Sequential transport for density estimation and its applications

Benjamin Zanger[1]

Joint work with **Olivier Zahm**[1], **Tiangang Cui**[3], and **Martin Schreiber**[1,2]

[1]LJK, INRIA Grenoble [2]Université Grenoble Alpes [3]the University of Sydney

April 22, 2025

The task of density estimation

There are generally two distinct tasks of density estimation:

From a model

Access to unnormalized density:

$$\pi(\boldsymbol{x}) \propto \mathcal{L}(\boldsymbol{x})\pi_0(\boldsymbol{x})$$

From data

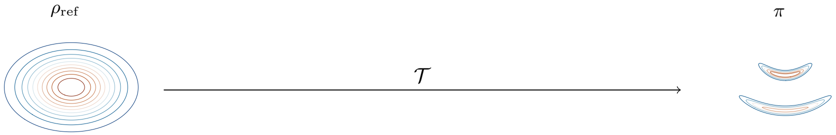
Access to data:

$$x^{(i)} \sim \pi$$

Goal

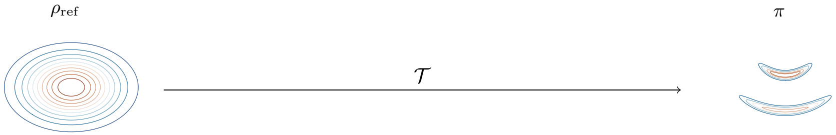
Sample from estimated distribution and/or evaluate its density

Measure transport



Building a transport map $\mathcal{T}$ from ρ_{ref} to π, so that $\mathcal{T}_\sharp\, \rho_{\mathsf{ref}} = \pi$ enables to efficiently sample from π.

Measure transport

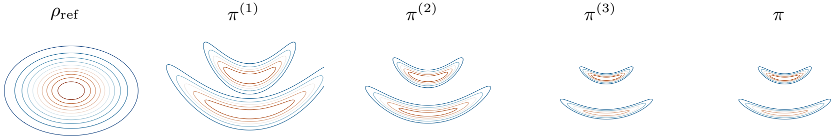


Building a transport map $\mathcal{T}$ from ρ_{ref} to π, so that $\mathcal{T}_\sharp\, \rho_{\mathsf{ref}} = \pi$ enables to efficiently sample from π.

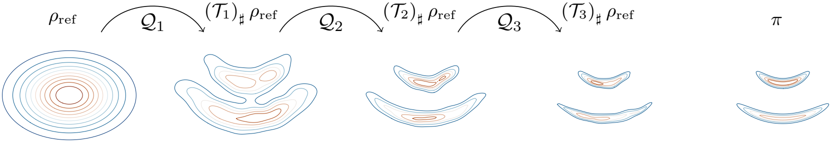
Pushforward of density: $\quad \mathcal{T}_\sharp\, \rho_{\mathsf{ref}} := \rho(\mathcal{T}^{-1}(\boldsymbol{x}))\det \nabla \mathcal{T}^{-1}(\boldsymbol{x}) = \tilde{\pi}$

Pushforward of samples: $\quad \mathcal{T}(\boldsymbol{z}) \sim \tilde{\pi} \quad \text{with } \boldsymbol{z} \sim \rho_{\mathsf{ref}}\,.$

Sequential measure transport

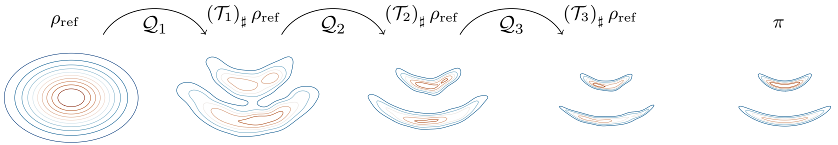


Sequential measure transport



sequential estimation $+$ Measure transport $=$ sequential Measure transport (**SMT**)

Sequential measure transport



sequential estimation $+$ Measure transport $=$ sequential Measure transport (**SMT**)

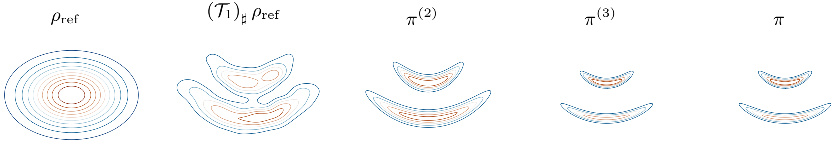
Examples of combination of these two methods:

- Denoising diffusion
- Normalizing flow based as Grenioux et al. 2023; Rezende and Mohamed 2016
- Others like Cui et al. 2024; Marzouk et al. 2016

Table of contents

Outline

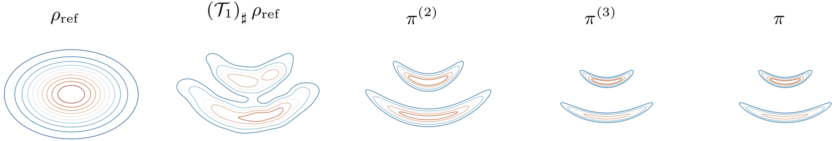
Our proposed SMT method



$\boxed{\text{VDE}}$ Variation density estimation of $\pi^{(1)}$

$$\tilde{f}^{(1)} = \arg\min_{\tilde{f} \in \mathcal{M}} \mathrm{D}_\alpha \left(\pi^{(1)} || \tilde{f} \right)$$

Our proposed SMT method



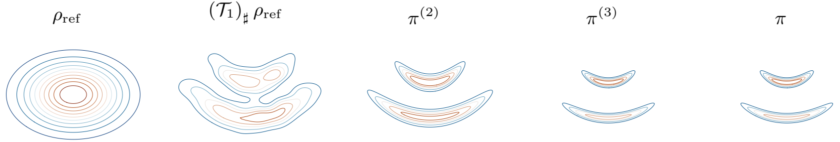
$\boxed{\text{VDE}}$ Variation density estimation of $\pi^{(1)}$

$$\tilde{f}^{(1)} = \arg\min_{\tilde{f} \in \mathcal{M}} \mathrm{D}_\alpha \left(\pi^{(1)} || \tilde{f}\right)$$

SoS-densities for $\mathcal{M}$

$$\begin{aligned} f_A(x) &= \sum \left(\Phi(x)^\top \boldsymbol{a}_i\right)^2 \\ &= \Phi(x)^\top A \Phi(x) \quad A \succeq 0 \end{aligned}$$

Our proposed SMT method



$\boxed{\text{VDE}}$ Variation density estimation of $\pi^{(1)}$

$$\tilde{f}^{(1)} = \arg\min_{\tilde{f} \in \mathcal{M}} \mathrm{D}_\alpha \left(\pi^{(1)} || \tilde{f} \right)$$

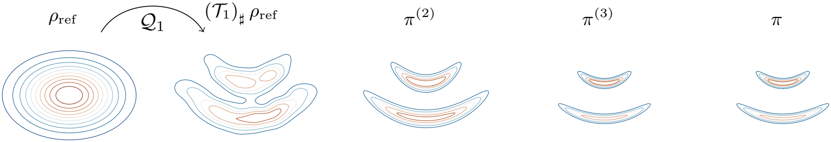
SoS-densities for $\mathcal{M}$

$$f_A(x) = \sum \left(\Phi(x)^\top a_i \right)^2$$
$$= \Phi(x)^\top A \Phi(x) \quad A \succeq 0$$

α-divergences D_α

- Includes: (reversed) KL-divergence, χ^2 divergence, ...
- Extension to unnormalized densities

Our proposed SMT method



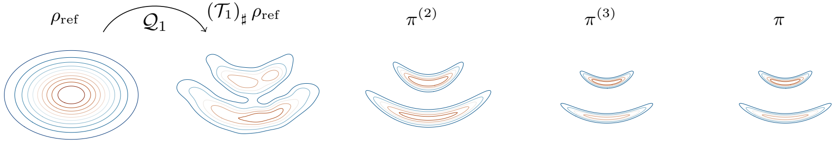
$\boxed{\text{KR}}$ Build the Knothe-Rosenblatt map $\mathcal{Q}_1$ from $\tilde{\pi}^{(1)}$

1. Normalize $\tilde{f}^{(1)}$ to get $\tilde{\pi}^{(1)}$

With $\Phi(\boldsymbol{x})$ orthonormal in L^2_μ,

$$\pi_A(\boldsymbol{x}) = \frac{\Phi(\boldsymbol{x})A\Phi(\boldsymbol{x})}{\mathrm{tr}A}\mu(\boldsymbol{x}).$$

Our proposed SMT method

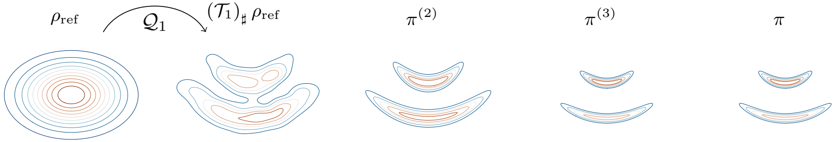


$\boxed{\text{KR}}$ Build the Knothe-Rosenblatt map $\mathcal{Q}_1$ from $\tilde{\pi}^{(1)}$

1. Normalize $\tilde{f}^{(1)}$ to get $\tilde{\pi}^{(1)}$
2. Build marginals $\pi(\boldsymbol{x}_{1:k})$ and CDFs of marginals

$$\Pi(\boldsymbol{x}_k|\boldsymbol{x}_{1:k-1}) = \frac{\Pi(\boldsymbol{x}_{1:k-1}, \boldsymbol{x}_k)}{\pi(\boldsymbol{x}_{1:k-1})}$$

Our proposed SMT method

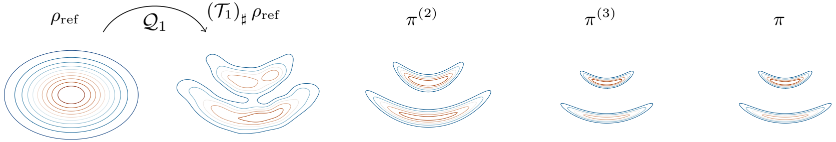


$\boxed{\text{KR}}$ Build the Knothe-Rosenblatt map $\mathcal{Q}_1$ from $\tilde{\pi}^{(1)}$

1. Normalize $\tilde{f}^{(1)}$ to get $\tilde{\pi}^{(1)}$
2. Build marginals $\pi(\boldsymbol{x}_{1:k})$ and CDFs of marginals
3. Map $\overline{\mathcal{Q}}$ from uniform distribution to $\tilde{\pi}^{(1)}$

$$\overline{\mathcal{Q}}(\xi_1,\dots,\xi_d) = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} \Pi^{-1}(\xi_1) \\ \Pi^{-1}(\xi_2|x_1) \\ \vdots \\ \Pi^{-1}(\xi_d|x_1,\dots,x_{d-1}) \end{pmatrix}$$

Our proposed SMT method

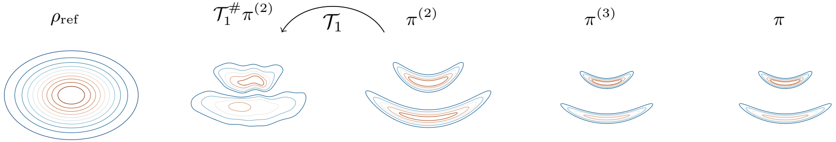


$\boxed{\text{KR}}$ Build the Knothe-Rosenblatt map $\mathcal{Q}_1$ from $\tilde{\pi}^{(1)}$

1. Normalize $\tilde{f}^{(1)}$ to get $\tilde{\pi}^{(1)}$
2. Build marginals $\pi(\boldsymbol{x}_{1:k})$ and CDFs of marginals
3. Map $\overline{\mathcal{Q}}$ from uniform distribution to $\tilde{\pi}^{(1)}$
4. Map $\mathcal{Q}$ from ρ_{ref} to $\tilde{\pi}^{(1)}$

$$\mathcal{R}_\sharp\, \rho_{\text{ref}} = U \quad \Rightarrow \quad (\mathcal{R} \circ \overline{\mathcal{Q}})_\sharp\, \rho_{\text{ref}} = \tilde{\pi}^{(1)}$$

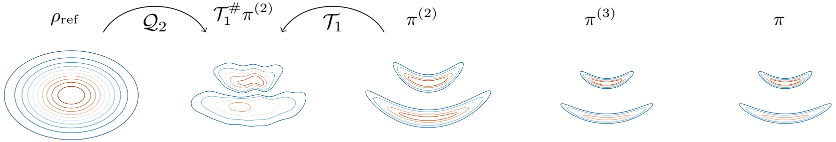
Our proposed SMT method



$\boxed{\text{PB}}$ Use $\mathcal{T}_1 = \mathcal{Q}_1$ to simplify $\pi^{(2)}$ using a *pullback* (idea originally from Cui and Dolgov 2021)

$$\tau^{(2)} = \mathcal{T}_1^{\sharp}\pi^{(2)}$$

Our proposed SMT method



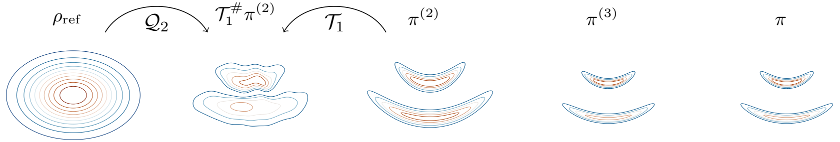
$\boxed{\text{VDE}}$ on density $\tau^{(2)}$ to get $\tilde{\tau}^{(2)}$

$$\tilde{\tau}^{(2)} = \underset{\tilde{\pi}\in\mathcal{M}}{\arg\min}\, \mathrm{D}_\alpha\left(\mathcal{T}_1^{\sharp}\pi^{(2)}||\tilde{\tau}\right)$$

$\boxed{\text{KR}}$ map $\mathcal{Q}_2$ from $\tilde{\tau}^{(2)}$

$$\begin{aligned}\Rightarrow \tilde{\pi}^{(2)} &= (\mathcal{T}_1)_\sharp \tilde{\tau}^{(2)} \\ &= \underbrace{(\mathcal{T}_1 \circ \mathcal{Q}_2)_\sharp}_{\mathcal{T}_2} \rho_{\mathrm{ref}}\end{aligned}$$

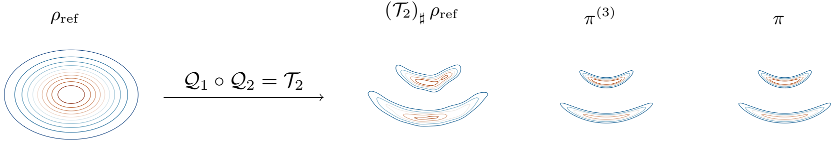
Our proposed SMT method



Algorithm SMT with KR maps

$\mathcal{T}_0 = \mathrm{id}$
for $\ell = 1$ to L **do**
    $\boxed{\text{PB}}$ $\tau^{(\ell)} = \mathcal{T}_{\ell-1}^{\sharp}\pi^{(\ell)}$
    $\boxed{\text{VDE}}$ to estimate $\tau^{(\ell)}$ by $\tilde{\tau}^{(\ell)}$
    $\boxed{\text{KR}}$ map $\mathcal{Q}_\ell$ from $\tilde{\tau}^{(\ell)}$
    $\mathcal{T}_\ell = \mathcal{T}_{\ell-1} \circ \mathcal{Q}_\ell$
end for

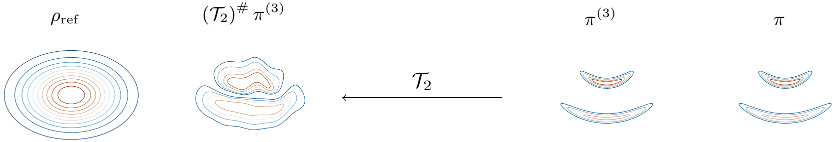
Our proposed SMT method



Algorithm SMT with KR maps

1: $\mathcal{T}_0 = \mathrm{id}$
2: **for** $\ell = 1$ to L **do**
3: $\boxed{\text{PB}}$ $\tau^{(\ell)} = \mathcal{T}_{\ell-1}^{\sharp}\pi^{(\ell)}$
4: $\boxed{\text{VDE}}$ to estimate $\tau^{(\ell)}$ by $\tilde{\tau}^{(\ell)}$
5: $\boxed{\text{KR}}$ map $\mathcal{Q}_\ell$ from $\tilde{\tau}^{(\ell)}$
6: $\mathcal{T}_\ell = \mathcal{T}_{\ell-1} \circ \mathcal{Q}_\ell$
7: **end for**

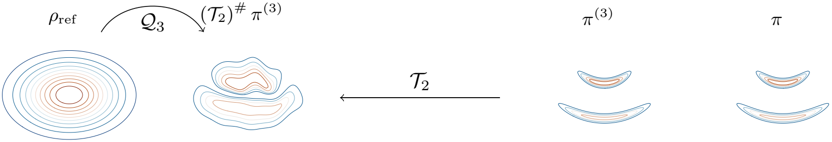
Our proposed SMT method



Algorithm SMT with KR maps

1: $\mathcal{T}_0 = \mathrm{id}$
2: **for** $\ell = 1$ to L **do**
3: $\boxed{\text{PB}}$ $\tau^{(\ell)} = \mathcal{T}_{\ell-1}^{\sharp}\pi^{(\ell)}$
4: $\boxed{\text{VDE}}$ to estimate $\tau^{(\ell)}$ by $\tilde{\tau}^{(\ell)}$
5: $\boxed{\text{KR}}$ map $\mathcal{Q}_\ell$ from $\tilde{\tau}^{(\ell)}$
6: $\mathcal{T}_\ell = \mathcal{T}_{\ell-1} \circ \mathcal{Q}_\ell$
7: **end for**

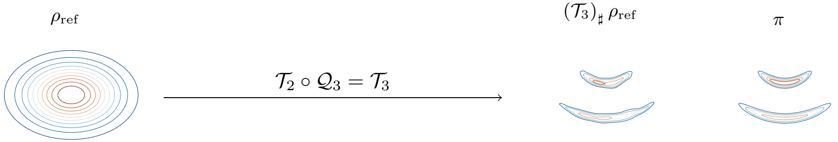
Our proposed SMT method



Algorithm SMT with KR maps

1: $\mathcal{T}_0 = \mathrm{id}$
2: **for** $\ell = 1$ to L **do**
3: $\boxed{\text{PB}}$ $\tau^{(\ell)} = \mathcal{T}_{\ell-1}^{\sharp}\pi^{(\ell)}$
4: $\boxed{\text{VDE}}$ to estimate $\tau^{(\ell)}$ by $\tilde{\tau}^{(\ell)}$
5: $\boxed{\text{KR}}$ map $\mathcal{Q}_\ell$ from $\tilde{\tau}^{(\ell)}$
6: $\mathcal{T}_\ell = \mathcal{T}_{\ell-1} \circ \mathcal{Q}_\ell$
7: **end for**

Our proposed SMT method

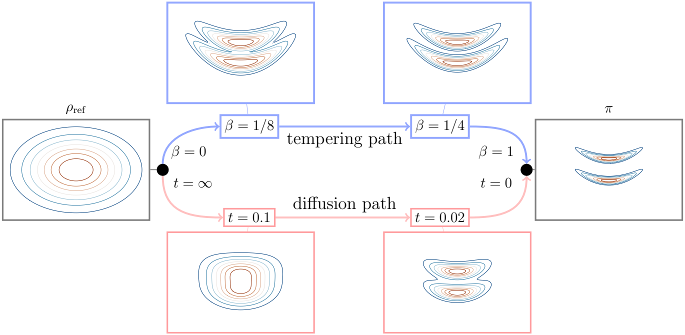


Algorithm SMT with KR maps

1: $\mathcal{T}_0 = \mathrm{id}$
2: **for** $\ell = 1$ to L **do**
3: $\boxed{\text{PB}}$ $\tau^{(\ell)} = \mathcal{T}^{\sharp}_{\ell-1}\pi^{(\ell)}$
4: $\boxed{\text{VDE}}$ to estimate $\tau^{(\ell)}$ by $\tilde{\tau}^{(\ell)}$
5: $\boxed{\text{KR}}$ map $\mathcal{Q}_\ell$ from $\tilde{\tau}^{(\ell)}$
6: $\mathcal{T}_\ell = \mathcal{T}_{\ell-1} \circ \mathcal{Q}_\ell$
7: **end for**

Outline

Bridging densities

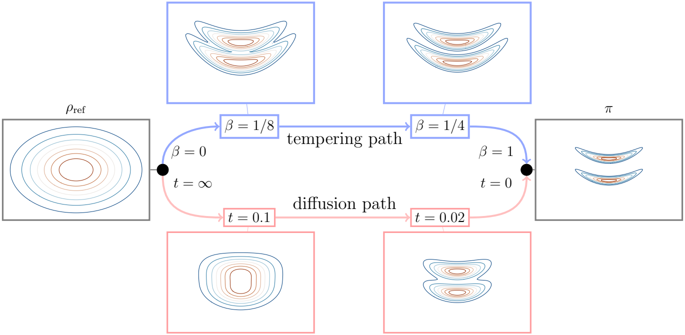


- *tempering* and *diffusion* among most popular bridging densities

Tempering

$$\pi^{(\ell)}(\boldsymbol{x}) = \rho_{\text{ref}}(\boldsymbol{x})^{1-\beta_\ell}\pi(\boldsymbol{x})^{\beta_\ell}$$

Bridging densities

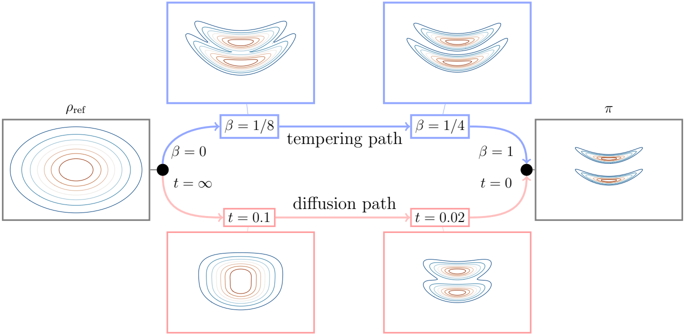


- *tempering* and *diffusion* among most popular bridging densities

Diffusion

$$X_{t_\ell}^{(i)} = \exp(-t_\ell)X^{(i)} + \sqrt{1 - \exp(-2t_\ell)}Z^{(i)} \qquad X^{(i)} \sim \pi, Z^{(i)} \sim \mathcal{N}(0, I)$$

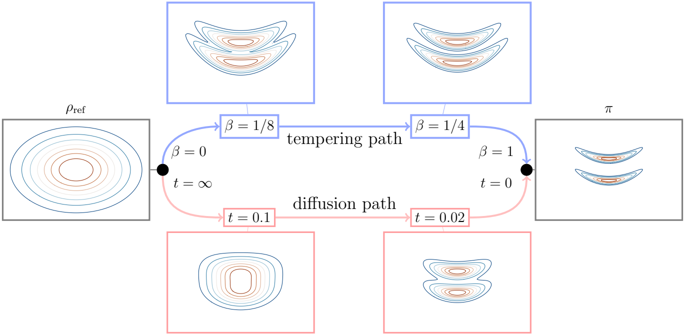
Bridging densities



- *tempering* and *diffusion* among most popular bridging densities
- How to choose hyperparamters, $\beta_1, \beta_2, \ldots$ and $t_1, t_2, \ldots$?

  $\Rightarrow$ Scheduling problem (Kingma et al. 2023)

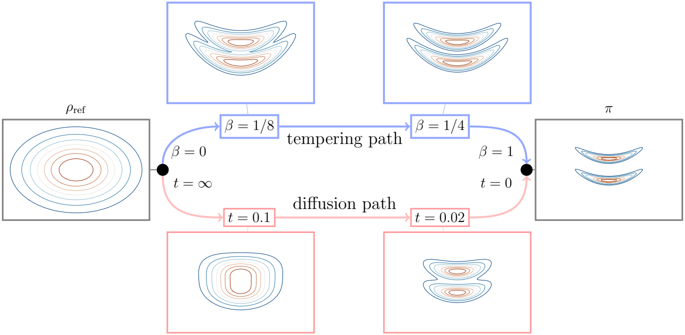
Bridging densities



Let η be so that

$$\mathrm{D}_\alpha(\pi^{(\ell+1)} || \pi^{(\ell)}) \leq \eta \qquad \text{for all } \ell = 1, \dots, L.$$

Bridging densities



Let η be so that

$$\mathrm{D}_\alpha(\pi^{(\ell+1)}||\pi^{(\ell)}) \leq \eta \qquad \text{for all } \ell = 1, \dots, L.$$

Sequences of β_ℓ and t_ℓ exists which satisfy $\eta = \mathcal{O}(1/L^2)$

Bridging densities – scheduling

Let ω be

$$\mathrm{D}_\alpha\left(\pi^{(\ell)}||\tilde{\pi}^{(\ell)}\right) \leq \omega\, \mathrm{D}_\alpha\left(\pi^{(\ell)}||\tilde{\pi}^{(\ell-1)}\right) \qquad \text{for all } \ell = 1, \ldots, L.$$

Bridging densities – scheduling

Let ω be

$$\mathrm{D}_\alpha\left(\pi^{(\ell)}||\tilde{\pi}^{(\ell)}\right) \leq \omega\, \mathrm{D}_\alpha\left(\pi^{(\ell)}||\tilde{\pi}^{(\ell-1)}\right) \qquad \text{for all } \ell = 1, \dots, L.$$

Proposition (informal, Zanger et al. 2024)

If $\omega < 1$, then the estimation of $\pi^{(L)}$ is bounded with

$$\mathrm{D}_\alpha\left(\pi^{(L)}||\widetilde{\pi}^{(L)}\right) \leq \frac{\omega(1+\epsilon)}{1-\omega}\eta.$$

Bridging densities – scheduling

Let ω be

$$\mathrm{D}_\alpha\left(\pi^{(\ell)}||\tilde{\pi}^{(\ell)}\right) \leq \omega\, \mathrm{D}_\alpha\left(\pi^{(\ell)}||\tilde{\pi}^{(\ell-1)}\right) \qquad \text{for all } \ell = 1, \ldots, L.$$

Proposition (informal, Zanger et al. 2024)

If $\omega < 1$, then the estimation of $\pi^{(L)}$ is bounded with

$$\mathrm{D}_\alpha\left(\pi^{(L)}||\widetilde{\pi}^{(L)}\right) \leq \frac{\omega(1+\epsilon)}{1-\omega}\eta.$$

- **Our strategy:** Equidistant changes to minimize η.
- **Different approach:** Best for the given approximation tool to minimize ω, see Marzouk et al. 2024.

Outline

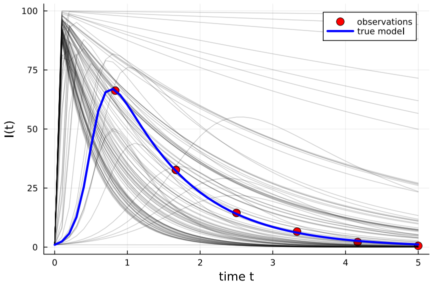
Numerical example - SIR model

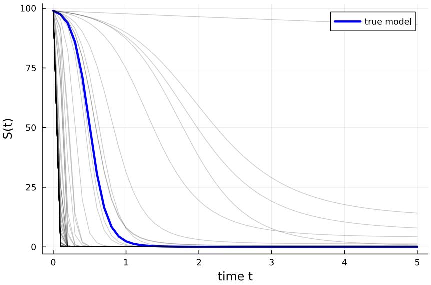
Model from epidemiology:

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}S = -\beta IS \\ \frac{\mathrm{d}}{\mathrm{d}t}I = \beta SI - \gamma I \\ \frac{\mathrm{d}}{\mathrm{d}t}R = \gamma I \end{cases}$$

Given observations of I, determine β and γ with $\pi_{\text{prior}} = U([0,2])^2$.

$$\mathcal{L}(\boldsymbol{y}|\boldsymbol{x}) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_i (I_{t_i} - y_i)^2\right)$$

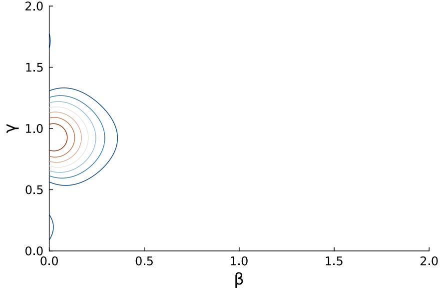




Numerical example - SIR

- Legendre polynomials of $\deg(\Phi) \leq 6$
- 1000 density evaluations for each VDE

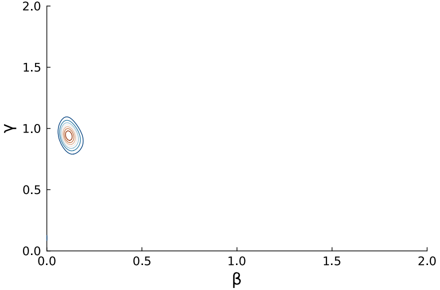
$(T_1)_\sharp \, \rho_{\text{ref}}$



Numerical example - SIR

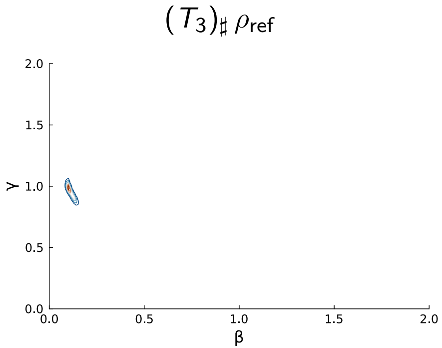
- Legendre polynomials of $\deg(\Phi) \leq 6$
- 1000 density evaluations for each VDE

$(T_2)_\sharp\, \rho_{\text{ref}}$



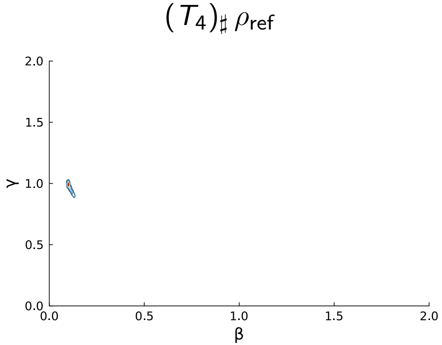
Numerical example - SIR

- Legendre polynomials of $\deg(\Phi) \leq 6$
- 1000 density evaluations for each VDE



Numerical example - SIR

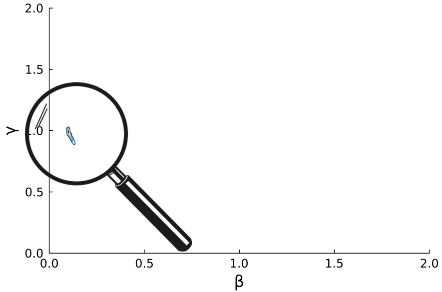
- Legendre polynomials of $\deg(\Phi) \leq 6$
- 1000 density evaluations for each VDE



Numerical example - SIR

- Legendre polynomials of $\deg(\Phi) \leq 6$
- 1000 density evaluations for each VDE

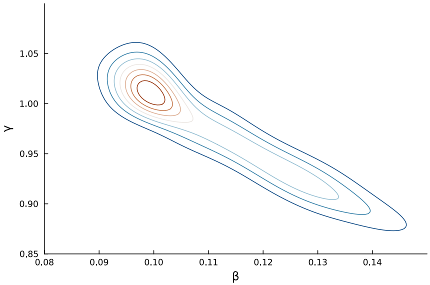
$(T_4)_\sharp\, \rho_{\text{ref}}$



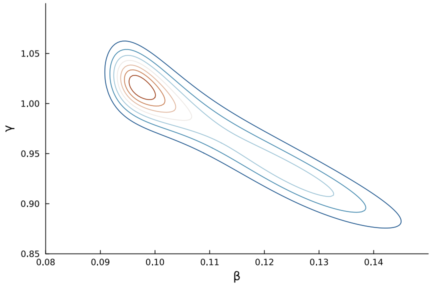
Numerical example - SIR

- Legendre polynomials of $\deg(\Phi) \leq 6$
- 1000 density evaluations for each VDE

$(T_4)_\sharp\, \rho_{\text{ref}}$

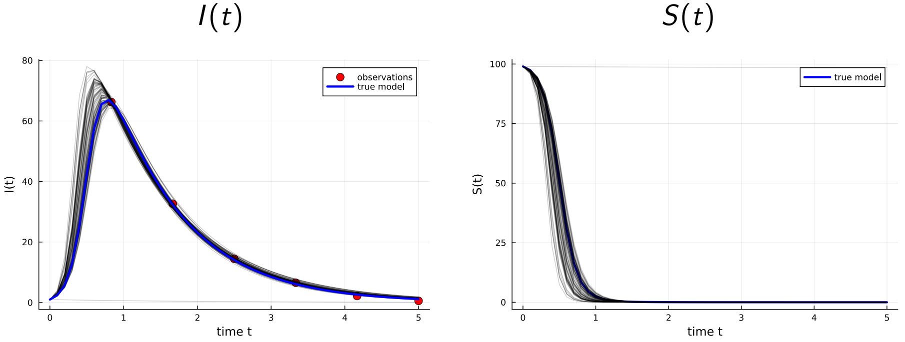


Unnormalized posterior



Numerical example - SIR

- Legendre polynomials of $\deg(\Phi) \leq 6$
- 1000 density evaluations for each VDE



Learning from datasets - Conditional density estimation

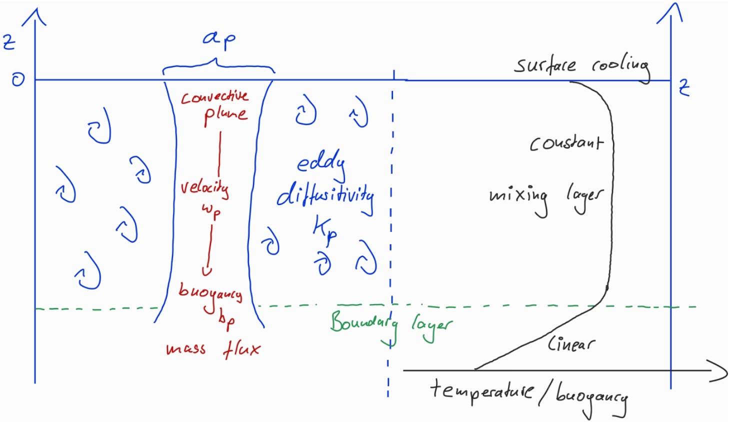
- 10 runs with different split of training/validation data
- Comparison with Baptista et al. 2023
- SoS polynomials of dimension 8, degree ≤ 3, use of random lazy maps

Table: Comparison of negative log likelihood function for different UCI datasets and conditional density estimation methods.

Dataset	(d, N)	SoS (1 sample)	SoS (4 samples)	ATM	# seq.
Boston	(12, 506)	2.8 ± 0.2	$\mathbf{2.5 \pm 0.2}$	$\mathbf{2.6 \pm 0.2}$	27.4 ± 4.2
Concrete	(9, 1030)	3.4 ± 0.3	$\mathbf{3.1 \pm 0.1}$	$\mathbf{3.1 \pm 0.1}$	44.0 ± 4.8
Energy	(10, 768)	2.2 ± 0.2	$\mathbf{1.7 \pm 0.2}$	$\mathbf{1.5 \pm 0.1}$	34.8 ± 6.1
Yacht	(7,308)	3.4 ± 1.1	2.0 ± 0.6	$\mathbf{0.5 \pm 0.2}$	40.8 ± 12.7

Calibration of Ocean-Surface parametrization

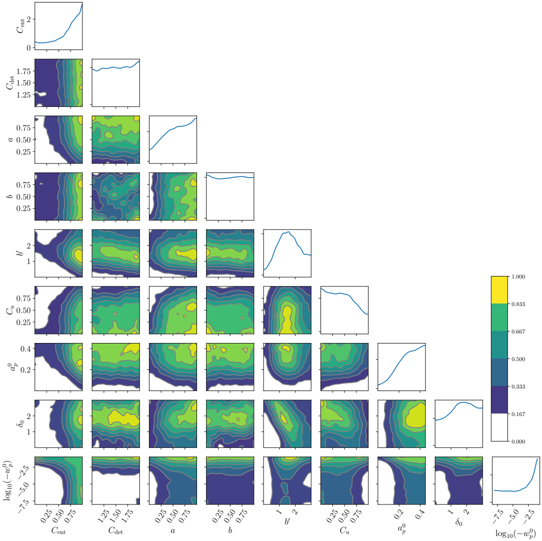
Joint work with Manolis Perrot and Florian Lemarié (Perrot et al. 2025).



⇒ Parametrization of the ocean surface using 9 free parameters which need to be calibrated.

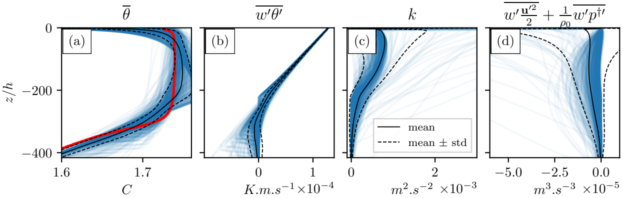
Calibration of Ocean-Surface parametrization

- 2 reference trajectories
- Calibration using the temperature at different depths
- 5 bridging densities
- 5000 density evaluations per estimation
- Polynomials with $\deg(\Phi) \leq 3$ and Putinar's Positivstellensatz

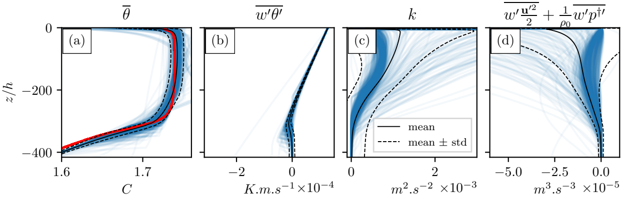


Calibration of Ocean-Surface parametrization

Prior:



Posterior:



Outline

Regularized Optimal Transport

Shortest path between p and q with cost c.

Kantorovich formulation:

$$\pi^* = \underset{\pi \in \Pi(p,q)}{\arg\min} \int_{\mathcal{X} \times \mathcal{Y}} c(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{y}) + \epsilon \cdot \mathrm{D}_{\varphi}(\pi || \pi_{\text{ref}})$$

$\epsilon = 10$ $\epsilon = 0.5$ $\epsilon = 0.03$ $\epsilon = 0.001$

Regularized Optimal Transport

Shortest path between p and q with cost c.
Kantorovich formulation:

$$\pi^* = \underset{\pi \in \Pi(p,q)}{\arg\min} \int_{\mathcal{X}\times\mathcal{Y}} c(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{y}) + \epsilon \cdot \mathrm{D}_{\varphi}(\pi || \pi_{\mathsf{ref}})$$

Our idea:
Model the transport plan π using sequential measure transport:

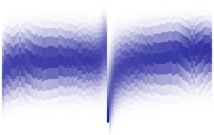
$$\pi^{(\ell)} = (\mathcal{T}_{\ell-1})_\sharp \, \tau^{(\ell)} = (\mathcal{T}_\ell)_\sharp \, \rho_{\mathsf{ref}}$$

with $\pi^{(\ell)}$ solution to regularized OT with ϵ_ℓ.

Regularized Optimal Transport

Shortest path between p and q with cost c.
Kantorovich formulation:

$$\tau^{(\ell)} = \underset{(\mathcal{T}_{\ell-1})_\sharp \tau \in \Pi(p,q)}{\arg\min} \int_{\mathcal{X}\times\mathcal{Y}} c \circ \mathcal{T}_{\ell-1}(\boldsymbol{x},\boldsymbol{y}) \mathrm{d}\tau(\boldsymbol{x},\boldsymbol{y}) + \epsilon_\ell \cdot \mathrm{D}_\varphi\left(\tau || \mathcal{T}^\sharp_{\ell-1}\pi_{\mathrm{ref}}\right)$$



$\xleftarrow{\mathcal{T}_1^\sharp}$

$\tilde{\pi}^{(1)}$ $\mathcal{T}_1^\sharp \pi^{(2)}$ $\pi^{(2)}$

Regularized Optimal transport

Shortest path between p and q with cost c.

Kantorovich formulation:

$$\pi^{(\ell)} = \underset{\pi \in \Pi(p,q)}{\arg\min} \int_{\mathcal{X}\times\mathcal{Y}} c(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{y}) + \epsilon_\ell \cdot \mathrm{D}_\varphi(\pi || \pi_{\text{ref}})$$

Dual formulation:

$$\underset{u,v}{\arg\max} \int_{\mathcal{X}} u(\boldsymbol{x}) \mathrm{d}p(\boldsymbol{x}) + \int_{\mathcal{Y}} v(\boldsymbol{y}) \mathrm{d}q(\boldsymbol{y}) - \epsilon_\ell \cdot \int_{\mathcal{X}\times\mathcal{Y}} \varphi^\dagger \left(\frac{u(\boldsymbol{x}) + v(\boldsymbol{y}) - c(\boldsymbol{x}, \boldsymbol{y})}{\epsilon_\ell} \right) \mathrm{d}\pi_{\mathrm{ref}}.$$

Regularized Optimal transport

Shortest path between p and q with cost c.

Kantorovich formulation:

$$\pi^{(\ell)} = \underset{\pi \in \Pi(p,q)}{\arg\min} \int_{\mathcal{X}\times\mathcal{Y}} c(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{y}) + \epsilon_\ell \cdot \mathrm{D}_\varphi \left(\pi || \pi_{\mathsf{ref}}\right)$$

Dual formulation:

$$\underset{u,v}{\arg\max} \int_{\mathcal{X}} u(\boldsymbol{x}) \mathrm{d}p(\boldsymbol{x}) + \int_{\mathcal{Y}} v(\boldsymbol{y}) \mathrm{d}q(\boldsymbol{y}) - \epsilon_\ell \cdot \int_{\mathcal{X}\times\mathcal{Y}} \varphi^\dagger \left(\frac{u(\boldsymbol{x}) + v(\boldsymbol{y}) - c(\boldsymbol{x}, \boldsymbol{y})}{\epsilon_\ell} \right) \frac{\pi_{\mathrm{ref}}}{(\mathcal{T}_{\ell-1})_\sharp \rho} \cdot \mathrm{d}(\mathcal{T}_{\ell-1})_\sharp \rho.$$

Regularized Optimal transport

Shortest path between p and q with cost c.

Kantorovich formulation:

$$\pi^{(\ell)} = \underset{\pi \in \Pi(p,q)}{\arg\min} \int_{\mathcal{X}\times\mathcal{Y}} c(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{y}) + \epsilon_\ell \cdot \mathrm{D}_\varphi (\pi || \pi_{\text{ref}})$$

Dual formulation:

$$\underset{u,v}{\arg\max} \int_{\mathcal{X}} u(\boldsymbol{x}) \mathrm{d}p(\boldsymbol{x}) + \int_{\mathcal{Y}} v(\boldsymbol{y}) \mathrm{d}q(\boldsymbol{y})$$
$$- \epsilon_\ell \cdot \int_{\mathcal{X}\times\mathcal{Y}} \varphi^\dagger \left(\frac{u(\boldsymbol{x}) + v(\boldsymbol{y}) - c(\boldsymbol{x}, \boldsymbol{y})}{\epsilon_\ell} \right) \frac{\pi_{\text{ref}}}{(\mathcal{T}_{\ell-1})_\sharp \rho} \cdot \mathrm{d}(\mathcal{T}_{\ell-1})_\sharp \rho.$$

Application of this idea in Sinkhorn algorithm: Schmitzer 2019 and Li et al. 2023

Outline

Summary & Future work

We presented the framework of **Sequential measure transport** which includes state-of-the art density estimation methods.

- KR maps with SoS densities
- Problem of scheduling
- Unified convergence analysis for α-divergences

Challenges and open problems:

- **Amount of density evaluations:** In Westermann and Zech 2023 and Cui et al. 2023b sampling is clear (least squares problem). Combination of SoS with α-divergences less clear $\rightarrow$ cross-validation.
- **Curse of dimensionality:** Scaling to high dimension requires adaptivity.

Link to preprint

https://arxiv.org/abs/2402.17943

Code

Generic implementation of sequential measure transport mechanism and SoS maps

Many thanks to my supervisors, collaborators, and for helpful discussions with Ricardo Baptista.

References I

Baptista, Ricardo, Youssef Marzouk, and Olivier Zahm (2023). "On the representation and learning of monotone triangular transport maps". In: *Foundations of Computational Mathematics*, pp. 1–46.

Cui, Tiangang and Sergey Dolgov (Sept. 2021). "Deep composition of tensor-trains using squared inverse Rosenblatt transports". en. In: *Foundations of Computational Mathematics*. arXiv:2007.06968 [cs, math, stat]. ISSN: 1615-3375, 1615-3383. DOI: 10.1007/s10208-021-09537-5. URL: http://arxiv.org/abs/2007.06968 (visited on 07/19/2022).

Cui, Tiangang, Sergey Dolgov, and Robert Scheichl (2024). "Deep Importance Sampling Using Tensor Trains with Application to a Priori and a Posteriori Rare Events". In: *SIAM Journal on Scientific Computing* 46.1, pp. C1–C29.

References II

Cui, Tiangang, Sergey Dolgov, and Olivier Zahm (2023a). "Scalable conditional deep inverse Rosenblatt transports using tensor trains and gradient-based dimension reduction". In: *Journal of Computational Physics* 485, p. 112103. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2023.112103. URL: https://www.sciencedirect.com/science/article/pii/S0021999123001985.

— (Mar. 2023b). *Self-reinforced polynomial approximation methods for concentrated probability densities*. en. arXiv:2303.02554 [cs, math, stat]. URL: http://arxiv.org/abs/2303.02554 (visited on 03/20/2023).

Grenioux, Louis, Alain Durmus, Éric Moulines, and Marylou Gabrié (2023). "On Sampling with Approximate Transport Maps". In: *arXiv preprint arXiv:2302.04763*.

References III

Kingma, Diederik P., Tim Salimans, Ben Poole, and Jonathan Ho (2023). *Variational Diffusion Models*. arXiv: 2107.00630 [cs.LG]. URL: https://arxiv.org/abs/2107.00630.

Li, Mengyu, Jun Yu, Tao Li, and Cheng Meng (2023). "Importance Sparsification for Sinkhorn Algorithm". en. In: *Journal of Machine Learning Research* 24, pp. 1–44.

Marzouk, Youssef, Tarek Moselhy, Matthew Parno, and Alessio Spantini (2016). "An introduction to sampling via measure transport". en. In: arXiv:1602.05023 [math, stat], pp. 1–41. DOI: 10.1007/978-3-319-11259-6_23-1. URL: http://arxiv.org/abs/1602.05023 (visited on 07/20/2022).

Marzouk, Youssef, Zhi Ren, Sven Wang, and Jakob Zech (2024). "Distribution Learning via Neural Differential Equations: A Nonparametric Statistical Perspective". en. In: *Journal of Machine Learning Research*.

References IV

- Perrot, M., F. Lemarié, and T. Dubos (Jan. 2025). "Energetically Consistent Eddy-Diffusivity Mass-Flux Convective Schemes: 1. Theory and Models". en. In: *Journal of Advances in Modeling Earth Systems* 17.1, e2024MS004273. ISSN: 1942-2466, 1942-2466. DOI: 10.1029/2024MS004273. URL: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004273 (visited on 04/21/2025).
- Rezende, Danilo Jimenez and Shakir Mohamed (June 2016). *Variational Inference with Normalizing Flows*. en. arXiv:1505.05770 [cs, stat]. URL: http://arxiv.org/abs/1505.05770 (visited on 06/13/2023).
- Schmitzer, Bernhard (Feb. 2019). *Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems*. en. arXiv:1610.06519 [math]. DOI: 10.48550/arXiv.1610.06519. URL: http://arxiv.org/abs/1610.06519 (visited on 03/20/2025).

References V

Westermann, Josephine and Jakob Zech (Nov. 2023). *Measure transport via polynomial density surrogates*. en. arXiv:2311.04172 [cs, math, stat]. URL: http://arxiv.org/abs/2311.04172 (visited on 12/01/2023).

Zanger, Benjamin, Olivier Zahm, Tiangang Cui, and Martin Schreiber (Oct. 2024). *Sequential transport maps using SoS density estimation and $*

alpha$-divergences. en. arXiv:2402.17943 [stat]. URL: http://arxiv.org/abs/2402.17943 (visited on 11/17/2024).

Convergence analysis - general idea

General proof idea from Cui et al. 2023a

$$
\begin{aligned}
\mathrm{D}\left(\pi^{(L)}||\widetilde{\pi}^{(L)}\right) &\leq \omega \mathrm{D}\left(\pi^{(L)}||\widetilde{\pi}^{(L-1)}\right) \\
&\leq \omega \mathrm{D}\left(\pi^{(L)}||\pi^{(L-1)}\right) + \omega \mathrm{D}\left(\pi^{(L-1)}||\widetilde{\pi}^{(L-1)}\right)
\end{aligned}
$$

recursion...

Convergence analysis - general idea

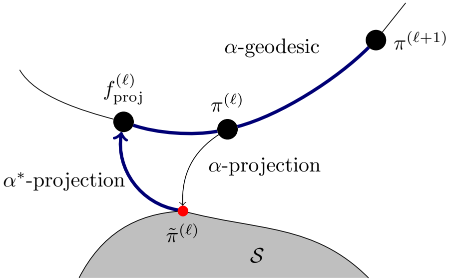
General proof idea from Cui et al. 2023a

$$\begin{aligned}
\mathrm{D}\left(\pi^{(L)}||\widetilde{\pi}^{(L)}\right) &\leq \omega \mathrm{D}\left(\pi^{(L)}||\widetilde{\pi}^{(L-1)}\right) \\
&\leq \omega \mathrm{D}\left(\pi^{(L)}||\pi^{(L-1)}\right) + \omega \mathrm{D}\left(\pi^{(L-1)}||\widetilde{\pi}^{(L-1)}\right)
\end{aligned}$$

recursion...

Problem: triangle inequality does not hold for α-divergences!

Convergence analysis by geometric properties of α-divergences



Assumption

We assume that there exists an $\epsilon \geq 0$ *so that*

$$\mathrm{D}_\alpha\left(\pi^{(\ell+1)} || f_{\text{proj}}^{(\ell)}\right) \leq (1+\epsilon)\, \mathrm{D}_\alpha\left(\pi^{(\ell+1)} || \pi^{(\ell)}\right) \qquad \forall l \in [L].$$

Convergence analysis by geometric properties of α-divergences

Pythagorean theorem instead of the triangle inequality:

$$\begin{aligned}
\mathsf{D}_\alpha\left(\pi^{(L)}||\widetilde{\pi}^{(L)}\right) &\leq \omega\,\mathsf{D}_\alpha\left(\pi^{(L)}||\widetilde{\pi}^{(L-1)}\right) \\
&= \omega\,\mathsf{D}_\alpha\left(\pi^{(L)}||f_{\mathrm{proj}}^{(L-1)}\right) + \omega\,\mathsf{D}_\alpha\left(f_{\mathrm{proj}}^{(L-1)}||\widetilde{\pi}^{(L-1)}\right) \\
&\leq \omega(1+\epsilon)\,\mathsf{D}_\alpha\left(\pi^{(L)}||\pi^{(L-1)}\right) + \omega\,\mathsf{D}_\alpha\left(f_{\mathrm{proj}}^{(L-1)}||\widetilde{\pi}^{(L-1)}\right) \\
&\leq \omega(1+\epsilon)\,\mathsf{D}_\alpha\left(\pi^{(L)}||\pi^{(L-1)}\right) + \omega\,\mathsf{D}_\alpha\left(\pi^{(L-1)}||\widetilde{\pi}^{(L-1)}\right)
\end{aligned}$$

Convergence analysis by geometric properties of α-divergences

Proposition

Furthermore, let $\mathrm{D}_\alpha(\pi^{(0)}||\widetilde{\pi}^{(0)}) = 0$*, for example by* $\pi^{(0)} = \rho_{ref}$*. If* $\omega < 1$*, than the estimation of* $\pi^{(L)}$ *is bounded with*

$$\mathrm{D}_\alpha\left(\pi^{(L)}||\widetilde{\pi}^{(L)}\right) \leq \frac{\omega(1+\epsilon)}{1-\omega}\eta(L).$$

Entropic regularized optimal transport

The entropic regularized optimal transport problem between distributions p and q with cost c can be written as

$$\tilde{\pi} = \arg\min_{\tilde{\pi}} \mathrm{D}_{\mathrm{KL}}\left(\pi || \xi_\epsilon\right) \qquad \text{s.t.} \quad \int_{\mathcal{X}} \pi(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} = q(\boldsymbol{y})$$
$$\int_{\mathcal{Y}} \pi(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{y} = p(\boldsymbol{x})$$

$\epsilon = 10$ $\qquad$ $\epsilon = 0.5$ $\qquad$ $\epsilon = 0.03$ $\qquad$ $\epsilon = 0.001$

Entropic regularized optimal transport

The entropic regularized optimal transport problem between distributions p and q with cost c can be written as

$$\tilde{\pi}^{(2)} = \arg\min_{\pi} \mathrm{D}_{\mathrm{KL}}\left(\pi || \mathcal{T}_1^{\sharp}\xi_{\epsilon}\right) \qquad \text{s.t.} \quad \int_{\mathcal{X}} (\mathcal{T}_1)_{\sharp}\pi(\boldsymbol{x}, \boldsymbol{y})\mathrm{d}\boldsymbol{x} = q(\boldsymbol{y})$$

$$\int_{\mathcal{Y}} (\mathcal{T}_1)_{\sharp}\pi(\boldsymbol{x}, \boldsymbol{y})\mathrm{d}\boldsymbol{y} = p(\boldsymbol{x})$$

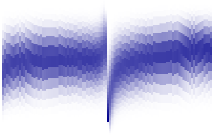
$\epsilon = 10$ $\qquad$ $\epsilon = 0.5$ $\qquad$ $\epsilon = 0.03$ $\qquad$ $\epsilon = 0.001$

Entropic regularized optimal transport

The entropic regularized optimal transport problem between distributions p and q with cost c can be written as

$$\tilde{\pi}^{(2)} = \arg\min_{\pi} \mathrm{D}_{\mathrm{KL}} \left(\pi || \mathcal{T}_1^{\sharp} \xi_{\epsilon} \right) \qquad \text{s.t.} \quad \begin{aligned} \int_{\mathcal{X}} (\mathcal{T}_1)_{\sharp} \pi(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{x} &= q(\boldsymbol{y}) \\ \int_{\mathcal{Y}} (\mathcal{T}_1)_{\sharp} \pi(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{y} &= p(\boldsymbol{x}) \end{aligned}$$

$\tilde{\pi}^{(1)}$



$\mathcal{T}_1^{\sharp} \pi^{(2)}$

$\xleftarrow{\mathcal{T}_1^{\sharp}}$

$\pi^{(2)}$