Computing high-dimensional optimal transport by flow neural networks

Chen Xu [1] Xiuyuan Cheng [2] Yao Xie [1]

[1]Georgia Institute of Technology [2]Duke University

Introduction

The problem of finding an optimal transport (OT) map between two general distributions P and Q in high dimension is essential in statistics, optimization, and machine learning. Such an OT map can be beneficial for problems including transfer learning, density ratio estimation, image-to-image translation, and so on.

This work focuses on a continuous-time formulation of the problem where we are to find an invertible transport map $T_t : \mathbb{R}^d \to \mathbb{R}^d$ continuously parametrized by time $t \in [0, 1]$ and satisfying that $T_0 = \mathrm{Id}$ and $(T_1)_{\#}P = Q$. Here we denote by $T_{\#}P$ the push-forward of distribution P by a mapping T, such that $(T_{\#}P)(\cdot) = P(T^{-1}(\cdot))$.

While different parametrizations of T_t are possible, we choose to adopt the neural Ordinary Differential Equation (ODE) approach [2], where we represent T_t as the solution map of an ODE. The velocity field in the neural ODE will be optimized to minimize the transport cost so as to approximate the optimal velocity in dynamic optimal transport (OT) formulation, i.e. Benamou-Brenier equation.

Our approach

We aim to learn $v(x(t), t)$ in the Benamou-Brenier equation [1]:

$$\inf_{\rho, v} \mathcal{T} := \int_0^1 \mathbb{E}_{x(t)\sim\rho(\cdot,t)} \|v(x(t), t)\|^2 dt \quad s.t. \quad \partial_t \rho + \nabla \cdot (\rho v) = 0, \quad \rho(x, 0) = p(x), \quad \rho(x, 1) = q(x). \tag{1}$$

We assume access only to finite samples $\boldsymbol{X} = \{X_i\} \sim P$ and $\tilde{\boldsymbol{X}} = \{\tilde{X}_i\} \sim Q$. We relax the terminal condition $\rho(\cdot, 1) = q$ by a KL divergence and consider a neural network parametrizaztion $f(x(t), t; \theta)$ of $v(x(t), t)$. We then define the solution map of the ODE from time s to t as

$$T_s^t(x; \theta) = x(s) + \int_s^t f(x(t'), t'; \theta) dt'. \tag{2}$$

Given a time discretization $0 = t_0 < t_1 < \ldots < t_K = 1$ of the unit interval, we train θ by minimizing the following loss:

$$\min_\theta \mathcal{L}^{P\to Q}(\theta) = \mathcal{L}_{\mathrm{KL}}^{P\to Q}(\theta) + \gamma \mathcal{L}_T^{P\to Q}(\theta), \tag{3}$$

where the two terms are defined as

$$\mathcal{L}_{\mathrm{KL}}^{P\to Q}(\theta) = -\frac{1}{N} \sum_{i=1}^N r_1(T_0^1(X_i; \theta); \hat{\varphi}_r). \tag{4}$$

$$\mathcal{L}_T^{P\to Q}(\theta) = \sum_{k=1}^K \frac{1}{h_k} \left(\frac{1}{N} \sum_{i=1}^N \|X_i(t_k; \theta) - X_i(t_{k-1}; \theta)\|^2 \right). \tag{5}$$

In (4), the function r_1 with parameters $\hat{\varphi}_r$ is a trained classifier between $\{T_0^1(X_i)\}$ and $\{\tilde{X}_i\}$, where r_1 aims to estimate $\log(p_1/q)$ for $p_1 = (T_0^1)_{\#}p$. Meanwhile, (5) can be viewed as a time discretization of $\mathcal{T}$ in (1), where $X_i(t_k; \theta) = T_0^{t_k}(X_i; \theta)$ and $h_k = t_k - t_{k-1}$. To increase numerical stability, we would also consider a symmetric version of (3) through $\mathcal{L}^{Q\to P}$.

In practice, we use (3) to *refine* any initialized flow between P and Q, as illustrated in Figure 1. Such an initialization can be constructed via concatenating two CNFs [5] or flow matching [3]. Given a trained OT $f(x(t), t; \hat{\theta})$, we further develop new DRE approaches leveraging the telescopic idea [4] (details in paper).

Main contributions

1. Propose *Q-flow*, a **flow-based OT map** between arbitrary P and Q in $\mathbb{R}^d$, using only finite samples from the distributions. The end-to-end training of the model refines *any* initial flow that may not attain the optimal transport.
2. Propose a new **density ratio estimator** (DRE) using classification losses along the time grid, upon leveraging the OT trajectory given by the trained Q-flow.
3. Show **effectiveness** on simulated and real-data experiments, including training energy-based generative models and image-to-image translation.

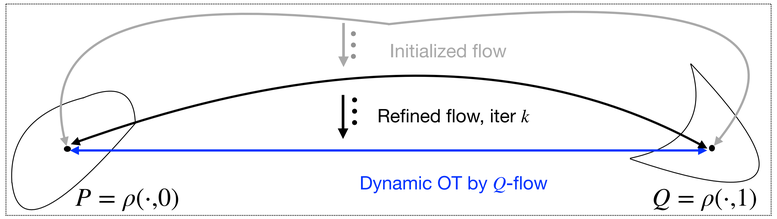
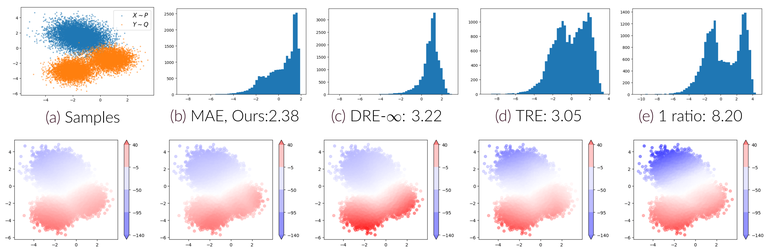


Figure 1. Illustration of learning the dynamic OT using our Q-flow (blue), which invertibly transports between P and Q over the interval $[0, 1]$ with the least transport cost. Taking any initial flow (grey) between P and Q, we iteratively refine flow trajectories to obtain flows with smaller transport cost (black), converging gradually to the dynamic OT between these two distributions.



(a) Samples (b) MAE, Ours:2.38 (c) DRE-∞: 3.22 (d) TRE: 3.05 (e) 1 ratio: 8.20

(f) Left to right: True ratio $r(x)$, $\hat{r}(x)$ by Ours, DRE-∞, TRE, and 1 ratio.

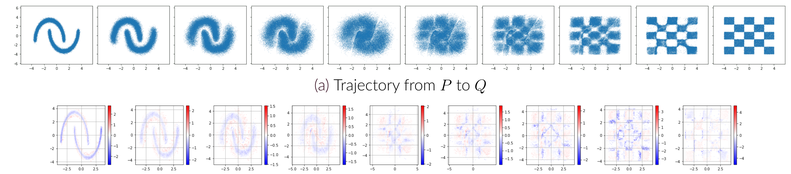
Figure 2. Estimated log density ratio between 2D Gaussian mixture distributions P (three components) and Q (two components). **Top:** (a) training samples from P and Q. (b)-(d) histograms of errors $\log(|r(x) - \hat{r}(x)|)$ computed at 10K test samples shown in log-scale. The MAE are shown in the captions. **Bottom:** true and estimated $\log(q/p)$ from different models shown under shared colorbars.

Table 1. DRE performance on the energy-based modeling task for MNIST, reported in BPD and lower is better. Results for DRE-∞ and TRE are from the original papers.

Choice of Q	RQ-NSF				Copula				Gaussian			
Method	Ours	DRE-∞	TRE	1 ratio	Ours	DRE-∞	TRE	1 ratio	Ours	DRE-∞	TRE	1 ratio
BPD (↓)	**1.05**	1.09	1.09	1.09	**1.14**	1.21	1.24	1.33	**1.31**	1.33	1.39	1.96

Numerical results

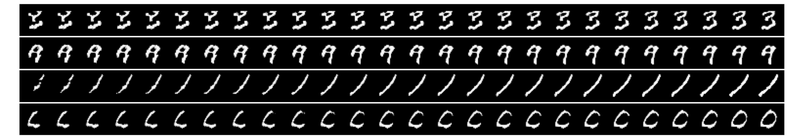
We showcase the effectiveness of our OT+DRE approach on simulated and real data. We further perform an image-to-image translation task in Figure 4b.



(a) Trajectory from P to Q

(b) Estimated log-ratio between $P_{t_{k-1}}$ and P_{t_k} by the trained Q-flow-ratio net.

Figure 3. Q-flow trajectory between arbitrary 2D distributions and corresponding log-ratio estimation. **Top**: intermediate distributions by Q-flow net. **Bottom**: corresponding log-ratio estimated by Q-flow-ratio net. Bluer color indicates smaller estimates of the difference $\log(p(x, t_k)/p(x, t_{k-1}))$ evaluated at the common support of the neighboring densities.



(a) MNIST

(b) Handbag → shoes

Figure 4. The trajectory of samples (in rows) from intermediate distributions of the Q-flow, as it pushes forward the base distribution (leftmost column) to the target distribution (rightmost column). Figure (a) shows the improvement of generated digits using the Q-flow. Figure (b) shows the image-to-image translation from handbag to shoes.

References

[1] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[2] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[3] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=PqvMRDCJT9t`.

[4] Benjamin Rhodes, Kai Xu, and Michael U. Gutmann. Telescoping density-ratio estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4905–4916. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/33d3b157ddc0896addfb22fa2a519097-Paper.pdf`.

[5] Chen Xu, Xiuyuan Cheng, and Yao Xie. Normalizing flow neural networks by JKO scheme. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=ZQM1fNijY5`.