^1(\Omega), \tau) d\tau \\
&\leqslant R + \int_0^R H^{1/2}(\Phi^{d+1,d}(L,W,S,B), C^1(\Omega), \tau/c) d\tau \\
&\lesssim R + \int_0^R \sqrt{N(\log \tau^{-1} + \log(N))} d\tau \\
&\lesssim R + \sqrt{N} \int_0^1 \sqrt{\log \tau^{-1} + \log N} d\tau + \sqrt{N} \int_1^R \sqrt{\log \tau^{-1} + \log N} d\tau \\
&\lesssim R + \sqrt{N} \int_0^1 \sqrt{\log \tau^{-1}} d\tau + \sqrt{N} \int_0^1 \sqrt{\log(N)} d\tau \\
&+ \sqrt{N} \int_1^R \sqrt{\log \tau^{-1} + \log N} d\tau \\
&\lesssim R + \sqrt{N}\frac{\pi}{2} + \sqrt{N \log N} + \sqrt{N \log N}(R-1) \lesssim \sqrt{N \log N} R =: \Psi(R)
\end{aligned}
$$

With this choice of upper bound $\Psi(R)$, it is clear that $\Psi(R)/R^2$ is non-increasing in R. Then, we can re-write the condition 2.9 as $\Psi(\delta_n) \lesssim \sqrt{n}\delta_n^2$, which is equivalent to

$$
\delta_n \gtrsim \sqrt{\frac{N \log N}{n}}.
$$

Step 3: Balancing terms. In order to balance the approximation error with the metric entropy term, we will choose N such that

$$
N^{-\frac{k-1}{d+1}} \simeq \sqrt{\frac{N \log N}{n}}.
$$

Up to the $\sqrt{\log N}$ factor, this is achieved by choosing $N \simeq n^{\frac{d+1}{d+1+2(k-1)}}$. Now, applying the general Theorem 2 with this choice yields the convergence rate

$$
\mathbf{E}^n_{P_0}[h^2((T^{\hat{f}})^\sharp \rho, p_0)] \lesssim n^{-\frac{2(k-1)}{d+1+2(k-1)}} \log n.
$$

■

Remark 21 (On the rate from Theorem 20) The final rate obtained in Theorem 20, up to a logarithmic factor, equals the optimal minimax rate

$$
n^{-\frac{2(k-1)}{d+1+2(k-1)}}
$$

for nonparametric estimation of a $(k-1)$-smooth function or density on a $(d+1)$-dimensional domain, in L^2 or Hellinger loss. The presence of $d+1$ (in place of d) in our rate is due to the fact that we are considering time-dependent velocity fields: given any transport map, there are infinitely many velocity fields whose time-one flow map matches this transport, and maximum likelihood estimation does not impose restrictions on the intermediate ODE trajectories between $t = 0$ and $t = 1$. Some recent work, e.g., Finlay et al. (2020); Onken et al. (2021); Marzouk et al. (2023), considers neural ODEs with *regularized* trajectories. In such settings, one might be able to improve the $d+1$ term to d; see below for further discussion.

One may also wonder why the smoothness index appearing in the final rate is $k-1$, rather than k. Indeed, this is due to the fact that for the given k-smooth reference and target densities from $\mathcal{M}(k, L_1, L_2)$, the velocity field whose time-one flow realizes the corresponding KR map also belongs to C^k. Considering the C^1 metric entropy then yields the index $k-1$. This sub-optimality could possibly be resolved by using additional information about the coupling velocity field: as observed in Irons et al. (2022) and Wang and Marzouk (2022), KR maps between C^k densities actually possess *anisotropic* regularity—specifically, higher regularity in their 'diagonal' input variables. It can be shown that the corresponding velocity field also satisfies this property (see Appendix B.1). With this additional smoothness, one might be able to improve the convergence index from $k-1$ to k. However, even with this additional knowledge, it is unclear how to construct neural network classes with such anisotropic regularity, rendering this observation less relevant for practical settings; we have thus omitted a generalization to this setting.

The second term appearing in the final rate, $\log n$, appears due to the metric entropy integral of the neural network class. This $\log n$ factor is commonly present in statistical theory for neural networks; see, e.g., Schmidt-Hieber (2020), which studies nonparametric regression using ReLU networks; Lu et al. (2021), which studies the problem of learning PDE solution fields with neural networks; and more recently Oko et al. (2023), which studies the statistical convergence of diffusion models.

5. Discussion and Future Work

We have developed the first statistical finite-sample guarantees for likelihood-based distribution learning with neural ODEs. Our results show that neural ODE models are efficient distribution estimators, under relatively mild assumptions. We obtained these results by first developing a broader framework for analyzing ODE-parameterized maximum likelihood density estimators. This framework is applicable to any class of velocity field, and characterizes the impact of the chosen class on statistical performance. We then specialized this theory to C^k velocity fields and to specific spaces of velocity fields described by neural networks, obtaining concrete minimax rates.

Our work suggests many important avenues for further work. First, our analysis exposes an interesting impact of the time-dependent construction intrinsic to neural ODEs, i.e., the fact that one seeks a velocity field f that depends on both space ($x \in \mathbb{R}^d$) and time ($t \in [0, 1]$). While this construction confers several advantages (e.g., invertibility of maps, computational tractability of maps and densities), as noted in Remark 21, the additional degree of freedom t raises the dimension-dependence of the minimax convergence rate to

$d+1$, from the optimal value of d. Several regularization schemes (Finlay et al., 2020; Onken et al., 2021; Marzouk et al., 2023) have recently been proposed to control this "extra" freedom by promoting smooth or even straight-line ODE trajectories, with good empirical success. These regularization methods take the form of penalty terms added to the log-likelihood training objective, and it is desirable to understand their impact on statistical rates. To that end, Wang and Marzouk (2022) develop convergence theory for *penalized* nonparametric density estimation using transport maps, and it would be fruitful to integrate such results with the ODE framework developed in this paper.

Second, we note that our work only considers density estimation on the hypercube $[0,1]^d$. Indeed, some of our arguments—for example, the construction of a suitable neural network ansatz space for velocity fields in Section 4.3, satisfying the no-flow boundary condition; and the lower bounds for densities used in the proof of Theorem 9—rely crucially on this fact. In future work, however, it would be useful to extend the present statistical convergence analysis to more general bounded domains and to unbounded domains. The latter will require a more refined understanding of the tail properties of the associated ODE flow maps.

To our knowledge, the question of *computational* guarantees for neural ODE training is quite open. It remains challenging to characterize the loss landscape and its interaction with optimization algorithms; here one must also assess the impact of ODE time discretization, and the potential impact of different ways of computing gradients in this setting, e.g., "discretize-then-optimize" versus "optimize-then-discretize" approaches that use continuous adjoints (Gholami et al., 2019).

We also note that several recently proposed generative modeling methods, e.g., flow matching (Lipman et al., 2022), rectified flow (Liu et al., 2022), and stochastic interpolants (Albergo et al., 2023), produce deterministic ODEs but depart from the maximum likelihood training approach considered in this paper. It is of interest to elucidate the statistical performance of such methods as well; some results in this direction have appeared very recently, after the submission of the current work (Gao et al., 2024; Fukumizu et al., 2024).

Acknowledgments

SW and YM acknowledge support from the US Air Force Office of Scientific Research (AFOSR) MURI, Analysis and Synthesis of Rare Events, award number FA9550-20-1-0397. ZR and YM acknowledge support from the US Department of Energy (DOE), Office of Advanced Scientific Computing Research, under grants DE-SC0021226 and DE-SC0023187. ZR also acknowledges support from a US National Science Foundation Graduate Research Fellowship.

Appendix A. Remaining Proofs for Section 2

A.1 Proofs from Statements in the Main Text

Proof [Proof of Lemma 1] We divide the proof into two steps. To simplify notation we drop the index f and write $X = X^f$. Recall that this is a map from $D = [0,1]^d \times [0,1] \to \mathbb{R}^d$. Its components are denoted by $X = (X_1, \dots, X_d)$.

Step 1: Trajectories remain in D. By definition

$$X(x,t) = x + \int_0^t f(X(x,s),s)ds, \qquad t \in [0,1]. \tag{A.1}$$

For an interior point $x \in (0,1)^d$, we show that $X(x,t) \in (0,1)^d$ for all $t \in (0,1)$, i.e., $0 < X_j(x,t) < 1$ for all $j = 1,\dots,d$. By symmetry, it suffices to show $X_1(x,t) > 0$.

Consider a point $y = (0, y_2, \dots, y_d) \in \{0\} \times (0,1)^{d-1} \subseteq \partial D$. Then, the normal outer vector $\nu_y = (-1, 0, \dots, 0) \in \mathbb{R}^d$ at y is well-defined. Since $f \in \mathcal{V}$, the definition of $\mathcal{V}$ in (2.4) yields $f(y) \cdot \nu_y = 0$, and thus

$$f_1(y,t) = 0 \qquad \forall y \in \{0\} \times (0,1)^{d-1}.$$

Moreover, $f \in C^1(\Omega)$ implies $\|\partial_{y_1} f\|_{C(D)} < \infty$. Hence, using the fundamental theorem of calculus

$$|f_1(y,t)| \leqslant y_1 \|\partial_{y_1} f\|_{C(D)} \leqslant y_1 \|f\|_{C^1(D)} \qquad \text{for all } y \in [0,1] \times (0,1)^{d-1},$$

and by continuity of f_1 the inequality extends to all $y \in D$.

Thus for $x \in (0,1)^d$, $t \in [0,1]$

$$\frac{d}{dt} X_1(x,t) = f_1(X(x,t),t) \geqslant -X_1(x,t)\|f\|_{C^1}, \qquad X_1(x,0) = x_1 > 0,$$

or equivalently $-\frac{d}{dt} X_1(x,t) \leqslant -\|f\|_{C^1}(-X_1(x,T))$. Applying Grönwall's inequality (in its differential form), we obtain that

$$-X_1(x,t) \leqslant -x_1 \exp(-t\|f\|_{C^1}) \iff X_1(x,t) \geqslant x_1 \exp(-t\|f\|_{C^1}) > 0.$$

Step 2: Bijectivity and differentiability. For any interior point $x \in (0,1)$, by Step 1 and the Picard-Lindelöf theorem, there exists a unique solution $t \mapsto (X(x,t),t) : [0,1] \to (0,1)^d \times [0,1]$ of (2.2) (or (A.1)). Consider the time-reversed ODE

$$Y'(y,s) = -f(Y(y,s), 1-s), \qquad Y(y,0) = y. \tag{A.2}$$

Clearly $\tilde{f}(z,\delta) := -f(z, 1-\delta)$ also belongs to $\mathcal{V}$, cp. (2.4), and hence for any interior point $y \in (0,1)^d$, by Step 1 and the Picard-Lindelöf theorem, there exists a unique solution $t \mapsto (Y(y,t),t) : [0,1] \to (0,1)^d \times [0,1]$ of (A.2). In either case, since the trajectories cannot cross, both maps $x \mapsto X(x,1) : (0,1)^d \to (0,1)^d$ and $y \mapsto Y(y,1) : (0,1)^d \to (0,1)^d$ are injective. Furthermore $X(Y(y,1),1) = y$ for all $y \in (0,1)^d$. By a continuity argument, we conclude that $X(\cdot,1) : D \to D$ is bijective. The same argument yields that $x \mapsto X(x,t) : D \to D$ is bijective for any $t \in [0,1]$.

Finally, using Corollary 3.1 in Hartman (2002) as well as the subsequent remark (which are applicable since $f \in C^1$ by assumption), we see that in fact $X(\cdot,t) \in C^1((0,1)^d)$ and $\det(\nabla X(x,t)) \neq 0$. Once more by symmetry in the forward and backward in time ODEs, also $X(\cdot,t)^{-1} \in C^1((0,1)^d)$. Thus $X(\cdot,t) : D \to D$ is a C^1-diffeomorphism. ■

Proof [Proof of Theorem 5] The proof can be seen from making quantitative the arguments underlying Theorem 10.13 in van de Geer (2000), in combination with several straightforward modifications of the assumptions there.

Let c be the constant from (10.70) in van de Geer (2000). Then, since $p_n^* \geqslant c$ is lower bounded and $p_0 \leqslant K$ is upper bounded, it holds that $p_0/p_n* \leqslant Kc^{-1}$, whence the assumption (10.69) in van de Geer (2000) is clearly fulfilled. Next, we notice that the relevant entropy integral in Theorem 10.13 of van de Geer (2000) is given by the expression (for some constant $c > 0$)

$$\max\Big\{R, \int_{R^2/c}^{R} H_B^{1/2}\Big(\big\{p \in \mathcal{Q}_n^* : h\big(\frac{p+p_n^*}{2}, p_n^*\big) \leqslant \delta\big\}, L^2(p_n^*), \tau\Big) d\tau\Big\}, \tag{A.3}$$

which is clearly upper bounded by our entropy integral $I_B(\mathcal{P}_n, R, p_n^*)$ from (2.12). Thus, any choice Ψ fulfilling the hypotheses of our theorem automatically also represents a desired upper bound for the entropy integral (A.3). It follows that Theorem 10.13 in van de Geer (2000) is applicable, and we obtain the convergence in probability

$$h((T^{\hat{f}})^\sharp \rho, p_0) = \mathcal{O}_{P_0^n}(\delta_n + h(p^*, p_0)).$$

It remains to show the non-asymptotic concentration inequality from (2.10), which is a stronger statement than mere convergence in probability. This follows from an inspection of the proof of Theorem 10.13 of van de Geer (2000), which we now detail. Indeed, the last step of the latter proof is based on the following case distinction for the terms I and II defined on p. 191 of van de Geer (2000).

Case 1: $I \leqslant II$. In this case, denoting $\hat{p} := (T^{\hat{f}})^\sharp \rho$, one obtains

$$h^2\Big(\frac{\hat{p}+p_n^*}{2}, p_n^*\Big) \leqslant 4(1+c_0)h(p_n^*, p_0).$$

Here, c_0 can be any constant such that $p_0/p_n^* \leqslant c_0^2$ (cf. (10.69) in van de Geer (2000)); in particular we may set $c_0 := \sqrt{Lc^{-1}}$. Using Lemma 4.2 from van de Geer (2000), it follows that

$$h^2(\hat{p}, p^*) \leqslant 16 h^2\Big(\frac{\hat{p}+p^*}{2}, p^*\Big) \leqslant 64(1+c_0)h(p^*, p_0).$$

Case 2: $I < II$. In this case, one obtains that

$$h^2\Big(\frac{\hat{p}+p^*}{2}, p^*\Big) \leqslant \int \log\Big(\frac{\hat{p}+p^*}{2p^*}\Big) d(P_n - P_0),$$

where P_n denotes the empirical measure and P_0 is the data-generating law. In this case, using the same concentration arguments as in Theorem 7.4 of van de Geer (2000), one obtains that for any $\delta \geqslant \delta_n$, and some $C > 0$ only depending on c_0 (and thus only depending on c, K),

$$P_0^n\big(h(\hat{p}, p^*) \geqslant \delta\big) \leqslant C \exp\Big(-\frac{n\delta^2}{C}\Big).$$

Then, using the triangle inequality

$$h(\hat{p}, p_0) \leqslant h(\hat{p}, p^*) + h(p^*, p_0)$$

completes the proof. ■

Proof [Proof of Lemma 6]

For notational convenience let us write $\|f - g\|_{C^1(\Omega)} = \epsilon$ for some $\epsilon > 0$. Then, for any $x \in D \subset \mathbb{R}^d$, we have that

$$\begin{aligned}\left|\frac{dX^f(x,t)}{dt} - \frac{dX^g(x,t)}{dt}\right| &= |f(X^f(x,t),t) - g(X^g(x,t),t)| \\ &\leqslant |f(X^f(x,t),t) - f(X^g(x,t),t)| \\ &+ |f(X^g(x,t),t) - g(X^g(x,t),t)|.\end{aligned}$$

By assumption, we have $\sup_{t\in[0,1]} \max_{i,j} \|(\nabla_x f(\cdot,t))_{i,j}\|_{C(D)} \leqslant r$, and it follows that

$$\begin{aligned}\sup_{t\in[0,1]} \|\nabla_x f(\cdot,t)\|_{C(D,\mathbb{R}^{d\times d})} &= \sup_{t\in[0,1],x\in D} \|\nabla_x f(x,t)\|_2 \leqslant \sup_{t\in[0,1],x\in D} \|\nabla_x f(x,t)\|_F \\ &\leqslant d \sup_{t\in[0,1]} \max_{i,j} \|(\nabla_x f(\cdot,t))_{i,j}\|_{C(D)} \leqslant dr,\end{aligned}$$

where we have equipped $\mathbb{R}^{d\times d}$ with the usual operator norm for matrices. Therefore, we can conclude that $|f(X^f(x,t),t) - f(X^g(x,t),t)| \leqslant dr|X^f(x,t) - X^g(x,t)|$. Next, we have

$$\begin{aligned}&|X^f(x,t) - X^g(x,t)| = \left|\int_0^t f(X^f(x,s),s) - g(X^g(x,s),s)ds\right| \\ &\leqslant \int_0^t |f(X^f(x,s),s) - g(X^g(x,s),s)|ds \\ &\leqslant \int_0^t |f(X^f(x,s),s) - f(X^g(x,s),s)|ds + \int_0^t |f(X^g(x,s),s) - g(X^g(x,s),s)|ds \\ &\leqslant dr\int_0^t |X^f(x,s) - X^g(x,s)|ds + t\epsilon\end{aligned}$$

Using Grönwall's inequality (integral form), we get

$$|T^f(x) - T^g(x)| = |X^f(x,1) - X^g(x,1)| \leqslant \epsilon e^{dr}, \quad \forall x.$$

Therefore $\max_j \|X_j^f(\cdot,1) - X_j^f(\cdot,1)\|_{C(D)} \leqslant \epsilon e^{dr}$.

Now, it remains to bound $\max_{i,j} \|(\nabla_x T^f(\cdot) - \nabla_x T^g(\cdot))_{i,j}\|_{C(D)}$, which could be achieved by bounding the Frobenius norm of the difference in Jacobian $\|\nabla_x T^f(x) - \nabla_x T^g(x)\|_F$ by equivalence of norms. Similarly as above, we can write for $t \in [0,1]$

$$
\begin{aligned}
&\|\nabla_x X^f(x,t) - \nabla_x X^g(x,t)\|_F = \|\int_0^t \nabla_x (f(X^f(x,s),s) - g(X^g(x,s),s))ds\|_F \\
&\leqslant \int_0^t \|\nabla_x (f(X^f(x,s),s) - g(X^g(x,s),s))\|_F ds \\
&= \int_0^t \Big\| \Big(\nabla_X f(X^f(x,s),s)\nabla_x X^f(x,s) - \nabla_X g(X^g(x,s),s)\nabla_x X^g(x,s)\Big) \Big\|_F ds \\
&\leqslant \int_0^t \|(\nabla_X f(X^g(x,s),s) - \nabla_X g(X^g(x,s),s))\nabla_x X^g(x,s)\|_F ds \\
&\quad + \int_0^t \|(\nabla_X f(X^f(x,s),s) - \nabla_X f(X^g(x,s),s))\nabla_x X^g(x,s)\|_F ds \\
&\quad + \int_0^t \|\nabla_X f(X^f(x,s),s)(\nabla_x X^f(x,s) - \nabla_x X^g(x,s))\|_F ds \\
&=: I + II + III,
\end{aligned}
$$

where $\nabla_X f(X^f(x,s),s)$ and $\nabla_X g(X^g(x,s),s))$ denote the spatial derivative of f and g *evaluated* at $(X^f(x,s),s)$ and $(X^g(x,s),s))$ respectively.

To bound term I, note that $\|f-g\|_{C^1(\Omega)} = \epsilon$ gives

$$
|(\nabla_X f(X^g(x,t),t) - \nabla_X g(X^g(x,t),t))_{i,j}| \leqslant \epsilon \quad \text{ for all } (x,t) \in \Omega,\ (i,j) \in [d]^2.
$$

To establish bounds on $\nabla_x X^g(x,t)$, we note that

$$
\begin{aligned}
\nabla_x X^g(x,t) &= I_{d\times d} + \int_0^t \nabla_x \big[g(X^g(x,s),s)\big] ds \\
&= I_{d\times d} + \int_0^t (\nabla_X g(X^g(x,s),s))\nabla_x X^g(x,s) ds,
\end{aligned}
$$

where $I_{d\times d}$ is the identity matrix of dimension d. Using the standard multiplication inequality $\|M_1 M_2\|_F \leqslant \|M_1\|_F \|M_2\|_F$ for the Frobenius norm and $\|g\|_{C^1(\Omega)} \leqslant r$, it follows that for all points (x,t),

$$
\begin{aligned}
\|\nabla_x X^g(x,t)\|_F &\leqslant \sqrt{d} + \int_0^t \|\nabla_X g(X^g(x,s),s)\|_F \|\nabla_x X^g(x,s)\|_F ds \\
&\leqslant \sqrt{d} + dr \int_0^t \|\nabla_x X^g(x,s)\|_F ds.
\end{aligned}
$$

By Grönwall's inequality, it follows that $\|\nabla_x X^g(x,t)\|_F \leqslant \sqrt{d} e^{drt}$ and in particular $\|\nabla_x X^g(x,1)\|_F \leqslant \sqrt{d} e^{dr}$. Therefore,

$$
\begin{aligned}
&\|(\nabla_X f(X^g(x,s),s) - \nabla_X g(X^g(x,s),s))\nabla_x X^g(x,s)\|_F \\
&\leqslant \|\nabla_x X^g(x,s)\|_F \|\nabla_X f(X^g(x,s),s) - \nabla_X g(X^g(x,s),s)\|_F \leqslant d^{\frac{3}{2}} \epsilon e^{drs},
\end{aligned}
$$

and term I may be bounded as

$$\begin{aligned} I &= \int_0^t \|(\nabla_X f(X^g(x,s),s) - \nabla_X g(X^g(x,s),s))\nabla_x X^g(x,s)\|_F ds \\ &\leqslant \int_0^t d^{\frac{3}{2}} \epsilon e^{drs} ds = \frac{(e^{drt}-1)\sqrt{d}\epsilon}{r} \leqslant \frac{(e^{dr}-1)\sqrt{d}\epsilon}{r}. \end{aligned}$$

To bound II, by the Lipschitz property, we have at any (x,t),

$$\|\nabla_X f(X^f(x,t),t) - \nabla_X f(X^g(x,t),t)\|_F \leqslant r|X^f(x,t) - X^g(x,t)|.$$

Since $|X^f(x,s) - X^g(x,s)| \leqslant \epsilon e^{dr}$ at all point (x,s) from the previous part, we obtain that

$$\begin{aligned} \|(\nabla_X f(X^f(x,s),s) - &\nabla_X f(X^g(x,s),s))\nabla_x X^g(x,s)\|_F \\ &\leqslant \|\nabla_x X^g(x,s)\|_F \|\nabla_X f(X^f(x,s),s) - \nabla_X f(X^g(x,s),s)\|_F \\ &\leqslant \sqrt{d} e^{drs} r\epsilon e^{drs} = \sqrt{d} r\epsilon e^{2drs}. \end{aligned}$$

Then, we have

$$\begin{aligned} II &= \int_0^t \|(\nabla_X f(X^f(x,s),s) - \nabla_X f(X^g(x,s),s))\nabla_X X^g(x,s)\|_F ds \\ &\leqslant \int_0^t \sqrt{d} r\epsilon e^{2drs} ds \leqslant \frac{(e^{2dr}-1)r\epsilon}{2\sqrt{d}r} \end{aligned}$$

Finally, to bound III, we have

$$\begin{aligned} \|\nabla_X f(X^f(x,s),s)&(\nabla_x X^f(x,s) - \nabla_x X^g(x,s))\|_F \\ &\leqslant \|\nabla_X f(X^f(x,s),s)\|_F \|\nabla_x X^f(x,s) - \nabla_x X^g(x,s)\|_F, \end{aligned}$$

where we can bound $\|\nabla_X f(X^f(x,s),s)\|_F$ by dr.

Combining all the terms, we obtain

$$\begin{aligned} &\|\nabla_x (X^f(x,t) - X^g(x,t))\|_F \\ &\quad \leqslant \frac{(e^{dr}-1)\sqrt{d}\epsilon}{r} + \frac{(e^{2dr}-1)r\epsilon}{2\sqrt{d}r} + dr\int_0^t \|\nabla_x X^f(x,s) - \nabla_x X^g(x,s)\|_F ds. \end{aligned}$$

Grönwall's inequality then gives

$$\|\nabla_x (X^f(x,1) - X^g(x,1))\|_F \leqslant \epsilon \frac{(e^{2dr}-1)r + 2(e^{dr}-1)d}{2\sqrt{d}r} e^{dr} \leqslant \frac{re^{3dr} + 2de^{2dr}}{2\sqrt{d}r}\epsilon$$

.

Since the above inequality holds for all x, considering pointwise entries in the Jacobian gives

$$\max_{i,j} \|(\nabla_x X^f(\cdot,1) - \nabla_x X^g(\cdot,1))_{i,j}\|_{C(D)} \leqslant \frac{re^{3dr} + 2de^{2dr}}{2\sqrt{d}r}\epsilon.$$

Using the C^1 norm, we conclude that

$$\begin{aligned}\|T^f(x) - T^g(x)\|_{C^1(D)} &= \|X^f(x,1) - X^g(x,1)\|_{C^1(D)} \\ &\leqslant \max\left\{\epsilon e^{dr}, \frac{re^{3dr} + 2de^{2dr}}{2\sqrt{dr}}\epsilon\right\} \leqslant C\|f-g\|_{C^1(\Omega)},\end{aligned}$$

where $C = \max\{e^{dr}, \frac{re^{3dr}+2de^{2dr}}{2\sqrt{dr}}\}$. ∎

Proof [Proof of Theorem 7] For notational convenience, let us write $\|T - G\|_{C^1(D)} = \epsilon$.

$$\begin{aligned}\|T^\sharp\rho - G^\sharp\rho\|_{C(D)} &= \|\rho(T)\det\nabla T - \rho(G)\det\nabla G\|_{C(D)} \\ &\leqslant \|\rho(T)\det\nabla T - \rho(G)\det\nabla T\|_{C(D)} + \|\rho(G)\det\nabla T - \rho(G)\det\nabla G\|_{C(D)} \\ &\leqslant \|\rho(T) - \rho(G)\|_{C(D)}\|\det\nabla T)\|_{C(D)} + \|\rho(G)\|_{C(D)}\|\det\nabla T - \det\nabla G\|_{C(D)}.\end{aligned}$$

We bound these two terms separately. By the Lipschitz continuity of ρ, we have $\|\rho(T) - \rho(G)\|_{C(D)} \leqslant \|\rho\|_{\mathrm{Lip}}\|T-G\|_{C(D)} \leqslant |\rho|_{\mathrm{Lip}}\epsilon$.

Moreover, by (Zech and Marzouk, 2022a, Lemma E.1),

$$\begin{aligned}|\det\nabla T(x) - \det\nabla G(x)| &= \left|\prod_{i=1}^d \lambda_i(x) - \prod_{i=1}^d \eta_i(x)\right| \\ &\leqslant \frac{\exp\left(\sum_{i=1}^d \frac{|\lambda_i(x)-\eta_i(x)|}{\lambda_d(x)}\right)\prod_{i=1}^d \lambda_i(x)}{\min\{\lambda_d(x), \eta_d(x)\}}\sum_{i=1}^d |\lambda_i(x) - \eta_i(x)|\end{aligned}$$

which is upper bounded by $\tilde{C}\sum_{i=1}^d |\lambda_i(x) - \eta_i(x)|$.

By Weyl's theorem, $\max_i |\lambda_i(x) - \eta_i(x)| \leqslant \|\nabla T(x) - \nabla G(x)\|_2$. Furthermore, $\forall x \in D, \|\nabla T(x) - \nabla G(x)\|_2 \leqslant d\epsilon$. Thus we can conclude that $|\det\nabla T(x) - \det\nabla G(x)| \leqslant \tilde{C}d^2\epsilon$ at all $x \in D$, from which it follows that $\|\det\nabla T - \det\nabla G\|_{C(D)} \leqslant \tilde{C}d^2\epsilon$.

Putting everything together, we have

$$\|T^\sharp\rho - G^\sharp\rho\|_{C(D)} \leqslant \left(|\rho|_{\mathrm{Lip}}\|\det\nabla T(x)\|_{C(D)} + \tilde{C}d^2\|\rho\|_{C(D)}\right)\epsilon.$$

Finally, using the fact that for a d-dimensional matrix A, $|\det A| \leqslant (\frac{\mathrm{tr}A}{d})^d$, we have

$$\|T_\sharp\rho - G_\sharp\rho\|_{C(D)} \leqslant \left(|\rho|_{\mathrm{Lip}}\|T\|_{C^1(D)}^d + \tilde{C}d^2\|\rho\|_{C(D)}\right)\epsilon.$$

∎

Proof [Proof of Theorem 8] For notational convenience we write $X = X^f$. The map $(T^f) : D \to D$ is obtained by integrating the ODE (2.2) forward in time, i.e., $(T^f)(x) =$

$x + \int_0^1 f(X(x,t),t)dt$. Taking the operator norm of the Jacobian, we obtain

$$\begin{aligned}
\|\nabla_x(T^f)(x)\|_2 &= \left\| I_{d\times d} + \int_0^1 \nabla_y f(X(y,t),t)dt \right\|_2 \\
&= \left\| I_{d\times d} + \int_0^1 \nabla_X f(X(x,t),t)\nabla_x X(x,t)dt \right\|_2 \\
&\leqslant 1 + \int_0^1 \|\nabla_X f(X(x,t),t)\|_2 \|\nabla_x X(x,t)\|_2 dt
\end{aligned}$$

Since $\|f\|_{C^1(D\times[0,1])} \leqslant M$, we have $\|\nabla_X f(X(x,s),s)\|_2 \leqslant dM, \forall s \in [0,1]$. On the other hand,

$$\begin{aligned}
\|\nabla_x X(x,t)\|_2 &= \|I_{d\times d} + \int_0^t \nabla_X f(X(x,s),t)\nabla_x X(x,s)ds\|_2 \\
&\leqslant 1 + \int_0^t \|\nabla_X f(X(x,s),s)\|_2 \|\nabla_x X(x,s)\|_2 ds \\
&\leqslant 1 + dM \int_0^t \|\nabla_x X(x,s)\|_2 ds.
\end{aligned}$$

It follows from Grönwall's inequality that $\|\nabla_x X(x,t)\|_2 \leqslant e^{dMt} \leqslant e^{dM}$, $\forall t \in [0,1]$. Putting things together, we get $\|\nabla_x T^f(x)\|_2 \leqslant 1 + dMe^{dM}, \forall x \in D$, from which it follows that $\lambda_1^f(x) \leqslant 1 + dMe^{dM}, \forall x \in D$.

On the other hand, consider $\lambda_d^f(x)$, the smallest singular value of $\nabla_x T^f(x)$. By the inverse function theorem, for all $x \in D$, writing $y = T^f(x)$, we have that $\nabla_y (T^f)^{-1}(y) = (\nabla_x T^f(x))^{-1}$. It follows that

$$\frac{1}{\lambda_d^f(x)} = \left\| [\nabla_x T^f(x)]^{-1} \right\|_2 = \left\| \nabla_y (T^f)^{-1}(y) \right\|_2.$$

Observe that the inverse transport $(T^f)^{-1}$ is given by integrating the ODE backwards in time. For this purpose, consider the following reverse ODE. For $y \in D$ so that $y = x + \int_0^1 f(X(x,t),t)dt$ and $Y(y,t) = X(x,1-t)$, we have

$$\begin{cases} \frac{dY(y,t)}{dt} &= -f(Y(y,t),1-t), \\ Y(y,0) &= y. \end{cases} \tag{A.4}$$

Then, by a similar argument as above, we can show $\|\nabla_y (T^f)^{-1}(y)\|_2 \leqslant 1 + dMe^{dM}$. Thus we have shown that $\lambda_d^f(x) \geqslant \frac{1}{1+dMe^{dM}}, \forall x \in D$. ∎

A.2 Auxiliary Results

We show three elementary lemmas. The first two provide bounds on the (bracketing) metric entropy.

Lemma 22 *Let μ be a measure on $D = [0,1]^d$ with positive Lebesgue density and let $\mathcal{F} \subseteq C(D, \mathbb{R}^d)$. Then for all $\tau > 0$ it holds that*

$$N_B(\mathcal{F}, L^2(D,\mu), \tau) \leqslant N\Big(\mathcal{F}, C(D), \frac{\tau}{2\sqrt{\mu(D)}}\Big).$$

Proof Let $N := N(\mathcal{F}, C(D), \tau)$. By the definition of metric entropy, there exist functions $f_1, \dots, f_N$ on D such that for each $f \in \mathcal{F}$ exists $i \in \{1, \dots, N\}$ with $\|f - f_i\|_{C(D)} \leqslant \tau$. For each $i \leqslant N$, set $f_{i,L} := f_i - \tau$ and $f_{i,U} := f_i + \tau$. Then $f_{i,L} \leqslant f \leqslant f_{i,U}$ on D. Since $\|f_{i,L}(x) - f_{i,U}(x)\|_{L^2(\mu)} \leqslant 2\tau\sqrt{\mu(D)}$, this implies

$$N_B(\mathcal{F}, L^2(\mu), 2\tau\sqrt{\mu(D)}) \leqslant N(\mathcal{F}, C(D), \tau)$$

for all $\tau > 0$ (cp. Definition 4). ■

Similarly:

Lemma 23 *Let $(X, \|\cdot\|_X)$, $(Y, \|\cdot\|_Y)$ be two normed spaces and $A \subseteq X$. Let $\Phi : A \to Y$ be Lipschitz continuous with Lipschitz constant L. Then for all $\tau > 0$*

$$N(\Phi(\mathcal{F}), Y, \tau) \leqslant N(\mathcal{F}, X, \frac{\tau}{2L}).$$

Proof Fix $\tau > 0$ and set $\tilde{\tau} := \frac{\tau}{2L}$ and $N := N(\mathcal{F}, X, \tilde{\tau})$. Then we can find $x_1, \dots, x_N \in X$ such that for each $x \in \mathcal{F}$ exists $i \in \{1, \dots, N\}$ with $\|x - x_i\|_X \leqslant \tilde{\tau}$. In particular, we can find $\tilde{x}_1, \dots, \tilde{x}_N \in \mathcal{F}$ such that for each $x \in \mathcal{F}$ exists $i \in \{1, \dots, N\}$ with $\|x - \tilde{x}_i\|_X \leqslant 2\tilde{\tau}$.

Let $y \in \Phi(\mathcal{F})$ arbitrary, i.e., $y = \Phi(x)$ for some $x \in \mathcal{F}$. Then there exists $i \in \{1, \dots, N\}$ such that

$$\|y - \Phi(\tilde{x}_i)\|_Y = \|\Phi(x) - \Phi(\tilde{x}_i)\|_Y \leqslant L\|x - \tilde{x}_i\|_X \leqslant 2L\tilde{\tau} = \tau.$$

This shows the claim. ■

The next lemma states that the Hellinger distance is bounded by the L^∞-distance whenever the maximum of both densities is bounded from below.

Lemma 24 *Let $L > 0$ and $D \subseteq \mathbb{R}^d$ measurable with Lebesgue measure bounded by one. Then for all probability densities $p_1(x)$, $p_2(x)$ on D with $\operatorname{ess\,inf}_{x\in D} \max\{p_1(x), p_2(x)\} \geqslant L$, it holds*

$$h(p_1, p_2) \leqslant \frac{1}{\sqrt{2L}} \|p_1 - p_2\|_{C(D)}.$$

Proof We have

$$\begin{aligned} h(p_1,p_2)^2 &= \frac{1}{2}\int_D (\sqrt{p_1(x)} - \sqrt{p_2(x)})^2 dx = \frac{1}{2}\int_D \left(\frac{p_1(x) - p_2(x)}{\sqrt{p_1(x)} + \sqrt{p_2(x)}}\right)^2 dx \\ &\leqslant \frac{1}{2}\int_D \left(\frac{\|p_1 - p_2\|_{C(D)}}{\sqrt{L}}\right)^2 dx = \frac{1}{2L}\|p_1 - p_2\|^2_{C(D)}. \end{aligned}$$

■

Appendix B. Proofs for Section 3

B.1 Proof of Theorem 9

In order to prove Theorem 9, we need three auxiliary Lemmas 25, 26, and 27. These auxiliary statements regard certain anisotropic regularity classes which describe Knothe–Rosenblatt maps between C^k-smooth densities, as was observed in Wang and Marzouk (2022). For any $k \geqslant 1$, we write $[k] = \{1, \dots, k\}$.

Step 1: Anisotropic regularity classes

For $k \geqslant 1$ integer, let us define the following classes of triangular functions on D with anisotropic regularity :

$$C^k_{diag}(D, \mathbb{R}^d) = \left\{ f \in C^k(D, \mathbb{R}^d) \text{ triangular} : \forall j \in [d] : \ \partial_j f_j \in C^k(D) \right\},$$

with norm

$$\|f\|_{C^k_{diag}(D)} := \sum_{j=1}^{d} \|f_j\|_{C^k([0,1]^j)} + \|\partial_j f_j\|_{C^k([0,1]^j)}.$$

We further introduce the class of bijective, monotone triangular maps with such anisotropic regularity:

$$\mathcal{A}^k_{diag} := \Big\{ S : D \to D \text{ triangular and bijective}, \\ S \in C^k_{diag}(D, \mathbb{R}^d), \ \forall j \in [d] : \partial_j S_j > 0 \Big\}.$$

For any constants $0 < c_{mon} < L < \infty$, we will also need the sub-classes with bounded norm

$$\mathcal{A}^k_{diag}(c_{mon}, L) := \Big\{ S \in \mathcal{A}^k_{diag}, \ \|S\|_{C^k_{diag}(D,\mathbb{R}^d)} \leqslant L, \inf_{x \in [0,1]^k} \partial_k S_k(x) \geqslant c_{mon} \Big\}.$$

The following lemma shows that the above classes are closed under composition and inversion.

Lemma 25 ***(i)*** *If $S, R \in \mathcal{A}^k_{diag}$, then $S \circ R \in A^k_{diag}$. Moreover, for any $c_{mon}, L > 0$ there exist $c'_{mon}, L' > 0$ such that for any $S, R \in \mathcal{A}^k_{diag}(c_{mon}, L)$, it holds that $S \circ R \in \mathcal{A}^k_{diag}(c'_{mon}, L')$.*

(ii) *If $S \in \mathcal{A}^k_{diag}$, then also $S^{-1} \in A^k_{diag}$. Moreover, for any $c_{mon}, L > 0$ there exist $c'_{mon}, L' > 0$ such that for any $S \in \mathcal{A}^k_{diag}(c_{mon}, L)$, it holds that $S^{-1} \in \mathcal{A}^k_{diag}(c'_{mon}, L')$.*

Proof We begin by proving part **(i)**. First, we observe that $S \circ R$ is still bijective and triangular. To see the triangularity, we observe that

$$[S \circ R](x) = \begin{bmatrix} S_1(R_1(x_1)) \\ \vdots \\ S_j(R_1(x_1), \dots, R_j(x_{[j]})) \\ \vdots \\ S_d(R_1(x_1), \dots\dots\dots, R_d(x)) \end{bmatrix}.$$

Thus, the j-th component map only depends on the first j coordinates of x. Next, it is also clear that since $S, R \in C^k(D, \mathbb{R}^d)$, also $S \circ R \in C^k(D, \mathbb{R}^d)$. It remains to assert the regularity of the 'diagonal derivatives' of $S \circ R$. For any $j \in [d]$, denoting $y(x) = [R_1(x_1), \ldots, R_j(x_{[j]})]$, using the chain rule and the fact that $R_l,\ l < j$ is independent of x_j, we obtain that

$$\begin{aligned}\frac{\partial}{\partial x_j}(S \circ R)_j &= \frac{\partial}{\partial x_j}\big[S_j(R_1(x_1), \ldots, R_k(x_{[j]}))\big] \\ &= \sum_{l=1}^{j}\Big(\frac{\partial}{\partial y_l}S_j\Big)(y(x))\Big(\frac{\partial}{\partial x_j}R_l\Big)(x_{[l]}) \\ &= \Big(\frac{\partial}{\partial y_j}S_j\Big)(y(x))\Big(\frac{\partial}{\partial x_j}R_j\Big)(x_{[j]}).\end{aligned}$$

Since $\frac{\partial}{\partial y_j}S_j : (0,1)^j \to \mathbb{R}$ is a C^k function, $y : x \mapsto y(x),\ (0,1)^j \to (0,1)^j$ is C^k function, and finally also $\frac{\partial}{\partial x_j}R_j : (0,1)^j \to \mathbb{R}$ is C^k function, we overall obtain that $\frac{\partial}{\partial x_j}(S \circ R)_j$ is also C^k. It is clear from the chain rule and standard multiplication inequalities for C^k-norms that we may choose the upper bound L' for the norm of $S \circ R$ just depending on L. Moreover, the preceding calculation clearly implies that for any $j = 1, \ldots, d$,

$$\frac{\partial}{\partial x_j}(S \circ R)_j \geqslant c_{mon}^2 =: c'_{mon} > 0,$$

which completes the proof of part **(i)**.

Let us now turn to part **(ii)**. Let ρ be the uniform density on D. Using Proposition 2.1 in Wang and Marzouk (2022), we know that for any $S \in \mathcal{A}^k_{diag}(c_{mon}, L)$, the pullback distribution $p_S := S^\sharp \rho$ is an upper and lower bounded $C^k(D)$ density, where the C^k norm, and the upper and lower bounds only depend on c_{mon} and L. Moreover, clearly S^{-1} is again triangular and bijective. Moreover, it satisfies $(S^{-1})^\sharp p_S = \rho$. By uniqueness of the KR-transport map, S^{-1} consitutes the unique Knothe–Rosenblatt transport map between two C^k densities. Thus, again using Proposition 2.1 in Wang and Marzouk (2022), we see that $S^{-1} \in \mathcal{A}^k_{diag}(c_{mon'}, L')$ for some $c'_{mon}, L > 0$. This concludes the proof of the lemma. ■

Step 2: A Hardy-type inequality for functions with anisotropic regularity

Lemma 26 *Let $d,\ k \in \mathbb{N}$, $D = [0,1]^d$, and $f \in C^k(D)$ such that $\partial_{x_d} f \in C^k(D)$ and additionally $f(x) = 0$ whenever $x_d = 0$. Then $g(x) := \frac{f(x)}{x_d} \in C^k(D)$ and there exists $C = C(d,k)$ such that*

$$\|g\|_{C^k(D)} \leqslant C\|\partial_{x_d} f\|_{C^k(D)}. \tag{B.1}$$

Proof Denote by $x_{[j]} = (x_1, \ldots, x_j)$ the first j coordinates of x. Since $f(x_1, \ldots, x_{d-1}, 0) = 0$ and $x_d \mapsto f(x) \in C^{k+1}([0,1])$ for all $x_{[d-1]} \in [0,1]^{d-1}$, it follows that for all $x \in D$ and any $l \in \{0, \ldots, k\}$,

$$f(x) = \sum_{j=1}^{l} \partial_{x_d}^j f(x_1, \ldots, x_{d-1}, 0)\frac{x_d^j}{j!} + \int_0^{x_d} \partial_{x_d}^{l+1} f(x_1, \ldots, x_{d-1}, t)\frac{(x_d - t)^l}{l!}\,\mathrm{d}t$$

and thus for any $l \in \{0, \dots, k\}$

$$\frac{f(x)}{x_d} = \underbrace{\sum_{j=0}^{l-1} \partial_{x_d}^{j+1} f(x_1, \dots, x_{d-1}, 0) \frac{x_d^j}{j!}}_{=:g_1(x)} + \underbrace{\frac{1}{x_d} \int_0^{x_d} \partial_{x_d}^{l+1} f(x_1, \dots, x_{d-1}, t) \frac{(x_d - t)^l}{k!} \, \mathrm{d}t}_{=:g_2(x)} .$$

Now, fix a multiindex $\boldsymbol{v} \in \mathbb{N}^d$ such that $|\boldsymbol{v}| \leqslant k$. To prove the lemma, we need to show that $\sup_{x \in D} |\partial^{\boldsymbol{v}} f(x)|$ is bounded by the right-hand side of (B.1) with some C solely depending on k and d. Set $l := k - v_d \geqslant 0$. Clearly

$$|\partial^{\boldsymbol{v}} g_1(x)| \leqslant \sum_{j=0}^{l-1} \left| \partial^{\boldsymbol{v}} \left(\partial_{x_d}^{j+1} f(x_1, \dots, x_{d-1}, 0) \frac{x_d^j}{j!} \right) \right| \leqslant C \|\partial_{x_d} f\|_{C^k(D)}.$$

For g_2, we first observe that with the change of variables $t = x_d s$, we obtain

$$g_2(x) = \int_0^1 \partial_{x_d}^{k+1} f(x_1, \dots, x_{d-1}, x_d s) \frac{x_d^l (1-s)^l}{l!} \, \mathrm{d}s.$$

Exchanging the integral with the derivative and repeatedly applying the product rule we find

$$|\partial^{\boldsymbol{v}} g_2(x)| \leqslant \int_0^1 \left| \partial^{\boldsymbol{v}} \left(\partial_{x_d}^{k+1} f(x-1, \dots, x_{d-1}, x_d s) \frac{x_d^k (1-s)^k}{k!} \right) \right| \mathrm{d}s \leqslant C \|\partial_{x_d} f\|_{C^k(D)}.$$

■

Lemma 27 *Consider the setting of Lemma 26, and additionally assume that $f(x) = 0$ whenever $x_d = 1$. Then, $g(x) := \frac{f(x)}{x_d(1-x_d)} \in C^k(D)$ and there exists $C = C(k, d)$ such that $\|g\|_{C^k(D)} \leqslant C \|\partial_{x_d} f\|_{C^k(D)}$.*

Proof We already know from the preceding lemma that the map $x \mapsto f(x)/x_d$ belongs to C^k. In order to show that $g \in C^k$, we only need to prove that the restrction of g to the 'half-cube' $\tilde{D} = \{x \in D : x_d \geqslant 1/2\}$ belongs to C^k. To this end, let us define

$$\tilde{f}(x) := \frac{f(x_1, \dots, x_{d-1}, 1 - x_d)}{1 - x_d}.$$

Clearly, showing that $g \in C^k(\tilde{D})$ is equivalent to showing that $x \mapsto \tilde{f}(x)/x_d$, restricted to the 'other half-cube' $\{x \in D : x_d \leqslant 1/2\}$. For this, we just need to show that $\tilde{f}$ satisfies the conditions of Lemma 26. Since $1 - x_d$ is bounded below when $x_d \leqslant 1/2$, clearly $\tilde{f}$ has the needed regularity. Moreover, for any x with $x_d = 0$, $\tilde{f}(x) = f(x_1, \dots, x_{d-1}, 1) = 0$ by assumption. We may thus apply Lemma 26 and the proof is complete. ■

Step 3: The main argument

With the previous lemmas in hand, we are now ready to prove Theorem 9.

Proof [Proof of Theorem 9] The first assertion **(i)** of the Theorem is proven in Marzouk et al. (2023), we thus only need to show the second part.

Let $p_0 \in \mathcal{M}(k, L_1, L_2)$. Then, it is proven in Wang and Marzouk (2022) that the unique KR map T between p_0 and ρ belongs to the anisotropic regularity class $\mathcal{A}^k_{diag}(k, L', c'_{mon})$ for some L', c'_{mon}. Since the identity map $\mathrm{Id} : x \mapsto x$ also belongs to $\mathcal{A}^k_{diag}(k, L', c'_{mon})$, and since $\mathcal{A}^k_{diag}(k, L', c'_{mon})$ is a convex set, we know that for any $t \in [0, 1]$, $G_t = tT + (1-t)\mathrm{Id}$ also belongs to $\mathcal{A}^k_{diag}(k, L', c'_{mon})$. By Lemma 25, it follows that for any $t \in [0, 1]$, $F(\cdot, t)$ also satisfies the same isotropic regularity. As a result, we know that the triangular velocity field $f^{\Delta}_{p_0}$ belongs to the class $C^k_{diag}(D, \mathbb{R}^d)$. Moreover, since $f^{\Delta}_{p_0}$ is the difference between two bijective triangular maps $D \to D$, we know that for every $j \in [d]$ and for every $x_{[j-1]} \in [0, 1]^{j-1}$ each component map satisfies $(f^{\Delta}_{p_0})_j(x_{[j-1]}, 0) = (f^{\Delta}_{p_0})_j(x_{[j-1]}, 1) = 0$. Thus, all the assumptions of Lemma 27 are satisfied, and it follows that for every $j \in [d]$, the function

$$g_j(x) = \frac{(f^{\Delta}_{p_0})_j(x)}{x_j(1 - x_j)}$$

belongs to $C^k(D)$. The corresponding norm bound for g_j also follows from Lemma 27. ■

B.2 Proof of Theorem 10

Metric entropy bounds for Hölder-Zygmund spaces

To prove Theorem 10, we will begin by deriving the necessary metric entropy bounds for $\mathcal{F}(r)$. For non-integer $s > 0$ we denote by C^s the standard Hölder spaces of $\lfloor s \rfloor$-times differentiable functions with $s - \lfloor s \rfloor$-Hölder continuous s-th partial derivatives, normed by

$$\|f\|_{C^s(\Omega)} = \|f\|_{C^{\lfloor s \rfloor}(\Omega)} + \max_{|\boldsymbol{\alpha}| = \lfloor s \rfloor} \sup_{x \neq y \in \Omega} \frac{|\partial^{\boldsymbol{\alpha}} f(x) - \partial^{\boldsymbol{\alpha}} f(y)|}{|x - y|^{s - \lfloor s \rfloor}}.$$

For $s \geqslant 0$, we will further denote by $B^s_{\infty\infty}(\Omega)$ the classical Besov spaces with indices $p = q = \infty$; see Triebel (2008) for definitions. It is well known that those spaces are equal to the Hölder-Zygmund spaces $\mathcal{C}^s(\Omega)$, $B^s_{\infty\infty}(\Omega) = \mathcal{C}^s(\Omega)$. Moreover, for non-integer $s > 0$, they are equivalent to Hölder spaces,

$$B^s_{\infty\infty}(\Omega) = \mathcal{C}^s(\Omega) = C^s(\Omega).$$

For any $s > 0$ and $R > 0$, let us denote the closed ball with radius R in $B^s_{\infty\infty}(\Omega)$ by

$$A^s(R) := \{f \in L^2(\Omega) : \|f\|_{B^s_{\infty\infty}(\Omega)} \leqslant R\}, \qquad R > 0.$$

The following lemma on metric entropies of Besov spaces is based on classical results which can be found e.g., in Triebel (2008).

Lemma 28 *Let $s_1, s_2 > 0$ and $s_1 > s_2$. Then, there exists some constant $C = C(d, s_1, s_2) > 0$ such that for any $R > 0$, $\tau > 0$,*

$$H(A^{s_1}(R), B^{s_2}_{\infty\infty}(\Omega), \tau) \leqslant C\big(R/\tau\big)^{\frac{d}{s_1 - s_2}}.$$

Proof This result follows from Theorem 4.33 in Triebel (2008) with $p_0, p_1, q_0, q_1 = \infty$ and s_1, s_2 in place of s_0, s_1 there. Note that with those choices, the requirement (4.126) in Triebel (2008) is satisfied. Indeed, the theorem in Triebel (2008) implies that for any k, the unit ball $A^{s_1}(1)$ in $B^{s_1}_{\infty\infty}$ can be covered by 2^k many balls of $\|\cdot\|_{B^{s_2}_{\infty\infty}}$-radius at most $ck^{-\frac{s_1-s_2}{d}}$, where $c > 0$ is some constant. Therefore, given any $\tau > 0$, setting $k_\tau = \lfloor (\tau/c)^{-\frac{d}{s_1-s_2}} \rfloor + 1$, we obtain that the τ-covering number of $A^s(1)$ is upper bounded by

$$H(A^{s_1}(1), B^{s_2}_{\infty\infty}(\Omega), \tau) \leqslant \log(2^{k_\tau}) = (\lfloor (\tau/c)^{-\frac{d}{s_1-s_2}} \rfloor + 1) \log 2 \lesssim \tau^{-\frac{d}{s_1-s_2}}$$

for any $\tau \leqslant 1$. [Note that for $\tau \geqslant 1$, we have that the left hand side is 0, so that the upper bound in the lemma trivially holds true for $R = 1$.] Then, the result for covering $A^{s_1}(R)$ follows from noting that $H(A^{s_1}(R), B^{s_2}_{\infty\infty}(\Omega), \tau) = H(A^{s_1}(1), B^{s_2}_{\infty\infty}(\Omega), \tau/R)$. ∎

Proof of Theorem 10

Proof Let $p_0 \in \mathcal{M}(k, L_1, L_2)$. Since $k \geqslant 2$, clearly the Assumptions 2.1, 2.2 are fulfilled. By Theorem 9, the velocity field $f^{\Delta}_{p_0}$ coupling p_0 and ρ belongs to $C^k \cap \mathcal{V}$, and we have that

$$\sup_{p_0 \in \mathcal{M}(k, L_1, L_2)} \|f^{\Delta}_{p_0}\|_{C^k(\Omega)} =: \bar{L} < \infty.$$

Thus, by choosing $r > \bar{L}$ we can ensure that $f^{\Delta}_{p_0} \in \mathcal{F}(r)$. We will now employ Theorem 2. By what precedes, we may choose $f^* = f^{\Delta}_{p_0}$, so that $(T^{f^*})^\sharp \rho = p^* = p_0$. We now calculate the metric entropy integral of $\mathcal{F}(r)$ in the $C^1(\Omega)$-norm. To do so, let us fix $\gamma \in (0,1) \cap k - d/2 - 3/2$, we have that

$$B^{1+\gamma}_{\infty\infty}(\Omega) = \mathcal{C}^{1+\gamma}(\Omega) = C^{1+\gamma}(\Omega) \subseteq C^1(\Omega),$$

where the last inclusion is a continuous embedding.

Combining the preceding bounds and using Lemma 28, it follows that for all $R > 0$ and some constants $0 < C_1, C_2, C_3 < \infty$,

$$\begin{aligned}
\int_0^R H^{1/2}(\mathcal{F}(r), C^1(\Omega), \tau) d\tau &\leqslant \int_0^R H^{1/2}(\mathcal{F}(r), B^{1+\gamma}_{\infty\infty}(\Omega), C_1\tau) d\tau \\
&\leqslant \int_0^R H^{1/2}(A^k_{\infty\infty}(C_2 r), B^{1+\gamma}_{\infty\infty}(\Omega), C_1\tau) d\tau \\
&\leqslant C_3 \int_0^R \left(\frac{C_2 r}{C_1 \tau}\right)^{\frac{d+1}{2(k-1-\gamma)}} d\tau \\
&\lesssim R^{1-\frac{d+1}{2(k-1-\gamma)}}.
\end{aligned}$$

Thus, the requirement (2.9) from Theorem 2 simplifies to

$$\sqrt{n}\delta_n^2 \gtrsim \delta_n + \delta_n^{1-\frac{d+1}{2(k-1-\gamma)}}.$$

This is satisfied if both $\delta_n \gtrsim n^{-1/2}$ as well as

$$\sqrt{n} \gtrsim \delta_n^{-1-\frac{d+1}{2(k-1-\gamma)}}, \qquad \text{which is equivalent to} \qquad \delta_n \gtrsim n^{-\frac{k-1-\gamma}{2(k-1-\gamma)+d+1}}.$$

The desired result now follows directly from Theorem 2. ■

Appendix C. Auxiliary Results for Section 4.1

In this appendix, we prove some auxiliary results about the uniform boundedness and Lipschitz properties of the ReLU^2 neural network class $\Phi^{d_1,1}(L,W,S,B)$ and its gradient space $\nabla\Phi^{d_1,1}(L,W,S,B)$, which will be used in the proof of Theorem 14. Our arguments are similar to those in Schmidt-Hieber (2020), Suzuki (2019), and Lu et al. (2021) with two key differences: (i) to ensure smoothness of the gradient space, we consider ReLU^2 networks, whereas Schmidt-Hieber (2020) and Suzuki (2019) consider ReLU networks, and Lu et al. (2021) considers ReLU^3 networks; and (ii) to obtain the C^1 metric entropy rate in Theorem 14, we construct a covering of both the NN function space and its gradient space.

Lemma 29 *For any $1 \leqslant l \leqslant L$, the following inequality holds for the class of ReLU^2 networks $\Phi^{d_1,1}(L,W,S,B)$:*

$$\sup_{x\in D, F_l\in\Phi_l^{d_1,1}(L,W,S,B)} \|F_l(x)\|_\infty \leqslant C_l W^{2^{l-1}-1}(B\vee d_1)^{2^l-1},$$

where C_l is a constant independent of W, B, d_1, depending only on l.

Proof [Proof of Lemma 29] We prove the lemma by induction. First note for any matrix $A\in\mathbb{R}^{d\times d}$, $\|A\|_{\infty,\infty}\leqslant B$ implies $\|A\|_\infty \leqslant dB$. When $l=1$, we have for all $x\in D$,

$$\|F_1(x)\|_\infty = \|W_F^{(1)}x + b_F^{(1)}\|_\infty \leqslant \|W_F^{(1)}\|_\infty\|x\|_\infty + \|b_F^{(1)}\|_\infty \leqslant d_1B + B \leqslant 2(B\vee d_1)^2.$$

Assuming the claim holds for $l-1$, where $l\geqslant 2$, we have that

$$\begin{aligned}
\|F_l(x)\|_\infty &= \|W_F^{(l)}\eta_2(F_{l-1}(x)) + b_F^{(l)}\|_\infty \leqslant WB\|F_{l-1}(x)\|_\infty^2 + B\\
&\leqslant W(B\vee d_1)\left(C_{l-1}W^{2^{l-2}-1}(B\vee d_1)^{2^{l-1}-1}\right)^2 + B\\
&\leqslant C_{l-1}^2 W^{2^{l-1}-2+1}(B\vee d_1)^{2^l-2+1} + (B\vee d_1)\\
&\leqslant (C_{l-1}^2+1)W^{2^{l-1}-1}(B\vee d_1)^{2^l-1} = C_l W^{2^{l-1}-1}(B\vee d_1)^{2^l-1}.
\end{aligned}$$

Hence the claim follows from induction. ■

Lemma 30 *For any $1 \leqslant l \leqslant L$, suppose that a pair of two different ReLU^2 networks $F_l, G_l \in \Phi_l^{d_1,1}(L,W,S,B)$ are given by*

$$F_l(x) = (W_F^{(l)}\eta_2(\cdot) + b_F^{(l)}) \circ \cdots \circ (W_F^{(1)}\eta_2(\cdot) + b_F^{(1)}),$$

$$G_l(x) = (W_G^{(l)}\eta_2(\cdot) + b_G^{(l)}) \circ \cdots \circ (W_G^{(1)}\eta_2(\cdot) + b_G^{(1)}).$$

Assume that the l_∞ norm between the neural network weights is uniformly upper bounded by δ, i.e., $\|W_F^{(l')} - W_G^{(l')}\|_{\infty,\infty} \leqslant \delta, \|b_F^{(l')} - b_G^{(l')}\|_\infty \leqslant \delta$, for all $1 \leqslant l' \leqslant l$. Then we have

$$\sup_{x\in D} \|F_l(x) - G_l(x)\|_\infty \leqslant A_l \delta W^{2^{l-1}-1}(B \vee d_1)^{2^l},$$

for some constant A_l that only depends on l.

Proof [Proof of Lemma 30] We prove the lemma by induction. For any $x \in D$ and $F_1, G_1 \in \Phi_1^{d_1,1}(L, S, W, B)$, it holds that

$$\begin{aligned}
\|F_1(x) - G_1(x)\|_\infty &= \|W_F^{(1)}x + b_F^{(1)} - W_G^{(1)}x - b_G^{(1)}\|_\infty \\
&\leqslant \|W_F^{(1)} - W_G^{(1)}\|_\infty \|x\|_\infty + \|b_F^{(1)} - b_G^{(1)}\|_\infty \\
&\leqslant \delta d_1 + \delta = \delta(d_1 + 1) \leqslant 2\delta(B \vee d_1) \leqslant 2\delta(B \vee d_1)^2.
\end{aligned}$$

Now suppose the claim holds for $l-1$. For the induction step, we will use that $\eta_2(x) = x^2$ satisfies $|\eta_2(x) - \eta_2(y)| \leqslant 2\max\{|x|, |y|\}|x-y|$. Thus, for any $x \in D$ and $F_l, G_l \in \Phi_l^{d_1,1}(L, W, S, B)$, we have

$$\begin{aligned}
\|F_l(x) - G_l(x)\|_\infty &= \|W_F^{(l)}\eta_2(F_{l-1}(x)) + b_F^{(l)} - W_G^{(l)}\eta_2(G_{l-1}(x)) - b_G^{(l)}\|_\infty \\
&\leqslant \|W_F^{(l)}\eta_2(F_{l-1}(x)) - W_G^{(l)}\eta_2(G_{l-1}(x))\|_\infty + \|b_F^{(l)} - b_G^{(l)}\|_\infty \\
&\leqslant \|W_F^{(l)}\eta_2(F_{l-1}(x)) - W_G^{(l)}\eta_2(F_{l-1}(x))\|_\infty \\
&+ \|W_G^{(l)}\eta_2(F_{l-1}(x)) - W_G^{(l)}\eta_2(G_{l-1}(x))\|_\infty + \delta \\
&\leqslant \|W_F^{(l)} - W_G^{(l)}\|_\infty \|\eta_2(F_{l-1}(x))\|_\infty \\
&+ \|W_G^{(l)}\|_\infty \|\eta_2(F_{l-1}(x)) - \eta_2(G_{l-1}(x))\|_\infty + \delta \\
&\leqslant W\delta \|F_{l-1}(x)\|_\infty^2 \\
&+ WB(2 \sup_{F_{l-1}\in\Phi_{l-1}^{d_1,1}(L,W,S,B)} \|F_{l-1}(x)\|_\infty) \|F_{l-1}(x) - G_{l-1}(x)\|_\infty + \delta \\
&\leqslant W\delta(C_{l-1}W^{2^{l-2}-1}(B \vee d_1)^{2^{l-1}-1})^2 \\
&+ 2WB(C_{l-1}W^{2^{l-2}-1}(B \vee d_1)^{2^{l-1}-1})(A_{l-1}\delta W^{2^{l-2}-1}(B \vee d_1)^{2^{l-1}}) + \delta \\
&\leqslant \delta C_{l-1}^2 W^{2^{l-1}-1}(B \vee d_1)^{2^l-2} \\
&+ 2\delta C_{l-1}A_{l-1}W^{2^{l-1}-1}(B \vee d_1)^{1+2^{l-1}-1+2^{l-1}} + \delta \\
&\leqslant \delta W^{2^{l-1}-1}\left(C_{l-1}^2(B \vee d_1)^{2^l-2} + 2C_{l-1}A_{l-1}(B \vee d_1)^{2^l} + 1\right) \\
&\leqslant A_l\delta W^{2^{l-1}-1}(B \vee d_1)^{2^l},
\end{aligned}$$

for some constant A_l that only depends on l. Hence the claim follows from induction. ■

Lemma 31 *For any $1 \leqslant l \leqslant L$, the following inequality holds for the class of ReLU^2 networks:*

$$\sup_{x \in D, F_l \in \Phi_l^{d_1,1}(L,W,S,B)} \|\nabla F_l(x)\|_\infty \leqslant M_k W^{2^{l-1}-1}(B \vee d_1)^{2^l},$$

for some constant M_l that only depends on l.

Proof We prove the lemma by induction. When $l = 1$, it holds that for all $x \in D$

$$\|\nabla F_1(x)\|_\infty \leqslant \|W_F^{(1)}\|_\infty \leqslant d_1 B \leqslant (B \vee d_1)^2.$$

Suppose the claim holds for $l-1$. Then, we may compute

$$\begin{aligned}\|\nabla F_l(x)\|_\infty &= \|W_F^{(l)} \nabla[\eta_2 \circ F_{l-1}](x)\|_\infty \leqslant \|W_F^{(l)}\|_\infty \|\nabla[\eta_2 \circ F_{l-1}](x)\|_\infty \\ &\leqslant WB\|\nabla[\eta_2 \circ F_{l-1}](x)\|_\infty \leqslant W(B \vee d_1)\|\nabla[\eta_2 \circ F_{l-1}](x)\|_\infty.\end{aligned}$$

Since the operator ∞-norm of a matrix equals the maximum row sum, we have

$$\begin{aligned}&\|\nabla[\eta_2 \circ F_{l-1}](x)\|_\infty = \sup_{1 \leqslant j \leqslant W} \sum_{i=1}^{d_1} |\eta_2'(F_{l-1,j}(x)) \frac{\partial F_{l-1,j}}{\partial x_i}| \\ &\leqslant 2\|F_{l-1}\|_\infty \sup_{1 \leqslant j \leqslant W} \sum_{i=1}^{d_1} |\frac{\partial F_{l-1,j}}{\partial x_i}| \\ &\leqslant 2C_{l-1} W^{2^{l-2}-1}(B \vee d_1)^{2^{l-1}-1} \|\nabla F_{l-1}(x)\|_\infty.\end{aligned}$$

Then, we get

$$\begin{aligned}&\|\nabla F_l(x)\|_\infty \leqslant W(B \vee d_1)\|\nabla[\eta_2 \circ F_{l-1}](x)\|_\infty \\ &\leqslant W(B \vee d_1) 2C_{l-1} W^{2^{l-2}-1}(B \vee d_1)^{2^{l-1}-1} \|\nabla F_{l-1}(x)\|_\infty \\ &\leqslant W(B \vee d_1) 2C_{l-1} W^{2^{l-2}-1}(B \vee d_1)^{2^{l-1}-1} W^{2^{l-2}-1}(B \vee d_1)^{2^{l-1}} \\ &\leqslant M_l W^{2^{l-1}-1}(B \vee d_1)^{2^l},\end{aligned}$$

if we absorb all the constants into M_l. The claim then follows from induction. ∎

Lemma 32 *For any $1 \leqslant l \leqslant L$, suppose that a pair of two different ReLU^2 networks $F_l, G_l \in \Phi_l^{d_1,1}(L,W,S,B)$ are given by*

$$F_l(x) = (W_F^{(l)} \eta(\cdot) + b_F^{(l)}) \circ \cdots \circ (W_F^{(1)} \eta(\cdot) + b_F^{(1)}),$$

$$G_l(x) = (W_G^{(l)} \eta(\cdot) + b_G^{(l)}) \circ \cdots \circ (W_G^{(1)} \eta(\cdot) + b_G^{(1)}).$$

Assume that the l_∞ norm between the neural network weights is uniformly upper bounded by δ, i.e., $\|W_F^{(l')} - W_G^{(l')}\|_{\infty,\infty} \leqslant \delta, \|b_F^{(l')} - b_G^{(l')}\|_\infty \leqslant \delta, 1 \leqslant l' \leqslant l$. Then we have

$$\sup_{x \in D} \|\nabla F_l(x) - \nabla G_l(x)\|_\infty \leqslant \delta N_l W^{2^{l-1}-1}(B \vee d_1)^{2^l+1},$$

where N_l is a constant the only depends on l.

Proof [Proof of Lemma 32] We prove this lemma by induction. When $l = 1$, it holds that for all $x \in D$,

$$\|\nabla F_1(x) - \nabla G_1(x)\|_\infty = \|W_F^{(1)} - W_G^{(1)}\|_\infty \leqslant \delta d_1 \leqslant \delta(B \vee d_1) \leqslant \delta(B \vee d_1)^3.$$

Assume that the claim holds for $l-1$. Then, for any $x \in D$, and $F_l, G_l \in \Phi_l^{d_1,1}(L, S, W, B)$ satisfying the conditions in the lemma, we can bound $\|\nabla F_l(x) - \nabla G_l(x)\|_\infty$ using the chain rule and triangular inequality as follows:

$$\begin{aligned}
&\|\nabla F_l(x) - \nabla G_l(x)\|_\infty = \|W_F^{(l)}\nabla[\eta_2 \circ F_{l-1}](x) - W_G^{(l)}\nabla[\eta_2 \circ G_{l-1}](x)\|_\infty \\
&\leqslant \|W_F^{(l)}\nabla[\eta_2 \circ F_{l-1}](x) - W_G^{(l)}\nabla[\eta_2 \circ F_{l-1}](x)\|_\infty \\
&+ \|W_G^{(l)}\nabla[\eta_2 \circ F_{l-1}](x) - W_G^{(l)}\nabla[\eta_2 \circ G_{l-1}](x)\|_\infty \\
&\leqslant \|W_F^{(l)} - W_G^{(l)}\|_\infty\|\nabla[\eta_2 \circ F_{l-1}](x)\|_\infty \\
&+ \|W_G^{(l)}\|_\infty\|\nabla[\eta_2 \circ F_{l-1}](x) - \nabla[\eta_2 \circ G_{l-1}](x)\|_\infty \\
&\leqslant \delta W\|\nabla[\eta_2 \circ F_{l-1}](x)\|_\infty + BW\|\nabla[\eta_2 \circ F_{l-1}](x) - \nabla[\eta_2 \circ G_{l-1}](x)\|_\infty \\
&:= I + II.
\end{aligned}$$

From Lemma 29 and 31, we can bound I by

$$\begin{aligned}
I &= \delta W\|\nabla[\eta_2 \circ F_{l-1}](x)\|_\infty \leqslant \delta W 2\|F_{l-1}\|_\infty\|\nabla[F_{l-1}](x)\|_\infty \\
&\leqslant 2\delta W C_{l-1} W^{2^{l-2}-1}(B \vee d_1)^{2^{l-1}-1} M_{l-1} W^{2^{l-2}-1}(B \vee d_1)^{2^{l-1}} \\
&= 2C_{l-1}M_{l-1}\delta W^{2^{l-1}-1}(B \vee d_1)^{2^l-1}
\end{aligned}$$

To bound II, note that

$$\begin{aligned}
II &= BW\|\nabla[\eta_2 \circ F_{l-1}](x) - \nabla[\eta_2 \circ G_{l-1}](x)\|_\infty \\
&= BW \sup_{1 \leqslant j \leqslant W^{(l-1)}} \Big(\sum_{i=1}^{d_1} |\eta_2'(F_{l-1,j})\frac{\partial F_{l-1,j}}{\partial x_i} - \eta_2'(G_{l-1,j})\frac{\partial G_{l-1,j}}{\partial x_i}|\Big),
\end{aligned}$$

and it holds that for all $j, 1 \leqslant j \leqslant W^{(l-1)}$,

$$\begin{aligned}
&\sum_{i=1}^{d_1} |\eta_2'(F_{l-1,j})\frac{\partial F_{l-1,j}}{\partial x_i} - \eta_2'(G_{l-1,j})\frac{\partial G_{l-1,j}}{\partial x_i}| \\
&\leqslant \sum_{i=1}^{d_1} |\eta_2'(F_{l-1,j})\frac{\partial F_{l-1,j}}{\partial x_i} - \eta_2'(G_{l-1,j})\frac{\partial F_{l-1,j}}{\partial x_i}| \\
&+ \sum_{i=1}^{d_1} |\eta_2'(G_{l-1,j})\frac{\partial F_{l-1,j}}{\partial x_i} - \eta_2'(G_{l-1,j})\frac{\partial G_{l-1,j}}{\partial x_i}| := III + IV.
\end{aligned}$$

III can be bounded as follows:

$$\begin{aligned}
III &= \sum_{i=1}^{d_1} |\eta_2'(F_{l-1,j})\frac{\partial F_{l-1,j}}{\partial x_i} - \eta_2'(G_{l-1,j})\frac{\partial F_{l-1,j}}{\partial x_i}| \\
&\leqslant \sum_{i=1}^{d_1} |\eta_2'(F_{l-1,j}) - \eta_2'(G_{l-1,j})||\frac{\partial F_{l-1,j}}{\partial x_i}| \\
&\leqslant 2\|F_{l-1} - G_{l-1}\|_\infty \sum_{i=1}^{d_1} |\frac{\partial F_{l-1,j}}{\partial x_i}| \leqslant 2\|F_{l-1} - G_{l-1}\|_\infty \|\nabla[F_{l-1}](x)\|_\infty \\
&\leqslant A_{l-1}\delta W^{2^{l-2}-1}(B \vee d_1)^{2^{l-1}} M_{l-1} W^{2^{l-2}-1}(B \vee d_1)^{2^{l-1}} \\
&= \delta A_{l-1} M_{l-1} W^{2^{l-1}-2}(B \vee d_1)^{2^l},
\end{aligned}$$

where the last inequality follows from Lemma 30 and 31.

Applying the inductive hypothesis and Lemma 29, IV can be bounded as follows:

$$\begin{aligned}
IV &= \sum_{i=1}^{d_1} |\eta_2'(G_{l-1,j})\frac{\partial F_{l-1,j}}{\partial x_i} - \eta_2'(G_{l-1,j})\frac{\partial G_{l-1,j}}{\partial x_i}| \\
&\leqslant 2\sup_{x\in D} \|G_{l-1}(x)\|_\infty \sum_{i=1}^{d_1} |\frac{\partial F_{l-1,j}}{\partial x_i} - \frac{\partial G_{l-1,j}}{\partial x_i}| \\
&\leqslant 2\sup_{x\in D} \|G_{l-1}(x)\|_\infty \|\nabla[F_{l-1}](x) - \nabla[G_{l-1}](x)\|_\infty \\
&\leqslant 2C_{l-1} W^{2^{l-2}-1}(B \vee d_1)^{2^{l-1}-1}(\delta N_{l-1} W^{2^{l-2}-1}(B \vee d_1)^{2^{l-1}+1})
\end{aligned}$$

Putting everything together, $\|\nabla F_l(x) - \nabla G_l(x)\|_\infty$ is upper bounded by:

$$\begin{aligned}
&\delta 2C_{l-1} M_{l-1} W^{2^{l-1}-1}(B \vee d_1)^{2^l-1} + \delta B A_{l-1} M_{l-1} W^{2^{l-1}-1}(B \vee d_1)^{2^l} + \\
&BW(2C_{l-1} W^{2^{l-2}-1}(B \vee d_1)^{2^{l-1}-1}(\delta N_{l-1} W^{2^{l-2}-1}(B \vee d_1)^{2^{l-1}+1})) \\
&\leqslant 2\delta C_{l-1} M_{l-1} W^{2^{l-1}-1}(B \vee d_1)^{2^l-1} + \delta A_{l-1} M_{l-1} W^{2^{l-1}-1}(B \vee d_1)^{2^l+1} \\
&+ 2\delta C_{l-1} N_{l-1} W^{2^{l-1}-1}(B \vee d_1)^{2^l+1} \leqslant \delta N_l W^{2^{l-1}-1}(B \vee d_1)^{2^l+1},
\end{aligned}$$

where N_l is a constant that only depends on l. Hence the claim follows by induction.

■

Appendix D. Neural Network Approximation Theory

In the following we work with the standard normalized one-dimensional B-spline of order $m \geqslant 1$ with equidistant knots, see e.g., (Schumaker, 2007, (4.46)–(4.47)):

$$B^m(x) := \sum_{i=0}^{m} (-1)^i \frac{\binom{m}{i} \max\{0, x-i\}^{m-1}}{(m-1)!} \in W^{m-1,\infty}(\mathbb{R}) \tag{D.1a}$$

where $0^0 := 0$. Additionally, for $n \in \mathbb{N}$ we[2] consider the stretched and shifted versions (Schumaker, 2007, (4.49))

$$B^m_{n,j}(x) := B^m(nx - j) \in W^{m-1,\infty}(\mathbb{R}), \qquad j \in \mathbb{Z}. \tag{D.1b}$$

Note that $B^m_{n,j}|_{[0,1]} \in C^{m-2}([0,1])$ is a piecewise polynomial of degree $m-1$ on the intervals $[\frac{j}{n}, \frac{j+1}{n}]$, and thus the function is C^∞ on

$$M_n := [0,1] \backslash \left\{ \frac{j}{n} \,\middle|\, 1 \leqslant j \leqslant n-1 \right\}.$$

Moreover $\mathrm{supp}(B^m_{n,j}) \subseteq [\frac{j}{n}, \frac{m+j}{n}]$.

D.1 One Dimensional Spline Approximation

It is well-known that one can construct continuous linear functionals $\lambda^m_{n,j} : C([0,1]) \to \mathbb{R}$ such that

$$Q^m_n[f](x) := \sum_{j=-m+1}^{n-1} \lambda^m_{n,j}[f] B^m_{n,j}(x) \tag{D.2}$$

yields an approximation to f that converges at a rate depending on the regularity $k \in \mathbb{N}$ of the target function $f \in C^k([0,1])$ as long as the the order $m \in \mathbb{N}$ of the spline is larger or equal to $k+1$.[3] While various approximation results for Sobolev or Besov spaces have been established in the literature, e.g., Oswald (1990), for our purposes approximation of C^k functions as stated in the following variant[4] of (Schumaker, 2007, Theorem 6.20) is sufficient:

Theorem 33 *Let $k \in \mathbb{N}_0$, $m \in \mathbb{N}$ and $k+1 \leqslant m$. Then there exists $C = C(k,m)$ such that for every $n \in \mathbb{N}$, there exist continuous (w.r.t. the topology of pointwise convergence) linear functionals $\lambda^m_{n,j} : C([0,1]) \to \mathbb{R}$, $j \in \{-m+1, \ldots, n-1\}$, such that*

1. *for all $n \in \mathbb{N}$, $j \in \{-m+1, \ldots, n-1\}$, $f \in C([0,1])$*

$$|\lambda^m_{n,j}[f]| \leqslant C \|f\|_{C([0,1])}, \tag{D.3}$$

2. *for all $r \in \{0, \ldots, k\}$, $f \in C^k([0,1])$ and with Q^m_n as in* (D.2)

$$\sup_{x \in M_n} \left| \frac{d^r}{dx^r} (f - Q^m_n[f]) \right| \leqslant C n^{-(k-r)} |f|_{C^k([0,1])}. \tag{D.4}$$

Proof We proceed in three steps: In step 1 we show an extension result for functions in $C^k([0,1])$, in step 2 we verify the error bound (D.4) and in step 3 we show continuity of the $\lambda^m_{n,j}$ and (D.3).

---

2. In Appendix D only, n denotes a stretching parameter rather than the sample size in the maximum likelihood estimation problem (2.5).
3. Here and throughout Appendix D, we always interpret continuity of a functional from $C([0,1]^d) \to \mathbb{R}$ w.r.t. to the topology of pointwise convergence on $C([0,1]^d)$.
4. The main difference to the presentation in Schumaker (2007) is our treatment of the boundary, which avoids the use of different spline basis functions near the endpoints 0 and 1 of the interval.

Step 1. Using standard techniques, we wish to define a bounded linear extension operator $E : C([0,1]) \to C([-m, 1+m])$ that additionally is stable between $C^r([0,1]) \to C^r([-m, 1+m])$ for each $r \in \{0, \dots, k\}$.

Fix distinct numbers $-\frac{1}{m} < \gamma_0 < \cdots < \gamma_k < 0$ and let $g \in C^k([0,1])$. Set $\tilde{g}(x) := g(x)$ if $x \in [0,1]$ and

$$\tilde{g}(x) := \sum_{j=0}^{k} \alpha_j g(\gamma_j x) \qquad \forall x \in [-m, 0], \tag{D.5}$$

for certain $\alpha_j \in \mathbb{R}$ that remain to be determined. It holds $\tilde{g} \in C^k([-m,1])$ iff $\tilde{g}^{(r)}(0) = g^{(r)}(0)$ for all $r \in \{0, \dots, k\}$, i.e.,

$$g^{(r)}(0) = g^{(r)}(0) \sum_{j=0}^{k} \alpha_j \gamma_j^r \qquad \forall r \in \{0, \dots, k\}.$$

This condition being satisfied for arbitrary $g \in C^k([0,1])$ is equivalent to

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ \gamma_0^1 & \gamma_1^1 & \cdots & \gamma_k^1 \\ \vdots & & \ddots & \vdots \\ \gamma_0^k & \gamma_1^k & \cdots & \gamma_k^k \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_k \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{D.6}$$

Since the matrix on the left-hand side is a Vandermonde matrix with distinct nodes $\gamma_0, \dots, \gamma_k$, it is regular. Hence there exists a unique set of numbers $(\alpha_j)_{j=0}^k$ satisfying (D.6).

In the same fashion $\tilde{g}(x)$ can be extended to $x \in [1, 1+m]$. This yields a linear map $E : C([0,1]) \to C([-m, 1+m])$ that evidently (cp. (D.5)) satisfies

$$\|Eg\|_{C^r([-m,1+m])} \leqslant C\|g\|_{C^r([0,1])} \qquad \forall g \in C^k([0,1]),\ \forall r \in \{0, \dots, k\}, \tag{D.7}$$

for some constant C depending on $(\gamma_j)_{j=0}^k$ and $(\alpha_j)_{j=0}^k$ (and hence on k and m) but independent of g.

Step 2. According to (Schumaker, 2007, Theorem 6.20), there exist bounded linear functionals $\tilde{\lambda}_{n,j}^m : C([-m, m+1]) \to \mathbb{R}$ such that for each $l \in \{0, \dots, n-1\}$ and $r \in \{0, \dots, k\}$ it holds[5]

$$\left\| \frac{d^r}{dx^r} \left(f - \sum_{j=l-m+1}^{l} \tilde{\lambda}_{n,j}^m (Ef) B_{n,j}^m(x) \right) \right\|_{L^\infty([\frac{l}{n}, \frac{l+1}{n}])} \leqslant C n^{-(k-r)} \omega\Big((Ef)^{(k)}, \frac{1}{n} \Big)_{C[\frac{l+1-m}{n}, \frac{l+m}{n}]}, \tag{D.8}$$

where $C = C(m)$ is independent of f, l and n, and

$$\omega\Big((Ef)^{(k)}, \frac{1}{n} \Big)_{C[\frac{l+1-m}{n}, \frac{l+m}{n}]} = \sup_{\substack{x,y \in [\frac{l+1-m}{n}, \frac{l+m}{n}] \\ |x-y| \leqslant \frac{1}{n}}} |(Ef)^{(k)}(x) - (Ef)^{(k)}(y)|$$

5. In the notation of (Schumaker, 2007, Theorem 6.20), we use equidistant knots "$y_l := \frac{l}{n}$" for $l \in \{-mn, \dots, (1+m)n\}$ on the interval "$[a,b] := [-m, 1+m]$" with "$\sigma := k+1$" and "$q := \infty$".

denotes the modulus of continuity for the kth derivative of Ef. Using (D.7) this term can be bounded by $2|Ef|_{C^k([-m,1+m])} \leqslant 2C|f|_{C^k([0,1])}$.

With

$$\lambda^m_{n,j}[f] := \tilde{\lambda}^m_{n,j}[Ef] \qquad j \in \{-m+1, \dots, n-1\} \tag{D.9}$$

we obtain by (D.2)

$$Q^m_n[f] = \sum_{j=-m+1}^{n-1} \lambda^m_{n,j}[f] B^m_{n,j}(x) = \sum_{j=-m+1}^{n-1} \tilde{\lambda}^m_{n,j}[Ef] B^m_{n,j}(x).$$

Since $\mathrm{supp}(B^m_{n,j}) \subseteq [\frac{j}{n}, \frac{j+m}{n}]$ as pointed out earlier, (D.8) shows the error bound (D.4) on the interval $[\frac{l}{n}, \frac{l+1}{n}]$. Because $l \in \{0, \dots, n-1\}$ was arbitrary, this shows (D.4).

Step 3. It remains to argue continuity of $\lambda^m_{n,j}$ and the bound (D.3). By construction of $\tilde{\lambda}^m_{n,j}$, see[6] (Schumaker, 2007, (6.39)), for $j \in \{-m+1, \dots, n-1\}$ the term $\tilde{\lambda}^m_{n,j}[f]$ is a linear combination of finitely many point evaluations of f in $[-m+1, m]$. Hence $\tilde{\lambda}^m_{n,j} : C([-m+1, m]) \to \mathbb{R}$ is continuous w.r.t. the topology of pointwise convergence. Now suppose that $(g_i)_{i\in\mathbb{N}} \subseteq C([0,1])$ is a sequence of functions converging pointwise to $g \in C([0,1])$. Then the construction of E (cp. (D.5)) implies that $Eg_i \to Eg \in C([-m+1, m])$ pointwise, and thus by definition of $\lambda^m_{n,j}$ in (D.9)

$$\lambda^m_{n,j}[g] = \tilde{\lambda}^m_{n,j}[Eg_i] \to \tilde{\lambda}^m_{n,j}[Eg] \in \mathbb{R} \qquad \text{as } i \to \infty,$$

which shows the claimed continuity of $\lambda^m_{n,j} : C([0,1]) \to \mathbb{R}$.

Moreover, as shown in the proof of (Schumaker, 2007, Theorem 6.22)

$$|\tilde{\lambda}^m_{n,j}[g]| \leqslant (2m)^m \|g\|_{C([-m,1+m])} \qquad \forall g \in C([-m, 1+m]),$$

so that for any $f \in C([0,1])$

$$|\lambda^m_{n,j}[f]| = |\tilde{\lambda}^m_{n,j}[Ef]| \leqslant (2m)^m \|Ef\|_{C([-m,1+m])} \leqslant C(2m)^m \|f\|_{C([0,1])}$$

for some C depending on k and m but independent of n, j and f. ∎

D.2 Multidimensional Spline Approximation

We next extend Theorem 33 to the multidimensional case. In principle such a statement is provided in (Schumaker, 2007, Theorem 12.7), however this result requires mixed regularity of the target function, which we wish to avoid.

To give the statement, we first introduce some notation. Fix m, $n \in \mathbb{N}$. With $\lambda^m_{n,j} : C([0,1]) \to \mathbb{R}$ as in Theorem 33, for $f \in C([0,1]^d)$ and a multiindex $\boldsymbol{\nu} = (\nu_1, \dots, \nu_d) \in \{-m+1, \dots, n-1\}$ define

$$\boldsymbol{\lambda}^m_{n,\boldsymbol{\nu}}[f] := \lambda^{m,x_d}_{n,\nu_d} \dots \lambda^{m,x_1}_{n,\nu_1}[f] \in \mathbb{R}. \tag{D.10}$$

6. We use the notation $\tilde{\lambda}^m_{n,i}$ for "λ_i" in (Schumaker, 2007, Chapter 6).

Here $\lambda_{n,\nu_i}^{m,x_i} : C([0,1]) \to \mathbb{R}$ is understood to act on the x_i variable only. Additionally with $B_{n,j}^m$ in (D.1b)

$$\boldsymbol{B}_{n,\boldsymbol{\nu}}^m(x_1,\ldots,x_d) := \prod_{i=1}^d B_{n,\nu_i}^m(x_i),$$

and

$$\boldsymbol{Q}_n^m[f](x_1,\ldots,x_d) := \sum_{-m+1\leqslant\nu_1,\ldots,\nu_d\leqslant n-1} \boldsymbol{\lambda}_{n,\boldsymbol{\nu}}^m[f]\boldsymbol{B}_{n,\boldsymbol{\nu}}(x_1,\ldots,x_d).$$

Lemma 34 *Let m, $n \in \mathbb{N}$, $\boldsymbol{\nu} \in \{-m+1,\ldots,n-1\}^d$ and $\nu \in \{-m+1,\ldots,n-1\}$.*

1. *Equation* (D.10) *defines a continuous (w.r.t. the topology of pointwise convergence) linear functional $\lambda_{n,\boldsymbol{\nu}}^m : C([0,1]^d) \to \mathbb{R}$.*
2. *There exists $C = C(m,d)$ independent of n and $\boldsymbol{\nu}$ such that*
$$|\boldsymbol{\lambda}_{n,\boldsymbol{\nu}}^m[f]| \leqslant C\|f\|_{C([0,1]^d)} \qquad \forall f \in C([0,1]^d).$$
3. *If $f \in C^k([0,1]^d)$ then for all $j \in \{1,\ldots,d\}$ and $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ with $|\boldsymbol{\alpha}| \leqslant k$ and $\alpha_j = 0$ it holds*
$$\partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}} \lambda_{n,\nu}^{m,x_j}[f] = \lambda_{n,\nu}^{m,x_j}[\partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}} f] \in C^{k-|\boldsymbol{\alpha}|}([0,1]^{d-1}). \tag{D.11}$$

Proof Throughout fix $\nu \in \{-m+1,\ldots,n-1\}$, $\boldsymbol{\nu} \in \{-m+1,\ldots,n-1\}^d$ and $f \in C([0,1]^d)$ arbitrary.

We first show that $\lambda_{n,\nu}^{m,x_j}[f] \in C([0,1]^{d-1})$ for each $j \in \{1,\ldots,d\}$. This then implies that $\boldsymbol{\lambda}_{n,\boldsymbol{\nu}}^m[f] \in \mathbb{R}$ in (D.10) is well-defined. Wlog $j = 1$. Let $\boldsymbol{x}_i \in [0,1]^{d-1}$, $i \in \mathbb{N}$, be a sequence of points converging to $\boldsymbol{x}^* \in [0,1]^{d-1}$. Then $g_i(x_1) := f(x_1, \boldsymbol{x}_i)$, $i \in \mathbb{N}$, defines a sequence of functions in $C([0,1])$ converging pointwise to $g(x_1) := f(x_1, \boldsymbol{x}^*)$. By Theorem 33 we thus have $\lambda_{n,\nu}^{m,x_1}[g_i] \to \lambda_{n,\nu}^{m,x_1}[g] \in \mathbb{R}$ as $i \to \infty$, i.e., $\lambda_{n,\nu}^{m,x_1}[f](\boldsymbol{x}_i) \to \lambda_{n,\nu}^{m,x_1}[f](\boldsymbol{x}^*)$ as $i \to \infty$. This shows continuity of $\boldsymbol{x} \mapsto \lambda_{n,\nu}^{m,x_1}[f](\boldsymbol{x})$ for $\boldsymbol{x} \in [0,1]^d$.

Next we claim that $\lambda_{n,\nu}^{m,x_j} : C([0,1]^d) \to C([0,1]^{d-1})$ is continuous w.r.t. the topologies of pointwise convergence on both spaces for all $j \in \{1,\ldots,d\}$. This then immediately yields that $\boldsymbol{\lambda}_{n,\boldsymbol{\nu}}^m : C([0,1]^d) \to \mathbb{R}$ in (D.10) (obtained by repeated application of such operators) is continuous. Wlog $j = 1$. Let $f_i \in C([0,1]^d)$, $i \in \mathbb{N}$, be a sequence of functions converging pointwise to $f \in C([0,1]^d)$ and fix $\boldsymbol{x}^* \in [0,1]^{d-1}$. Then $g_i(x_1) := f_i(x_1, \boldsymbol{x}^*)$, $i \in \mathbb{N}$, is a sequence of functions in $C([0,1])$ that converges pointwise to $g(x_1) := f(x_1, \boldsymbol{x}^*) \in C([0,1])$. Thus by Theorem 33

$$\lambda_{n,\nu}^{m,x_1}[f_i](\boldsymbol{x}^*) = \lambda_{n,\nu}^{m,x_1}[g_i] \to \lambda_{n,\nu}^{m,x_1}[g] = \lambda_{n,\nu}^{m,x_1}[f](\boldsymbol{x}^*) \qquad \text{as } i \to \infty,$$

which shows the claimed continuity and concludes the proof of 1.

Next, 2 follows directly by d fold application of (D.3) to the definition (D.10) of $\boldsymbol{\lambda}_{n,\boldsymbol{\nu}}^m$.

Finally we show (D.11) and assume $d \geqslant 2$. Wlog $j = 1$. Fix $x_2 \in [0,1]$ and $\boldsymbol{x}^* \in [0,1]^{d-2}$. Then

$$\begin{aligned}\lambda_{n,\nu}^{m,x_1}[\partial_{x_2} f](x_2, \boldsymbol{x}^*) &= \lambda_{n,\nu}^{m,x_1}\Big[\lim_{h\to 0} \frac{f(x_1, x_2+h, \boldsymbol{x}^*) - f(x_1, x_2, \boldsymbol{x}^*)}{h}\Big] \\ &= \lim_{h\to 0} \frac{\lambda_{n,\nu}^{m,x_1}[f](x_2+h, \boldsymbol{x}^*) - \lambda_{n,\nu}^{m,x_1}[f](x_2, \boldsymbol{x}^*)}{h}.\end{aligned}$$

The second equality follows by the fact that the difference quotient defines a family of pointwise convergent functions in $C([0,1]^d)$ indexed over h, and the operator $\lambda_{n,\nu}^{m,x_1} : C([0,1]^d) \to C([0,1]^{d-1})$ is continuous w.r.t. the topology of pointwise convergence as shown above. Hence the last limit converges pointwise for all $(x_2, x^*) \in [0,1]^{d-1}$, which shows that $\lambda_{n,\nu}^{m,x_1}[f](x_2,\dots,x_d)$ is indeed differentiable in x_2 and the derivative in x_2 may be exchanged with $\lambda_{n,\nu}^{m,x_1}$. Repeatedly applying this argument yields the claim. ■

Theorem 35 *Let $k \in \mathbb{N}_0$, d, $m \in \mathbb{N}$ and $k+1 \leqslant m$. Then there exists $C = C(d,k,m)$ such that for all $r \in \{0,\dots,k\}$, $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ with $|\boldsymbol{\alpha}| = r$, $f \in C^k([0,1]^d)$, and $n \geqslant 1$,*

$$\sup_{\boldsymbol{x} \in M_n^d} |\partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}}(f(\boldsymbol{x}) - \boldsymbol{Q}_n^m[f](\boldsymbol{x}))| \leqslant C n^{-(k-r)} |f|_{C^k([0,1]^d)}. \tag{D.12}$$

Proof In the following we use the notation

$$Q_n^{m,x_j}[f] := \sum_{j=-m+1}^{n-1} \lambda_{n,j}^{m,x_j}[f],$$

so that

$$\boldsymbol{Q}_n^m[f] = Q_n^{m,x_d} \dots Q_n^{m,x_1}[f]. \tag{D.13}$$

In this proof we will use the following facts:

- By Lemma 34, for any $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ with $|\boldsymbol{\alpha}| \leqslant k$ and $\alpha_j = 0$ holds

$$\partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}} Q_n^{m,x_j}[f] = \sum_{i=-m+1}^{n-1} \partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}} \lambda_{n,i}^{m,x_j}[f] B_{n,i}^m(x_j) = \sum_{i=-m+1}^{n-1} \lambda_{n,i}^{m,x_j}[\partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}} f] B_{n,i}^m(x_j) = Q_n^{m,x_j}[\partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}} f], \tag{D.14}$$

  i.e., Q_n^{m,x_j} commutes with $\partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}}$.

- From (D.4) (with "$k = r$") we conclude that for any $g \in C^{\alpha_j}([0,1]^d)$ and $0 \leqslant \alpha_j \leqslant m-1$

$$\sup_{x_j \in M_n} |\partial_{x_j}^{\alpha_j} Q_n^{m,x_j}[g](x_1,\dots,x_d)| \leqslant C \sup_{x_j \in [0,1]} |\partial_{x_j}^{\alpha_j} g(x_1,\dots,x_d)| \tag{D.15}$$

  where $x_i \in [0,1]$ is arbitrary for all $i \neq j$, and $C = C(m,d)$ is independent of g.

- Again by (D.4), for $g \in C^r([0,1]^d)$, $0 \leqslant \alpha_j \leqslant r \leqslant m-1$ and $x_i \in [0,1]$ arbitrary for all $i \neq j$,

$$\sup_{x_j \in M_n} |\partial_{x_j}^{\alpha_j}(Q_n^{m,x_j}[g](x_1,\dots,x_d) - g(x_1,\dots,x_d))| \leqslant C n^{-(r-\alpha_j)} \sup_{x_j \in [0,1]} |\partial_{x_j}^r g(x_1,\dots,x_d)|. \tag{D.16}$$

Now fix $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ with $|\boldsymbol{\alpha}| \leqslant k$. Then for any $\boldsymbol{x} = (x_1,\dots,x_d) \in M_n^d$ (cp. (D.13))

$$|\partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}}(f(\boldsymbol{x}) - \boldsymbol{Q}_n^m[f](\boldsymbol{x}))| \leqslant \sum_{j=1}^d \left|\partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}}(Q_n^{m,x_d} \dots Q_n^{m,x_{j+1}}[f](\boldsymbol{x}) - Q_n^{m,x_d} \dots Q_n^{m,x_j}[f](\boldsymbol{x}))\right|, \tag{D.17}$$

where for $j = d$ the term $Q_n^{m,x_d} \dots Q_n^{m,x_{j+1}}[f](\boldsymbol{x})$ is understood as $f(\boldsymbol{x})$. Fix $j \in \{1, \dots, d\}$ and denote

$$\boldsymbol{\alpha}_- := (\alpha_1, \dots, \alpha_{j-1}, 0, \dots, 0)^\top \qquad \text{and} \qquad \boldsymbol{\alpha}_+ := (0, \dots, 0, \alpha_{j+1}, \dots, \alpha_d)^\top.$$

With (D.14) and (D.15) we get

$$\begin{aligned}
&\left|\partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}}\big(Q_n^{m,x_d} \dots Q_n^{m,x_{j+1}}[f](\boldsymbol{x}) - Q_n^{m,x_d} \dots Q_n^{m,x_j}[f](\boldsymbol{x})\big)\right| \\
&\quad = |\partial_{x_d}^{\alpha_d} Q_n^{m,x_d} \cdots \partial_{x_{j+1}}^{\alpha_{j+1}} Q_n^{m,x_{j+1}}[\partial_{x_j}^{\alpha_j} \partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}_-} f - \partial_{x_j}^{\alpha_j} Q_n^{m,x_j}[\partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}_-} f]](\boldsymbol{x})| \\
&\quad \leqslant C \sup_{x_d \in [0,1]} |\partial_{x_d}^{\alpha_d} \partial_{x_{d-1}}^{\alpha_{d-1}} Q_n^{m,x_d} \cdots \partial_{x_{j+1}}^{\alpha_{j+1}} Q_n^{m,x_{j+1}}[\partial_{x_j}^{\alpha_j} \partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}_-} f - \partial_{x_j}^{\alpha_j} Q_n^{m,x_j}[\partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}_-} f]](\boldsymbol{x})| \\
&\quad \leqslant \dots \leqslant C \sup_{x_d, \dots, x_{j+1} \in [0,1]} |\partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}_+}(\partial_{x_j}^{\alpha_j} \partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}_-} f - \partial_{x_j}^{\alpha_j} Q_n^{m,x_j}[\partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}_-} f])(\boldsymbol{x})| \\
&\quad = C \sup_{x_d, \dots, x_{j+1} \in [0,1]} |\partial_{x_j}^{\alpha_j} \partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}_+} \partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}_-} f(\boldsymbol{x}) - \partial_{x_j}^{\alpha_j} Q_n^{m,x_j}[\partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}_+} \partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}_-} f](\boldsymbol{x})|.
\end{aligned}$$

By assumption $\partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}_+} \partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}_-} f \in C^{k-|\boldsymbol{\alpha}_+ + \boldsymbol{\alpha}_-|}([0,1]^d)$ and thus by (D.16) the last term is bounded by

$$C n^{-(k-|\boldsymbol{\alpha}|)} \sup_{x_d, \dots, x_j \in [0,1]} |\partial_{x_j}^{k-|\boldsymbol{\alpha}_+ + \boldsymbol{\alpha}_-|} \partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}_+} \partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}_-} f(\boldsymbol{x})|.$$

Applying this estimate to (D.17) and taking the supremum over $\boldsymbol{x} \in M_n^d$ we find for any $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ with $|\boldsymbol{\alpha}| \leqslant k$

$$\sup_{\boldsymbol{x} \in M_n^d} |\partial_{\boldsymbol{x}}^{\boldsymbol{\alpha}}(f(\boldsymbol{x}) - \boldsymbol{Q}_n^m[f](\boldsymbol{x}))| \leqslant C n^{-(k-|\boldsymbol{\alpha}|)} |f|_{C^k([0,1]^d)}$$

for some $C = C(k, m, d)$ as claimed. ■

D.3 Translating Spline Approximation to Neural Networks

Proof [Proof of theorem 16] We wish to express the function

$$\tilde{f} := \boldsymbol{Q}_n^m[f] = \sum_{-m+1 \leqslant \nu_1, \dots, \nu_{d_1} \leqslant n-1} \boldsymbol{\lambda}_{n,\boldsymbol{\nu}}^m[f] B_{n,\boldsymbol{\nu}}^m(\boldsymbol{x}) \tag{D.18}$$

by a ReLU^{m-1} network. To this end we use the following facts:

- According to (Li et al., 2020, Theorem 2.5), there exists a network of finite width and depth that exactly expresses the square function x^2 on $\mathbb{R}$. It is now a standard observation, that using the polarization formula $xy = \frac{(x+y)^2 - x^2 - y^2}{2}$, we may also express the product of two numbers as a neural network. Repeatedly stacking such networks, we conclude that there exists a neural network $\tilde{p}$ of finite width and depth that takes $(x_1, \dots, x_{d_1}) \in \mathbb{R}^{d_1}$ as input and outputs $\tilde{p}(x_1, \dots, x_{d_1}) = \prod_{i=1}^{d_1} x_i \in \mathbb{R}$. That is, for some fixed $C_{\tilde{p}} = C_{\tilde{p}}(d_1)$ holds $\tilde{p} \in \Phi^{d_1,1}(L, W, S, B)$ with L, W, S, $B \leqslant C_{\tilde{p}}$.

- For each $n \in \mathbb{N}$ and each $j \in \{-m+1, \ldots, n-1\}$ the spline (cp. (D.1))

$$B_{n,j}^m(x) = \sum_{i=0}^{m} (-1)^m \frac{\binom{m}{i} \max\{0, nx - (ni + j)\}^{m-1}}{(m-1)!}$$

  corresponds to a ReLU^{m-1} network in $\Phi^{d_1,1}(L, W, S, B)$ with $L = 2$, $W = m+1$, $S = 3(m+1)$ and $B = nm + n - 1$. For the bound on B we used that the maximum bias occurs in the term $ni + j$ with $i = m$ and $j = n-1$.

We first compute in parallel the terms

$$B_{n,j}^m(x_i) \qquad \forall j \in \{-m+1, \ldots, n-1\},\ i \in \{1, \ldots, d_1\}.$$

This can be achieved by a network $\tilde{f}_1 : \mathbb{R}^{d_1} \to \mathbb{R}^{(n+m-1)d_1}$ of depth 2, width $(m+1)(n+m-1)d_1$, and sparsity $O(3(m+1)(n+m-1)d_1)$. Additionally all weights and biases are upper bounded by $nm + n - 1$.

Next, given the output of $\tilde{f}_1$, we consider a network $\tilde{f}_2 : \mathbb{R}^{(n+m-1)d_1} \to \mathbb{R}^{(n+m-1)^{d_1}}$ consisting of $(n+m-1)^{d_1}$ parallel product networks $\tilde{p}$, such that $\tilde{f}_2 \circ \tilde{f}_1$ produces the outputs

$$B_{n,\boldsymbol{\nu}}^m(\boldsymbol{x}) = \tilde{p}(B_{n,\nu_1}^m(x_1), \ldots, B_{n,\nu_{d_1}}^m(x_{d_1})) \qquad -m+1 \leqslant \nu_1, \ldots, \nu_{d_1} \leqslant n-1.$$

Then $\tilde{f}_2$ has depth at most $C_{\tilde{p}}$, width at most $C_{\tilde{p}}(m+n-1)^{d_1}$, sparsity at most $C_{\tilde{p}}(m+n-1)^{d_1}$, and all weights and biases are bounded by $C_{\tilde{p}}$.

Given the output of $\tilde{f}_2 \circ \tilde{f}_1$, a network $\tilde{f}_3 : \mathbb{R}^{(n+m-1)^{d_1}} \to \mathbb{R}$ consisting of only one linear transformation is used to produce the function in (D.18). This network has depth 1, width $(m+n-1)^{d_1}$, sparsity $(m+n-1)^{d_1}$, and upper bound $C\|f\|_{C([0,1]^{d_1})}$ for the modulus of all weights and biases. The last bound holds according to Lemma 34.

Finally, to combine all three networks we use the so-called "sparse-concatenation" denoted by $\odot$, which was first introduced for ReLU networks in (Petersen and Voigtlaender, 2018, Definition 2.5), but which can be extended to ReLU^{m-1} networks, see (Opschoor et al., 2022, Section 2.2.3). That is, we set

$$\tilde{f} := \tilde{f}_3 \odot \tilde{f}_2 \odot \tilde{f}_1.$$

It is a consequence of the properties of sparse concatenation (see Opschoor et al. (2022)) that this defines a network realizing the function $\tilde{f}(x_1, \ldots, x_{d_1}) = \tilde{f}_3(\tilde{f}_2(\tilde{f}_1(x_1, \ldots, x_{d_1})))$ such that the depth and width are bounded up to a multiplicative and additive constant by the sum of the depth and sparsity of the three subnetworks. An upper bound on the modulus of the network's weights and biases is obtained, up to an additive constant, by the maximal bound of the three subnetworks for this quantity. Finally, the sparsity of $\tilde{f}$ is bounded by the summed sparsity of $\tilde{f}_1$, $\tilde{f}_2$ and $\tilde{f}_3$ together with the number of required connections between $\tilde{f}_1$ and $\tilde{f}_2$, as well as between $\tilde{f}_2$ and $\tilde{f}_3$. Since each of the $(m+n-1)^{d_1}$ networks $\tilde{p}$ in $\tilde{f}_2$ gets exactly d_1 inputs, the former is bounded by $O((m+n-1)^{d_1} d_1)$. Since $\tilde{f}_3$ merely computes a linear combination of the $(n+m-1)^{d_1}$ outputs of $\tilde{f}_2$, the latter is bounded by $O((n+m-1)^{d_1})$. Absorbing some terms into the constant, for the network $\tilde{f}$ this leads to the bounds

$$
\begin{aligned}
L &\leqslant C \\
W &\leqslant C(1+m+n)^{d_1} = O(n^{d_1}) \\
S &\leqslant C(1+d(m+n)^{d_1} + m(n+m)d_1) = O(n^{d_1}) \\
B &\leqslant C(1+\|f\|_{C([0,1]^{d_1})} + nm) = O(n)
\end{aligned}
$$

for some $C = C(d_1, m)$ independent of n and f, and where the constants in the $O(\cdot)$ notation only depend on m and d_1. Substituting $N(n) := C(1+(m+n)^{d_1} + m(n+m)d_1) = O(n^{d_1})$ yields (4.1), and Theorem 35 implies (4.2). ∎

Proof [Proof of corollary 17] Denote $\sigma_m(x) = \max\{0, x\}^m$, $p := \lceil \log_m(\max\{2, k\}) \rceil \geqslant 1$ and $\tilde{m} := (m)^p$. Then

$$
\underbrace{\sigma_m \circ \cdots \circ \sigma_m}_{p \text{ times}} = \sigma_{\tilde{m}} \tag{D.19}
$$

and by definition $\tilde{m} \geqslant k$. Fix $N \in \mathbb{N}$.

According to Theorem 16, for each $j \in \{1, \ldots, d_2\}$ there exists a ReLU$^{\tilde{m}}$ network $\tilde{f}_j \in \Phi^{d_1,1}(L_{j,1}, W_{j,1}, S_{j,1}, B_{j,1})$ such that

$$
L_{j,1} \leqslant C, \qquad W_{j,1} \leqslant N, \qquad S_{j,1} \leqslant N, \qquad B_{j,1} \leqslant C\|f_j\|_{C([0,1]^{d_1})} + N^{1/d_1},
$$

for some $C = C(d_1, k, \tilde{m})$ independent of j and

$$
\|f_j - \tilde{f}_j\|_{W^{r,\infty}([0,1]^{d_1})} \leqslant C N^{-\frac{k-r}{d_1}} |f_j|_{C^k([0,1]^{d_1})} \qquad \forall r \in \{0, \ldots, k\}. \tag{D.20}
$$

Replacing each activation function $\sigma_{\tilde{m}}$ with the composition (D.19), we may interpret $\tilde{f}_j$ as a ReLU^m network in $\Phi^{d_1,1}(L_{j,2}, W_{j,2}, S_{j,2}, B_{j,2})$ with

$$
L_{j,2} \leqslant pC, \qquad W_{j,2} \leqslant N, \qquad S_{j,2} \leqslant pN, \qquad B_{j,2} \leqslant C\|f_j\|_{C([0,1]^{d_1})} + N^{1/d_1},
$$

i.e., the depth and sparsity increase by the multiplicative k and m dependent factor p, but the width and bound on the weights are not affected.

Next observe that $x^m = \sigma_m(x) + (-1)^m \sigma_m(-x)$. Since $x^m, (x+1)^m, \ldots, (x+m)^m$ are linearly independent functions, we can find coefficients $c_0, \ldots, c_m$ such that $x = \sum_{j=0}^m c_j \sigma_m(x) + (-1)^m \sigma_m(-x)$, i.e., the identity is expressible by a network of width $2(m+1)$ and with one hidden layer. By concatenating $\tilde{f}_j$ with $L_{j,2} - \lceil pC \rceil$ such identity networks, we may assume that all $\tilde{f}_j$ have the same depth $\lceil pC \rceil$, i.e., $\tilde{f}_j \in \Phi^{d_1,1}(L_{j,3}, W_{j,3}, S_{j,3}, B_{j,3})$ with

$$
\begin{aligned}
&L_{j,2} = \lceil pC \rceil, \qquad W_{j,2} \leqslant \max\{N, 2d_1(m+1)\}, \qquad S_{j,2} \leqslant (pN + K) \\
&B_{j,2} \leqslant \max\{C\|f_j\|_{C([0,1]^{d_1})} + N^{1/d_1}, \max_{j=1,\ldots,m} c_j\},
\end{aligned}
$$

where K is an absolute constant representing the size of the identity network of depth $\lceil pC \rceil$.

Parallelizing these networks of the same depth, yields one big ReLU^{m}network $(\tilde{f}_j)_{j=1}^{d_2} \in \Phi^{d_1,d_2}(L,W,S,B)$ with

$$L = \lceil pC \rceil, \qquad W \leqslant d_2 \max\{N, 2d_1(m+1)\}, \qquad S \leqslant d_2(pN+K)$$
$$B \leqslant \max\{C\|f_j\|_{C([0,1]^{d_1})} + N^{1/d_1}, \max_{j=1,\ldots,m} c_j\}.$$

Setting $\tilde{N}(d_2,N) := d_2(pN+K)$ yields the claimed bounds (4.3), and (4.4) follows by (D.20) and $N = \frac{\tilde{N}}{d_2} - O(1)$. ■

References

Sergios Agapiou and Sven Wang. Laplace priors and spatial inhomogeneity in Bayesian inverse problems. *arXiv:2112.05679*, 2021.

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv:2303.08797*, 2023.

Ethan Anderes and Marc A Coram. Two-dimensional density estimation using smooth invertible transformations. *Journal of Statistical Planning and Inference*, 141(3):1183–1193, 2011.

V.I. Arnold and R.A. Silverman. *Ordinary Differential Equations*. London, 1978. ISBN 9780262510189. URL `https://books.google.com/books?id=5NumQgAACAAJ`.

R. Baptista, Y. Marzouk, and O. Zahm. On the representation and learning of monotone triangular transport maps. *arXiv 2009.10303v2*, 2022.

Ricardo Baptista, Bamdad Hosseini, Nikola B Kovachki, Youssef M Marzouk, and Amir Sagiv. An approximation theory framework for measure-transport sampling algorithms. *arXiv:2302.13965*, 2023.

Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv:2111.13606*, 2021.

Lucien Birgé. On estimating a density using hellinger distance and some other strange facts. *Probability Theory and Related Fields*, 71(2):271–291, 1986. doi: 10.1007/BF00332312. URL `https://doi.org/10.1007/BF00332312`.

Vladimir Igorevich Bogachev, Aleksandr Viktorovich Kolesnikov, and Kirill Vladimirovich Medvedev. Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309, 2005.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. *arXiv:2302.07194*, 2023a.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31, 2018.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ODE is provably fast. *arXiv:2305.11798*, 2023b.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

R. DeVore, R. Howard, and C. Micchelli. Optimal nonlinear approximation. *Manuscripta mathematica*, 63(4):469–478, 1989. URL `http://eudml.org/doc/155392`.

R.A. DeVore and G.G. Lorentz. *Constructive Approximation*, volume 303. Springer, Berlin, 1993.

Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL `http://arxiv.org/abs/1410.8516`.

Vincent Divol, Jonathan Niles-Weed, and Aram-Alexandre Pooladian. Optimal transport map estimation in general function spaces. *arXiv:2212.03722*, 2022.

Chris Finlay, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam Oberman. How to train your neural ODE: the world of Jacobian and kinetic regularization. In *International conference on machine learning*, pages 3154–3164. PMLR, 2020.

Kenji Fukumizu, Taiji Suzuki, Noboru Isobe, Kazusato Oko, and Masanori Koyama. Flow matching achieves minimax optimal convergence. *arXiv:2405.20879*, 2024.

Yuan Gao, Jian Huang, Yuling Jiao, and Shurong Zheng. Convergence of continuous normalizing flows for learning probability distributions. *arXiv:2404.00551*, 2024.

Amir Gholami, Kurt Keutzer, and George Biros. ANODE: unconditionally accurate memory-efficient gradients for neural odes. *arXiv:1902.10298*, 2019.

E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional statistical models*. Cambridge University Press, 2016.

M. Giordano and R. Nickl. Consistency of Bayesian inference with Gaussian process priors in an elliptic inverse problem. *Inverse Problems*, 36(8):1–35, 2020.

A. Goldenshluger and O. Lepski. On adaptive minimax density estimation on rd. *Probability Theory and Related Fields*, 159(3):479–543, 2014.

Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. *arXiv:1810.01367*, 2018.

Philip Hartman. *Ordinary Differential Equations.* Society for Industrial and Applied Mathematics, second edition, 2002.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv:2210.02303*, 2022.

Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron C. Courville. Neural autoregressive flows. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2083–2092. PMLR, 2018. URL `http://proceedings.mlr.press/v80/huang18d.html`.

Chin-Wei Huang, Ricky TQ Chen, Christos Tsirigotis, and Aaron Courville. Convex potential flows: Universal probability distributions with optimal transport and convex optimization. *arXiv:2012.05942*, 2020.

Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166 – 1194, 2021. doi: 10.1214/20-AOS1997. URL `https://doi.org/10.1214/20-AOS1997`.

N. Irons, M. Scetbon, S. Pal, and Z. Harchaoui. Triangular flows for generative modeling: Statistical consistency, smoothness classes, and fast rates. *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 151:10161–10195, 2022.

Isao Ishikawa, Takeshi Teshima, Koichi Tojo, Kenta Oono, Masahiro Ikeda, and Masashi Sugiyama. Universal approximation property of invertible neural networks. *arXiv:2204.07415*, 2022.

Ki-Tae Kim, Umberto Villa, Matthew Parno, Youssef Marzouk, Omar Ghattas, and Noemi Petra. hIPPYlib-MUQ: a Bayesian inference software framework for integration of data with complex predictive models under uncertainty. *ACM Transactions on Mathematical Software*, 2023.

Diederik P. Kingma, Tim Salimans, Rafal Józefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational autoencoders with inverse autoregressive flow. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4736–4744, 2016. URL `https://proceedings.neurips.cc/paper/2016/hash/ddeebdeefdb7e7e7a697e1c3e3d8ef54-Abstract.html`.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

Ivan Kobyzev, Simon Prince, and Marcus Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine*

Intelligence, page 1–1, 2020. ISSN 1939-3539. doi: 10.1109/tpami.2020.2992934. URL http://dx.doi.org/10.1109/TPAMI.2020.2992934.

Bo Li, Shanshan Tang, and Haijun Yu. Better approximations of high dimensional smooth functions by deep neural networks with rectified power units. *Commun. Comput. Phys.*, 27(2):379–411, 2020. ISSN 1815-2406. doi: 10.4208/cicp.oa-2019-0168. URL https://doi.org/10.4208/cicp.oa-2019-0168.

Qianxiao Li, Ting Lin, and Zuowei Shen. Deep learning via dynamical systems: An approximation perspective. *Journal of the European Mathematical Society*, 25(5):1671–1709, 2022.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv:2209.03003*, 2022.

Yiping Lu, Haoxuan Chen, Jianfeng Lu, Lexing Ying, and Jose H. Blanchet. Machine learning for elliptic PDEs: Fast rate generalization bound, neural scaling law and minimax optimality. *CoRR*, abs/2110.06897, 2021. URL https://arxiv.org/abs/2110.06897.

Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *arXiv:2107.12364*, 2021.

Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini. *Sampling via Measure Transport: An Introduction*, pages 1–41. Springer International Publishing, Cham, 2016. ISBN 978-3-319-11259-6. doi: 10.1007/978-3-319-11259-6_23-1. URL https://doi.org/10.1007/978-3-319-11259-6_23-1.

Youssef Marzouk, Robert Ren, and Jakob Zech. Distribution learning for neural differential equations: minimal energy regularization and approximation theory. *preprint*, 2023.

Robert J McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.

H. N. Mhaskar. Approximation properties of a multilayered feedforward artificial neural network. *Adv. Comput. Math.*, 1(1):61–80, 1993. ISSN 1019-7168. doi: 10.1007/BF02070821. URL https://doi.org/10.1007/BF02070821.

H.N Mhaskar and Charles A Micchelli. Approximation by superposition of sigmoidal and radial basis functions. *Advances in Applied Mathematics*, 13(3):350–373, 1992. ISSN 0196-8858. doi: https://doi.org/10.1016/0196-8858(92)90016-P. URL https://www.sciencedirect.com/science/article/pii/019688589290016P.

F. Monard, R. Nickl, and G. Paternain. Efficient nonparametric Bayesian inference for X-ray transforms. *Annals of Statistics*, 2019.

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

Benjamin Nachman and David Shih. Anomaly detection with density estimation. *Physical Review D*, 101(7):075042, 2020.

R. Nickl, S. van de Geer, and S. Wang. Convergence rates for penalised least squares estimators in PDE-constrained regression problems. *SIAM J. Uncert. Quant.*, 8, 2020.

Richard Nickl. *Bayesian Non-linear Statistical Inverse Problems*. EMS press, 2023.

Richard Nickl and Edriss S. Titi. On posterior consistency of data assimilation with Gaussian process priors: the 2D Navier–Stokes equations, 2023.

Richard Nickl and Sven Wang. On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms. *Journal of the European Mathematical Society*, 2022.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. *CoRR*, abs/2303.01861, 2023. doi: 10.48550/arXiv.2303.01861. URL `https://doi.org/10.48550/arXiv.2303.01861`.

Derek Onken, Samy Wu Fung, Xingjian Li, and Lars Ruthotto. OT-Flow: Fast and accurate continuous normalizing flows via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9223–9232, 2021.

J. A. A. Opschoor, Ch. Schwab, and J. Zech. Exponential ReLU DNN expression of holomorphic maps in high dimension. *Constr. Approx.*, 55(1):537–582, 2022. ISSN 0176-4276. doi: 10.1007/s00365-021-09542-5. URL `https://doi.org/10.1007/s00365-021-09542-5`.

P. Oswald. On the degree of nonlinear spline approximation in Besov-Sobolev spaces. *J. Approx. Theory*, 61(2):131–157, 1990. ISSN 0021-9045. doi: 10.1016/0021-9045(90)90001-7. URL `https://doi.org/10.1016/0021-9045(90)90001-7`.

Houman Owhadi and Gene Ryan Yoo. Kernel flows: From learning kernels from data into the abyss. *Journal of Computational Physics*, 389:22–47, 2019.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680, 2021.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, 2018. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2018.08.019. URL `https://www.sciencedirect.com/science/article/pii/S0893608018302454`.

Allan Pinkus. Approximation theory of the MLP model in neural networks. In *Acta Numerica, 1999*, volume 8 of *Acta Numer.*, pages 143–195. Cambridge Univ. Press, Cambridge, 1999. doi: 10.1017/S0962492900002919. URL `https://doi.org/10.1017/S0962492900002919`.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1530–1538. JMLR.org, 2015. URL `http://proceedings.mlr.press/v37/rezende15.html`.

Daniel Rudolf and Björn Sprungk. On a generalization of the preconditioned Crank–Nicolson metropolis algorithm. *Foundations of Computational Mathematics*, 18:309–343, 2018.

Domènec Ruiz-Balet and Enrique Zuazua. Control of neural transport for normalizing flows. *arXiv:2307.07817*, 2023a.

Domènec Ruiz-Balet and Enrique Zuazua. Neural ODE control for classification, approximation, and transport. *SIAM Review*, 65(3):735–773, 2023b. doi: 10.1137/21M1411433.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Springer, 2015. URL `https://www.math.u-psud.fr/~filippo/OTAM-cvgmt.pdf`.

Claudia Schillings and Christoph Schwab. Scaling limits in computational Bayesian inversion. *ESAIM: Mathematical Modelling and Numerical Analysis*, 50(6):1825–1856, 2016.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4), 2020. doi: 10.1214/19-aos1875. URL `https://doi.org/10.1214%2F19-aos1875`.

Larry L. Schumaker. *Spline functions: basic theory*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, third edition, 2007. ISBN 978-0-521-70512-7. doi: 10.1017/CBO9780511618994. URL `https://doi.org/10.1017/CBO9780511618994`.

Yuyang Shi, Valentin De Bortoli, George Deligiannidis, and Arnaud Doucet. Conditional simulation using diffusion Schrödinger bridges. In *Uncertainty in Artificial Intelligence*, pages 1792–1802. PMLR, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv:2011.13456*, 2020b.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34: 1415–1428, 2021.

Taiji Suzuki. Adaptivity of deep ReLU network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL `https://openreview.net/forum?id=H1ebTsActm`.

Esteban G Tabak and Eric Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

Takeshi Teshima, Koichi Tojo, Masahiro Ikeda, Isao Ishikawa, and Kenta Oono. Universal approximation property of neural ordinary differential equations. *CoRR*, abs/2012.02414, 2020. URL `https://arxiv.org/abs/2012.02414`.

Hans Triebel. *Function spaces and wavelets on domains*, volume 7 of *EMS Tracts in Mathematics*. European Mathematical Society (EMS), Zürich, 2008. ISBN 978-3-03719-019-7. doi: 10.4171/019. URL `https://mathscinet.ams.org/mathscinet-getitem?mr=2455724`.

Sara van de Geer. *Empirical Processes in M-Estimation*. Cambridge U. Press, 2000.

Sara van de Geer. Least squares estimation with complexity penalties. *Mathematical Methods of statistics*, (10):355–374, 2001.

Cédric Villani. *Optimal transport. Old and new.* Springer-Verlag, Berlin, 2009. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9. URL `https://doi.org/10.1007/978-3-540-71050-9`.

Sven Wang and Youssef Marzouk. On minimax density estimation via measure transport. *arXiv:2207.10231*, 2022. doi: 10.48550/ARXIV.2207.10231. URL `https://arxiv.org/abs/2207.10231`.

Antoine Wehenkel and Gilles Louppe. Unconstrained monotonic neural networks. *Advances in neural information processing systems*, 32, 2019.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv:2209.00796*, 2022.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017. doi: 10.1016/j.neunet.2017.07.002. URL `https://doi.org/10.1016/j.neunet.2017.07.002`.

Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/979a3f14bae523dc5101c52120c535e9-Abstract.html`.

Jakob Zech and Youssef Marzouk. Sparse approximation of triangular transports. Part I: The finite-dimensional case. *Constructive Approximation*, 55:919–986, 2022a.

Jakob Zech and Youssef Marzouk. Sparse approximation of triangular transports. Part II: The infinite-dimensional case. *Constructive Approximation*, 55(3):987–1036, 2022b.