Optimal Transport
Back to comfortable ground!
BIASED

...toward my own research!
Understand geometry from a “softened” probabilistic standpoint.

Secondary goal: Application of machinery from previous lectures (vector fields, geodesics, metric spaces, optimization...)}
“Somewhere over here.”
“Exactly here.”
Probalistic Geometry

\[ \rho(x) \]

Superposition

“One of these two places.”
Which is closer, 1 or 2?
Motivating Question

Which is closer, 1 or 2?
Fuzzy Version

Which is closer, 1 or 2?

$p(x, y)$

$p_1(x, y)$

$p_2(x, y)$

Query

1

2

Which is closer, 1 or 2?
Typical Measurement

\[ p_1(x) \quad \rightarrow \quad p_1(x) - p_2(x) \quad \rightarrow \quad d(p_1, p_2) \]

\[ p_2(x) \]

Lp norm

KL divergence
Which is closer, 1 or 2?
Returning to the Question

Query

\[ p(x, y) \]

\[ p_1(x, y) \]

\[ p_2(x, y) \]

Neither! Equidistant.
What’s Wrong?

Measured overlap, not displacement.
Smaller bins worsen histogram distances
The Root Cause

Permuting histogram bins has no effect on these distances.
Optimal Transport

\[ \mathcal{P}(\Omega) \]

[McCann’95]
Interpolant

Over a metric space

Geometric theory of probability

Image courtesy M. Cuturi
Compare in **this** direction

Not in **this** direction
Match mass from the distributions
Earth Mover’s Distance

Cost to move mass $m$ from $x$ to $y$:

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

Match mass from the distributions
Transportation Matrix

- **Supply** distribution $p_0$
- **Demand** distribution $p_1$

\[
T \geq 0
\]

\[
T^1 = p_0
\]

\[
T^\top 1 = p_1
\]
Earth Mover’s Distance

\[
\begin{align*}
\min_T & \sum_{i,j} T_{i,j} d(x_i, x_j) \\
\text{s.t.} & \sum_j T_{i,j} = p_i \\
& \sum_i T_{i,j} = q_j \\
& T \geq 0
\end{align*}
\]

- Starts at \( p \)
- Ends at \( q \)
- Positive mass

\[ m \cdot d(x, y) \]
EMD is a 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)
Comparing histogram descriptors
Discrete Perspective

Sources

Matching in disguise?

Min-cost flow

Sinks
Algorithm for Small-Scale Problems

- **Step 1:** Compute $D_{ij}$

- **Step 2:** Solve linear program
  - Simplex
  - Interior point
  - Hungarian algorithm
  - ...
Transportation Matrix Structure

Matches bins

Underlying map!
Useful conclusions:

1. Practical

Can do better than generic solvers.
Useful conclusions:

1. Practical
   Can do better than generic solvers.

2. Theoretical
   \( T \in [0, 1]^{n \times n} \) usually contains \( O(n) \) nonzeros.

"Complementary slackness"

Min-cost flow
Challenge for Large-Scale Problems

\[ \min_{T} \sum_{i,j} T_{ij} d(x_i, x_j) \]

s.t.
\[ \sum_{j} T_{ij} = p_i \]
\[ \sum_{i} T_{ij} = q_j \]
\[ T \geq 0 \]

Matrix \( T_{ij} \) is too big!

Precompute \( d(x_i, x_j) \) for all \( i, j \)!
Today’s Questions

- Can we optimize faster?
- Is there a continuum interpretation?
- What properties does this model exhibit?

We’ll answer them in parallel!
First example:
Linear Transportation on Graphs

Supply -1

Demand +2

Supply -1
Linear Transportation on Graphs

Supply

Demand

1
1
1

-1

-1

+2
Routing Supply to Meet Demand

\[ \text{min} \ \langle T, D \rangle \]
\[ \text{s.t.} \ T \geq 0 \]
\[ T1 = p_0 \]
\[ T^\top 1 = p_1 \]

* D contains shortest path length.
Simplification for Linear Cost

\[
\begin{align*}
\min_T \quad & \sum_{ij} T_{ij} d(x_i, x_j) \\
\text{s.t.} \quad & \sum_j T_{ij} = p_i \\
& \sum_i T_{ij} = q_j \\
& T \geq 0
\end{align*}
\]
Simplification for Linear Cost

\[
\min_T \sum_{i,j} T_{ij} d(x_i, x_j)
\]
\[
s.t. \quad \sum_j T_{ij} = p_i
\]
\[
\sum_i T_{ij} = q_j
\]
\[
T \geq 0
\]
Orient edges arbitrarily

D_{ev} := \begin{cases} 
-1 & \text{if } E_{e1} = v \\
1 & \text{if } E_{e2} = v \\
0 & \text{otherwise}
\end{cases}
In computer science:

**Beckmann Formulation**

Better scaling for sparse graphs!

\[
\begin{align*}
\min_T & \quad \sum_e c_e |J_e| \\
\text{s.t.} & \quad D^\top J = p_1 - p_0
\end{align*}
\]

In computer science:

**Network flow problem**
We used the structure of $D$. 

$$
\begin{align*}
\text{min}_T \quad & \sum_{i,j} T_{ij} d(x_i, x_j) \\
\text{s.t.} \quad & \sum_j T_{ij} = p_i \\
\quad & \sum_i T_{ij} = q_j \\
\quad & T \geq 0
\end{align*}
$$
Think of probabilities like a fluid. 

Probabilities *advect* along the surface. 

“Eulerian” 

Application of our vector field lectures! 

**Solomon, Rustamov, Guibas, and Butscher.**  
“Earth Mover’s Distances on Discrete Surfaces.”  
SIGGRAPH 2014
Alternative Formulation for $W_1$

$W_1(\rho_0, \rho_1) = \left\{ \begin{array}{c} \inf \int_{M} \| J(x) \| \, dx \\ \text{s.t. } \nabla \cdot J(x) = \rho_1(x) - \rho_0(x) \\ J(x) \cdot n(x) = 0 \quad \forall x \in \partial M \end{array} \right\}$

“Beckmann problem”

Total work

Advects from $\rho_0$ to $\rho_1$

Scales linearly
Recall:
Helmholtz-Hodge Decomposition

Divergence free

Harmonic

Curl free
Hodge Decomposition of $J$

$$J(x) = \nabla f(x) + \mathcal{R} \nabla g(x)$$

Ignoring harmonic part!

$$\nabla \cdot J = \Delta f = \rho_1 - \rho_0$$

Curl-free

Div-free
Fast Optimization

1. \[ \Delta f = \rho_1 - \rho_0 \] Sparse SPD linear solve for \( f \)

2. \[ \inf_g \int_M \| \nabla f(x) + R \cdot \nabla g(x) \| \, dx \] Unconstrained and convex optimization for \( g \)
Pointwise Distance

\[ W(\rho_0, \rho_1) \rightarrow d(p_0, p_1) \]
Proposition: Satisfies triangle inequality.
No “displacement interpolation”
More General Formulation

Monge-Kantorovich Problem
Probability Measure

\[ \mu(X) = 1 \]

\[ \mu(S \subseteq X) \in [0, 1] \]

\[ \mu \left( \bigcup_{i \in I} E_i \right) = \sum_{i \in I} \mu(E_i) \]

“Prob(X)”

when \( E_i \) disjoint, \( I \) countable

Function from sets to probability
Measure Coupling

\[ \mu, \nu \in \text{Prob}(X) \]

\[ \Pi(\mu, \nu) := \left\{ \pi \in \text{Prob}(X \times X) : \begin{pmatrix} \pi(U \times X) = \mu(U) \\ \pi(X \times V) = \nu(V) \end{pmatrix} \right\} \]

Analog of transportation matrix
$p$-Wasserstein Distance

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

General cost: “Monge-Kantorovich problem”

Geodesic distance $d(x,y)$

Continuous analog of EMD
Monge Formulation

\[ \inf_{T \# \mu = \nu} \int_X D(x, T(x)) \, d\mu(x) \]

Image courtesy M. Cuturi

Not always well-posed!
In One Dimension

PDF $\rightarrow$ [CDF] $\rightarrow$ CDF$^{-1}$

$W_1(\mu, \nu) = \|CDF(\mu) - CDF(\nu)\|_1$

$W_2(\mu, \nu) = \|CDF^{-1}(\mu) - CDF(\nu)\|_2$
What Goes Wrong: Median Problems

\[ \min_{x \in \mathbb{R}^2} (\|x - x_0\|_2 + \|x - x_1\|_2) \]

\( W_1 \) ineffective for averaging tasks
Displacement Interpolation

$W_2$

$t = 0$
$t = 1/4$
$t = 1/2$
$t = 3/4$
$t = 1$

“Explains” shortest path.

Image from “Optimal Transport with Proximal Splitting” (Papadakis, Peyré, and Oudet)

Mass moves along shortest paths
Frustrating Issue

\[
\min_{T} \sum_{i,j} T_{ij} d(x_i, x_j) \\
\text{s.t.} \quad \sum_{j} T_{ij} = 1 \\
\sum_{i} T_{ji} = 0 \\
T \geq 0
\]
Entropic Regularization

$$\min_T \sum_{i,j} T_{ij} d(x_i, x_j) - \gamma H(T)$$

s.t. $$\sum_j T_{ij} = p_i$$
$$\sum_i T_{ij} = q_j$$
$$T \geq 0$$

$$H(T) := - \sum_{i,j} T_{ij} \log T_{ij}$$

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

Prove on the board:

\[ T = \text{diag}(u) K \text{diag}(v), \]

where \( K_{ij} := e^{-D_{ij} / \gamma} \)

\[
\min_T \sum_{ij} T_{ij} d(x_i, x_j) - \gamma H(T) \\
\text{s.t.} \quad \sum_j T_{ij} = p_i \\
\sum_i T_{ij} = q_j \\
T \geq 0
\]

\[ H(T) := -\sum_{ij} T_{ij} \log T_{ij} \]
Sinkhorn Algorithm

\[ T = \text{diag}(u) K \text{diag}(v), \]

where \( K_{ij} := e^{-D_{ij}/\gamma} \)

\[ u \leftarrow p/Kv \]

\[ v \leftarrow q/K^\top u \]

Ingredients for Sinkhorn

1. Supply vector $p$
2. Demand vector $q$
3. Multiplication by $K$

$$K_{ij} = e^{-\frac{D_{ij}}{\gamma}}$$
On a Grid: Fast $K$ Product

$$(Kv)_{ij} = \sum_{k\ell} g_\sigma(\|(i,j) - (k,\ell)\|_2)u_{k\ell}$$

Gaussian convolution
Sinkhorn on a Grid

\[ u \leftarrow \frac{p}{K} v \]

\[ v \leftarrow \frac{q}{K^\top} u \]

No need to store K
Sinkhorn on a Grid

\[ u \leftarrow p / K v \]
\[ v \leftarrow q / K^\top u \]

What about surfaces?

No need to store K
Recall:

Geodesic Distances

\[ d_g (p, q) = \lim_{t \to 0} \sqrt{-4t \log k_{t,p}(q)} \]

“Varadhan’s Theorem”

“Geodesics in heat”
Crane, Weischedel, and Wardetzky; TOG 2013
Approximate Sinkhorn


Replace $K$ with heat kernel
Curious Observation

$W_1$  $W_2$

Similar problems, different algorithms
Flow-Based $W_2$

Critical theoretical idea, computationally challenging

$$W_2^2(\rho_0, \rho_1) = \inf_{\rho, v} \int \int_{M \times [0,1]} \frac{1}{2} \rho(x, t) \|v(x, t)\|^2 \, dx \, dt$$

s.t. $\nabla \cdot (\rho(x, t)v(x, t)) = \frac{\partial \rho(x,t)}{\partial t}$

$v(x, t) \cdot \hat{n}(x) = 0 \ \forall x \in \partial M$

$\rho(x, 0) = \rho_0(x)$

$\rho(x, 1) = \rho_1(x)$

Benamou & Brenier
“A computational fluid mechanics solution of the Monge-Kantorovich mass transfer problem”
\[ \langle V, W \rangle_\mu := \int \langle V(x), W(x) \rangle \, d\mu(x) \]
Aside:
Parallel to Information Geometry

- Consider set of distributions as a manifold
- **Tangent spaces** from advection
- **Geodesics** from displacement interpolation
Only Scratching the Surface

Topics in Optimal Transportation
Villani, 2003

Giant field in modern math
Many Other Approaches


Example: Semi-discrete transport
Derived Problems

\[ \min_{\mu \in \mathcal{P}(\Omega)} \sum_{i=1}^{N} \lambda_i W_p^p(\mu, \nu_i) \]

\[ \mathcal{P}(\Omega) \]

Wasserstein Barycenter [Agueh’11]

\[ \nu_1 \]

\[ \nu_2 \]

\[ \nu_3 \]

Slide courtesy M. Cuturi
Any (ML) problem involving a **KL** or **L2** loss between (parameterized) histograms or probability measures can be easily *Wasserstein-ized* if we can differentiate $W$ efficiently.
“Wasserstein Propagation for Semi-Supervised Learning” (Solomon et al.)

“Fast Computation of Wasserstein Barycenters” (Cuturi and Doucet)

Learning
Computational Applications

“Displacement Interpolation Using Lagrangian Mass Transport” (Bonneel et al.)

“An Optimal Transport Approach to Robust Reconstruction and Simplification of 2D Shapes” (de Goes et al.)

Morphing and registration
Computational Applications

“Earth Mover’s Distances on Discrete Surfaces” (Solomon et al.)

“Blue Noise Through Optimal Transport” (de Goes et al.)
Computational Applications

“Geodesic Shape Retrieval via Optimal Mass Transport” (Rabin, Peyré, and Cohen)

“Adaptive Color Transfer with Relaxed Optimal Transport” (Rabin, Ferradans, and Papadakis)

Vision and image processing
Optimal Transport

Justin Solomon
MIT, Spring 2017