Optimal Transport via ICNNs

Daran Xu and Garrett Mulcahy

November 8, 2022

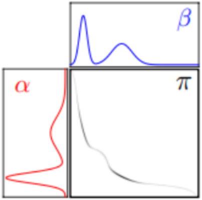
UNIVERSITY *of* WASHINGTON

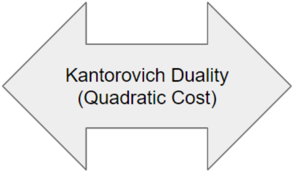
Contents

Introduction

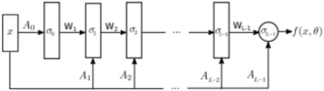
Introduction

OPTIMAL
TRANSPORT





INPUT CONVEX
NEURAL NETWORKS



OT Theory Results

Problem Setting

- P and Q be two probability distribution on $\mathbb{R}^d$ with finite second order moments
- Q admits a density in $\mathbb{R}^d$

Monge Problem

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

Kantorovich Relaxation

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

where $\Pi(P,Q)$ is set of couplings whose marginal distributions are P and Q.

Kantorovich Duality

Kantorovich Duality (Theorem 1.3, Villani (2003))

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

where

$$\Phi_c = \left\{ (f,g) \in L^1(P) \times L^1(Q) : f(x) + g(y) \leq \frac{1}{2}||x-y||_2^2, \forall (x,y)\ dP \otimes dQ\ a.e. \right\}$$

Since $\frac{1}{2}||x-y||_2^2 = \frac{1}{2}||x||_2^2 + \frac{1}{2}||y||_2^2 - \langle x, y \rangle$, rewrite

$$\begin{aligned}\Phi_C = \{(f,g) \in L^1(P) \times L^1(Q) : \frac{1}{2}||x||_2^2 - f(x) + \frac{1}{2}||y||_2^2 - g(y) \geq \langle x, y \rangle, \\ \forall (x,y)\ dP \otimes dQ\ a.e. \}\end{aligned}$$

Kantorovich Duality

With $\phi(\cdot) = \frac{1}{2}||\cdot||^2 - f(\cdot)$ and $\psi(\cdot) = \frac{1}{2}||\cdot||^2 - g(\cdot)$:

$$
\begin{aligned}
W_2^2(P,Q) &= \sup_{\phi,\psi \in \tilde{\Phi}_c} \mathbb{E}_P[\frac{1}{2}||X||_2^2 - \phi(X)] + \mathbb{E}_Q[\frac{1}{2}||Y||_2^2 - \psi(Y)] \\
&= \frac{1}{2}\mathbb{E}[|X|^2 + |Y|^2] - \inf_{\phi,\psi \in \tilde{\Phi}_c} \{\mathbb{E}_P[\phi(X)] + \mathbb{E}_Q[\psi(Y)]\} \\
&= C_{P,Q} - \inf_{\phi,\psi \in \tilde{\Phi}_c} \{\mathbb{E}_P[\phi(X)] + \mathbb{E}_Q[\psi(Y)]\} \\
\tilde{\Phi}_c &= \{(\phi,\psi) \in L^1(P) \times L^1(Q) : \phi(x) + \psi(y) \geq \langle x, y \rangle, \\
&\quad \forall (x,y)\ dP \otimes dQ\ a.e.\}
\end{aligned}
$$

Observe that $C_{P,Q} := \frac{1}{2}\mathbb{E}[|X|^2 + |Y|^2]$ is independent of (ϕ,ψ), so

$$
\begin{aligned}
W_2^2(P,Q) &= C_{P,Q} - \inf_{f,g \in \tilde{\Phi}_c} \{\mathbb{E}_P[f(X)] + \mathbb{E}_Q[g(Y)]\} \\
\tilde{\Phi}_c &= \{(f,g) \in L^1(P) \times L^1(Q) : f(x) + g(y) \geq \langle x, y \rangle, \\
&\quad \forall (x,y)\ dP \otimes dQ\ a.e.\}
\end{aligned}
$$

Convex Conjugate

$f^*(y) = \sup_x\{\langle x, y\rangle - f(x)\}$ is the convex conjugate of a convex function $f(\cdot)$.

For a convex lower semi-continuous function f, $f^{**}(\cdot) = f(\cdot)$

Theorem

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

$$\langle x, y\rangle \leq f(x) + f^*(y),$$

and

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

We define

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

Remark: f differentiable at $x \Rightarrow \partial f(x) = \{\nabla f(x)\}$.
Remark: When f^* differentiable at $y \Rightarrow \langle \nabla f^*(y), y\rangle = f(\nabla f^*(y)) + f^*(y)$.

Kantorovich Duality (Quadratic Cost)

We set

$$\mathrm{CVX}(P) := \left\{ f \in L^1(P) : f \text{ is convex} \right\}$$

Via "double convexification" trick (Villani (2003), Theorem 2.9), transform general duality statement

$$W_2^2(P,Q) = C_{P,Q} - \inf_{f,g \in \tilde{\Phi}_c} \{\mathbb{E}_P[f(X)] + \mathbb{E}_Q[g(Y)]\}$$
$$\tilde{\Phi}_c = \{(f,g) \in L^1(P) \times L^1(Q) : f(x) + g(y) \geq \langle x, y \rangle, \forall (x,y)\ dP \otimes dQ\ a.e.\}$$

to get

Kantorovich Duality (Quadratic Cost)

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

Brenier Theorem

Brenier Theorem (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, $\pi = (\nabla f^* \times \mathrm{Id})_{\#}Q$, 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.

Main Results

Consistency

Theorem

(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.

Consistency Proof

Want to show

$$\sup_{\substack{f\in \mathrm{CVX}(P),\\ f^*\in L^1(Q)}} \inf_{g\in \mathrm{CVX}(Q)} \mathcal{V}(f,g) + C_{P,Q} = C_{P,Q} - \left(\inf_{f\in \mathrm{CVX}(P)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[f^*(Y)] \right)$$

First need to show $g, f^* \in L^1(Q) \Rightarrow g, f^*$ are differentiable Q-a.e.

- As $Q(\{g=\infty\}) = 0$, $Q(\operatorname{Dom} g) = 1$
- $\operatorname{Dom} g$ convex $\Rightarrow m(\partial \operatorname{Dom} g) = 0$ (Lang (1986)) $\Rightarrow Q(\operatorname{Int}\operatorname{Dom} g) = 1$
- g convex $\Rightarrow g$ is differentiable Q-a.e. on $\operatorname{Int}\operatorname{Dom} g$
- Same logic applies to f^*

Consistency Proof

For Q-a.e. $y \in \mathbb{R}^d$, $\nabla g(y)$ exists and $\partial f^*(y) = \{\nabla f^*(y)\}$, giving

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

Taking expectation gives

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

with equality if $g = f^*$. This gives

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

Stability

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

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

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

Theorem

(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

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

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.

Stability

The sketch of proof

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

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

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

Stability

f is α-strongly convex, f^* is $\frac{1}{\alpha}$ smooth (by Kakade and Shalev-Shwartz (2009) Theorem 6): $\forall y, z \in \mathbb{R}^d$, $f^*(z) \leq f^*(y) + \langle \nabla f^*(y), z - y \rangle + \frac{1}{2\alpha}||z - y||^2 \triangleq h_y(z)$

$$f^*(z) \leq h_y(z)$$
$$f(x) = f^{**}(x) = \sup_z(\langle z, x \rangle - f^*(z)) \geq \sup_z(\langle z, x \rangle - h_y(z)) = h_y^*(x)$$

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

$$f(x) + f^*(y) - \langle y, x \rangle \geq \frac{\alpha}{2}||x - \nabla f^*(y)||^2$$

Stability

Recall Theorem 1: When f^* differentiable at $y \Rightarrow \langle \nabla f^*(y), y \rangle = f(\nabla f^*(y)) + f^*(y)$.

$$f(x) + f^*(y) - \langle y, x \rangle \geq \frac{\alpha}{2} ||x - \nabla f^*(y)||^2 \tag{9}$$

$$\begin{aligned}
\epsilon_1(f, g) &= \mathcal{V}(f, g) - \inf_{\tilde{g} \in CVX(Q)} \mathcal{V}(f, \tilde{g}) \\
&= \mathcal{V}(f, g) - \mathcal{V}(f, f^*) \text{ (by Theorem 2)} \\
&= -\mathbb{E}_P[f(X)] - \mathbb{E}_Q[\langle Y, \nabla g(Y) \rangle - f(\nabla g(Y))] + \\
&\mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla f^*(Y) \rangle - f(\nabla f^*(Y))] \\
&= -\mathbb{E}_Q[\langle Y, \nabla g(Y) \rangle - f(\nabla g(Y))] + \mathbb{E}_Q[\langle Y, \nabla f^*(Y) \rangle - f(\nabla f^*(Y))] \\
&= -\mathbb{E}_Q[\langle Y, \nabla g(Y) \rangle - f(\nabla g(Y))] + \mathbb{E}_Q[f^*(Y)] \text{ (by Theorem 1)} \\
&= \mathbb{E}_Q[f(\nabla g(Y)) + f^*(Y) - \langle Y, \nabla g(Y) \rangle] \\
&\geq \frac{\alpha}{2} \mathbb{E}_Q[||\nabla g(Y) - \nabla f^*(Y)||^2] \text{ (by 9, with } x = \nabla g(y))
\end{aligned}$$

Stability

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

$$\begin{aligned}
\epsilon_2(f) &= \sup_{\tilde{f}\in CVX(P)} \{\inf_{\tilde{g}\in CVX(Q)} \mathcal{V}(\tilde{f},\tilde{g})\} - \inf_{\tilde{g}\in CVX(Q)} \mathcal{V}(f,\tilde{g}) \\
&= \mathcal{V}(f_0, f_0^*) - \inf_{\tilde{g}\in CVX(Q)} \mathcal{V}(f,\tilde{g}) \\
&= \mathcal{V}(f_0, f_0^*) - \mathcal{V}(f, f^*) \text{ (by Theorem 2)} \\
&= -\mathbb{E}_P[f_0(X)] - \mathbb{E}_Q[\langle Y, \nabla f_0^*(Y)\rangle - f_0(\nabla f_0^*(Y))] + \\
&\mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla f^*(Y)\rangle - f(\nabla f^*(Y))]
\end{aligned}$$

Stability

Recall Theorem 1: When f^* differentiable at $y \Rightarrow \langle \nabla f^*(y), y \rangle = f(\nabla f^*(y)) + f^*(y)$.

$$f(x) + f^*(y) - \langle y, x \rangle \geq \frac{\alpha}{2} ||x - \nabla f^*(y)||^2$$

$$\begin{aligned}
\epsilon_2(f) = &-\mathbb{E}_P[f_0(X)] - \mathbb{E}_Q[\langle Y, \nabla f_0^*(Y) \rangle - f_0(\nabla f_0^*(Y))] + \\
&\mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla f^*(Y) \rangle - f(\nabla f^*(Y))] \\
&= -\mathbb{E}_P[f_0(X)] - \mathbb{E}_Q[f_0^*(Y)] + \mathbb{E}_P[f(X)] + \mathbb{E}_Q[f^*(Y)] \text{ (by Theorem 1)} \\
&= -\mathbb{E}_Q[f_0(\nabla f_0^*(Y))] - \mathbb{E}_Q[f_0^*(Y)] + \mathbb{E}_Q[f(\nabla f_0^*(Y))] + \mathbb{E}_Q[f^*(Y)] \\
&\ (\nabla f_0^*(\cdot) \text{ pushforwards from } Q \text{ to } P) \\
&= -\mathbb{E}_Q[\langle Y, \nabla f_0^*(Y) \rangle] + \mathbb{E}_Q[f(\nabla f_0^*(Y))] + \mathbb{E}_Q[f^*(Y)] \text{ (by Theorem 1)} \\
&= \mathbb{E}_Q[f(\nabla f_0^*(Y)) + f^*(Y) - \langle Y, \nabla f_0^*(Y) \rangle] \\
&\geq \frac{\alpha}{2} \mathbb{E}_Q[||Y - \nabla f_0^*(Y)||^2] \text{ (by 9, with } x = \nabla f_0^*(y))
\end{aligned}$$

Algorithm

Parameterization & Objective

- Parametrize convex functions f and g using the same ICNN architecture with parameters θ_f, θ_g
- For θ_f, enforce $W_l \geq 0$
- For θ_g, can relax non-negative constraint and regularize with

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

to obtain faster convergence
- The objective function is

$$\max_{\theta_f : W_l \geq 0, \forall l \in [L-1]} \min_{\theta_g} J(\theta_f, \theta_g) + R(\theta_g) \tag{10}$$

where

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

Statement

Algorithm 1: Algorithm: the numerical procedure to solve the optimization (10)

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 Loss $J(\theta_f, \theta_g) + R(\theta_g)$ using Adam

  end

  Update θ_f to maximize Loss $J(\theta_f, \theta_g) + R(\theta_g)$ using Adam

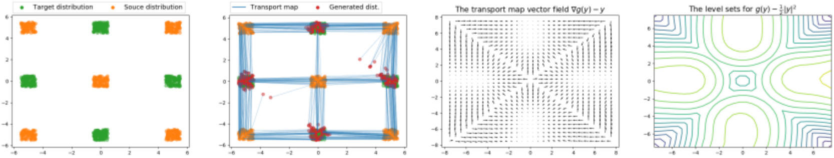
  Projection: $w \leftarrow \max(w, 0)$, for all $w \in \{W^l\} \subseteq \theta_f$

end

return: optimizer θ_f^*, θ_g^*, optimal transport $\nabla(g^{\theta_g^*})$

Experimental Results

Performance of Algorithm



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

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

Discontinuities

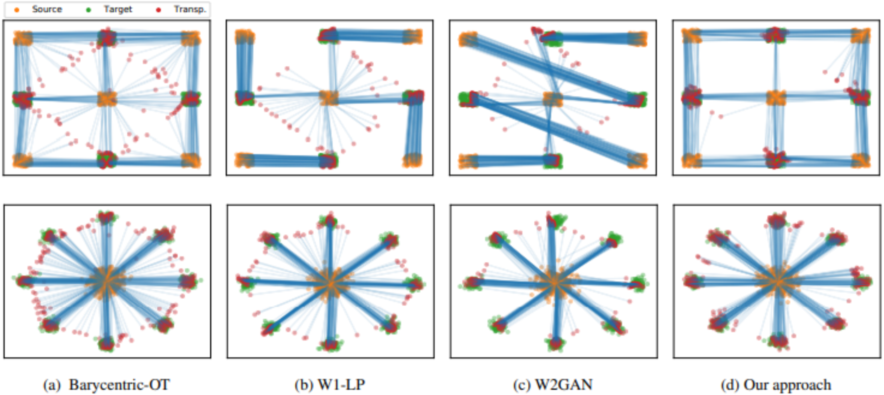


Figure: From Makkuva et al. (2020), the transport maps learned under different approaches for "checkerboard" data set (top row) and mixture of eight Gaussians (bottom rows)

Robustness to Initialization

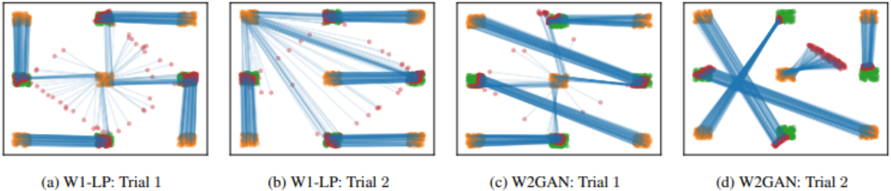


Figure: From Makkuva et al. (2020), transport maps learned by different approaches under different initializations

References

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.