Algorithmic Problems in Optimal Transportation

Justin Solomon
justin.solomon@stanford.edu

September 12, 2014 (USC Theory Group Lunch)

In this document, I outline a few challenging algorithmic problems arising in my research on optimal transportation. I have attempted to pose these problems in "discrete" language. It easily could be the case that they admit straightforward solutions—which is great news!—or that you can do little more than the obvious algorithm. Any and all guidance from members of the theory community is welcome.

1 Earth Mover's Distances

I will begin with a review of the *earth mover's distance* (EMD) between discrete probability distributions. This distance is the basic tool for many techniques in vision, graphics, and learning; efficient optimization of this distance wider and higher-dimensional applications of this machinery.

We will denote the n-dimensional probability simplex as $\mathcal{S}_n$, defined as follows:

$$\mathcal{S}_n \equiv \{x \in [0,1]^n : \mathbb{1}^\top x = 1\}.$$

Suppose we are given two distributions $p, q \in \mathcal{S}_n$ as well as a matrix of distances between bins $D \in (\mathbb{R}^+)^{n \times n}$. Then, the *earth mover's distance* between p and q is given by the optimal value of the following linear program:

$$\begin{array}{rll} \text{EMD}_D(p,q) \equiv \min_{T \in \mathbb{R}^{n \times n}} & \sum_{ij} D_{ij} T_{ij} \\ \text{such that} & T\mathbb{1} = p \\ & T^\top \mathbb{1} = q \\ & T \geq 0. \end{array}$$

The matrix T is known as a *transportation matrix* and should be thought of as the amount of mass moved from bin i of p to bin j of q. This problem is nothing more than an instance of bipartite matching or multi-commodity flow. See [3] for the paper that coined the term "earth mover's distance" and for a proof that it satisfies the triangle inequality under certain conditions on D.

The theory of *optimal transportation* deals with a continuous analog of EMD for distributions p, q on $\mathbb{R}^n$ or on a manifold. Of particular interest is the *Wasserstein distance* $\mathcal{W}_r$, defined by taking $D(x,y)$ to be the r-th power of the geodesic distance between x and y; the r-th root of this quantity is a distance for any $r \geq 1$.

EMD can be computed in polynomial time using any of a number of well-known techniques. To maximize efficiency of methods in this domain, however, we will ask questions of the following flavor:

- If we know D is structured (e.g. contains pairwise Euclidean distances between points or shortest-path distances on a graph), can we evaluate EMD_D more efficiently?

- How can EMD be incorporated into larger optimization problems?
- What conditions on D can help characterize the behavior of EMD?

Each of these questions leads to challenging algorithmic and mathematical problems whose resolution may have practical impact.

2 Graph EMD with Quadratic Ground Distance

Suppose we are given a (connected) graph $G = (V, E)$ and take $n = |V|$. Then, we can think of $p, q \in \mathcal{S}_n$ as distributions over the vertices of G.

If D_{vw} for $v, w \in V$ is given by the shortest-path distance between v and w along G, then EMD can be evaluated using an alternative linear program:

$$\mathcal{W}_1(p,q) = \begin{array}{ll} \min_{F \in \mathbb{R}^{|E|}} & \|F\|_1 \\ \text{such that} & p_v - \left[\sum_{(v \to w) \in E} F_{v \to w}\right] + \left[\sum_{(w \to v) \in E} F_{w \to v}\right] = q_v \; \forall v \in V. \end{array}$$

We denote this distances as $\mathcal{W}_1$, since it is analogous to the 1-Wasserstein distance in the continuous theory. This linear program puts one signed flow F_e on each directed edge $e = (v \to w) \in E$. The objective is the total amount of flow along the entire graph, and the constraints specify that F transforms p into q. This alternative formulation is preferable for many applications because the number of unknowns scales with $|E|$ rather than with $|V|^2$. This change can represent considerable savings when the graph is sparse, that is, $|E| \ll |V|^2$.

The theory of Wasserstein distances, however, is best understood when D contains ground distances *squared*. In our case, this would imply the form $D_{vw} = d(v, w)^2$, if $d(\cdot, \cdot)$ represents shortest-path distances along G. This new optimization problem is more likely to be strictly convex, ensuring stronger uniqueness for the transportation matrix and related properties.

Squaring the ground distance, however, loses the simplified flow-based formulation discussed above. This leads to our first question:

> **Can EMD with ground distance that is *quadratic* in graph distance be computed in a way that scales like $|E|$?**

It may be advisable to approach this problem first by assuming a particular form for G, e.g. star or line graphs. Even these cases admit nontrivial structure.

3 Spectral/Hodge Approximation of $\mathcal{W}_1$

Let us return to the case of computing EMD when D contains shortest-path distances (*not* squared) on a graph. We can define a difference operator $\nabla \in \{-1, 0, 1\}^{|E| \times n}$ as follows:

$$\nabla_{ev} \equiv \begin{cases} -1 & \text{when } e = (v \to w) \\ 1 & \text{when } e = (w \to v) \\ 0 & \text{otherwise.} \end{cases}$$

This matrix allows us to simplify notation considerably:

$$\mathcal{W}_1(p,q) = \begin{array}{ll} \min_{F \in \mathbb{R}^{|E|}} & \|F\|_1 \\ \text{such that} & p - q = \nabla^\top F. \end{array}$$

As a differential geometer, my intuition for ∇ is that it acts as a *discrete gradient* operator. That is, if $g \in \mathbb{R}^n$ is a per-vertex "function," then ∇g is its per-edge gradient. If this analogy carries through, the transpose $\nabla^\top$—or adjoint, in continuous language—can be thought of as a *discrete divergence* operator acting on vector fields. Notice that unlike the case of vector fields on $\mathbb{R}^n$, it is not clear what the proper operator should be to measure curl on a graph.

In [5], we are able to approximate EMD on triangulated surfaces using a continuous version of the divergence-based formulation explained in the previous paragraphs. Our approximation has the favorable property of satisfying the triangle inequality even after aggressive approximation. It would be interesting if we can carry out a similar trick on graphs.

To approximate EMD, we use the *Hodge decomposition* of a vector field. This decomposition shows that any vector field on a surface can be written as a sum of three parts: a divergence-free part, a curl-free part, and a harmonic part (for surfaces with holes). This decomposition plays a fundamental role in the differential topology of smooth manifolds (search term: "De Rham cohomology").

Returning to the discrete problem, using simple linear algebra arguments rather than differential topology, we can always write F as:

$$F = \nabla f + G,$$

where $f \in \mathbb{R}^n$, $G \in \mathbb{R}^{|E|}$, and $\nabla^\top G = 0$. Then, we can simplify the constraint for $\mathcal{W}_1$ by substitution:

$$\nabla^\top F = \nabla^\top (\nabla f + G) = \nabla^\top \nabla f + 0 = \Delta f,$$

where we have defined $\Delta \in \mathbb{R}^{n \times n}$ to be the graph Laplacian $\Delta \equiv \nabla^\top \nabla$.

Substituting this formula back changes our optimization somewhat:

$$\mathcal{W}_1(p,q) = \begin{array}{rl} \min_{f \in \mathbb{R}^n, G \in \mathbb{R}^{|E|}} & \|G + \nabla f\|_1 \\ \text{such that} & p - q = \Delta f \\ & \nabla^\top G = 0. \end{array}$$

An interesting thing happens here: The matrix Δ is invertible (up to constant shift), so we can compute the unknown variable f using matrix (pseudo-)inverses as:

$$f = \Delta^+ (p - q).$$

This linear solve can be carried out very efficiently using either direct sparse solvers or an iterative method like conjugate gradients. Hence, we only need to optimize with respect to G.

This leads to our first question in this section:

Is the optimization with respect to G any easier than the original problem?

Beyond this point, the connection to continuous mathematics begins to degrade. On a manifold, we can write G in a spectral basis for divergence-free vector fields. This basis consists of large circulating vector fields (corresponding to low eigenvalues) followed by vector fields with smaller and smaller eddies. Writing G in such a basis allows us to remove the constraint altogether and solve an L_1 optimization problem. The next question is:

Is there a spectral basis for curl-free edge vector fields G (satisfying $\nabla^\top G = 0$)?

4 Advective Models for $\mathcal{W}_2$

Now, we return to the case of quadratic ground distances.

Suppose M is a manifold or subset of $\mathbb{R}^k$. Then, given two probability distribution functions $\rho_0, \rho_1 \in \text{Prob}(M)$, there are two completely equivalent but contrasting ways to compute the quadratic Wasserstein distance $\mathcal{W}_2$. The essentially mimics the linear programming formulation above. The second, introduced by Benamou and Brenier in [1], is to use a PDE-based formulation:

$$\mathcal{W}_2(\rho_0, \rho_1) = \begin{array}{rl} \min_{\rho(x,t), J(x,t)} & \frac{1}{2}\int_0^1 \int_M \frac{\|J(x,t)\|^2}{\rho(x,t)}\, dx\, dt \\ \text{such that} & \frac{\partial \rho}{\partial t} = \nabla \cdot J \\ & \rho(\cdot, 0) = \rho_0 \\ & \rho(\cdot, 1) = \rho_1 \\ & \rho(\cdot, t) \in \text{Prob}(M)\ \forall t \in [0,1]. \end{array}$$

This convex optimization intuitively can be thought of as follows:

- $\rho(x,t)$ is a probability distribution over M at each time t. At time $t = 0$, $\rho = \rho_0$ and at time $t = 1$, $\rho = \rho_1$. It should be identified with *mass*.
- $J(x,t)$ is a vector field on M that can change with time that *advects* ρ from ρ_0 to ρ_1 (the first constraint). It should be thought of as the *momentum* of ρ as it moves along M.
- Over all the possible ways to advect from ρ_0 to ρ_1, we wish to choose the one that does so with minimal *work* (momentum squared divided by mass—equivalent to "$\frac{1}{2}mv^2$").

The original formulation in [1] uses slightly different variables by writing the problem above in terms of mass ρ and velocity $v \equiv J/\rho$.

This formulation can be more easy to work with in a computational setting, because it scales linearly instead of quadratically in M (but now we must subdivide time $t \in [0,1]$ as well!). It also has led to important developments in the theory of optimal transportation. Hence, an interesting research direction might be the following:

Does a similar advective distance exist for distributions over graph vertices? Is it computable without approximation?

We provide a partial resolution in [4], which works backward from Benamou and Brenier's formulation to a transportation distance on $\mathcal{S}_n$ but in doing so introduces a continuous time variable that must be discretized (and hence approximated) using numerical ODE.

5 Barycenters and Propagation Problems

Our final problems involve optimizations in which EMD is incorporated into a larger energy functional. Suppose we are given a set of distributions $p_1, p_2, \ldots, p_m \in \mathcal{S}_n$. Then, the *barycenter problem* can be defined as:

$$\min_{p \in \mathcal{S}_n} \quad \sum_k \text{EMD}(p, p_k).$$

See [2] for a recent paper with a continuous optimization method for computing these barycenters.

Our paper [6] uses a similar optimization for a semi-supervised learning problem. Suppose $G = (V, E)$ is a graph with a given subset of vertices $V_0 \subseteq V$. For each $v \in V_0$, we are given a probability vector (over some other space with a fixed ground distance matrix D) $p_0(v)$. Then, this paper proposes the following optimization for filling in the missing distributions associated with vertices in $V \backslash V_0$:

$$\begin{array}{rl} \min_{p(v): V \to \mathcal{S}_n} & \sum_{(v,w) \in E} \text{EMD}_D(p(v), p(w)) \\ \text{such that} & p(v) = p_0(v)\ \forall v \in V_0 \\ & p(v) \in \mathcal{S}_n\ \forall v \in V. \end{array}$$

Our paper focuses on properties of this optimization in a continuous context, in particular characterizing the means, variances, and sparsities of the interpolated p's.

These large-scale optimizations contain EMD as a subproblem and hence can scale very poorly! For this reason, we simultaneously would like to understand the best optimization techniques as well as what happens at the minima of the two energies above:

> **What is the fastest way to solve the barycenter and semi-supervised learning problems above? Can the speed be improved if we know more about the pairwise distance matrix D in the formula for EMD?**
>
> **Is it possible to solve the barycenter problem in time proportional to $|V|$ rather than $|E|$ or $|V|^2$? For this question, I am concerned with scaling in the "source" graph size rather than the number of bins n (which may not equal either $|V|$ or $|E|$).**
>
> **Can the sparsity of the computed p's be bounded by the sparsity of the fixed p_k's (or $p_0(v)$'s)?**

The last problem is likely to require some conditions on the matrix D in the optimization for EMD.

References

[1] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[2] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *Proc. ICML*, pages 685–693, 2014.

[3] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40(2):99–121, Nov. 2000.

[4] J. Solomon, R. Rustamov, L. Guibas, and A. Butscher. Continuous-flow transportation distances. *Under review*.

[5] J. Solomon, R. Rustamov, L. Guibas, and A. Butscher. Earth mover's distances on discrete surfaces. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 33(4):67:1–67:12, July 2014.

[6] J. Solomon, R. Rustamov, G. Leonidas, and A. Butscher. Wasserstein propagation for semi-supervised learning. In *Proc. ICML*, pages 306–314, 2014.