Optimal Transport via Input Convex Neural Networks

Daran Xu & Garrett Mulcahy

November 1 2022

Introduction

Makkuva et al. (2020) propose a minimax formulation for the Kantorovich duality problem under Wasserstein-2 cost, eliminating the constraints in the Kantorovich duality by restricting search to the set of of convex functions. They also provide a numerical scheme to learn the Kantorovich potentials that involves learning two convex functions with the help of input convex neural networks (ICNNs).

Preliminaries

Let P and Q be two probability distribution on $\mathbb{R}^d$ with finite second order moments. Throughout this paper, we assume Q admits a density in $\mathbb{R}^d$. The Monge optimal transportation problem is to find a map transport probability mass under Q to P with the least amount of cost:

$$\min_{T:T_{\#}Q=P} \frac{1}{2}\mathbb{E}_{X\sim Q}\|X - T(X)\|^2$$

where $T_{\#}Q = P$ means T pushforwards Q to P, i.e., for any Borel subset B of $\mathbb{R}^d$, $Q(T^{-1}(B)) = P(B)$.

Kantorovich introduced a relaxation of the problem:

$$W_2^2(P,Q) = \inf_{\pi\in\Pi(P,Q)} \frac{1}{2}\mathbb{E}_{(X,Y)\sim\pi}\|X - Y\|^2,$$

where $\Pi(P,Q)$ stands for the set of all couplings whose marginal distributions are P and Q. The optimal value of it is the 2-Wasserstein distance $W_2(P,Q)$, and the coupling π reaching the infimum is called the optimal coupling.

Kantorovich duality is a dual formulation for this problem:

$$W_2^2(P,Q) = \sup_{f,g\in\Phi_c} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[g(Y)],$$

where Φ_c denotes the constrained space of functions, $\{(f,g) \in L^1(P)\times L^1(Q) : f(x) + g(y) \leq \frac{1}{2}||x-y||_2^2, \forall(x,y)\ dP\otimes dQ\ a.e.\}$

By writing $\frac{1}{2}||x-y||_2^2$ as $\frac{1}{2}||x||_2^2 + \frac{1}{2}||y||_2^2 - \langle x,y\rangle$, Φ_C is equivalent to $\{(f,g) \in L^1(P)\times L^1(Q) : \frac{1}{2}||x||_2^2 - f(x) + \frac{1}{2}||y||_2^2 - g(y) \geq \langle x,y\rangle, \forall(x,y)\ dP\otimes dQ\ a.e.\}$ Reparametrizing $\frac{1}{2}||\cdot||^2 - f(\cdot)$ and

$\frac{1}{2}||\cdot||^2 - g(\cdot)$ by ϕ, ψ, respectively, we yield

$$\begin{aligned} W_2^2(P,Q) &= \sup_{\phi,\psi\in\tilde{\Phi}_c} \mathbb{E}_P[\frac{1}{2}||X||_2^2 - \phi(X)] + \mathbb{E}_Q[\frac{1}{2}||Y||_2^2 - \psi(Y)] \\ &= \frac{1}{2}\mathbb{E}[|X|^2 + |Y|^2] - \inf_{\phi,\psi\in\tilde{\Phi}_c} \{\mathbb{E}_P[\phi(X)] + \mathbb{E}_Q[\psi(Y)]\} \\ &= C_{P,Q} - \inf_{\phi,\psi\in\tilde{\Phi}_c} \{\mathbb{E}_P[\phi(X)] + \mathbb{E}_Q[\psi(Y)]\} \end{aligned}$$

where $C_{P,Q} = \frac{1}{2}\mathbb{E}[|X|^2 + |Y|^2]$ is independent of (ϕ,ψ), and $\tilde{\Phi}_c = \{(\phi,\psi) \in L^1(P) \times L^1(Q) : \phi(x) + \psi(y) \geq \langle x, y\rangle, \forall (x,y)\ dP \otimes dQ\ a.e.\}$

To keep notation consistent with Makkuva et al. (2020), we denote (f, g) for (ϕ,ψ):

$$W_2^2(P,Q) = C_{P,Q} - \inf_{f,g\in\tilde{\Phi}_c} \{\mathbb{E}_P[f(X)] + \mathbb{E}_Q[g(Y)]\}$$

where $\tilde{\Phi}_c = \{(f,g) \in L^1(P) \times L^1(Q) : f(x) + g(y) \geq \langle x, y\rangle, \forall (x,y)\ dP \otimes dQ\ a.e.\}$

Thanks to the double convexification trick ((Villani (2003), Theorem 2.9)), the constrained optimization problem can be formulated as:

$$W_2^2(P,Q) = C_{P,Q} - \left(\inf_{f\in \mathrm{CVX}(P)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[f^*(Y)] \right) \tag{1}$$

where $f^*(y) = \sup_x\{\langle x, y\rangle - f(x)\}$ is the convex conjugate of convex function $f(\cdot)$. For a convex and lower semicontinuous function f, the convex conjugate of its convex conjugate is itself (i.e. $f^{**} = f$). Additionally, another useful property of the convex conjugate in optimal transport is as follows:

Theorem 1. *(Villani (2003), Proposition 2.4) Let f be a proper lower semi-continuous convex function on $\mathbb{R}^n$. Then, for all $x, y \in \mathbb{R}^n$,*

$$\langle x, y\rangle = f(x) + f^*(y) \Leftrightarrow y \in \partial f(x) \Leftrightarrow x \in \partial f^*(y)$$

We let $\partial f(x)$ and $\partial f^*(y)$ denote subdifferentials. Namely, we have

$$\partial f(x) := \{y \in \mathbb{R}^n : f(z) \geq f(x) + \langle y, z - x\rangle \text{ for all } z \in \mathbb{R}^n\}.$$

Importantly, if f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$.

The function pair (f, f^*) reaching the infimum in (1) is called Kantorovich potentials. Brenier's work provides a way to find the optimal solutions:

Theorem 2. *(Villani (2003),Theorem 2.12) If Q admits a density with respect to the Lebesgue measure on $\mathbb{R}^d$, then there is a unique optimal coupling π for the primal problem under the Kantorovich's formulation, $\pi = (\nabla f^* \times Id)_\# Q$, and the convex pair (f, f^*) reaches the minimum in the dual problem under the Kantorovich's formulation. Moreover, ∇f^* is the unique solution to Monge transportation problem from Q to P.*

Now we are ready to propose the main results of Makkuva et al. (2020), which transforms the problem in (1) into a minimax problem.

Main Results

Consistency

As Theorem 2 demonstrates, the problem of computing W_2^2 can be solved by considering the space of convex functions. However, it also requires computing the convex conjugate. Makkuva et al. (2020) overcome this limitation by instead introducing another convex function and transforming the computation of W_2^2 into a minimax problem. The motivation for such a transformation is that ICNNs are able to approximate any convex function over a compact domain with respect to the supremum norm (Chen et al. (2018)). Thus, this formulation provides the theoretical foundation for using ICNNs to learn W_2^2 distance and optimal transport maps. To be precise, Makkuva et al. (2020) prove:

Theorem 3. *(Makkuva et al. (2020), Theorem 3.3) Whenever Q admits a density in $\mathbb{R}^d$, we have*

$$W_2^2(P,Q) = \sup_{\substack{f\in \mathrm{CVX}(P),\\ f^*\in L^1(Q)}} \inf_{g\in \mathrm{CVX}(Q)} \mathcal{V}(f,g) + C_{P,Q} \tag{2}$$

where $V_{P,Q}(f,g)$ is a functional of f,g defined as

$$V_{P,Q}(f,g) = -\mathbb{E}_P[f(X)] - \mathbb{E}_Q[\langle Y, \nabla g(Y)\rangle - f(\nabla g(Y))]. \tag{3}$$

In addition, there exists an optimal pair (f_0, g_0) achieving the infimum and supremum respectively, where ∇g_0 is the optimal transport map from Q to P.

Proof. Let $g \in \mathrm{CVX}(Q)$ and $f \in \mathrm{CVX}(P)$ with $f^* \in L^1(Q)$, we will first establish that g and f^* are differentiable Q-a.e. As $\int g dQ < \infty$, we must have $Q(\{g = \infty\}) = 0$. As $\mathrm{Dom}\, g = \{g \neq \infty\}$, this gives that $Q(\mathrm{Dom}\, g) = 1$ and thus $Q(\overline{\mathrm{Dom}\, g}) = 1$. Since $\mathrm{Dom}\, g$ is a convex set, we have $m(\partial \,\mathrm{Dom}\, g) = 0$ (Lang (1986)). As Q is given by a density, this gives that $Q(\partial\, \mathrm{Dom}\, g) = 0$. These observations combine to give $Q(\mathrm{Int}\,\mathrm{Dom}\, g) = 1$. Since a convex function is differentiable on the interior of its domain (Chapter 2, Villani (2003)), we have that g is differentiable Q-a.e. As we have insisted that $f^* \in L^1(Q)$, it follows that f^* is also differentiable Q-a.e.

As the union of two Q-null sets is still Q-null, we have for Q-a.e. $y \in \mathbb{R}^d$ that $\nabla g(y)$ exists and f^* is differentiable (so $\partial f^*(y) = \{\nabla f^*(y)\}$). Hence, by (1) we have for Q-a.e. $y \in \mathbb{R}^d$ that

$$\langle \nabla g(y), y\rangle - f(\nabla g(y)) \leq \langle \nabla f^*(y), y\rangle - f(\nabla f^*(y)) = f^*(y).$$

Taking expectation with respect to Q then gives

$$\mathbb{E}_Q[\langle \nabla g(Y), Y\rangle - f(\nabla g(Y))] \leq \mathbb{E}_Q[f^*(Y)]$$

with equality if and only if $g = f^*$. Hence, we have that

$$\begin{aligned}\inf_{g\in \mathrm{CVX}(Q)} V_{P,Q}(f,g) &= \inf_{g\in \mathrm{CVX}(Q)} -\mathbb{E}_P[f(X)] - \mathbb{E}_Q[\langle Y, \nabla g(Y)\rangle - f(\nabla g(Y))] \\ &= -\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f^*(Y)].\end{aligned}$$

This gives that

$$\begin{aligned}\sup_{\substack{f\in \mathrm{CVX}(P),\\ f^*\in L^1(Q)}} \inf_{g\in \mathrm{CVX}(Q)} \mathcal{V}(f,g) + C_{P,Q} &= \sup_{f\in \mathrm{CVX}(P), f^*\in L^1(Q)} \frac{1}{2}\mathbb{E}_P[X^2] + \frac{1}{2}\mathbb{E}_Q[Y^2] - \mathbb{E}_P[f(X)] - \mathbb{E}_Q[f^*(Y)] \\ &= W_2^2(P,Q),\end{aligned}$$

where the last equality follows from (1). In particular, we see that the optimal pair (f, f^*) from Theorem 2 will be the optimal pair achieving the infimum and supremum respectively, and ∇f^* is indeed the optimal transport map from Q to P. □

Stability

For any pair (f,g), we define the minimization gap ϵ_1 and maximization gap ϵ_2 as follows:

$$\epsilon_1(f,g) = \mathcal{V}(f,g) - \inf_{\tilde{g}\in \mathrm{CVX}(Q)} \mathcal{V}(f,\tilde{g}) \tag{4}$$

$$\epsilon_2(f) = \sup_{\tilde{f}\in \mathrm{CVX}(P)} \left\{ \inf_{\tilde{g}\in \mathrm{CVX}(Q)} \mathcal{V}(\tilde{f},\tilde{g}) \right\} - \inf_{\tilde{g}\in \mathrm{CVX}(Q)} \mathcal{V}(f,\tilde{g}) \tag{5}$$

Theorem 4. *(Makkuva et al. (2020), Theorem 3.6) In the optimization problem 2, suppose Q admits a density and let $\nabla g_0(\cdot)$ denote the optimal transport map from Q to P. Then for any pair (f,g) such that f is α-strongly convex, we have*

$$||\nabla g - \nabla g_0||^2_{L^2(Q)} \le \frac{2}{\alpha}\left(\epsilon_1(f,g) + \epsilon_2(f)\right) \tag{6}$$

where ϵ_1 and ϵ_2 are defined in 4 and 5, respectively, and $||\cdot||_{L^2(Q)}$ denotes the L^2-norm with respect to measure Q.

When we solve the min-max problem in (2) up to a small error, ϵ_1 can be interpreted as an error of the minimum problem in the min-max problem (2), while ϵ_2 can be interpreted as an error of the maximum problem in it. Theorem 4 provides a bound of the error between ∇g and the optimal transport map ∇g_0, as a function of ϵ_1 and ϵ_2.

Proof. The proof of (6) follows from the bounds proved below

$$||\nabla g - \nabla f^*||^2_{L^2(Q)} \le \frac{2}{\alpha}\epsilon_1(f,g) \tag{7}$$

$$||\nabla f^* - \nabla g_0||^2_{L^2(Q)} \le \frac{2}{\alpha}\epsilon_2(f) \tag{8}$$

With the triangle inequality,

$$||\nabla g - \nabla g_0||^2_{L^2(Q)} \le ||\nabla g - \nabla f^*||^2_{L^2(Q)} + ||\nabla f^* - \nabla g_0||^2_{L^2(Q)} \le \frac{2}{\alpha}\left(\epsilon_1(f,g) + \epsilon_2(f)\right)$$

Before we prove those two bounds, we need to derive a useful lemma. Since f is α-strongly convex, f^* is $\frac{1}{\alpha}$ smooth (by Kakade and Shalev-Shwartz (2009) Theorem 6), which means:

$$f^*(z) \le f^*(y) + \langle \nabla f^*(y), z-y\rangle + \frac{1}{2\alpha}||z-y||^2, \forall y,z \in \mathbb{R}^d.$$

Let $h_y(z)$ denote $f^*(y) + \langle \nabla f^*(y), z-y\rangle + \frac{1}{2\alpha}||z-y||^2$, the right-hand side of the inequality. From $f^*(z) \le h_y(z)$, it follows that the convex conjugate

$$f(x) = f^{**}(x) = \sup_z(\langle z,x\rangle - f^*(z)) \ge \sup_z(\langle z,x\rangle - h_y(z)) = h_y^*(x) \tag{9}$$

To obtain $h_y^*(x)$, we use the definition of the convex conjugate:

$$\begin{aligned} h_y^*(x) &= \sup_z(\langle z,x\rangle - h_y(z)) \\ &= \sup_z(\langle z,x\rangle - f^*(y) - \langle \nabla f^*(y), z-y\rangle - \frac{1}{2\alpha}||z-y||^2) \\ &= \sup_z(\langle z,x\rangle - \langle \nabla f^*(y), z\rangle - \frac{1}{2\alpha}||z-y||^2) - f^*(y) - \langle \nabla f^*(y), -y\rangle \end{aligned}$$

To find the supremum, we take the gradient:

$$
\begin{aligned}
0 &= \nabla_z(\langle z, x\rangle - \langle \nabla f^*(y), z\rangle - \frac{1}{2\alpha}\langle (z-y),(z-y)\rangle) \\
&= x - \nabla f^*(y) - \frac{1}{\alpha}(z-y)
\end{aligned}
$$

This gives us $z - y = \alpha x - \alpha \nabla f^*(y)$ or $z = y + \alpha x - \alpha \nabla f^*(y)$. We fill them back into $h_y^*(x)$:

$$
\begin{aligned}
h_y^*(x) &= \langle y + \alpha x - \alpha\nabla f^*(y), x\rangle - \langle \nabla f^*(y), y + \alpha x - \alpha \nabla f^*(y)\rangle \\
&\quad - \frac{1}{2\alpha}||\alpha x - \alpha \nabla f^*(y)||^2 - f^*(y) - \langle \nabla f^*(y), -y\rangle \\
&= \langle y, x\rangle + \alpha x^2 - \alpha\langle \nabla f^*(y), x\rangle - \langle \nabla f^*(y), y\rangle - \alpha\langle \nabla f^*(y), x\rangle + \alpha \nabla f^*(y)^2 \\
&\quad - \frac{\alpha}{2}||x - \nabla f^*(y)||^2 - f^*(y) - \langle \nabla f^*(y), -y\rangle \\
&= \langle y, x\rangle + \alpha||x - \nabla f^*(y)||^2 - \frac{\alpha}{2}||x - \nabla f^*(y)||^2 - f^*(y) \\
&= \langle y, x\rangle + \frac{\alpha}{2}||x - \nabla f^*(y)||^2 - f^*(y)
\end{aligned}
$$

With (9), we get:

$$
\begin{aligned}
f(x) &\geq \langle y, x\rangle + \frac{\alpha}{2}||x - \nabla f^*(y)||^2 - f^*(y) \\
f(x) + f^*(y) - \langle y, x\rangle &\geq \frac{\alpha}{2}||x - \nabla f^*(y)||^2 \qquad (10)
\end{aligned}
$$

Now we are ready to prove the two bounds claimed at the beginning.

$$
\begin{aligned}
\epsilon_1(f, g) &= \mathcal{V}(f, g) - \inf_{\tilde{g}\in \mathrm{CVX}(Q)} \mathcal{V}(f, \tilde{g}) \\
&= \mathcal{V}(f, g) - \mathcal{V}(f, f^*) \text{ (as proved in Theorem 3)} \\
&= -\mathbb{E}_P[f(X)] - \mathbb{E}_Q[\langle Y, \nabla g(Y)\rangle - f(\nabla g(Y))] + \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla f^*(Y)\rangle - f(\nabla f^*(Y))] \\
&= -\mathbb{E}_Q[\langle Y, \nabla g(Y)\rangle - f(\nabla g(Y))] + \mathbb{E}_Q[\langle Y, \nabla f^*(Y)\rangle - f(\nabla f^*(Y))] \\
&= -\mathbb{E}_Q[\langle Y, \nabla g(Y)\rangle - f(\nabla g(Y))] + \mathbb{E}_Q[f^*(Y)] \text{ (by Theorem 1)} \\
&= \mathbb{E}_Q[f(\nabla g(Y)) + f^*(Y) - \langle Y, \nabla g(Y)\rangle] \\
&\geq \frac{\alpha}{2}\mathbb{E}_Q[||\nabla g(Y) - \nabla f^*(Y)||^2] \text{ (by 10, with } x = \nabla g(y))
\end{aligned}
$$

Let (f_0, f_0^*) denotes the optimal pair achieving the infimum and supremum in problem 2. Then

$$
\begin{aligned}
\epsilon_2(f) &= \sup_{\tilde{f}\in \mathrm{CVX}(P)} \left\{ \inf_{\tilde{g}\in \mathrm{CVX}(Q)} \mathcal{V}(\tilde{f}, \tilde{g}) \right\} - \inf_{\tilde{g}\in \mathrm{CVX}(Q)} \mathcal{V}(f, \tilde{g}) \\
&= \mathcal{V}(f_0, f_0^*) - \inf_{\tilde{g}\in \mathrm{CVX}(Q)} \mathcal{V}(f, \tilde{g}) \\
&= \mathcal{V}(f_0, f_0^*) - \mathcal{V}(f, f^*) \text{ (by Theorem 3)} \\
&= -\mathbb{E}_P[f_0(X)] - \mathbb{E}_Q[\langle Y, \nabla f_0^*(Y)\rangle - f_0(\nabla f_0^*(Y))] + \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla f^*(Y)\rangle - f(\nabla f^*(Y))] \\
&= -\mathbb{E}_P[f_0(X)] - \mathbb{E}_Q[f_0^*(Y)] + \mathbb{E}_P[f(X)] + \mathbb{E}_Q[f^*(Y)] \text{ (by Theorem 1)} \\
&= -\mathbb{E}_Q[f_0(\nabla f_0^*(Y))] - \mathbb{E}_Q[f_0^*(Y)] + \mathbb{E}_Q[f(\nabla f_0^*(Y))] + \mathbb{E}_Q[f^*(Y)] \text{ } (\nabla f_0^*(\cdot) \text{ pushforwards from } Q \text{ to } P) \\
&= -\mathbb{E}_Q[\langle Y, \nabla f_0^*(Y)\rangle] + \mathbb{E}_Q[f(\nabla f_0^*(Y))] + \mathbb{E}_Q[f^*(Y)] \text{ (by Theorem 1)} \\
&= \mathbb{E}_Q[f(\nabla f_0^*(Y)) + f^*(Y) - \langle Y, \nabla f_0^*(Y)\rangle] \\
&\geq \frac{\alpha}{2}\mathbb{E}_Q[||Y - \nabla f_0^*(Y)||^2] \text{ (by 10, with } x = \nabla f_0^*(y))
\end{aligned}
$$

With these two bounds and the triangle inequality at the beginning, we finish the proof. □

Algorithm

Makkuva et al. (2020) parametrizes the convex function f, g using the same ICNN architecture. For θ_f, they enforce all weights W_l's to be non-negative, as maximization over g can be unbounded whenever f is non-convex. However, for θ_g, any function $g \in L^1(Q)$ reaching the infimum in the problem 2 equals f^* and is convex, so they relax the non-negative constraint on the weights and introduce a regularization term instead:

$$R(\theta_g) = \lambda \sum_{W_l \in \theta_g} || \max(W_l, 0)||_F^2$$

where $\lambda > 0$ is a regularization constant. Empirically, it is observed that this relaxation makes it converge faster.

The objective for optimization is an empirical counterpart of 2:

$$\max_{\theta_f : W_l \geq 0, \forall l \in [L-1]} \min_{\theta_g} J(\theta_f, \theta_g) + R(\theta_g) \tag{11}$$

where

$$J(\theta_f, \theta_g) = \frac{1}{M} \sum_{i=1}^{M} \{f(\nabla g(Y_i)) - \langle Y_i, \nabla(Y_i)\rangle - f(X_i)\}$$

The algorithm is summarized below:

Algorithm 1: Algorithm: the numerical procedure to solve the optimization (11)

Input: Source distribution Q, Target distribution P, Batch size M, Generator iterations K, Total iterations T

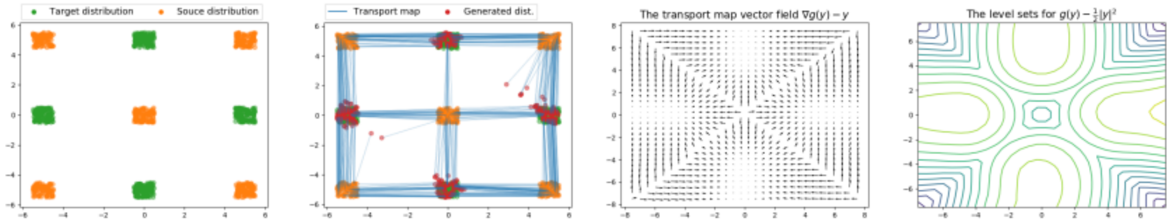
for *t=1,...,T* **do**
  Sample batch $\{X_i\}_{i=1}^M \sim P$
  for k = *1,...K* **do**
    Sample batch $\{Y_i\}_{i=1}^M \sim Q$
    Update θ_g to minimize (11) using Adam
  end
  Update θ_f to maximize (11) using Adam
  Projection: $w \leftarrow \max(w, 0)$, for all $w \in \{W^l\} \in \theta_f$
end
return: optimizer θ_f^*, θ_g^*, optimal transport $\nabla(g^{\theta_g^*})$

Experimental Results

To illustrate the strength of the algorithm proposed in the previous section, Makkuva et al. (2020) compare their method to other standard algorithms that learn optimal transport maps, namely Barycentric OT, W1-LP, and W2GAN. The there are two main benefits to the approach of Makkuva et al. (2020) when compared to these methods

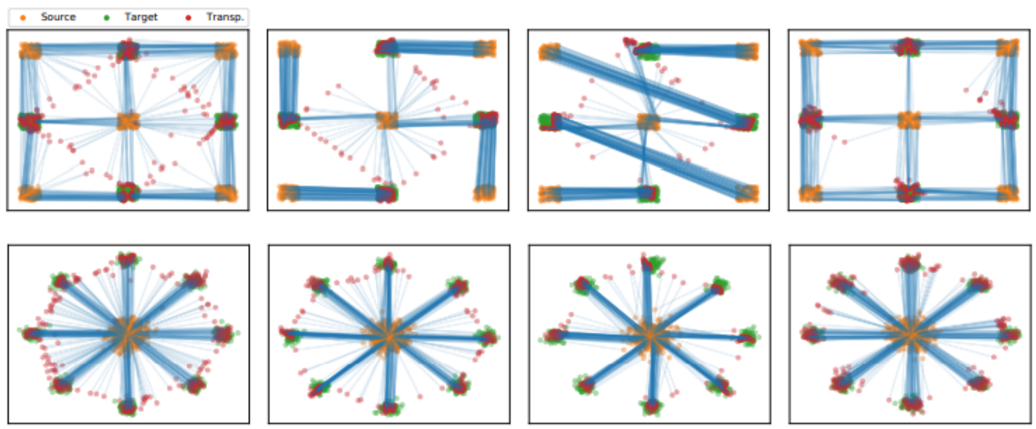
(1) Ability to learn discontinuous maps

(2) Robustness of learned maps to initialization



(a) Data samples (b) Our transport map (c) Displacement vector field (d) Level sets

Figure 1: From Makkuva et al. (2020), behavior of proposed algorithm on "checkerboard" data set



(a) Barycentric-OT (b) W1-LP (c) W2GAN (d) Our approach

Figure 2: From Makkuva et al. (2020), the transport maps learned under different approaches for "checkerboard" data set (top row) and mixture of eight Gaussians (bottom rows)

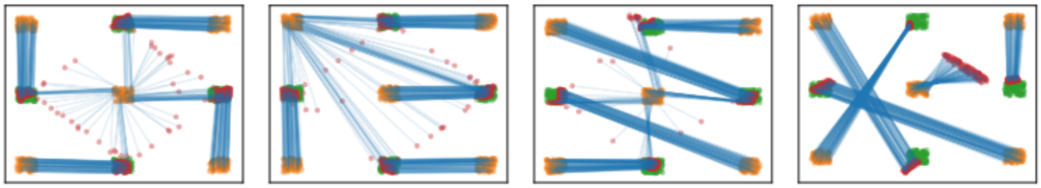
These feature are readily apparent in Figure 1, in which Makkuva et al. (2020) apply their algorithm to a "checkerboard" data set. We note that the support of both the source and target distributions of these data sets can be given by a disjoint union of sets with sharp boundaries. Panel (b) shows that the algorithm is indeed effective at learning a good transportation map, and the vertical tangents in panel (c) show that the learned map is discontinuous.

To show the desirability of these features and how they are not present in other approaches, Makkuva et al. (2020) perform several experiments. We will now summarize the experimental setup that led to these observations.

Discontinuities

We recall that the method of Makkuva et al. (2020) computes the optimal transport map by taking the gradient of a convex function. This differentiation steps allows for the possibility of obtaining discontinuous maps (for example, consider differentiating ReLU). This trait is desirable as optimal transport maps may often have to be discontinuous, such as if the target or source distribution is supported on disjoint separated sets. To see that this method learns discontinuous maps in practice, Makkuva et al. (2020) applied their algorithm and three others to two data sets: a "checkerboard data set" and a mixture of eight Gaussians. The learned maps are shown in Figure 2.

In Figure 2, the image of the source samples under the learned map are shown in red. In this case, it is clear that in both tasks the algorithm of Makkuva et al. (2020) had superior performance



(a) W1-LP: Trial 1 (b) W1-LP: Trial 2 (c) W2GAN: Trial 1 (d) W2GAN: Trial 2

Figure 3: From Makkuva et al. (2020), transport maps learned by different approaches under different initializations

to the other approaches. This is due precisely to the algorithm's ability to learn discontinuous maps, as an optimal transport map in both situations would have to be discontinuous due to the disjoint structure of the support of at least one of the target or source distributions. Moreover, the forced continuity in the other three methods is made apparent by the relative abundance of "trailing points" in panels (a)-(c) compared to panel (d).

Robustness to Initialization

Another benefit of the algorithm proposed by Makkuva et al. (2020) is that the optimal transport map learned is robust to the initialization of the ICNN. Other methods based on generative adversarial networks (GANs), such as W2GAN and W1-LP, only use the optimal transport metric as a distance. Thus, the transport maps learned from these methods are sensitive to intialization. This behavior is shown in Figure 3, where under different initializations both W1-LP and W2GAN learn very different mappings.

References

Chen, Y., Shi, Y., and Zhang, B. (2018). Optimal control via neural networks: A convex approach.

Kakade, S. M. and Shalev-Shwartz, S. (2009). On the duality of strong convexity and strong smoothness : Learning applications and matrix regularization.

Lang, R. (1986). A note on the measurability of convex sets. *Archiv der Mathematik*, 47(1):90–92.

Makkuva, A., Amirhossein, T., Lee, J., and Oh, S. (2020). Optimal transport mapping via input convex neural networks. In *the 37th International Conference on Machine Learning*. PMLR.

Villani, C. (2003). *Topics in optimal transportation*, volume 58. American Mathematical Soc.