Statistical convergence rates for transport- and ODE-based models

Sven Wang (HU Berlin)

Joint work with Y. Marzouk (MIT), R. Ren (MIT), J. Zech (Heidelberg)

WIAS Mathematical Statistics Seminar

HUMBOLDT-UNIVERSITÄT ZU BERLIN

November 8, 2023

Learning probability distributions & generative modelling

An important goal in statistics (and machine learning) is to learn complicated probability distributions π.

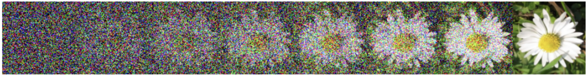
- **Generate samples** Z with approximate law $\mathcal{L}(Z) \approx \pi$.
- **Estimate** $\hat{\pi} \approx \pi$.

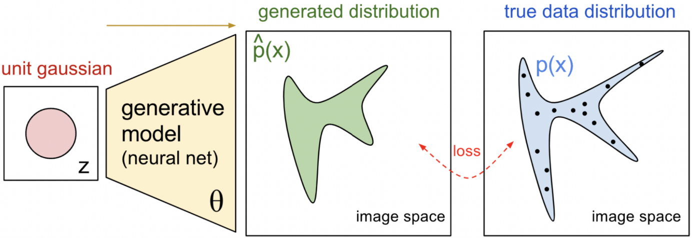
Depending on the context, we have access either

- Data $X_1, \dots, X_N$ (**Density estimation / generative modelling**).
- Evaluations of $\pi(\cdot)$ up to normalization constant or $\nabla \log \pi$ (**Bayesian computation / sampling**).

Recent strategies for generative modelling

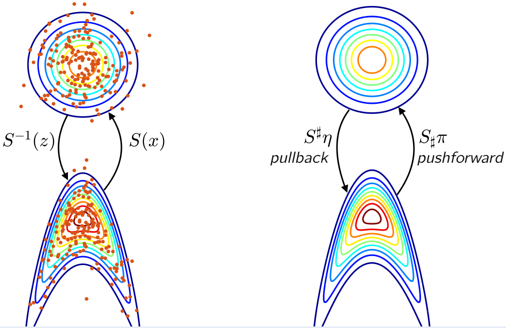
- Normalising flows: composition of bijective maps [Rezende & Mohamed 2015]
- Autoregressive flows: composition of triangular maps [Kingma et al. 2018]
- NeuralODE [Chen et al. 2018]
- Score-based diffusion models [Song et al. 2021]
- Stochastic interpolants [Albergo & Vanden-Eijnden 2022]





Transport methods

- A **transport map** S induces a *deterministic coupling* between a target distribution π and a reference distribution η
  - Choose η to be simple/tractable (standard normal, uniform)
  - Find an invertible S such that $S_\sharp \pi = \eta$
  - Estimate the target density: $\pi(\mathbf{x}) = S^\sharp \eta(\mathbf{x}) := \eta \circ S(\mathbf{x}) \, |\det \nabla S(\mathbf{x})|$
  - Generate cheap and independent samples: $\mathbf{Z} \sim \eta \;\Leftrightarrow\; S^{-1}(\mathbf{Z}) \sim \pi$
  - Bayesian computation via transport [Marzouk et al. (2016)]



Classes of transport maps

- There are many ways to couple π and η.
  - **Brenier maps:** Under mild assumptions on π, η, there exists a unique map $S^{OT}_{\pi,\eta}$ (gradient of some convex function) such that $(S^{OT}_{\pi,\eta})_\sharp \pi = \eta$ and
  $$W^2(\eta, \pi) = \int \|x - S^{OT}_{\pi,\eta}(x)\|^2 d\pi(x),$$
  i.e. $(Id \times \nabla S^{OT}_{\pi,\eta})_\sharp \pi$ is an optimal coupling.
  - **Knothe-Rosenblatt maps:** Triangular, partially monotone maps (equal to OT maps for $d = 1$)
  - **ODE flow maps**: Evolving π (at $t = 0$) to η (at $t = 1$) through an ODE flow,
  $$\frac{dX(t)}{dt} = f(X(t), t), \quad X(0) \sim \pi, \quad X(1) \sim \eta.$$
- Statistical performance of those methods?
- Can they achieve minimax rates?

Existing theory work

- **Approximation** of transport maps in high dimension using sparse polynomials or ReLU networks [Zech & Marzouk (2021)]
- **Statistical consistency** for triangular maps in KL-distance [Irons et al. (2022)]
- **Estimation of OT maps:** Minimax rate in d dimensions: $N^{-\frac{\alpha}{2\alpha-2+d}} \vee N^{-1}$. [Hütter & Rigollet (2021, AOS)]
- **Computational OT** [Peyré & Cuturi (2019)]
- **Density estimation** for Wasserstein loss [Weed & Berthet (2019), Hütter and Rigollet (2021)]
- **Diffusion models** [Oko, Akiyama, Suzuki (2023)]

Nonparametric density estimation via transport

- **Data.** We are given N i.i.d. observations on $[0,1]^d$,

$$X_1, \dots, X_N \sim P_0.$$

- **Goal.** Estimate unknown Lebesgue density p_0 within some class $\mathcal{P}$.

Likelihood objective

For some class $\mathcal{S}$ of bijective maps $[0,1]^d \to [0,1]^d$, take

$$\hat{S} \in \arg\min_{S \in \mathcal{S}} -\frac{1}{N} \sum_{i=1}^{N} \log \left(\eta \circ S(X_i) \det \nabla S(X_i) \right)$$

$$\left(\approx \arg\min_{S \in \mathcal{S}} KL(p_0 || S^{\#}\eta) + \text{const.} \right)$$

- How should one choose $\mathcal{S}$?

Monotone triangular transport maps

Let us now focus on **Knothe–Rosenblatt (KR) rearrangements** on unit cube $[0,1]^d$

$$S_{\pi,\eta}(x) \equiv S(x) = \begin{bmatrix} S_1(x_1) \\ S_2(x_1, x_2) \\ \vdots \\ S_d(x_1, x_2, \ldots, x_d) \end{bmatrix}, \quad S^{\#}\eta = \pi.$$

- Exists and is unique under mild assumptions on π and η (given a variable ordering)
- Invertibility is guaranteed by one-dimensional monotonicity $\partial_k S_k > 0$
- $\det \nabla S(\mathbf{x})$ simple to evaluate
- Components S_k characterize marginal conditionals of π:

$$\pi_{\mathbf{X}} = \pi_{\mathbf{X}_1} \pi_{\mathbf{X}_2|\mathbf{X}_1} \cdots \pi_{\mathbf{X}_d|\mathbf{X}_1,\ldots,\mathbf{X}_{d-1}}$$

Intuitive result

- If η, p_0 are C^α, then so is the KR-map S_{η,p_0} [Santambrogio '15].
- For **smoothness level** $\alpha > 0$ and $0 < c < B < \infty$, let

$$\mathcal{M}(\alpha, B, c) := \Big\{\nu \in C^\alpha([0,1]^d),\ \|\nu\|_{C^\alpha} \le B,\ \nu \ge c,\ \int \nu(x)dx = 1\Big\}.$$

- For $L, c_{min} > 0$, define the classes

$$\mathcal{S}(\alpha, L, c_{min}) := \Big\{S : [0,1]^d \to [0,1]^d \text{ bijective and triangular,}$$
$$\|S_k\|_{C^\alpha} \le L,\ \partial_k S_k \ge c_{min},\ 1 \le k \le d\Big\}.$$

Theorem (Wang and Marzouk 2022)

Suppose $p_0 \in \mathcal{M}(\alpha, B, c)$ (and that η is smooth). Then, for some $L, c_{min} > 0$, maximizers $\hat{S}$ over $\mathcal{S}(\alpha, L, c_{min})$ satisfy

$$E^N_{P_0}[h^2(\hat{S}^\sharp\eta, p_0)] \lesssim N^{-\frac{2(\alpha-1)}{2(\alpha-1)+d}}.$$

Why?

- The estimator $\hat{S}^{\#}\eta$ is equivalently an **MLE**:

$$\hat{S}^{\#}\eta \in \arg\max_{p\in\mathcal{P}} \sum_{i=1}^{N} \log p(X_i), \quad \mathcal{P} = \{S^{\#}\eta : S \in \mathcal{S}\}.$$

Proof ingredients

- If $S, \eta \in C^{\alpha}$, then $S^{\#}\eta \in C^{\alpha-1}$.
- Hellinger convergence theory for MLEs [e.g. van de Geer 2000]

Anisotropic smoothness and minimax rates

The previous result is **not** minimax-optimal. Define the anisotropic subclasses

$$\mathcal{AS}(\alpha, L, c_{min}) := \Big\{ S \in \mathcal{S}(\alpha, L, c_{min}), \forall 1 \le k \le d : \|\partial_k S_k\|_{C^\alpha([0,1]^k)} \le L \Big\}$$

Theorem (Wang and Marzouk 2022)

The transport map MLE $\hat{S}$ based on classes $\mathcal{AS}$ satisfies minimax optimal rates

$$E_{P_0^N}\big[h(\hat{S}^\# \eta, p_0)^2\big] \lesssim N^{-\frac{2\alpha}{2\alpha+d}}.$$

General penalized estimators

The sets $\mathcal{AS}(\alpha, L, c_{min})$ may be hard to optimize over. In practice, one may want to re-parameterize transport maps S_θ via a (e.g. Euclidean) parameter $\theta \in \Theta$.

- Let Θ be a set parameterising triangular maps $\mathcal{S} = \{S_\theta : [0,1]^d \to [0,1]^d,\ \theta \in \Theta\}$.
- Penalized objective

$$\mathcal{J}_{N,\lambda}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log\left[\eta(S_\theta(X_i)) \det \nabla S_\theta(X_i)\right] + \lambda^2 \mathrm{pen}(\theta)^2,$$

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} \mathcal{J}_{N,\lambda}(\theta).$$

General theorem (triangular maps)

Define $\|S\|_{C^1_{diag}} := \sum_{k=1}^d \|S_k\|_\infty + \|\partial_k S_k\|_\infty$. For any $\theta^* \in \Theta$, let

$$\mathcal{S}(\lambda, R) := \{S_\theta : h^2(S^\sharp_\theta \eta, S^\sharp_{\theta^*}\eta) + \lambda^2 \text{pen}(\theta)^2 \leq R^2\},$$
$$\mathcal{J}(\lambda, R) := R + \int_0^R H^{1/2}(\mathcal{S}(\lambda, R), \|\cdot\|_{C^1_{diag}}, \rho) d\rho.$$

Theorem (Wang and Marzouk 2022)

Suppose that $K^{-1} \leq p_0, \eta \leq K$, and that $\{S_\theta : \theta \in \Theta\}$ is uniformly bounded in C^1_{diag}. Then there exist $C, \gamma > 0$ such that for any $\lambda, \delta > 0$ satisfying

$$\delta^2 \geq \frac{C\mathcal{J}(\lambda, \delta)}{\sqrt{N}}, \quad \text{we have that}$$

$$\mathbb{E}_0^N \left[h^2(S^\sharp_{\hat\theta}\eta, p_0)\right] \lesssim h^2(S^\sharp_{\theta^*}\eta, p_0) + \lambda^2 \text{pen}(\theta^*)^2 + \delta^2.$$

Parameterisation of Knothe-Rosenblatt maps

Let $U = V = [0,1]^d$. Three required properties:

- Triangularity
- Monotonicity
- Range constraint

Example: Rational parameterization

For $x \in [0,1]^d$ and $1 \leq k \leq d$, let

$$S_{F,k}(x) := \frac{\int_0^{x_k} \Phi\big(F_k(x_{1:k-1}, y)\big)\, dy}{\int_0^1 \Phi\big(F_k(x_{1:k-1}, y)\big)\, dy}, \qquad x_{1:k} \in [0,1]^d.$$

- $\Phi : \mathbb{R} \to (K_{min}, K_{max})$ is a 'link function'
- $F : [0,1]^d \to \mathbb{R}^d$ is any (say, L^∞) function.

Implications & extensions

- Natural parameterizations of triangular maps with H^α Sobolev penalty, or with high-dimensional wavelet penalty, achieve the minimax rate.
- The general theorem also holds for general non-triangular classes of maps, and on general bounded domains.

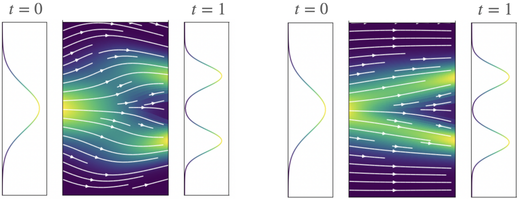
ODE couplings of probability distributions

Let $D = [0,1]^d$, and $\Omega = D \times [0,1]$. Let $f : \Omega \mapsto \mathbb{R}^d$ be a *velocity field* which governs the ODE

$$\begin{cases} \frac{d}{dt} u(t) = f(u(t), t), t \in (0,1), \\ u(0) = x. \end{cases}$$

This defines unique trajectories (flow)

$$X_f(x,t) = x + \int_0^t f(X_f(x,s), s) ds, \qquad t \in [0,1], \qquad x \in D.$$



Neural network parameterization of $f \in \mathcal{F} \equiv$ neural ODEs.

Parameterization of velocity fields

Minimal requirement:

$$\mathcal{V} = \Big\{ f : \Omega \to \mathbb{R}^d \;\Big|\; f \in C^1(\Omega), f \cdot \nu \equiv 0 \text{ on } \partial D \times [0,1] \Big\}.$$

Lemma

If $f \in \mathcal{V}$, then $X_f(\cdot, t) : D \to D$ is a diffeomorphism for any $t \in [0,1]$, and the pullback density is given by

$$(X_f(\cdot, t))^{\#} \rho(x) = \rho(X_f(x,t)) \det \left[\nabla_x X_f(x,t) \right].$$

- $\mathcal{F} \subseteq \mathcal{V}$ class of velocity fields
- Let $T^f = X_f(\cdot, 1)$ (time-one flow map)

Training objective

$$\hat{f}_{\mathrm{ODE}} := \arg\max_{f \in \mathcal{F}} \sum_{i \in [n]} \log \rho(T^f(X_i)) \det \left[\nabla_x T^f(X_i) \right].$$

General convergence theorem

- Suppose $(X_i : 1 \le i \le n) \sim^{i.i.d.} p_0 \le K$.
- $\rho \equiv 1$ uniform reference on $D = (0,1)^d$.
- For some constant $B > 0$,

$$\sup_{f \in \mathcal{F}} \Big(\|f\|_{C^1(\Omega)} + \sup_{t \in [0,1]} \|\nabla_x f(\cdot, t)\|_{Lip} \Big) \le B, \tag{1}$$

- Define C^1-metric entropy integral

$$I(\mathcal{F}, R) := R + \int_0^R \sqrt{\log N(\mathcal{F}, \|\cdot\|_{C^1}, \rho)}\, d\rho.$$

Theorem (Marzouk, Ren, W, Zech 2023)

There exists $C > 0$ such that for all $f^* \in \mathcal{F}$ and all $\delta_N > 0$ with

$$\sqrt{N}\delta_N^2 \ge C \cdot I(\mathcal{F}, \delta_N), \tag{2}$$

$$\mathbb{E}_{p_0}^N \big[h\big((T^{\hat{f}})^\sharp \rho, p_0 \big) \big] \lesssim \underbrace{h\big((T^{f^*})^\# \rho, p_0 \big)}_{\text{approximation term}} + \delta_N. \tag{3}$$

Proof idea

Lemma (Local Lipschitz parameterisation)

We have the local Lipschitz estimates (on $W^{2,\infty}$-bounded sets)

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

Lemma (Bounds for induced transport maps)

Suppose that $\sup_{f\in\mathcal{F}} \|f\|_{C^1(\Omega)} =: M < \infty$. *Then, for all* $f \in \mathcal{F}$, *we have*

$$\sup_{x\in D} \|\nabla(T^f)(x)\|_{\mathbb{R}^d\to\mathbb{R}^d} \leq 1 + dMe^{dM}.$$

The largest and smallest eigenvalues of $\nabla(T^f(x))$ *are bounded as*

$$\sup_{f\in\mathcal{F}} \sup_{x\in D} \lambda^f_{\max}(x) \leq 1 + dMe^{dM}, \qquad \inf_{f\in\mathcal{F}} \inf_{x\in D} \lambda^f_{\min}(x) \geq (1 + dMe^{dM})^{-1}.$$

Regularity theorem

Lemma (Existence of C^k coupling velocity field)

Let $p_0 \in C^k([0,1]^d)$. Then, there exists some $f^{\Delta}_{p_0}$ such that $(T^{f^{\Delta}_{p_0}})^{\#}\rho = p_0$, and such that $f^{\Delta}_{p_0} \in C^k(\Omega)$. Moreover, the velocity field g with components

$$[g(x,s)]_j := \frac{(f^{\Delta}_{p_0}(x,s))_j}{x_j(1-x_j)}, \quad j = 1, \dots, d, \tag{4}$$

also belongs to $C^k(\Omega)$. Consequently, $f^{\Delta}_{p_0} \in \mathcal{V} \cap C^k(\Omega)$.

Construction of C^k vector field

Let T be the KR-map pushing p_0 to ρ, and let

$$G_t(x) = tT(x) + (1-t)x$$

be the straight-line interpolation. Then let $F : D \times [0,1] \to D$, $F(x,t) = G_t^{-1}(x)$, and define

$$f_{p_0}^{\Delta}(y,s) = T(F(y,s)) - F(y,s).$$

- This vector field satisfies the desired regularity.

C^1-covering of C^k-spaces

For $0 < s_1, s_2 < \infty$, $R > 0$, $\tau > 0$,

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

Convergence rate for C^k-classes

Define the C^k classes

$$\mathcal{F}(B) := \Big\{ f \in C^k(\Omega, \mathbb{R}^d) : \|f\|_{C^k} \leq B, \\ f(x,t) \cdot \nu_x \equiv 0 \text{ for all } (t,x) \in [0,1] \times \partial D \Big\},$$

Theorem

Let $p_0 \in C^k$ and $\mathcal{F} = \mathcal{F}(B)$. Then, it holds that for all $n \geq 1$,

$$\mathbb{E}^N_{p_0}\big[h^2((T^{\hat{f}})^{\#}\rho, p_0)\big] \lesssim N^{-\eta}, \quad \textit{with } \eta = \frac{2(k-1-\gamma)}{2(k-1-\gamma)+d+1} > 0.$$

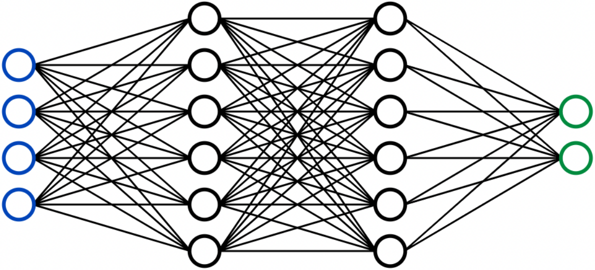
- Approximation of $\mathcal{F}(B)$ by wavelets, polynomials, trig. polynomials possible

Neural network classes

- Suppose $p_0 \in \mathcal{M}(k, L_1, L_2)$ is $C^k(D)$.
- Let

$$\mathcal{F}_{\mathrm{NN}}(L, W, S, B, R) = \Phi_{NN}^{d+1,d}(L, W, S, B) \cap \{f \in W^{2,\infty}(\Omega) : \|f\|_{W^{2,\infty}(\Omega)} \leq R\},$$

be the network class with ReLU^2 activation function, mapping from Ω to $\mathbb{R}^d$.



Theorem for neuralODE

Set

- $L = O(1)$ (depth)
- $W = O(N^{\frac{d+1}{d+1+2(k-1)}})$ (width),
- $S = O(N^{\frac{d+1}{d+1+2(k-1)}})$ (sparsity),
- $B = O(N^{\frac{d+1}{d+1+2(k-1)}})$ (l^∞ bound on weights)
- $R = O(1)$ (large enough).

Theorem

The neuralODE estimator over $\mathcal{F}_{NN}(L, W, S, B, R)$ satisfies

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

Proof ingredients I: metric entropy

Theorem

Consider the ReLU^2 *network space* $\Phi(L, W, S, B)$ *of networks* $\mathbb{R}^d \to \mathbb{R}$ *with* $L = \mathcal{O}(1)$, $W = \mathcal{O}(M)$, $S = \mathcal{O}(M)$ *and* $B = \mathcal{O}(M)$. *Then*

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

- Builds on existing bounds from Schmidt-Hieber (2020), Suzuki (2019) from the regression setting.
- Modifications of previous results to ReLU^2 activation functions.

Proof ingredients II: approximation

Theorem

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

$$L \leq C, \qquad W \leq M, \qquad S \leq M, \qquad B \leq C\|f\| + M^{1/d} \tag{5}$$

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

$$\|f - \tilde{f}\|_{W^{r,\infty}([0,1]^d)} \leq CM^{-\frac{k-r}{d}} \|f\|_{C^k([0,1]^d)} \qquad \forall r \in \{0, \dots, k\}. \tag{6}$$

- Adapts classical neural network approximation results (e.g. Pinkus 1999, Yarotsky 2017) to the setting with smoother activation functions.

Future directions

- **Brenier maps** are known to possess regularity properties
  - Statistical convergence rates?
  - How to parameterize Brenier maps?
- Extension to **high dimensions**
- Theoretical guarantees for **conditional sampling**
- Flow matching methods

References

S. Wang and Y. Marzouk: On minimax density estimation via measure transport. arXiv:2207.10231 (2022)

Y. Marzouk, R. Ren, S. Wang and J. Zech: Distribution learning via neural differential equations: a nonparametric statistical perspective. arXiv:2309.01043 (2023)

Thank you!