$L_1$ optimal transport with applications

Stanley Osher

February 17, 2018

Joint work with Wilfrid Gangbo, Wuchen Li, Ernest Ryu, Penghang Yin and Wotao Yin.

Partially supported by ONR and DOE.
Motivation

The metric among histograms (images intensities) is a key concept in many applications. E.g.

The theory of optimal transport provides a useful metric, which has been used in

- Image processing: Color transferring; Image segmentation; Image repairing, e.t.c.;
- Machine learning; Domain Adaptation;
- Shape optimization;
- Game theory, including Mean field games.
Introduction: Earth Mover’s distance (EMD)

Question: What is the optimal way to move (transport) some dirt with shape $X$, density $\rho^0(x)$ to another shape $Y$ with density $\rho^1(y)$?

The question leads to the definition of Earth Mover’s distance (also named Wasserstein metric, Monge-Kantorovich problem). It links to an $L_1$ minimization problem, which is highly related to the ones in image processing and compressed sensing.
EMD defines a metric on the probability space of a convex, compact set \( \Omega \subset \mathbb{R}^d \):

\[
EMD(\rho^0, \rho^1) := \inf_{\pi} \int_{\Omega \times \Omega} d(x, y)\pi(x, y) \, dx \, dy
\]

where \( d : \Omega \times \Omega \to \mathbb{R} \) is a given cost function, and the infimum is taken among all joint measures having \( \rho^0(x) \) and \( \rho^1(y) \) as marginals, i.e.

\[
\int_{\Omega} \pi(x, y) \, dy = \rho^0(x) , \quad \int_{\Omega} \pi(x, y) \, dx = \rho^1(y) , \quad \pi(x, y) \geq 0 .
\]

In this talk, we will present fast algorithms for EMD with two different choices of \( d \):

\[
d(x, y) = \|x - y\|_2 \quad \text{(Euclidean)} \quad \text{or} \quad \|x - y\|_1 \quad \text{(Manhattan)} .
\]
Introduction: Optimal transport

In fact, the above linear optimization (named Kantorovich problem) is a relaxed problem considered by Monge in 1781:

$$\inf_T \int_{\Omega} d(x, T(x)) \rho^0(x) dx$$

where the infimum is among all transport maps $T$, which transfers $\rho^0(x)$ to $\rho^1(x)$. E.g.

(a) $\rho^0$. (b) $\rho^1$. (c) Map $T$. 

5
Recall that the distance function can be formulated into an optimal control problem:

\[ d(x, T(x)) = \inf_{\gamma} \left\{ \int_0^1 \| \dot{\gamma}(t) \| dt : \gamma(0) = x, \gamma(1) = T(x) \right\}, \]

where \( \| \cdot \| \) is 1 or 2-norm. So the Monge problem can be reformulated into a fluid dynamics setting (Briener-Benamou 2000):

\[ \inf_{m, \rho} \int_0^1 \int_{\Omega} \| m(t, x) \| dx dt \]

where \( m(t, x) \) is a flux function satisfying zero flux condition, such that \( \rho^0 \) is transported to \( \rho^1 \) continuously in time, i.e.

\[ \frac{\partial \rho}{\partial t} + \nabla \cdot m = 0. \]

The above problem has many minimizers. One is

\[ \rho(t, x) = (1 - t)\rho^0(x) + t\rho^1(x). \]
Introduction: $L_1$ minimization

Thus EMD is equivalent to the following minimal flux formulation: problem:

$$\inf_m \left\{ \int_\Omega \| m(x) \| dx : \nabla \cdot m(x) + \rho^1(x) - \rho^0(x) = 0 \right\} .$$

It is an $L_1$ minimization, whose minimizer solves the following PDE pairs (Evans and Gangbo 1999)

$$\begin{cases} m(x) = a(x) \nabla \Phi(x) , \\ \| \nabla \Phi(x) \| = 1 , \\ \nabla \cdot (a(x) \nabla \Phi(x)) + \rho^1(x) - \rho^0(x) = 0 . \end{cases}$$

One may recover the optimal map $T$ by $m$ (under suitable conditions on $\rho^0$ and $\rho^1$):

$$\frac{d}{dt} z(t) = \frac{m(z)}{(1-t)\rho^0(z) + t\rho^1(z)} , \quad z(0) = x , \quad z(1) = T(x) .$$
From numerical purposes, the minimal flux formulation has two benefits

- The dimension is much lower than the one in the original linear optimization problem.
- It is an $L_1$-type minimization problem, which shares its structure with many problems in compressed sensing and image processing. We borrow a very fast and simple algorithm used there to solve it.
Current methods

Linear programming

P: Many tools;
C: Involves quadratic number of variables and does not use the structure of $L_1$ minimization.

Alternating direction method of multipliers (ADMM) \(^1\)

P: Fewer iterations;
C: Solves an inverse Laplacian at each iteration; Not easy to parallelize.

In this talk, we apply a Primal-Dual method (Chambolle and Pock).

\(^1\)(Benamou et.al, 2014), (Benamou et.al, 2016), (Solomon et.al, 2014)
Settings

Consider a uniform grid $G = (V, E)$ with spacing $\Delta x$ to discretize the spatial domain, where $V$ is the vertex set and $E$ is the edge set. $i = (i_1, \cdots, i_d) \in V$ represents a point in $\mathbb{R}^d$.

Consider a discrete probability set supported on all vertices:

$$\mathcal{P}(G) = \{(p_i)_{i=1}^N \in \mathbb{R}^n \mid \sum_{i=1}^N p_i = 1, \ p_i \geq 0, \ i \in V\},$$

and a discrete flux function defined on the edge of $G$:

$$m_{i+\frac{1}{2}} = (m_{i+\frac{1}{2}}e_v)_v^d,$$

where $m_{i+\frac{1}{2}}e_v$ represents a value on the edge $(i, i + e_v) \in E$, $e_v = (0, \cdots, \Delta x, \cdots, 0)^T$, $\Delta x$ is at the $v$-th column.
Minimization: Euclidean distance

We first consider EMD with the Euclidean distance. The discretized problem forms

$$\begin{align*}
\text{minimize} & \quad \|m\| = \sum_{i=1}^{N} \|m_i + \frac{1}{2}\|_2 \\
\text{subject to} & \quad \text{div}_G(m) + p^1 - p^0 = 0,
\end{align*}$$

which can be formulated explicitly

$$\begin{align*}
\text{minimize} & \quad \sum_{i=1}^{N} \sqrt{\sum_{v=1}^{d} |m_i + \frac{1}{2} e_v|^2} \\
\text{subject to} & \quad \frac{1}{\Delta x} \sum_{v=1}^{d} (m_i + \frac{1}{2} e_v - m_i - \frac{1}{2} e_v) + p^1_i - p^0_i = 0.
\end{align*}$$
Primal-dual algorithm

We solve the minimization problem by looking at its saddle point structure. Denote \( \Phi = (\Phi_i)_{i=1}^N \) as a Lagrange multiplier:

\[
\min_m \max_{\Phi} \|m\| + \Phi^T (\text{div}_G(m) + p^1 - p^0) .
\]

The iteration steps are as follows (using Chambolle and Pock):

\[
\begin{aligned}
m^{k+1} &= \arg \min_m \|m\| + (\Phi^k)^T \text{div}_G(m) + \frac{\|m-m^k\|_2^2}{2\mu} ; \\
\Phi^{k+1} &= \arg \max_{\Phi} \Phi^T \text{div}_G(m^{k+1} + \theta(m^{k+1} - m^k) + p^1 - p^0)
\end{aligned}
\]

where \( \mu, \tau \) are two small step sizes, \( \theta \in [0, 1] \) is a given parameter. These steps are alternating a gradient ascent in the dual variable \( \Phi \) and a gradient descent in the primal variable \( m \).
The primal-dual iteration can be solved by simple explicit formulae. First, notice

\[ \min_m \|m\| + (\Phi^k)^T \text{div}_G(m) + \frac{\|m - m^k\|_2^2}{2\mu} \]

\[= \sum_{i=1}^{N} \min_{m_{i+1/2}} \left( \|m_{i+1/2}\|_2 - (\nabla_G \Phi^k_{i+1/2} + \frac{1}{2})^T m_{i+1/2} + \frac{1}{2\mu} \|m_{i+1/2} - m^k_{i+1/2}\|_2^2 \right), \]

where \(\nabla_G \Phi^k_{i+1/2} := \frac{1}{\Delta x} (\Phi^k_{i+e_v} - \Phi^k_{i})_{v=1}^d\). So the first update in the algorithm becomes

\[m^k_{i+1/2} = \text{shrink}_2(m^k_{i+1/2} + \mu \nabla_G \Phi^k_{i+1/2}, \mu), \]

where we define

\[\text{shrink}_2(y, \alpha) := \frac{y}{\|y\|_2} \max\{\|y\|_2 - \alpha, 0\}, \quad \text{where } y \in \mathbb{R}^d.\]
Second, consider

\[
\max_{\Phi} \Phi^T \text{div}_G (m^{k+1} + \theta (m^{k+1} - m^k) + p^1 - p^0) - \frac{\|\Phi - \Phi^k\|_2^2}{2\tau}
\]

\[
= \sum_{i=1}^{N} \max_{\Phi_i} \{ \Phi_i [\text{div}_G (m_i^{k+1} + \theta (m_i^{k+1} - m_i^k)) + p_i^1 - p_i^0] - \frac{\|\Phi_i - \Phi_i^k\|_2^2}{2\tau} \} .
\]

Thus the second update in the algorithm becomes

\[
\Phi_i^{k+1} = \Phi_i^k + \tau \{ \text{div}_G (m_i^{k+1} + \theta (m_i^{k+1} - m_i^k)) + p_i^1 - p_i^0 \} .
\]
Algorithm

Primal-dual method for EMD-Euclidean metric

1. for $k = 1, 2, \cdots$ Iterates until convergence
2. $m_{i+\frac{1}{2}}^{k+1} = \text{shrink}_2(m_{i+\frac{1}{2}}^{k} + \mu \nabla_G \Phi_{i+\frac{1}{2}}^{k}, \mu)$;
3. $\Phi_{i}^{k+1} = \Phi_{i}^{k} + \tau \{ \text{div}_G(m_{i}^{k+1} + \theta(m_{i}^{k+1} - m_{i}^{k})) + p_{i}^{1} - p_{i}^{0} \}$;
4. End
Minimization: Manhattan distance

We consider EMD with the Manhattan distance:

\[
\inf_{\pi}\int_{\Omega \times \Omega} \|x - y\|_1 \pi(x, y) \, dx \, dy : \pi \text{ has marginals } \rho^0, \rho^1.
\]

Similarly, this problem is equivalent to

\[
\inf_{\mathbf{m}} \int_{\Omega} \|\mathbf{m}(x)\|_1 \, dx : \nabla \cdot \mathbf{m}(x) + \rho^1(x) - \rho^0(x) = 0.
\]

It is an exact \(L_1\) minimization problem, which can be solved even faster.
Minimization: Manhattan distance

The discretized problem forms

\[
\begin{align*}
\text{minimize} & \quad \|m\| = \sum_{i=1}^{N} \|m_i + \frac{1}{2}\|_1 \\
\text{subject to} & \quad \text{div}_G(m) + p^1 - p^0 = 0.
\end{align*}
\]

The optimization is not strictly convex, which means it may have multiple minimizers. We consider a quadratic modification: for a small \(\epsilon > 0\),

\[
\begin{align*}
\text{minimize} & \quad \|m\| + \frac{\epsilon}{2}\|m\|_2^2 = \sum_{i=1}^{N} \|m_i + \frac{1}{2}\|_1 + \frac{\epsilon}{2}\sum_{i=1}^{N} \|m_i + \frac{1}{2}\|_2^2 \\
\text{subject to} & \quad \text{div}_G(m) + p^1 - p^0 = 0.
\end{align*}
\]

We solve it similarly by Primal-Dual algorithm (Chambolle and Pock).
Algorithm

Prime-dual method for EMD-Manhattan distance

1. for $k = 1, 2, \cdots$ Iterates until convergence
2. $m_{i+e_v}^{k+1} = \frac{1}{1+\epsilon \mu} \text{shrink}(m_{i+e_v}^k + \mu \nabla_G \Phi_{i+e_v}^k, \mu);$ 
3. $\Phi_{i}^{k+1} = \Phi_{i}^k + \tau \{\text{div}_G (m_{i+1}^{k+1} + \theta (m_{i}^{k+1} - m_{i}^k)) + p_{i}^1 - p_{i}^0\};$
4. End

Here we use the conventional shrink operator for the Manhattan metric:

$$\text{shrink}(y, \alpha) := \frac{y}{|y|} \max\{|y| - \alpha, 0\}, \quad \text{where } y \in \mathbb{R}^1.$$
Comparison

<table>
<thead>
<tr>
<th>Grids size</th>
<th>EMD CUDA</th>
<th>EMD CPU</th>
<th>Ling</th>
<th>Pele</th>
</tr>
</thead>
<tbody>
<tr>
<td>32 × 32</td>
<td>0.012s</td>
<td>0.08s</td>
<td>0.007s</td>
<td>2.74s</td>
</tr>
<tr>
<td>64 × 64</td>
<td>0.063s</td>
<td>0.9s</td>
<td>0.009s</td>
<td>N/A</td>
</tr>
<tr>
<td>128 × 128</td>
<td>0.336s</td>
<td>12.9s</td>
<td>2.3s</td>
<td>N/A</td>
</tr>
<tr>
<td>256 × 256</td>
<td>6.8s</td>
<td>245.5s</td>
<td>80.8s</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Table: Runtime of algorithms.

In Table, we compare the wall-clock runtime of the parallel EMD algorithm with other state of art methods. Our method is faster whenever the number of nodes are large.
Optimal flux I

(d) EMD with Euclidean distance.
(e) EMD with Manhattan distance.
Optimal flux II

(f) EMD with Euclidean distance.

(g) EMD with Manhattan distance.
Importance of $\epsilon$

(h) $\epsilon = 0$.

(i) $\epsilon = 0$.

(j) $\epsilon$ small.

Two different minimizers above on left are for $\epsilon = 0$. 
Importance of $\epsilon$

It is worth mentioning that the minimizer of $\epsilon$ regularized problem

$$\inf_{m} \left\{ \int_{\Omega} \|m(x)\| + \frac{\epsilon}{2} \|m(x)\|^2 \, dx : \nabla \cdot m(x) + \rho^1(x) - \rho^0(x) = 0 \right\},$$

satisfies a nice (formal) system

$$\begin{align*}
m(x) &= \frac{1}{\epsilon} \left( \nabla \Phi(x) - \frac{\nabla \Phi(x)}{|\nabla \Phi(x)|} \right), \\
\frac{1}{\epsilon} \left( \Delta \Phi(x) - \nabla \cdot \frac{\nabla \Phi(x)}{|\nabla \Phi(x)|} \right) &= \rho^0(x) - \rho^1(x),
\end{align*}$$

where the second equation holds when $|\nabla \Phi| \geq 1$.

Notice that the term $\nabla \cdot \frac{\nabla \Phi(x)}{|\nabla \Phi(x)|}$ relates to the mean curvature formula.
Introduction: Unbalanced optimal transport

The original problem assumes that the total mass of given densities should be equal, which often does not hold in practice. E.g. the intensities of two images can be different.

Partial optimal transport seeks optimal plans between two measures $\rho^0$, $\rho^1$ with unbalanced masses, i.e.

$$\int_{\Omega} \rho^0(x)dx \neq \int_{\Omega} \rho^1(y)dy.$$

\[ (k) \quad \rho^0. \]

\[ (l) \quad \rho^1. \]
Unbalanced optimal transport

A particular example is the weighted average of Earth Mover’s metric and $L_1$ metric, known as Kantorovich-Rubinstein norm. One important formulation is

$$\inf_{u,m} \left\{ \int_{\Omega} \| m(x) \| dx : \nabla \cdot m(x) + \rho^0(x) - u(x) = 0, \quad 0 \leq u(x) \leq \rho^1(x) \right\}.$$  

Our method can solve the problem by 3 line codes.
Primal-dual method for Partial optimal transport

**Input:** Discrete probabilities $p^0, p^1$;
Initial guess of $m^0$, parameter $\epsilon > 0$, step size $\mu, \tau, \theta \in [0, 1]$.

**Output:** $m$ and $\|m\|$.

1. for $k = 1, 2, \cdots$ Iterates until convergence
2. $m_{i+\frac{e_y}{2}}^{k+1} = \frac{1}{1+\epsilon \mu} \text{shrink}(m_{i+\frac{e_y}{2}}^k + \mu \nabla \Phi_i^k, \mu)$ ;
3. $u_{i}^{k+1} = \text{Proj}_{C_i}(u_i^k - \mu \Phi_i^k)$ ;
4. $\Phi_i^{k+1} = \Phi_i^k + \tau \{ \text{div}(2m_{i}^{k+1} - m_i^k) + 2u_{i}^{k+1} - u_i^k) - p_i^0 \} $ ;
5. End
Examples: Optimal flux

(m) Euclidean distance.

(n) Manhanttan distance.

Figure: Unbalanced transportation from three delta measures concentrated at two points (red) to five delta measures (blue).
Examples

Figure: Hand written digit images and the computed distances between the top left image and the rest.
Discussion

Our method for solving $L_1$ Monge-Kantorovich related problems

- handles the sparsity of histograms;
- has simple updates and is easy to parallelize;
- introduces a novel PDE system (Mean curvature formula in Monge Kantorovich equation).
Main references

Antonin Chambolle and Thomas Pock.  
A first-order primal-dual algorithm for convex problems with applications to imaging, 2011.

Lawrence Evans and Wilfrid Gangbo.  

Wuchen Li, Ernest Ryu, Stanley Osher, Wotao Yin and Wilfrid Gangbo.  
A parallel method for Earth Mover’s distance, 2016.

Wuchen Li, Penghang Yin and Stanley Osher.  
A Fast algorithm for unbalanced $L_1$ Monge-Kantorovich problem.
Questions