Continuous optimal transport

Brandon Amos • Meta AI (FAIR) NYC

Entropic estimation of optimal transport maps

Aram-Alexandre Pooladian*, Jonathan Niles-Weed*†

*Center for Data Science, New York University
†Courant Institute of Mathematical Sciences, New York University
ap6599@nyu.edu, jnw@cims.nyu.edu

May 10, 2022

Abstract

We develop a computationally tractable method for estimating the optimal map between two distributions over $\mathbb{R}^d$ with rigorous finite-sample guarantees. Leveraging an entropic version of Brenier's theorem, we show that our estimator—the *barycentric projection* of the optimal entropic plan—is easy to compute using Sinkhorn's algorithm. As a result, unlike current approaches for map estimation, which are slow to evaluate when the dimension or number of samples is large, our approach is parallelizable and extremely efficient even for massive data sets. Under smoothness assumptions on the optimal map, we show that our estimator enjoys comparable statistical performance to other estimators in the literature, but with much lower computational cost. We showcase the efficacy of our proposed estimator through numerical examples. Our proofs are based on a modified duality principle for entropic optimal transport and on a method for approximating optimal entropic plans due to Pal (2019).

On amortizing convex conjugates for optimal transport

Brandon Amos
Meta AI

Abstract

This paper focuses on computing the convex conjugate operation that arises when solving Euclidean Wasserstein-2 optimal transport problems. This conjugation, which is also referred to as the Legendre-Fenchel conjugate or c-transform, is considered difficult to compute and in practice, Wasserstein-2 methods are limited by not being able to exactly conjugate the dual potentials in continuous space. I show that combining amortized approximations to the conjugate with a solver for fine-tuning is computationally easy. This combination significantly improves the quality of transport maps learned for the Wasserstein-2 benchmark by Korotin et al. (2021a) and is able to model many 2-dimensional couplings and flows considered in the literature. All of the baselines, methods, and solvers in this paper are available at: http://github.com/facebookresearch/w2ot

Background: optimal transport connects spaces

Studies how to **transport** mass **between probability measures**

- Yields a **geometry** for probability measures, e.g. to **interpolate** between them
- **Generalizes matching** and linear assignment problems

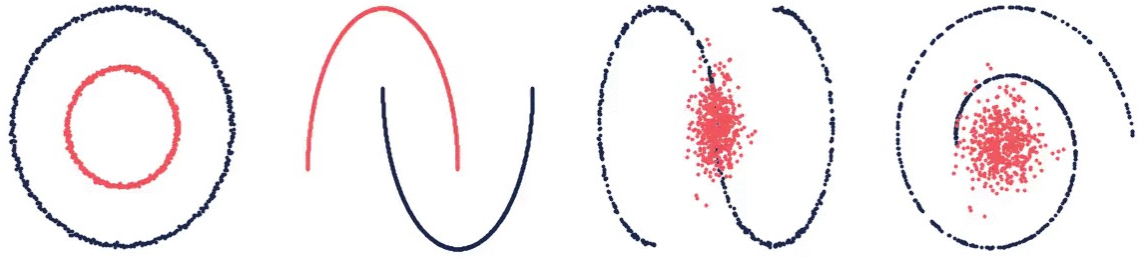
Monge's formulation between Euclidean spaces with a squared distance cost is:

$$T^\star(\alpha,\beta) \in \underset{T\in\mathcal{C}(\alpha,\beta)}{\operatorname{argmin}} \mathbb{E}_{x\sim\alpha}\|x - T(x)\|_2^2$$

where α, β are **measures**, $\mathcal{C}(\alpha,\beta)$ is a **coupling**, T is a **transport map** from α to β.

Most computational work takes α and β to be **discrete measures**

- Results in **convex formulations** with efficient solvers (e.g., linear programming, Sinkhorn)



2.1.1 Entropic OT between discrete measures with the Sinkhorn algorithm

Let $\alpha := \sum_{i=1}^m a_i \delta_{x_i}$ and $\beta := \sum_{i=1}^n b_i \delta_{y_i}$ be *discrete* measures, where δ_z is a Dirac at point z and $a \in \Delta_{m-1}$ and $b \in \Delta_{n-1}$ are in the *probability simplex* defined by

$$\Delta_{k-1} := \{x \in \mathbb{R}^k : x \geq 0 \text{ and } \sum_i x_i = 1\}. \quad (2)$$

Algorithm 1 Sinkhorn($\alpha, \beta, c, \epsilon, f_0 = 0, g_0 = 0$)

for iteration $i = 1$ to N **do**
  $f_i \leftarrow \epsilon \log a - \epsilon \log \left(K \exp\{g_{i-1}/\epsilon\}\right)$
  $g_i \leftarrow \epsilon \log b - \epsilon \log \left(K^\top \exp\{f_{i-1}/\epsilon\}\right)$
end for
Compute P_N from f_N, g_N using eq. (6)
return $P_N \approx P^\star$

Discrete OT. In the discrete setting, eq. (1) simplifies to the *linear program*

$$P^\star(\alpha, \beta, c) \in \underset{P \in U(a,b)}{\arg\min} \langle C, P \rangle \qquad U(a,b) := \{P \in \mathbb{R}_+^{n \times m} : P 1_m = a, \quad P^\top 1_n = b\} \quad (3)$$

where P is a *coupling matrix*, $P^\star(\alpha, \beta)$ is the *optimal* coupling, and the *cost* can be discretized as a matrix $C \in \mathbb{R}^{m \times n}$ with entries $C_{i,j} := c(x_i, y_j)$, and $\langle C, P \rangle := \sum_{i,j} C_{i,j} P_{i,j}$,

Entropic OT. The linear program above can be regularized adding the entropy of the coupling to smooth the objective as in Cominetti and Martín [1994], Cuturi [2013], resulting in:

$$P^\star(\alpha, \beta, c, \epsilon) \in \underset{P \in U(a,b)}{\arg\min} \langle C, P \rangle - \epsilon H(P) \quad (4)$$

where $H(P) := -\sum_{i,j} P_{i,j}(\log(P_{i,j}) - 1)$ is the discrete entropy of a coupling matrix P.

Entropic OT dual. As presented in Peyré et al. [2019, Prop. 4.4], the dual of eq. (4) is

$$f^\star, g^\star \in \underset{f \in \mathbb{R}^n, g \in \mathbb{R}^m}{\arg\max} \langle f, a \rangle + \langle g, b \rangle - \epsilon \left\langle \exp\{f/\epsilon\}, K \exp\{g/\epsilon\} \right\rangle, \quad K_{i,j} := \exp\{-C_{i,j}/\epsilon\}, \quad (5)$$

where $K \in \mathbb{R}^{m \times n}$ is the *Gibbs kernel* and the *dual variables* or *potentials* $f \in \mathbb{R}^n$ and $g \in \mathbb{R}^m$ are associated, respectively, with the marginal constraints $P 1_m = a$ and $P^\top 1_n = b$. The optimal duals depend on the problem, e.g. $f^\star(\alpha, \beta, c, \epsilon)$, but we omit this dependence for notational simplicity.

Discrete OT: permutations and matchings

LEARNING LATENT PERMUTATIONS WITH GUMBEL-SINKHORN NETWORKS

Gonzalo E. Mena *
Department of Statistics,
Columbia University
gem2131@columbia.edu

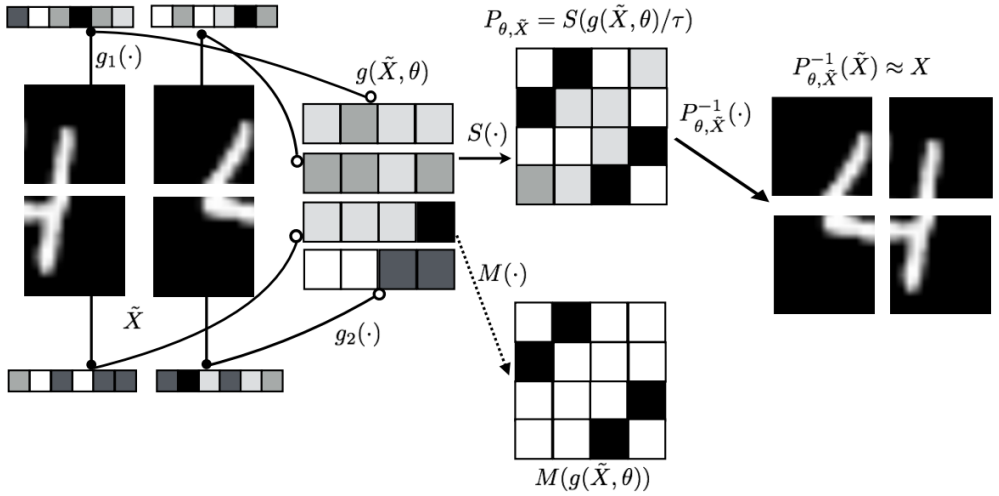
David Belanger
Google Brain

Scott Linderman
Department of Statistics,
Columbia University

Jasper Snoek
Google Brain

ABSTRACT

Permutations and matchings are core building blocks in a variety of latent variable models, as they allow us to align, canonicalize, and sort data. Learning in such models is difficult, however, because exact marginalization over these combinatorial objects is intractable. In response, this paper introduces a collection of new methods for end-to-end learning in such models that approximate discrete maximum-weight matching using the continuous Sinkhorn operator. Sinkhorn operator is attractive because it functions as a simple, easy-to-implement analog of the softmax operator. With this, we can define the Gumbel-Sinkhorn method, an extension of the Gumbel-Softmax method (Jang et al., 2016; Maddison et al., 2016) to distributions over latent matchings. We demonstrate the effectiveness of our method by outperforming competitive baselines on a range of qualitatively different tasks: sorting numbers, solving jigsaw puzzles, and identifying neural signals in worms.



Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains

Justin Solomon
Stanford University

Fernando de Goes
Pixar Animation Studios

Gabriel Peyré
CNRS & Univ. Paris-Dauphine

Marco Cuturi
Kyoto University

Adrian Butscher
Autodesk, Inc.

Andy Nguyen
Stanford University

Tao Du
Stanford University

Leonidas Guibas
Stanford University

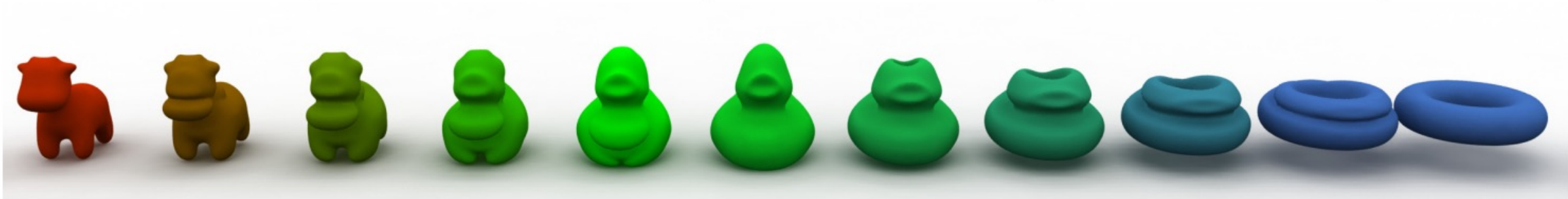


Figure 1: *Shape interpolation from a cow to a duck to a torus via convolutional Wasserstein barycenters on a* $100 \times 100 \times 100$ *grid, using the method at the beginning of* §7.

Limitations of discrete OT

1. **Computationally challenging** for high-dimensional measures (10k+ points)
   - Cost and coupling matrices may be large (100M+ entries when measures have 10k+ points)
   - Results in high-dimensional convex optimization problems
   - (Approximations may help reduce these burdens, e.g., hierarchies/slicing)
2. **Uses discretized geometries** that may come from truly continuous spaces
   - Ignores the continuous structure which may come up in practice
   - Not easy to obtain the true continuous transport map

Monge's Euclidean Wasserstein-2 formulation

$$T^\star(\alpha,\beta) \in \underset{T\in\mathcal{C}(\alpha,\beta)}{\operatorname{argmin}} \mathbb{E}_{x\sim\alpha}\|x - T(x)\|_2^2$$

Duality and continuous OT

Monge (primal)

$$T^\star(\alpha,\beta) \in \operatorname*{argmin}_{T\in\mathcal{C}(\alpha,\beta)} \mathbb{E}_{x\sim\alpha}\|x - T(x)\|_2^2$$

$T^\star = \nabla\hat{f}$

Kantorovich (dual)

$$\hat{f} \in \operatorname*{argmax}_{f\in\mathcal{L}^1(\alpha)} -\mathbb{E}_{x\sim\alpha}[f(x)] - \mathbb{E}_{y\sim\beta}[f^\star(y)]$$

$$f^\star(y) := -\inf_{x\in\mathcal{X}} J_f(x;y) \quad \text{with objective} \quad J_f(x;y) := f(x) - \langle x, y\rangle.$$

When α and β are discrete:

- T is a doubly stochastic matrix, f is a finite-dimensional vector, and the dual is convex w.r.t. f
- Expectations in the objectives are exactly computable

When α and β are continuous:

- T is a continuous map s.t. $T_\#\alpha = \beta$, in general hard to satisfy and optimize over
  - Density models (e.g., flows) only have to satisfy $T_\#\alpha = \beta$ without being optimal, still hard
- f is a continuous map. The dual no longer convex w.r.t. most parameterizations of f
- The expectations need to be approximated from samples

Wasserstein GAN: continuous Wasserstein-1 OT

Wasserstein-1:

$$\min_{T \in \mathcal{C}(\alpha,\beta)} \mathbb{E}_{x\sim\alpha}\|x - T(x)\|_2$$

$$\hat{f} \in \underset{f \in 1-\text{Lipschitz}}{\text{argmax}} \; -\mathbb{E}_{x\sim\alpha}[f(x)] - \mathbb{E}_{y\sim\beta}[f(y)]$$

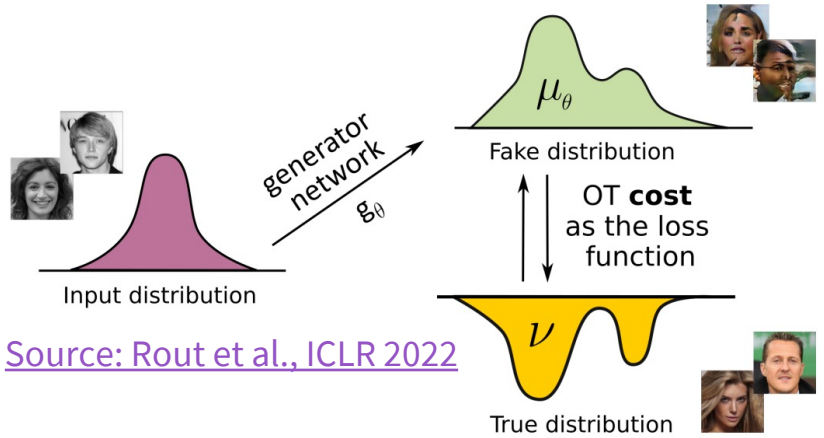
Potential f is 1-Lipschitz and self-conjugate

Most of this talk: computing Wasserstein-2
Potential f is convex, conjugate is the Fenchel

Algorithm 1 WGAN, our proposed algorithm. All experiments in the paper used the default values $\alpha = 0.00005$, $c = 0.01$, $m = 64$, $n_{\text{critic}} = 5$.

Require: : α, the learning rate. c, the clipping parameter. m, the batch size. n_{critic}, the number of iterations of the critic per generator iteration.
Require: : w_0, initial critic parameters. θ_0, initial generator's parameters.

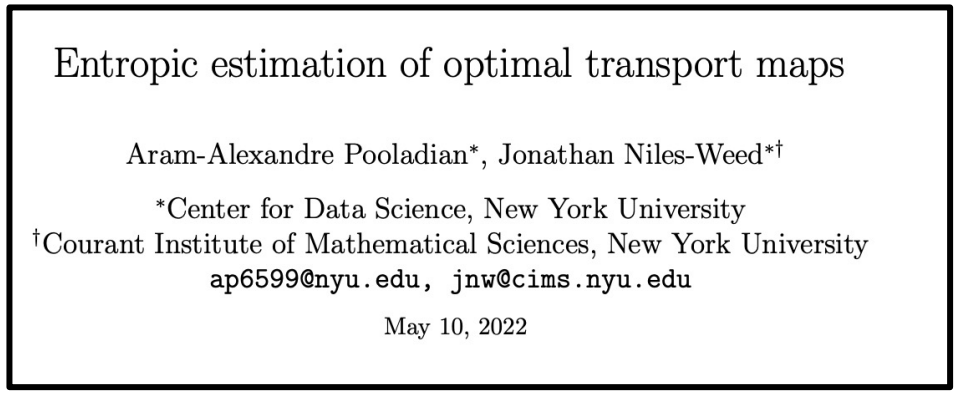
while θ has not converged **do**
  for $t = 0, \ldots, n_{\text{critic}}$ **do**
    Sample $\{x^{(i)}\}_{i=1}^{m} \sim \mathbb{P}_r$ a batch from the real data.
    Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
    $g_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^{m} f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))\right]$
    $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_w)$
    $w \leftarrow \text{clip}(w, -c, c)$
  end for
  Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
  $g_\theta \leftarrow -\nabla_\theta \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))$
  $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$
end while



Source: Rout et al., ICLR 2022

Computing continuous (Wasserstein-2) OT maps

Approach 1. solve a discrete OT problem with Sinkhorn and extend the solution to be continuous



Entropic estimation of optimal transport maps

Aram-Alexandre Pooladian*, Jonathan Niles-Weed*†

*Center for Data Science, New York University
†Courant Institute of Mathematical Sciences, New York University
ap6599@nyu.edu, jnw@cims.nyu.edu

May 10, 2022

We write $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ for the empirical distributions corresponding to the samples from P and Q, respectively. Our proposed estimator is $T_{\varepsilon,(n,n)}$, the entropic map between P_n and Q_n, which can be written explicitly as

$$T_{\varepsilon,(n,n)}(x) = \frac{\frac{1}{n}\sum_{i=1}^n Y_i e^{\frac{1}{\varepsilon}(g_{\varepsilon,(n,n)}(Y_i) - \frac{1}{2}\|x-Y_i\|^2)}}{\frac{1}{n}\sum_{i=1}^n e^{\frac{1}{\varepsilon}(g_{\varepsilon,(n,n)}(Y_i) - \frac{1}{2}\|x-Y_i\|^2)}}, \tag{14}$$

where $g_{\varepsilon,(n,n)}$ is the entropic potential corresponding to Q_n in the optimal entropic plan between P_n and Q_n, which can be obtained as part of the output of Sinkhorn's algorithm (see

Approach 2. use neural networks!

2-wasserstein approximation via restricted convex potentials. Taghvaei and Jalali, 2019.
Three-Player Wasserstein GAN via Amortised Duality. Nhan Dam et al., IJCAI 2019.
Optimal transport mapping via input convex neural networks. Makkuva et al., ICML 2020.
Wasserstein-2 generative networks. Korotin et al., ICLR 2020.
On amortizing convex conjugates for optimal transport. Amos, ICLR 2023.

Solving Kantorovich's dual with a neural net

2-wasserstein approximation via restricted convex potentials with application to improved training for GANs. Taghvaei and Jalali, 2019.

Parameterize the **dual potential** $f_\theta : \mathcal{X} \to \mathbb{R}$

Optimize the dual objective

$$\max_\theta \mathcal{V}(\theta) \quad \text{where} \quad \mathcal{V}(\theta) := -\mathop{\mathbb{E}}_{x\sim\alpha}[f_\theta(x)] - \mathop{\mathbb{E}}_{y\sim\beta}[f_\theta^\star(y)] = -\mathop{\mathbb{E}}_{x\sim\alpha}[f_\theta(x)] + \mathop{\mathbb{E}}_{y\sim\beta}[J_{f_\theta}(\breve{x}(y))]\,.$$

$$J_f(x;y) := f(x) - \langle x, y\rangle.$$

Assumes access to the **exact** conjugate is available

Differentiating and applying Danskin's envelope theorem gives:

$$\begin{aligned}\nabla_\theta \mathcal{V}(\theta) &= \nabla_\theta \left[-\mathop{\mathbb{E}}_{x\sim\alpha}[f_\theta(x)] + \mathop{\mathbb{E}}_{y\sim\beta}[J_{f_\theta}(\breve{x}(y))] \right] \\ &= -\mathop{\mathbb{E}}_{x\sim\alpha}[\nabla_\theta f_\theta(x)] + \mathop{\mathbb{E}}_{y\sim\beta}[\nabla_\theta f_\theta(\breve{x}(y))]\end{aligned}$$

Computing the conjugate $f^\star$

$$f^\star(y) := -\inf_{x \in \mathcal{X}} J_f(x; y) \quad \text{with objective} \quad J_f(x; y) := f(x) - \langle x, y \rangle.$$

Algorithm 2 CONJUGATE(f, y, x_{init})

$x \leftarrow x_{\text{init}}$
while unconverged **do**
  Update x with $\nabla_x J_f(x; y)$
end while
return optimal $\breve{x}(y) = x$

Numerically solving (e.g., with L-BFGS, SGD, or Adam)
2-wasserstein approximation via restricted convex potentials. Taghvaei and Jalali, 2019.

Amortization: parameterize and learn to predict the argmin (i.e., $\hat{\mathrm{x}}_\theta \approx x^\star$)

Three-Player Wasserstein GAN via Amortised Duality. Nhan Dam et al., IJCAI 2019.
Optimal transport mapping via input convex neural networks. Makkuva et al., ICML 2020.
Wasserstein-2 generative networks. Korotin et al., ICLR 2020.

Both: combine amortization with a numerical solve for fine-tuning
On amortizing convex conjugates for optimal transport. Amos, ICLR 2023.

Entropic estimation vs. neural networks

$$T_{\varepsilon,(n,n)}(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i e^{\frac{1}{\varepsilon}(g_{\varepsilon,(n,n)}(Y_i) - \frac{1}{2}\|x - Y_i\|^2)}}{\frac{1}{n}\sum_{i=1}^{n} e^{\frac{1}{\varepsilon}(g_{\varepsilon,(n,n)}(Y_i) - \frac{1}{2}\|x - Y_i\|^2)}},$$

Has convergence guarantees

Easy when Sinkhorn is tractable (~1k-10k samples)

Not done in high-dimensions

Evaluating T: sum over discrete solutions

Entropy may cause $T_{\#}\alpha \neq \beta$

$$\mathrm{T}(\mathrm{x}) = \nabla_{\mathrm{x}} f_\theta(x)$$

No convergence guarantees

Scales to 1M+ samples from the measures

SOTA benchmark results (in up to 10k dimensions)

Evaluating T: derivative of a neural network

No entropy, but $T_{\#}\alpha = \beta$ may still not be perfect

Learning flows via continuous OT

On amortizing convex conjugates for optimal transport. Amos, ICLR 2023.

Challenges for learning flows (with potentials or otherwise)
1. The model needs to be invertible
2. The likelihood of the base density is required

$$p_Y(y) = p_X(f^{-1}(y)) \left| \frac{\partial f^{-1}(y)}{\partial y} \right|$$

Optimizing the potential-based flow for the **Kantorovich dual** can help with both of these!
1. Often parameterize the model as a non-convex MLP, invertibility no longer matters
2. Only requires samples from the densities

$$\max_\theta \mathcal{V}(\theta) \quad \text{where} \quad \mathcal{V}(\theta) := - \mathop{\mathbb{E}}_{x\sim\alpha}[f_\theta(x)] - \mathop{\mathbb{E}}_{y\sim\beta}[f_\theta^\star(y)] = - \mathop{\mathbb{E}}_{x\sim\alpha}[f_\theta(x)] + \mathop{\mathbb{E}}_{y\sim\beta}[J_{f_\theta}(\breve{x}(y))]\,.$$

$$J_f(x; y) := f(x) - \langle x, y\rangle.$$

Beyond Euclidean Wasserstein-2 OT

Unpublished/active areas of research I'm thinking about:

1. Continuous Riemannian OT
2. Continuous Gromov-Wasserstein

Continuous Riemannian OT

Existing work getting close, but not computing OT

Riemannian Convex Potential Maps

Samuel Cohen [*1] **Brandon Amos** [*2] **Yaron Lipman** [2 3]

Abstract

Modeling distributions on Riemannian manifolds is a crucial component in understanding non-Euclidean data that arises, *e.g.*, in physics and geology. The budding approaches in this space are limited by representational and computational tradeoffs. We propose and study a class of flows that uses convex potentials from Riemannian optimal transport. These are universal and can model distributions on any compact Riemannian manifold without requiring domain knowledge of the manifold to be integrated into the architecture. We demonstrate that these flows can model standard distributions on spheres, and tori, on synthetic and geological data. Our source code is freely available online at github.com/facebookresearch/rcpm.

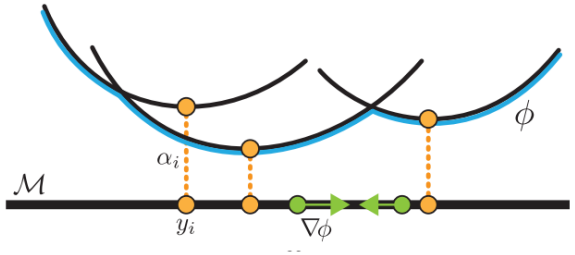
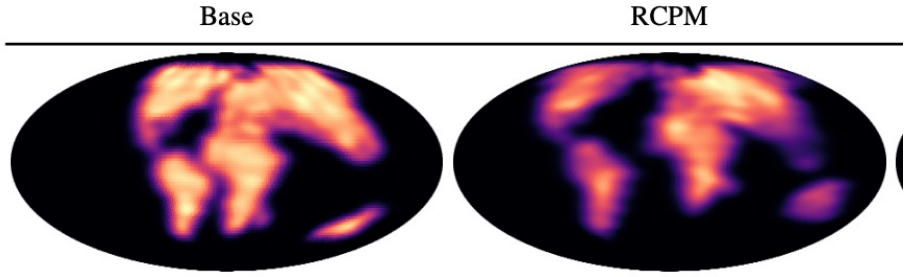


Figure 1. Illustration of a discrete c-concave function ϕ (blue) over a base manifold $\mathcal{M}$ (bold line). These consist of discrete components $\{\alpha_i, y_i\}$ and have a Riemannian gradient $\nabla\phi \in T_x\mathcal{M}$.

need to squeeze mass in zero volume subspaces. Moreover, knowledge of the space geometry can improve the learning



Br

Riemannian Flow Matching on General Geometries

Ricky T. Q. Chen [1] **Yaron Lipman** [1 2]

Abstract

We propose Riemannian Flow Matching (RFM), a simple yet powerful framework for training continuous normalizing flows on manifolds. Existing methods for generative modeling on manifolds either require expensive simulation, inherently cannot scale to high dimensions, or use approximations to limiting quantities that result in biased objectives. Riemannian Flow Matching bypasses these inconveniences and exhibits multiple benefits over prior approaches: It is completely simulation-free on simple geometries, it does not require divergence computation, and its target vector field is computed in closed form even on general geometries. The key ingredient behind RFM is the construction of a simple kernel function for defining per-sample vector fields, which subsumes existing Euclidean cases. Extending to general geometries, we rely on the use of spectral decompositions to efficiently compute kernel functions. Our method achieves state-of-the-art performance on real-world non-Euclidean datasets, and we showcase, for the first time, tractable training on general geometries, including on triangular meshes and maze-like manifolds with boundaries.

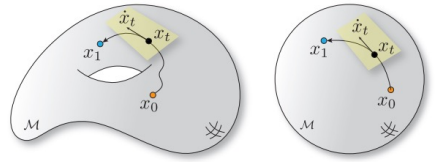
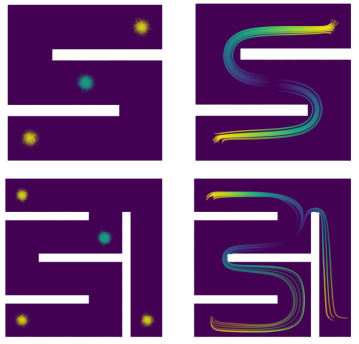
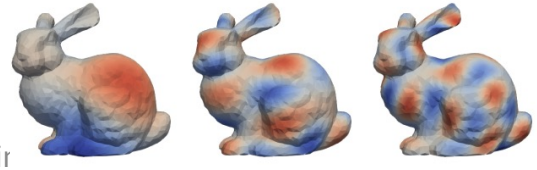


Figure 1: Riemannian Conditional Flow Matching (RCFM) regresses onto the vector field of flows x_t connecting a source $x_0 \sim p$ and and a target $x_1 \sim q$. (*Left*) On general geometries, x_t is obtained through solving an ODE. (*Right*) On simple geometries (*e.g.*, hypersphere), RCFM can set x_t as a geodesic path and is completely simulation-free.



(a) Distributions (b) Model trajectories

Figure 5: (a) Source (*green*) and target (*yellow*) distributions on a manifold with non-trivial boundaries. (b) Samples from a CNF model trained through Riemannian Flow Matching with the Biharmonic distance.



Contir

Riemannian Metric Learning via Optimal Transport

Christopher Scarvelis
MIT CSAIL
scarv@mit.edu

Justin Solomon
MIT CSAIL
jsolomon@mit.edu

Abstract

We introduce an optimal transport-based model for learning a metric tensor from cross-sectional samples of evolving probability measures on a common Riemannian manifold. We neurally parametrize the metric as a spatially-varying matrix field and efficiently optimize our model's objective using a simple alternating scheme. Using this learned metric, we can nonlinearly interpolate between probability measures and compute geodesics on the manifold. We show that metrics learned using our method improve the quality of trajectory inference on scRNA and bird migration data at the cost of little additional cross-sectional data.

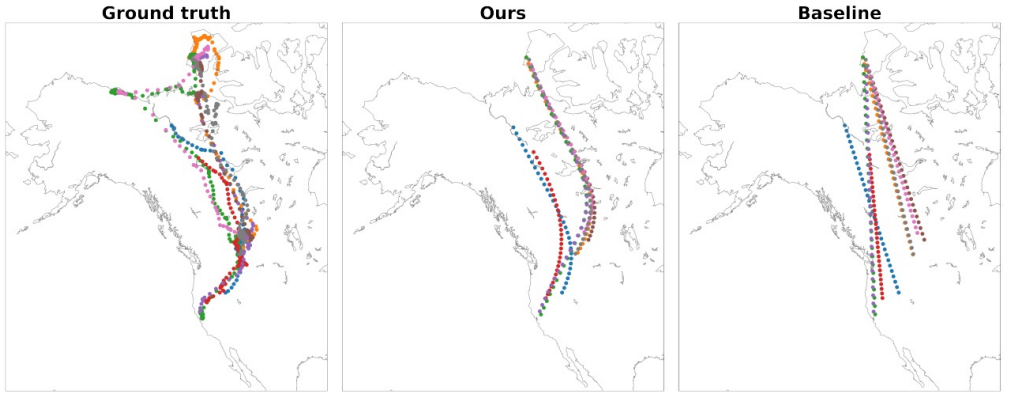


Figure 4: By using a metric $\hat{A}(x)$ learned from time-stamped bird sightings, we obtain inferred trajectories (center) that capture the curved structure of the ground truth migratory paths (left). Our method results in a 26.9% reduction in mean DTW distance between the inferred and ground truth trajectories relative to the Euclidean baseline (right).

$$\inf_{A:\mathbb{R}^D \to S^D_{++}} \sup_{\substack{\phi^k:\mathcal{M}\to\mathbb{R} \\ \|\nabla\phi^k(x)\|_{A^{-1}(x)} \leq 1}} \frac{1}{K}\sum_{k=1}^{K}\left(\int_{\mathcal{M}}\phi^k(x)d\rho_0^k(x) - \int_{\mathcal{M}}\phi^k(x)d\rho_1^k(x)\right) + \lambda R(A)\,. \tag{4}$$

Gromov-Wasserstein

Traditionally solved in entropic/discrete settings

Inputs: {(similarity/kernel matrix, histogram)}

$(d, \mu) \quad \mu = \sum_i \mu_i \delta_{x_i} \quad d_{i,i'} = d(x_i, x_{i'})$

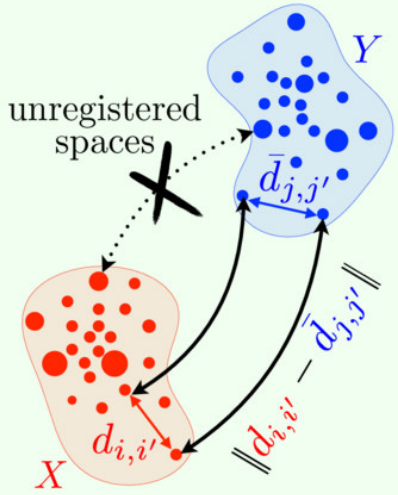
$(\bar{d}, \nu) \quad \nu = \sum_j \nu_j \delta_{y_j} \quad \bar{d}_{j,j'} = \bar{d}(y_j, y_{j'})$

Def. *Gromov-Wasserstein distance:*

$$\mathrm{GW}_p^p(d, \mu, \bar{d}, \nu) \overset{\text{def.}}{=} \min_{T \in C_{\mu,\nu}} \mathcal{E}^p_{d,\bar{d}}(T)$$

$$\mathcal{E}^p_{d,\bar{d}}(T) \overset{\text{def.}}{=} \sum_{i,i',j,j'} |d_{i,i'} - \bar{d}_{j,j'}|^p \, T_{i,j} T_{i',j'}$$

[Memoli 2011]



(Source: Gabriel Peyré, Justin Solomon, Marco Cuturi)

Entropic Metric Alignment for Correspondence Problems

Justin Solomon* MIT
Gabriel Peyré CNRS & Univ. Paris-Dauphine
Vladimir G. Kim Adobe Research
Suvrit Sra MIT

Abstract

Many shape and image processing tools rely on computation of correspondences between geometric domains. Efficient methods that stably extract "soft" matches in the presence of diverse geometric structures have proven to be valuable for shape retrieval and transfer of labels or semantic information. With these applications in mind, we present an algorithm for probabilistic correspondence that optimizes an entropy-regularized Gromov-Wasserstein (GW) objective. Built upon recent developments in numerical optimal transportation, our algorithm is compact, provably convergent, and applicable to any geometric domain expressible as a metric measure matrix. We provide comprehensive experiments illustrating the convergence and applicability of our algorithm to a variety of graphics tasks. Furthermore, we expand entropic GW correspondence to a framework for other matching problems, incorporating partial distance matrices, user guidance, shape exploration, symmetry detection, and joint analysis of more than two domains. These applications expand the scope of entropic GW correspondence to major shape analysis problems and are stable to distortion and noise.

Keywords: Gromov-Wasserstein, matching, entropy

Concepts: •**Computing methodologies** → **Shape analysis;**

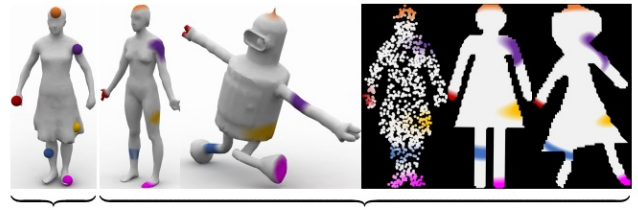


Figure 1: *Entropic GW can find correspondences between a source surface (left) and a surface with similar structure, a surface with shared semantic structure, a noisy 3D point cloud, an icon, and a hand drawing. Each fuzzy map was computed using the same code.*

are violated these algorithms suffer from having to patch together local elastic terms into a single global map.

In this paper, we propose a new correspondence algorithm that minimizes distortion of long- and short-range distances alike. We study an entropically-regularized version of the *Gromov-Wasserstein* (GW) mapping objective function from [Mémoli 2011] measuring the distortion of geodesic distances. The optimizer is a probabilistic matching expressed as a "fuzzy" correspondence matrix in the style of [Kim et al. 2012; Solomon et al. 2012]; we control sharpness of the correspondence via the weight of an entropic regularizer.

Gromov-Wasserstein for RL

Another application only using discrete GW, even though the spaces are truly continuous

CROSS-DOMAIN IMITATION LEARNING VIA OPTIMAL TRANSPORT

Arnaud Fickinger[13*] **Samuel Cohen**[23] **Stuart Russell**[1] **Brandon Amos**[3]
[1]Berkeley AI Research [2]University College London [3]Facebook AI

ABSTRACT

Cross-domain imitation learning studies how to leverage expert demonstrations of one agent to train an imitation agent with a different embodiment or morphology. Comparing trajectories and stationary distributions between the expert and imitation agents is challenging because they live on different systems that may not even have the same dimensionality. We propose *Gromov-Wasserstein Imitation Learning (GWIL)*, a method for cross-domain imitation that uses the Gromov-Wasserstein distance to align and compare states between the different spaces of the agents. Our theory formally characterizes the scenarios where GWIL preserves optimality, revealing its possibilities and limitations. We demonstrate the effectiveness of GWIL in non-trivial continuous control domains ranging from simple rigid transformation of the expert domain to arbitrary transformation of the state-action space. [1]

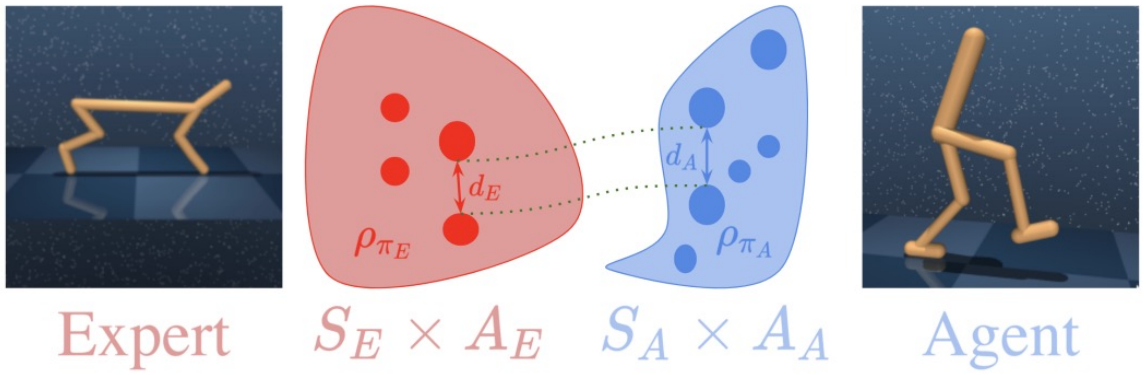


Figure 1: The Gromov-Wasserstein distance enables us to compare the stationary state-action distributions of two agents with different dynamics and state-action spaces. We use it as a pseudo-reward for cross-domain imitation learning.

Gromov-Wasserstein for single-cell multi-omics

Another application only using discrete GW, could have continuous extensions

Gromov-Wasserstein optimal transport to align single-cell multi-omics data

Pinar Demetci*[1,2], Rebecca Santorella*[3], Björn Sandstede[3], William Stafford Noble[4,5], and Ritambhara Singh[1,2]

[1]*Department of Computer Science, Brown University*
[2]*Center for Computational Molecular Biology, Brown University*
[3]*Division of Applied Mathematics, Brown University*
[4]*Department of Genome Sciences, University of Washington*
[5]*Paul G. Allen School of Computer Science and Engineering, University of Washington*
[*]*Equal Contribution*

Abstract

Data integration of single-cell measurements is critical for our understanding of cell development and disease, but the lack of correspondence between different types of single-cell measurements makes such efforts challenging. Several unsupervised algorithms are capable of aligning heterogeneous types of single-cell measurements in a shared space, enabling the creation of mappings between single cells in different data modalities. We present Single-Cell alignment using Optimal Transport (SCOT), an unsupervised learning algorithm that uses Gromov Wasserstein-based optimal transport to align single-cell multi-omics datasets. SCOT calculates a probabilistic coupling matrix that matches cells across two datasets. The optimization uses k-nearest neighbor graphs, thus preserving the local geometry of the data. We use the resulting coupling matrix to project one single-cell dataset onto another via a barycentric projection. We compare the alignment performance of SCOT with state-of-the-art algorithms on three simulated and two real datasets. Our results demonstrate that SCOT yields results that are comparable in quality to those of competing methods, but SCOT is significantly faster and requires tuning fewer hyperparameters. The code is available at `https://github.com/rsinghlab/SCOT`

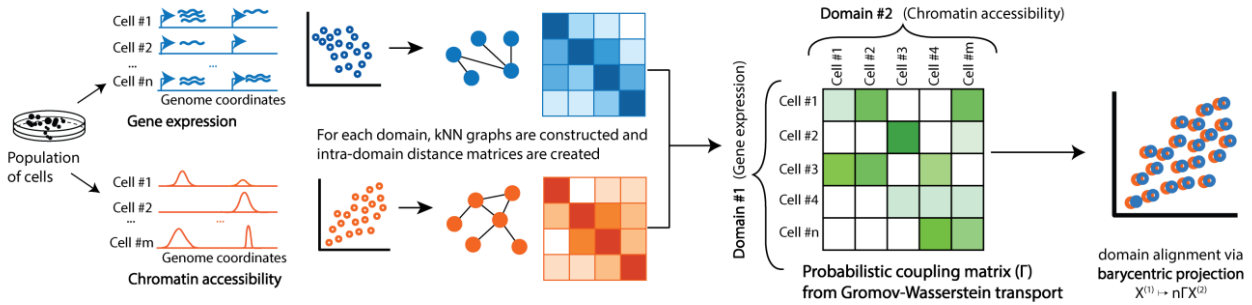


Figure 1: **Schematic of application of SCOT to single-cell multi-omics data alignment.** A population of cells is aliquoted for different single-cell sequencing assays in order to capture complementary aspects (e.g. gene expression and chromatin accessibility) of the molecular dynamics in single cells. Data obtained from these assays may exhibit different observed manifolds but share a common latent manifold. SCOT constructs k-NN graphs, where vertices represent cells, and Euclidean distances between them weigh the edges between the k-nearest neighbors. The SCOT algorithm finds a probabilistic coupling between the samples of each domain that will minimize the distance between the two intra-domain graph distance matrices. Barycentric projection uses this coupling matrix to project one domain onto another.

Gromov-Wasserstein Alignment of Word Embedding Spaces

Discretizes a continuous feature space, solves discrete GW

David Alvarez-Melis
CSAIL, MIT
dalvmel@mit.edu

Tommi S. Jaakkola
CSAIL, MIT
tommi@mit.edu

Abstract

Cross-lingual or cross-domain correspondences play key roles in tasks ranging from machine translation to transfer learning. Recently, purely unsupervised methods operating on monolingual embeddings have become effective alignment tools. Current state-of-the-art methods, however, involve multiple steps, including heuristic post-hoc refinement strategies. In this paper, we cast the correspondence problem directly as an optimal transport (OT) problem, building on the idea that word embeddings arise from metric recovery algorithms. Indeed, we exploit the *Gromov-Wasserstein* distance that measures how similarities between pairs of words relate across languages. We show that our OT objective can be estimated efficiently, requires little or no tuning, and results in performance comparable with the state-of-the-art in various unsupervised word translation tasks.

Algorithm 1 Gromov-Wasserstein Computation for Word Embedding Alignment

Input: Source and target embeddings $\mathbf{X}$, $\mathbf{Y}$. Regularization λ. Probability vectors $\mathbf{p}, \mathbf{q}$.
// Compute intra-language similarities
$\mathbf{C}_s \leftarrow \cos(\mathbf{X}, \mathbf{X}), \quad \mathbf{C}_t \leftarrow \cos(\mathbf{Y}, \mathbf{Y})$
$\mathbf{C}_{st} \leftarrow \mathbf{C}_s^2 \mathbf{p} \mathbb{1}_m^\top + \mathbb{1}_n \mathbf{q} (\mathbf{C}_t^2)^\top$
while not converged **do**
  // Compute pseudo-cost matrix (Eq. (9))
  $\hat{\mathbf{C}}_\Gamma \leftarrow \mathbf{C}_{st} - 2\mathbf{C}_s \Gamma \mathbf{C}_t^\top$
  // Sinkhorn iterations (Eq. (7))
  $\mathbf{a} \leftarrow \mathbb{1}, \quad \mathbf{K} \leftarrow \exp\{-\hat{\mathbf{C}}_\Gamma / \lambda\}$
  while not converged **do**
    $\mathbf{a} \leftarrow \mathbf{p} \oslash \mathbf{K}\mathbf{b}, \ \mathbf{b} \leftarrow \mathbf{q} \oslash \mathbf{K}^\top \mathbf{a}$
  end while
  $\Gamma \leftarrow \operatorname{diag}(\mathbf{a})\, \mathbf{K} \operatorname{diag}(\mathbf{b})$
end while
// Optional step: Learn explicit projection
$\mathbf{U}, \Sigma, \mathbf{V}^\top \leftarrow \text{SVD}(\mathbf{X}\Gamma\mathbf{Y}^\top)$
$\mathbf{P} = \mathbf{U}\mathbf{V}^\top$
return $\Gamma, \mathbf{P}$

Learning Generative Models across Incomparable Spaces

Discretizes continuous spaces, solves discrete GW

Charlotte Bunne [1] **David Alvarez-Melis** [2] **Andreas Krause** [1] **Stefanie Jegelka** [2]

Abstract

Generative Adversarial Networks have shown remarkable success in learning a distribution that faithfully recovers a reference distribution *in its entirety*. However, in some cases, we may want to only learn some aspects (e.g., cluster or manifold structure), while modifying others (e.g., style, orientation or dimension). In this work, we propose an approach to learn generative models across such *incomparable* spaces, and demonstrate how to steer the learned distribution towards target properties. A key component of our model is the *Gromov-Wasserstein* distance, a notion of discrepancy that compares distributions *relationally* rather than absolutely. While this framework subsumes current generative models in identically reproducing distributions, its inherent flexibility allows application to tasks in manifold learning, relational learning and cross-domain learning.

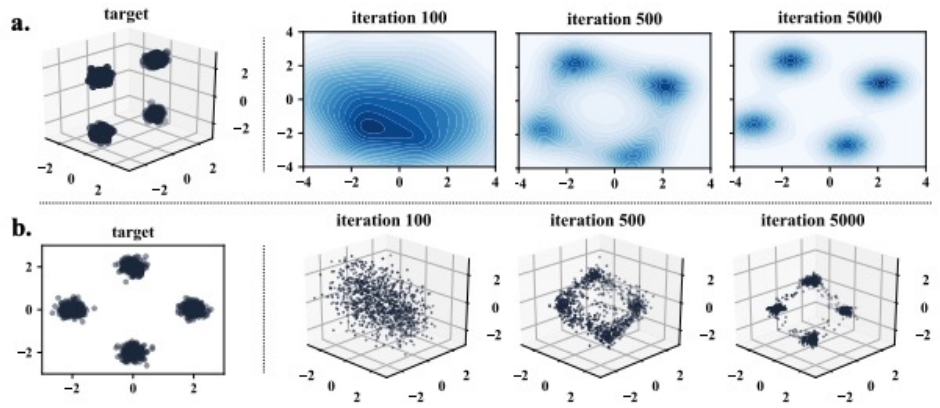


Figure 3. The GW GAN can be applied to generate samples of **a.** reduced and **b.** increased dimensionality compared to the target distribution. All plots show 1000 samples.

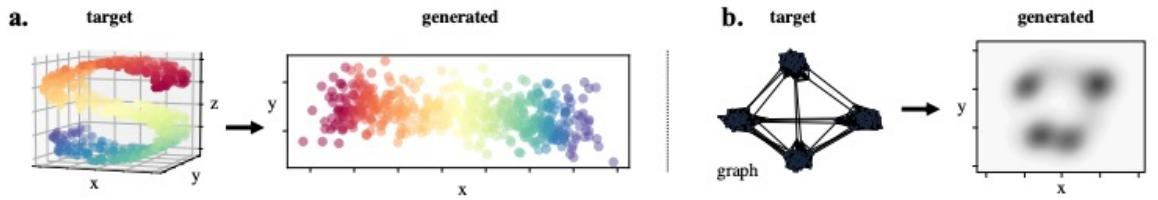


Figure 4. By learning from intra-space distances, the GW GAN learns the manifold structure of the data. **a.** The model can be applied to dimensionality reduction tasks and reproduce a three-dimensional S-curve in two-dimensions. Intra-space distances of the data samples are Floyd-Warshall shortest paths of the corresponding k-nearest neighbor graph. **b.** Similarly, it can map a graph into $\mathbb{R}^2$. The plots display 500 samples.

GW for cross-domain alignment

Discretizes a continuous feature space, solves discrete GW

Graph Optimal Transport for Cross-Domain Alignment

Liqun Chen [1] Zhe Gan [2] Yu Cheng [2] Linjie Li [2] Lawrence Carin [1] Jingjing Liu [2]

Abstract

Cross-domain alignment between two sets of entities (*e.g.*, objects in an image, words in a sentence) is fundamental to both computer vision and natural language processing. Existing methods mainly focus on designing advanced attention mechanisms to simulate soft alignment, with no training signals to *explicitly* encourage alignment. The learned attention matrices are also dense and lacks interpretability. We propose Graph Optimal Transport (GOT), a principled framework that germinates from recent advances in Optimal Transport (OT). In GOT, cross-domain alignment is formulated as a graph matching problem, by representing entities into a dynamically-constructed graph. Two types of OT distances are considered: (*i*) Wasserstein distance (WD) for node (entity) matching; and (*ii*) Gromov-Wasserstein distance (GWD) for edge (structure) matching. Both WD and GWD can be incorporated into existing neural network models, effectively acting as a drop-in regularizer. The inferred transport plan also yields *sparse* and *self-normalized* alignment, enhancing the interpretability of the learned model. Experiments show consistent outperformance of GOT over baselines across a wide range of tasks, including image-text retrieval, visual question answering, image captioning, machine translation, and text summarization.

tol et al., 2015), and machine translation (Bahdanau et al., 2015; Vaswani et al., 2017). Considering VQA as an example, in order to understand the contexts in the image and the question, a model needs to interpret the latent alignment between regions in the input image and words in the question. Specifically, a good model should: (*i*) identify entities of interest in both the image (*e.g.*, objects/regions) and the question (*e.g.*, words/phrases), (*ii*) quantify both intra-domain (within the image or sentence) and cross-domain relations between these entities, and then (*iii*) design good metrics for measuring the quality of cross-domain alignment drawn from these relations, in order to optimize towards better results.

CDA is particularly challenging as it constitutes a *weakly supervised learning* task. That is, only paired spaces of entity are given (*e.g.*, an image paired with a question), while the ground-truth relations between these entities are not provided (*e.g.*, no supervision signal for a "dog" region in an image aligning with the word "dog" in the question). State-of-the-art methods principally focus on designing advanced attention mechanisms to simulate soft alignment (Bahdanau et al., 2015; Xu et al., 2015; Yang et al., 2016b;a; Vaswani et al., 2017). For example, Lee et al. (2018); Kim et al. (2018); Yu et al. (2019) have shown that learned co-attention can model dense interactions between entities and infer cross-domain latent alignments for vision-and-language tasks. Graph attention has also been applied to relational reasoning for image captioning (Yao et al., 2018) and VQA (Li et al., 2019a), such as graph attention network

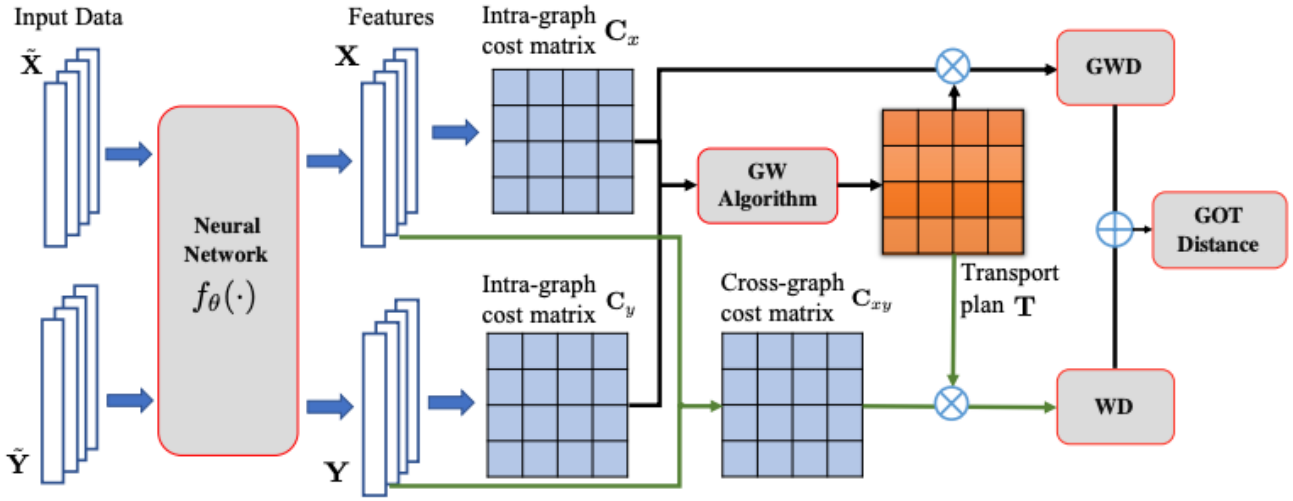


Figure 2. Schematic computation graph of the Graph Optimal Transport (GOT) distance used for cross-domain alignment. WD is short for Wasserstein Distance, and GWD is short for Gromov-Wasserstein Distance. See Sec. 2.1 and 2.4 for details.

Continuous Gromov-Wasserstein

Studied theoretically, but not many computational instantiations
Seems promising to use other continuous OT methods (entropic or neural) for subproblems here

Titouan VAYER

A contribution to Optimal Transport on incomparable spaces

Theorem 4.2.5. *Let $\mathcal{X}$ and $\mathcal{Y}$ be compact subset of respectively $\mathbb{R}^p$ and $\mathbb{R}^q$. Let $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$. Assume without loss of generality that $\mathbb{E}_{X\sim\mu}[X] = 0$ and $\mathbb{E}_{Y\sim\nu}[Y] = 0$. Then problems:*

$$\inf_{\pi\in\Pi(\mu,\nu)} \int (\|\mathbf{x}-\mathbf{x}'\|_2^2 - \|\mathbf{y}-\mathbf{y}'\|_2^2)^2 \mathrm{d}\pi(\mathbf{x},\mathbf{y})\mathrm{d}\pi(\mathbf{x}',\mathbf{y}') \qquad \text{(sqGW)}$$

and

$$\sup_{\pi\in\Pi(\mu,\nu)} \sup_{\mathbf{P}\in\mathbb{R}^{q\times p}} \int (\langle \mathbf{P}\mathbf{x},\mathbf{y}\rangle_q + \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2)\mathrm{d}\pi(\mathbf{x},\mathbf{y}) - \frac{1}{8}\|\mathbf{P}\|_{\mathcal{F}}^2 \qquad \text{(dual-sqGW)}$$

are equivalent.