Data-driven optimal transport

Esteban G. Tabak
NYU - Courant Institute of Mathematical Sciences

Giulio Trigila
Technische Universität München

Machine Learning Seminar, November 5 2013
The story in a picture
Optimal transport (Monge formulation)

Find a plan to transport material from one location to another that minimizes the total transportation cost. In terms of the probability densities $\rho(x)$ and $\mu(y)$ with $x, y \in \mathbb{R}^n$,

$$\min_{y(x)} M(y) = \int_{\mathbb{R}^n} c(x, y(x)) \rho(x) dx.$$  

subject to

$$\int_{y^{-1}(\Omega)} \rho(x) dx = \int_{\Omega} \mu(y) dy$$

for all measurable sets $\Omega$. If $y$ is smooth and one to one, this is equivalent to the point-wise relation

$$\rho(x) = \mu(y(x)) |J^y(x)|,$$
Kantorovich formulation

\[ \min_{\pi(x,y)} K(\pi) = \int c(x, y) \pi(x, y) dx dy \]

subject to

\[ \rho(x) = \int \pi(x, y) dy, \quad \mu(y) = \int \pi(x, y) dx \]

If \( c(x, y) \) is convex,

\[ \min M(y) = \max K(\pi) \]

and

\[ \pi(S) = \rho(\{x : (x, y(x)) \in S\}). \]
Dual formulation

Maximize

\[ D(u, v) = \int u(x)\rho(x)dx + \int v(x)\mu(y)dy \]

over all continuous functions \( u \) and \( v \) satisfying

\[ u(x) + v(y) \leq c(x, y). \]
The standard example: \( c(c, y) = \|y - x\|^2 \)

\[
\min \int c(x, y) \pi(x, y) dx dy \rightarrow \max \int x \cdot y \pi(x, y) dx dy,
\]

with dual

\[
\min \int u(x) \rho(x) dx + \int v(y) \mu(y) dy, \quad u(x) + v(y) \geq x \cdot y.
\]

\[
\begin{align*}
u(x) &= \max_y (x \cdot y - v(y)) \equiv v^*(x) \\
v(y) &= \max_x (x \cdot y - u(x)) \equiv u^*(y)
\end{align*}
\]

and the optimal map for Monge’s problem is given by

\[ y(x) = \nabla u(x), \]

with potential \( u(x) \) satisfying the Monge-Ampere equation

\[
\rho(x) = \mu(\nabla u(x)) \det(D^2 u)(x).
\]
A data-based formulation

In Kantorovich dual formulation, the objective function to maximize is a sum of two expected values:

\[ D(u, v) = \int u(x)\rho(x)dx + \int v(x)\mu(y)dy. \]

Hence, if \( \rho(x) \) and \( \mu(y) \) are known only through samples, it is natural to pose the problem in terms of empirical means:

Maximize

\[ D(u, v) = \frac{1}{m} \sum_{i=1}^{m} u(x_i) + \frac{1}{n} \sum_{j=1}^{n} v(y_j) \]

over functions \( u \) and \( v \) satisfying

\[ u(x) + v(y) \leq c(x, y). \]
A purely discrete reduction

Maximize

\[ D(u, v) = \frac{1}{m} \sum_{i=1}^{m} u_i + \frac{1}{n} \sum_{j=1}^{n} v_j \]

over vectors \( u \) and \( v \) satisfying

\[ u_i + v_j \leq c_{ij}. \]

This is dual to the *uncapacitated transportation problem*:

Minimize

\[ C(\pi) = \sum_{i,j} c_{ij} \pi_{ij} \]

subject to

\[ \sum_{j} \pi_{ij} = n \]

\[ \sum_{i} \pi_{ij} = m. \]
Fully constrained scenario

\[
D(u, v) = \frac{1}{m} \sum_{i=1}^{m} u(x_i) + \frac{1}{n} \sum_{j=1}^{n} v(y_j) = \int u(x)\rho(x)dx + \int v(x)\mu(y)dy
\]

with

\[
\rho(x) = \frac{1}{m} \sum_{i=1}^{m} \delta(x - x_i), \quad \mu(y) = \frac{1}{n} \sum_{j=1}^{n} \delta(y - y_j),
\]

which brings us back to the purely discrete case.
The Lagrangian

\[ L(\pi, u, v) = \int c(x, y) \pi(x, y) \, dx \, dy \]

\[ - \int \left[ \rho(x) - \frac{1}{m} \sum_{i=1}^{m} \delta(x - x_i) \right] u(x) \, dx \]

\[ - \int \left[ \mu(y) - \frac{1}{n} \sum_{j=1}^{n} \delta(y - y_j) \right] v(y) \, dy, \]

where \( \rho(x) \) and \( \mu(y) \) are shortcut notations for

\[ \rho(x) = \int \pi(x, y) \, dy \]

and

\[ \mu(y) = \int \pi(x, y) \, dx. \]
In terms of the Lagrangian, the problem can be formulated as a game:

\[ d : \max_{u(x), v(y)} \min_{\pi(x,y) \geq 0} \ L(\pi, u, v), \]

with dual

\[ p : \min_{\pi(x,y) \geq 0} \max_{u(x), v(y)} L(\pi, u, v). \]

If the functions \( u(x) \) and \( v(y) \) are unrestricted, the problem \( p \) becomes

\[ p : \min_{\pi(x,y) \geq 0} \int c(x, y) \ \pi(x, y) \ dx dy \]

subject to

\[ \rho(x) = \frac{1}{m} \sum_{i=1}^{m} \delta(x - x_i), \ \mu(y) = \frac{1}{n} \sum_{j=1}^{n} \delta(y - y_j), \]

as before.
A suitable relaxation

Instead, restrictict the space of functions $F$ from where $u$ and $v$ can be selected. If $F$ is invariant under dilations:

$$u \in F, \lambda \in R \rightarrow \lambda u \in F,$$

then the problem $p$ becomes

$$p : \min_{\pi(x,y) \geq 0} \int c(x, y) \pi(x, y) \, dx \, dy$$

$$\int \left[ \rho(x) - \frac{1}{m} \sum_{i=1}^{m} \delta(x - x_i) \right] u(x) \, dx = 0,$$

$$\int \left[ \mu(y) - \frac{1}{n} \sum_{j=1}^{n} \delta(y - y_j) \right] v(y) \, dy = 0$$

for all $u(x), v(y) \in F$, a weak formulation of the constraints.
Some possible choices for $F$

1. **Constants:** The constraints just guaranty that $\rho$ and $\mu$ integrate to one.

2. **Linear Functions:** The solution is a rigid displacement that moves the mean of $\{x_i\}$ into the mean of $\{y_j\}$.

3. **Quadratic Functions:** A linear transformation mapping the empirical mean and covariance matrix of $\{x_i\}$ into those of $\{y_j\}$.

4. **Smooth functions with appropriate local bandwidth,** neither to over nor under-resolve $\rho(x)$ and $\mu(y)$. 
A flow-based, primal-dual approach

\[ c = \frac{1}{2} \|x - y\|^2, \quad y(x) = \nabla u(x). \]

We think of the map \( y(x) \) as the endpoint of a time-dependent flow \( z(x, t) \), such that

\[ z(x, 0) = x, \quad \text{and} \quad z(x, \infty) = y(x). \]

\( z(x, t) \) follows the gradient flow

\[
\begin{cases}
\dot{z} = -\nabla_z \left[ \frac{\delta \tilde{D}_t}{\delta u} \right]_{u=\frac{1}{2}||x||^2} \\
z(x, 0) = x
\end{cases}
\]

where

\[ \tilde{D}_t = \int u(z) \rho_t(z) dz + \int u^*(y) \mu(y) dy \]

and \( \rho_t \) is the evolving probability distribution underlying the points \( z(x, t) \).
The variational derivative of $\tilde{D}$ adopts the form

$$\frac{\delta \tilde{D}_t}{\delta u} = \rho_t - |D^2 u| \mu(\nabla u)$$

which, applied at $u = \frac{1}{2} \| x \|^2$ (the potential corresponding to the identity map), yields the simpler expression

$$\frac{\delta \tilde{D}_t}{\delta u}(z) = \rho_t(z) - \mu(z),$$

so

$$\begin{cases}
\dot{z} = -\nabla_z [\rho_t(z) - \mu(z)] \\
z(x, 0) = x
\end{cases}$$
The probability density \( \rho_t \) satisfies the continuity equation

\[
\frac{\partial \rho_t(z)}{\partial t} + \nabla_z \cdot [\rho_t(z) \dot{z}] = 0,
\]

yielding

\[
\frac{\partial \rho_t(z)}{\partial t} - \nabla_z \cdot [\rho_t(z) \nabla_z (\rho_t(z) - \mu(z))] = 0,
\]

a nonlinear heat equation that converges to \( \rho_{t=\infty}(z) = \mu(z) \).
In terms of samples,

1. **Objective function:**

   \[
   \tilde{D}_t = \int u(z) \rho_t(z) dz + \int u^*(y) \mu(y) dy \rightarrow \frac{1}{m} \sum_i u(z_i) + \frac{1}{n} \sum_j u^*(y_j)
   \]

2. **Test functions:** General \( u(z) \in F \rightarrow \) localized \( u(z, \beta) \), with \( \beta \) a finite-dimensional parameter and appropriate bandwidth.

3. **Gradient descent:** \( -\dot{u} = \frac{\delta \tilde{D}_t}{\delta u}(z) = \rho_t(z) - \mu(z) \rightarrow \)

   \[
   -\dot{\beta} = \nabla_\beta \tilde{D}_t(z) = \frac{1}{m} \sum_i \nabla_\beta u(z_i) - \frac{1}{n} \sum_j \nabla_\beta u(y_j)
   \]
Left to discuss:

1. Form of $u(z, \beta)$, determination of bandwidth:

$$u(z, \beta) = \frac{1}{2} \|x\|^2 + \sum_k \beta_k F_k \left( \frac{\|z - z_k\|}{\alpha_k} \right),$$

$$\alpha_k \propto \left( \frac{1}{\rho_t(z_k)} \right)^{\frac{1}{d}}.$$

2. Enforcement of the map's optimality.
Some applications and extensions

1. Maps: fluid flow from tracers, effects of a medical treatment, resource allocation, ...
2. Density estimation: $\rho(x)$
3. Classification and clustering: $\rho_k(x)$
4. Regression: $\pi(x, y)$, $\text{dim}(x) \neq \text{dim}(y)$,
   
   $$c(x, y) = \sum_k w_k \frac{\|y - y_k\|^2}{\alpha_k^2 + \|x - x_k\|^2}$$

5. Determination of worst scenario under given marginals: $\pi(x_1, x_2, \ldots, x_n)$
Thanks!