Motivation

Theory

Practice
A geometric way to compare probability measures.

What is Optimal Transport?

Nobel prize
Monge  Kantorovich  Dantzig  Wasserstein  Brenier  Otto  McCann  Villani

Fields medal (and French politician)

1. Introduction to optimal transport
   • Construction
   • Many formulas

2. Applications

3. Discrete/discretized transport
   • Entropic regularization
   • Eulerian transport
   • Semidiscrete transport

4. Extensions & frontiers
Useful References

Shameless self-promotion:

Computational Optimal Transport

Justin Solomon

Optimal transport is the mathematical discipline of matching supply to demand while minimizing shipping costs. This matching problem becomes extremely challenging as the quantity of supply and demand points increases; modern applications must cope with thousands or millions of these at a time. Here, we introduce the computational optimal transport prob-
Plan For Today

1. **Introduction** to optimal transport
   - Construction
   - Many formulas

2. Applications

3. **Discrete/discretized** transport
   - Entropic regularization
   - Eulerian transport
   - Semidiscrete transport

4. Extensions & frontiers
“Somewhere over here.”
Probability as Geometry

\[ \rho(x) \]

\[ \delta_{x_0} \]

\[ x_0 \]

“Exactly here.”
"One of these two places."
How We Compute Distances

$p_1(x)$

$p_2(x)$

$p_1(x) - p_2(x)$

$L^p$ norm

KL divergence

$d(p_1, p_2)$
Equidistant!

\[ \|p - q\|_1 = \sum_i |p_i - q_i| \]

\[ \text{KL}(p || q) = -\sum_i p_i \log \frac{q_i}{p_i} \]
What’s Wrong?

Measured overlap, not displacement.
Cost to move mass $m$ from $x$ to $y$: 

$$m \cdot d(x, y)$$
Observation

Even the laziest shoveler must do some work.

Property of the distributions themselves!

My house!
Measure Coupling

\[ \pi(x, y) := \text{Amount moved from } x \text{ to } y \]

\[ \pi(x, y) \geq 0 \ \forall x \in X, y \in Y \]

\[ \int_Y \pi(x, y) \, dy = \rho_0(x) \ \forall x \in X \]

\[ \int_X \pi(x, y) \, dx = \rho_1(y) \ \forall y \in Y \]

Mass is positive

Must scoop everything up

Must cover the target
Kantorovich Problem

\[ \mathcal{OT}(\mu, \nu; c) := \min_{\pi \in \Pi(\mu, \nu)} \iint_{X \times Y} c(x, y) \, d\pi(x, y) \]
$p$-Wasserstein Distance

$\mathcal{W}_p(\mu, \nu) \equiv \min_{\pi \in \Pi(\mu, \nu)} \left( \int \int_{X \times X} d(x, y)^p \, d\pi(x, y) \right)^{1/p}$

- Shortest path distance
- Expectation

Geodesic distance $d(x, y)$
$\mathcal{W}_1(\rho_0, \rho_1) := \begin{cases} \min_{\pi} & \int \int_{\mathbb{R} \times \mathbb{R}} \pi(x, y) |x - y| \, dx \, dy \\ \text{s.t.} & \pi \geq 0 \, \forall x, y \in \mathbb{R} \\ & \int_{\mathbb{R}} \pi(x, y) \, dy = \rho_0(x) \, \forall x \in \mathbb{R} \\ & \int_{\mathbb{R}} \pi(x, y) \, dx = \rho_1(y) \, \forall y \in \mathbb{R} \end{cases}$

Minimize total work
Nonnegative mass
Starts from $\rho_0$
Ends at $\rho_1$
In One Dimension: Closed-Form

PDF \[\rightarrow\] [CDF] \[\rightarrow\] CDF\(^{-1}\)

\(\mathcal{W}_1(\mu, \nu) = \int_{-\infty}^{\infty} |\text{CDF}(\mu) - \text{CDF}(\nu)| \, d\ell\)

\(\mathcal{W}_2(\mu, \nu) = \int_{-\infty}^{\infty} (\text{CDF}^{-1}(\mu) - \text{CDF}^{-1}(\nu))^2 \, d\ell\)

Doesn’t extend past 1d!
Fully-Discrete Transport

\[ [\mathcal{W}_p(\mu_0, \mu_1)]^p = \begin{cases} \min_{T \in \mathbb{R}^{k_0 \times k_1}} & \sum_{ij} T_{ij} |x_{0i} - x_{1j}|^p \\ \text{s.t.} & T \geq 0 \\ & \sum_j T_{ij} = a_{0i} \\ & \sum_i T_{ij} = a_{1j} \end{cases} \]

Linear program: Finite number of variables

Algorithms: Simplex, interior point, auction, ...

"Empirical measures"
Semidiscrete Transport

\[ \mu_0 := \sum_{i=1}^{k_0} a_{0i} \delta_{x_{0i}} \]

\[ \mu_1(S) := \int_S \rho_1(x) \, dx \]

Never a reason to “leapfrog” mass!
Monge Formulation

\[ \inf_{\phi \# \rho_0 = \rho_1} \int_{-\infty}^{\infty} c(x, \phi(x)) \rho_0(x) \, dx \]

[Monte 1781]; image courtesy Marco Cuturi

Not always well-posed!
1. Introduction to optimal transport
   • Construction
   • Many formulas

2. Applications

3. Discrete/discretized transport
   • Entropic regularization
   • Eulerian transport
   • Semidiscrete transport

4. Extensions & frontiers
Example: Discrete Transport

\[ X = \{1, 2, \ldots, k_1\}, \quad Y = \{1, 2, \ldots, k_2\} \]

\[ OT(v, w; C) = \min_{T \in \mathbb{R}^{k_1 \times k_2}} \sum_{i,j} T_{ij} c_{ij} \quad \text{s.t.} \quad \begin{align*}
  T &\geq 0 \\
  \sum_j T_{ij} &= v_i \quad \forall i \in \{1, \ldots, k_1\} \\
  \sum_i T_{ij} &= w_j \quad \forall j \in \{1, \ldots, k_2\}.
\end{align*} \]

“Earth Mover’s Distance”

Metric when \( d(x,y) \) satisfies the triangle inequality.

“The Earth Mover's Distance as a Metric for Image Retrieval”

Revised in:
“Ground Metric Learning”
Cuturi and Avis; JMLR 15 (2014)
Kantorovich Duality

\[
\text{OT}(\mu, \nu; c) := \begin{cases} 
\min_{\pi} & \int_{X \times Y} c(x, y) \, d\pi(x, y) \\
\text{s.t.} & \pi \in \Pi(\mu, \nu) 
\end{cases} \quad \text{Primal} \\
= \begin{cases} 
\max_{\phi, \psi} & \int_X \phi(x) \, d\mu(x) + \int_Y \psi(y) \, d\nu(y) \\
\text{s.t.} & \phi(x) + \psi(y) \leq c(x, y) \text{ for a.e. } x \in X, y \in Y 
\end{cases} \quad \text{Dual}
\]
Flow-Based $W_2$

$$W_2^2(\rho_0, \rho_1) = \left\{ \begin{array}{l}
\inf_{\rho,v} \int_0^1 \int_M \frac{1}{2} \rho(x,t) \|v(x,t)\|^2 \, dx \, dt \\
\text{s.t. } \nabla \cdot (\rho(x,t)v(x,t)) = \frac{\partial \rho(x,t)}{\partial t} \\
v(x,t) \cdot \hat{n}(x) = 0 \quad \forall x \in \partial M \\
\rho(x,0) = \rho_0(x) \\
\rho(x,1) = \rho_1(x)
\end{array} \right. $$

[Benamou & Brenier 2000]

Tip of an iceberg: Manifold theory of transport!
Plan For Today

1. Introduction to optimal transport
   • Construction
   • Many formulas

2. Applications

3. Discrete/discretized transport
   • Entropic regularization
   • Eulerian transport
   • Semidiscrete transport

4. Extensions & frontiers
Wassersteinization


Introduction of optimal transport into a computational problem.

cf. least-squarification, L1ification, deep-netification, kernelization
We have tools to

- **Solve** optimal transport problems numerically
- **Differentiate** transport distances in terms of their input distributions

**Bonus:**

Transport cost from $\mu$ to $\nu$ is a **convex** function of $\mu$ and $\nu$. 
Histograms and Descriptors

Use word embeddings

Word Mover’s Distance (WMD)

[Kusner et al. 2015]
Registration and Reconstruction

Caveat:
Not a good model for deformation!

(a) Dataset  (b) OT fidelity  (c) RKHS fidelity

Fig. 2. First row: Matching of fibres bundles. Second row: Matching of two hand surfaces using a balanced OT fidelity. Target is in purple.

[Feudy, Charlier, Vialard, and Peyré 2017]
Engineering Design
Interpolation

Image from [Lavenant, Claici, Chien, & Solomon 2018]

Image from [Vaxman & Solomon 2019]
Blue Noise and Distribution Approximation

\[
\min_{x_1, \ldots, x_n} \mathcal{W}_2^2 \left( \mu, \frac{1}{n} \sum_i \delta_{x_i} \right)
\]

Zebra image courtesy F. de Goes; photo by F. Durand; distribution image courtesy S. Claici
**Statistical Estimation**

\[ \text{MLE} := \min_{\theta \in \Theta} \text{KL}(\nu_{\text{data}}|p_\theta) \]

\[ \rightarrow \text{MKE} := \min_{\theta \in \Theta} \mathcal{W}_2(\nu_{\text{data}}, p_\theta) \]


**Minimum Kantorovich Estimator**
1. **Estimate** transport map
2. **Transport** labeled samples to new domain
3. **Train** classifier on transported labeled samples

[Courty et al. 2017]
Generative Adversarial Networks (GANs)

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: $\alpha$, the learning rate. $c$, the clipping parameter. $m$, the batch size. $n_{\text{critic}}$, the number of iterations of the critic per generator iteration.

Require: $\theta_0$, initial critic parameters. $\delta_0$, initial generator’s parameters.

1: while $\theta$ has not converged do
2:   for $t = 0, \ldots, n_{\text{critic}}$ do
3:     Sample $\{x^{(i)}\}_{i=1}^{m} \sim P_r$, a batch from the real data.
4:     Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
5:     $g_\theta \leftarrow \nabla_w \left[ \frac{1}{m} \sum_{i=1}^{m} f_\theta(x^{(i)}) - \frac{1}{m} \sum_{i=1}^{m} f_r(y_0(z^{(i)})) \right]$
6:     $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_\theta)$
7:   end for
8:   Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
9:   $g_\theta \leftarrow \nabla_\theta \left[ \frac{1}{m} \sum_{i=1}^{m} f_r(y_0(z^{(i)})) \right]$
10:  $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$
11: end while

Figure 9: WGAN algorithm: generator and critic are DCGANs.

[Arjovsky et al. 2017]
1. Introduction to optimal transport
   • Construction
   • Many formulas

2. Applications

3. Discrete/discretized transport
   • Entropic regularization
   • Eulerian transport
   • Semidiscrete transport

4. Extensions & frontiers
Theme in computation:
Same in theory, but different in practice

Choose one of each:
- Formulation
- Discretization
“No Free Lunch”

<table>
<thead>
<tr>
<th></th>
<th>Sym</th>
<th>Loc</th>
<th>Lin</th>
<th>Pos</th>
<th>Psd</th>
<th>Con</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Value</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Intrinsic Del</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>?</td>
</tr>
<tr>
<td>Combinatorial</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>Cotan</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>

Observe that none of the Laplacians considered in graphics fulfill all desired properties. Even more: none of them satisfy the first four properties.

[Wardetzky et al. 2007]
1. Introduction to optimal transport
   • Construction
   • Many formulas

2. Applications

3. Discrete/discretized transport
   • Entropic regularization
   • Eulerian transport
   • Semidiscrete transport

4. Extensions & frontiers
Entropic Regularization

\[
\begin{align*}
\min_T & \quad \sum_{i,j} T_{ij} c_{ij} - \alpha H(T) \\
\text{s.t.} & \quad \sum_j T_{ij} = p_i \\
& \quad \sum_i T_{ij} = q_j
\end{align*}
\]

Cuturi. “Sinkhorn distances: Lightspeed computation of optimal transport” (NIPS 2013)

OK to drop nonnegative constraint!
Sinkhorn Algorithm

$$T = \text{diag}(u) K_\alpha \text{diag}(v),$$

where $$K_\alpha := \exp(-C/\alpha)$$

$$u \leftarrow p \otimes (K_\alpha v)$$

$$v \leftarrow q \otimes (K_\alpha^\top u)$$

Ingredients for Sinkhorn

1. Supply vector \( p \)
2. Demand vector \( q \)
3. Multiplication by \( K \)

\[
K_{ij} = e^{-c_{ij}/\alpha}
\]

Sinkhorn Divergences

\[ \overline{W}_c,\varepsilon(\mu, \nu) := 2W_c,\varepsilon(\mu, \nu) - W_c,\varepsilon(\mu, \mu) - W_c,\varepsilon(\nu, \nu) \]

- Debiases entropy-regularized transport near zero
- Easy to compute: Three calls to Sinkhorn
- Links optimal transport to maximum mean discrepancy (MMD)

\[ \begin{align*}
\overline{W}_c,\varepsilon(\mu, \nu) & \xrightarrow{\varepsilon \to 0} 2W_c(\mu, \nu) \\
\overline{W}_c,\varepsilon(\mu, \nu) & \xrightarrow{\varepsilon \to \infty} \text{MMD}_c(\mu, \nu)
\end{align*} \]

[Genevay et al. 2016]
1. Introduction to optimal transport
   - Construction
   - Many formulas

2. Applications

3. Discrete/discretized transport
   - Entropic regularization
   - Eulerian transport
   - Semidiscrete transport

4. Extensions & frontiers
Discretization

Unknown: $\mu : [0, 1] \times \mathcal{M} \rightarrow \mathbb{R}_+$

$$\min_{\mu, m} \left\{ \int_0^1 \int_{\mathcal{M}} \frac{|m|^2}{2\mu} \right\}$$

where $m = \mu v$ is the momentum, under the constraints

$$\begin{align*}
\partial_t \mu + \nabla \cdot m &= 0, \\
\mu_0 &= \beta_0, \\
\mu_1 &= \beta_1
\end{align*}$$

Images/math from [Lavenant, Claici, Chien, and Solomon 2018]
In computer science: Network flow problem

\[
\min_T \sum_e c_e |J_e| \\
\text{s.t. } D^\top J = \begin{cases} \ p_1 - p_0 \end{cases}_{f}
\]

Graph analog:

Beckmann Formulation

Better scaling for sparse graphs!

Smooth PDE analog: [Solomon et al. 2014]
1. **Introduction** to optimal transport
   - Construction
   - Many formulas

2. **Applications**

3. **Discrete/discretized** transport
   - Entropic regularization
   - Eulerian transport
   - Semidiscrete transport

4. **Extensions & frontiers**
Semidiscrete General Case

\[ \mu_0 := \sum_{i=1}^{k} a_i \delta_{x_i} \quad \text{and} \quad \nu(S) := \int_S \rho(x) \, dx \]

\[ \mathcal{W}_2^2(\mu, \nu) = \sup_{\phi \in \mathbb{R}^k} \sum_i \left[ a_i \phi_i + \int_{\text{Lag}_{\phi}(x_i)} \rho(y) [c(x_i, y) - \phi_i] \, dA(y) \right] \]

\[ \text{Lag}_{\phi}(x_i) := \{ y \in \mathbb{R}^n : c(x_i, y) - \phi_i \leq c(x_j, y) - \phi_j \ \forall \ j \neq i \} \]

https://www.jasondavies.com/power-diagram/
Simple algorithm: Gradient ascent  
Ingredients: Power diagram

More complex: Newton’s method  
Converges globally [de Goes et al. 2012; Kitagawa, Mérigot, & Thibert 2016]

ML setting: Stochastic optimization  
[Genevay et al. 2016; Staib et al. 2017; Claici et al. 2018]
<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entropic regularization</td>
<td>• Fast</td>
<td>• Blurry</td>
</tr>
<tr>
<td></td>
<td>• Easy to implement</td>
<td>• Becomes singular as $\alpha \to 0$</td>
</tr>
<tr>
<td></td>
<td>• Works on mesh using heat kernel</td>
<td></td>
</tr>
<tr>
<td>Eulerian optimization</td>
<td>• Provides displacement interpolation</td>
<td>• Hard to optimize</td>
</tr>
<tr>
<td></td>
<td>• Connection to PDE</td>
<td>• Triangle mesh formulation unclear</td>
</tr>
<tr>
<td>Semidiscrete optimization</td>
<td>• No regularization</td>
<td>• Expensive computational geometry algorithms</td>
</tr>
<tr>
<td></td>
<td>• Connection to &quot;classical&quot; geometry</td>
<td></td>
</tr>
</tbody>
</table>

Many others:
Stochastic transport, dual ascent, Monge-Ampère PDE, …
Plan For Today

1. Introduction to optimal transport
   • Construction
   • Many formulas

2. Applications

3. Discrete/discretized transport
   • Entropic regularization
   • Eulerian transport
   • Semidiscrete transport

4. Extensions & frontiers

Extra:
New methods in learning
Figure 1: For a given fixed set of samples \((z_1, \ldots, z_m)\), and input data \((y_1, \ldots, y_n)\), flow diagram for the computation of Sinkhorn loss function \(\theta \mapsto \hat{E}_\varepsilon^{(L)}(\theta)\). This function is the one on which automatic differentiation is applied to perform parameter learning. The display shows a simple 2-layer neural network \(g_\theta : z \mapsto x\), but this applies to any generative model.
Proposition 19. The dual of entropy-regularized OT between two probability measures \( \alpha \) and \( \beta \) can be rewritten as the maximization of an expectation over \( \alpha \otimes \beta \):

\[
W_\varepsilon^c(\alpha, \beta) = \max_{u,v \in C(X) \times C(Y)} \mathbb{E}_{\alpha \otimes \beta}[f_{\varepsilon}^{XY}(u,v)] + \varepsilon,
\]

where

\[
f_{\varepsilon}^{xy} \overset{\text{def}}{=} u(x) + v(y) - \varepsilon \exp \frac{u(x)+v(y)-c(x,y)}{\varepsilon} \quad \text{for } \varepsilon > 0.
\]

and when \( \beta = \sum_{j=1}^{m} \beta_j \delta_{y_j} \) is discrete, the potential \( v \) is a \( m \)-dimensional vector \((v_j)_j\)

Algorithm 4 Averaged SGD for Semi-Discrete OT

**Input:** step size \( C \in \mathbb{R}_+ \)

**Output:** dual potential \( \bar{v} \in \mathbb{R}^m \)

- \( v \leftarrow 0_m \) (iterates for SGD)
- \( \bar{v} \leftarrow v \) (dual potential obtained by averaging)

for \( k = 1, 2, \ldots \) do

  Sample \( x_k \) from \( \alpha \)

  \( v \leftarrow v + \frac{C}{\sqrt{k}} \nabla v g^{x_k}_\varepsilon(v) \) (gradient ascent step using \( v \))

  \( \bar{v} \leftarrow \frac{1}{k} v + \frac{k-1}{k} \bar{v} \) (averaging step to get faster convergence of \( v \))

end for

Parameterize dual potentials:
- Using RKHS [Genevay et al. 2016]
- Using neural networks [Seguy et al. 2017]
New Progress on the Monge Formulation

Input Convex Neural Networks

Brandon Amos, Lei Xu, J. Zico Kolter

Abstract

This paper presents the input convex neural network architecture. These are scalar-valued (potentially deep) neural networks with constraints on the network parameters such that the output of the network is a convex function of any of the inputs. The networks allow for efficient inference via optimization over some inputs to the network given others, and can be applied to settings including structured prediction, data imputation, reinforcement learning, and others. In this paper we lay the basic groundwork for these models, proposing methods for inference, optimization and learning, and analyze their representational power. We show that many existing neural network architectures can be made input-convex with a minor modification, and develop specialized optimization algorithms tailored to this setting. Finally, we highlight the performance of the methods on multi-label prediction, image completion, and reinforcement learning problems, where we show improvement over the existing state-of-the-art in many cases.

Optimal transport mapping via input convex neural networks

Ashok Vardhan Makkuva, Amirhossein Taghvaei, Jason D. Lee, Sewooong Oh

Abstract

In this paper, we present a novel and principled approach to learn the optimal transport between two distributions, from samples. Guided by the optimal transport theory, we learn the optimal Kantorovich potential which induces the optimal transport mapping.

1. Introduction

Finding a mapping that transports mass from one distribution $Q$ to another distribution $P$ is an important task in various machine learning applications, such as deep generative models (Goodfellow et al., 2014; Kingma & Welling, 2013) and domain adaptation (Gopalan et al., 2011; Bengio et al., 2013).
Plan For Today

1. Introduction to optimal transport
   • Construction
   • Many formulas

2. Applications

3. Discrete/discretized transport
   • Entropic regularization
   • Eulerian transport
   • Semidiscrete transport

4. Extensions & frontiers
Extension:

Wasserstein Barycenters

Wasserstein: \( \mu^* := \left[ \arg \min_{\mu \in \text{Prob}(\mathbb{R}^n)} \sum_i W_2^2(\mu, \mu_i) \right] \)

[Agueh and Carlier 2010]
Barycenters in Bayesian Inference

Wasserstein Subset Posterior (WASP)
[Srivastava et al. 2018]
Extension: Quadratic Matching

\[ GW^2_2((\mu_0, d_0), (\mu, d)) := \min_{\gamma \in \mathcal{M}(\mu_0, \mu)} \iint_{\Sigma_0 \times \Sigma} [d_0(x, x') - d(y, y')]^2 d\gamma(x, y) d\gamma(x', y') \]
Variety of Correspondence Tasks

[Solomon et al. 2016]
Extension:

**Gradient Flows**

“Entropic Wasserstein Gradient Flows” [Peyré 2015]
Extension:

Matrix Fields and Vector Measures

"Quantum Optimal Transport for Tensor Field Processing"
[Peyré et al. 2017]

Open problem: Dynamical version? Curved surfaces?
Extension: Sampling Problems

\[ \min_{\rho} \left[ \mathcal{W}_2^2(\rho_0, \rho) + \mathcal{W}_2^2(\rho_1, \rho) \right] \]

Somewhere between semidiscrete and smooth

Wasserstein barycenter