Wasserstein Soft Label Propagation on Hypergraphs: Algorithm and Generalization Error Bounds

Tingran Gao, Shahab Asoodeh, Yi Huang, James Evans
The University of Chicago
trg17@uchicago.edu, shahab@uchicago.edu, yhuang10@uchicago.edu, jevans@uchicago.edu

Abstract

Inspired by recent interests of developing machine learning and data mining algorithms on hypergraphs, we investigate in this paper the semi-supervised learning algorithm of propagating "soft labels" (e.g. probability distributions, class membership scores) over hypergraphs, by means of optimal transportation. Borrowing insights from Wasserstein propagation on graphs [Solomon et al. 2014], we re-formulate the label propagation procedure as a message-passing algorithm, which renders itself naturally to a generalization applicable to hypergraphs through Wasserstein barycenters. Furthermore, in a PAC learning framework, we provide generalization error bounds for propagating one-dimensional distributions on graphs and hypergraphs using 2-Wasserstein distance, by establishing the *algorithmic stability* of the proposed semi-supervised learning algorithm. These theoretical results also shed new lights upon deeper understandings of the Wasserstein propagation on graphs.

Introduction

Recent decades have witnessed a growing interest in developing machine learning and data mining algorithms on *hypergraphs* (Zhou, H., and Schölkopf 2007; Jost and Mulas 2018; Bulò and Pelillo 2009; Li and Ramchandran 2015; Li and Milenkovic 2017; Hein et al. 2013; Huang, Zhang, and Yu 2015). As a natural generalization of graphs, a hypergraph is a combinatorial structure consisting of vertices and hyperedges, where each hyperedge is allowed to connect any number of vertices. This additional flexibility facilitates capturing higher order interactions among objects; applications have been found in many fields such as computer vision (Govindu 2005), network clustering (Demir, Aykanat, and Cambazoglu 2008), folksonomies (Ghoshal et al. 2009), cellular networks (Klamt, Haus, and Theis 2009), and community detection (Kim, Bandeira, and Goemans 2018).

This paper develops a probably approximately correct (PAC) learning framework for *soft label propagation* or *Wasserstein propagation* (Solomon et al. 2014), a recently proposed semi-supervised learning algorithm based on optimal transport (Villani 2003; 2008), on graphs and hypergraphs. Different from the prototypical semi-supervised learning algorithm of *label propagation* (Belkin, Matveeva, and Niyogi 2004), in which labels of interest are typically numerical or categorical variables, Wasserstein propagation aims at inferring unknown *soft labels*, such as histograms or probability distributions, from known ones, based on pairwise similarities qualitatively characterized by edge connectivity and quantitatively measured using Wasserstein distances. Compared with traditional "hard labels," soft labels are built with extra flexibility and informativeness, rendering themselves naturally to applications where uncertainty or distributional information is crucial. For instance, the traffic density at routers in the Internet network or topic distributions in the co-authorship network are more naturally modeled as probability distributions.

Briefly speaking, semi-supervised learning is a paradigm that leverages unlabelled data to improve the generalization performance for supervised learning, under generic, unsupervised structural assumptions (e.g. the manifold assumption) on the dataset; see (Seeger 2001; Zhu 2008; Chapelle, Schölkopf, and Zien 2006) for an overview. Given a graph $G = (V, E)$ and a subset of vertices $V_0 \subset V$, label propagation is the procedure of extending an assignment of labels on V_0, denoted as a map $f_0 : V_0 \to \mathcal{D}$ valued in an arbitrary set $\mathcal{D}$, to a map $f : V \to \mathcal{D}$ on the entire vertex set V. Borrowing an analogy with classical heat equations, this extension procedure is reminiscent of heat propagation from "boundary" V_0 to the "entire domain" V. For soft label propagation, the label set $\mathcal{D}$ is the probability distributions $\mathcal{P}(N)$ modeled on a complete separable metric space (N, d_N).

Among the first works addressing semi-supervised learning with soft labels are (Corduneanu and Jaakkola 2005; Tsuda 2005; Subramanya and Bilmes 2011). In all these works, the similarity between two soft labels is quantitatively measured using the Kullback-Leibler (KL) divergence, which often incurs instability and discontinuity in the inferred soft labels. In (Solomon et al. 2014) the authors proposed to replace the KL divergence with 1- or 2-Wasserstein distance. The resulting soft label propagation algorithm is thus termed "Wasserstein propagation." Specifically, given a measure-valued map $f_0 : V_0 \to \mathcal{P}(N)$ defined on $V_0 \subset V$, Wasserstein propagation extends f_0 to $f : V \to \mathcal{P}(N)$ by solving the variational problem

$$\min_{f:V\to\mathcal{P}(N)} \sum_{(v,w)\in E} W_p^p\left(f(v), f(w)\right) \tag{1}$$

subject to the constraint $f \restriction V_0 = f_0$. Here $W_p(\mu, \nu)$ denotes the p-Wasserstein distance between probability distributions $\mu, \nu \in \mathcal{P}(N)$ defined as

$$W_p(\mu,\nu) \coloneqq \inf_{\pi \in \Pi(\mu,\nu)} \left[\iint_{N\times N} d_N^p(x,y)\, \mathrm{d}\pi(x,y) \right]^{\frac{1}{p}} \tag{2}$$

where $\Pi(\mu,\nu)$ is the set of all probabilistic couplings on $N \times N$ with μ and ν as marginals. When $p = 2$, the minimizer of (1) can be interpreted as a harmonic map, with boundary condition $f \restriction V_0 = f_0$, that takes value in a weak, metric-measure space sense (Otto 2001; Ambrosio, Gigli, and Savare 2005; Lott and Villani 2009; Lavenant 2017). Note that this is a nontrivial fact because in general harmonic maps (or minimizers of the Dirichlet energy) only exist when the target metric space $\mathcal{D}$ is negatively curved in the sense of Alexandrov (Jost 1994), but $\mathcal{P}(N)$ equipped with the 2-Wasserstein distance has positive Alexandrov curvature (Ambrosio, Gigli, and Savare 2005, §7.3). When $\mathcal{D}$ is the one-dimensional distributions on the real line equipped with the 2-Wasserstein distance, (Solomon et al. 2014) related (1) to a Dirichlet problem.

In this work, we first extend the framework of (Solomon et al. 2014) to hypergraphs using *Wasserstein barycenter* (Agueh and Carlier 2011; Asoodeh, Gao, and Evans 2018). For 2-Wasserstein distances this is equivalent to solving a *multi-marginal optimal transport* (Carlier and Ekeland 2010) problem with a naturally constructed cost function. The hypergraph extension of Wasserstein propagation is based on a novel interpretation of the original algorithm on graphs (Solomon et al. 2014) as a message-passing algorithm. Next, we take a deeper look at the statistical learning aspects of our proposed algorithm, and establish generalization error bounds for propagating one-dimensional distributions on graphs and hypergraphs using the 2-Wasserstein distance. One dimensional distributions such as histograms are among the most frequent application scenarios of soft label propagation. The main technical ingredient is *algorithmic stability* (Bousquet and Elisseeff 2002). To our knowledge, our generalization bound is the first of its type in the literature of Wasserstein distance based soft label propagation; on graphs these results generalize the generalization error bounds in (Belkin, Matveeva, and Niyogi 2004). As no general semi-supervised learning algorithm is available for large dataset (Petegrosso et al. 2017), the new connection between Wasserstein barycenter and semi-supervised learning might be of theoretical as well as computational interest.

In the supplemental material, we provide promising numerical results for both synthetic and real data. In particular, we apply our hypergraph soft label propagation algorithm to random uniform hypergraphs as well as several UCI datasets adopting hypergraph representations.

Notation

We denote an undirected simple graph as $G = (V, E)$ where $V = [n] \coloneqq \{1, \ldots, n\}$ is the vertex set and $E \in V \times V$ denotes edges. We use L to denote the (weighted) graph Laplacian associated with (weighted) graph G, which is a real square matrix of size n-by-n defined by $L \coloneqq D - W$, where $W \in \mathbb{R}^{n\times n}$ is the (weighted) adjacency matrix of G, and $D \in \mathbb{R}^{n\times n}$ is a diagonal matrix with the (weighted) degree of vertex j at its (j,j)-th entry. We use $H = (V, \mathcal{E})$ to denote a hypergraph where $\mathcal{E} \in 2^V$ is the set of hyperedges of H. Given $k \geq 2$ probability measures $\rho_1, \ldots, \rho_k$ in $\mathcal{P}(N)$, their *Wasserstein barycenter* is

$$\mathsf{bar}\left(\{\rho_i\}_{i=1}^k\right) \coloneqq \inf_{\nu\in\mathcal{P}(N)} \frac{1}{k}\sum_{i=1}^k W_2^2(\rho_i,\nu). \tag{3}$$

Fundamental properties of the minimizer in (3) are studied in (Agueh and Carlier 2011); similar results hold when the squared 2-Wasserstein distance are weighted differently. Given a hyperedge E of H, we use $\mathsf{bar}(E)$ to denote $\mathsf{bar}\left(\{\mu_i\}_{i=1}^{|E|}\right)$ where the probability measures $\mu_1, \ldots, \mu_{|E|}$ associated with each vertex i in E are clear from the context.

Message Passing and Label Propagation on Graph and Hypergraph

In this section, we formulate our hypergraph label propagation as a special case of belief propagation. To this end, we begin with a brief description of a slightly generalized version of Wasserstein label propagation (Solomon et al. 2014) from a message passing perspective.

A learning problem is specified by a probability distribution D on $X \times Y$ according to which labeled sample pairs $z_i = (x_i, y_i)$ are drawn and presented to a learning algorithm; the algorithm outputs a map from X to Y. In soft label propagation problems, the maps of interest take values in a space of probability distributions Y. From now on, we assume Y is the space of probability distributions on a complete metric space (N, d_N), i.e. $Y = \mathcal{P}(N)$. Since N is complete, the space Y equipped with Wasserstein distance is also a complete metric space (Villani 2003, Theorem 6.18).

Wasserstein Label Propagation on Graphs

Let X be a graph $G = (V, E)$, possibly with weights $\omega_{ij} \geq 0$ on each edge (i, j). Wasserstein label propagation is an extension of Tikhonov regularization framework on graphs (Belkin, Matveeva, and Niyogi 2004) from real-valued functions to measure-valued maps. Denote a measure-valued map from G to $\mathcal{P}(N)$ as $\mu : V \to \mathcal{P}(N)$. For simplicity, write $\mu_i \coloneqq \mu(i)$ for $i \in V$. A prototypical semi-supervised learning setting assumes $\mu_1, \cdots, \mu_m$ are known, where $1 \leq m \ll n$, and the goal is to determine $\mu_{m+1}, \cdots, \mu_n$ on the rest of the vertices. We will do so by minimizing the following objective function with Tikhonov regularization

$$\min_{f:V\to\mathcal{P}(N)} \frac{1}{m}\sum_{i=1}^m W_2^2(\mu_i, f_i) + \gamma \sum_{(i,j)\in E} \omega_{ij} W_2^2(f_i, f_j), \tag{4}$$

where $\gamma > 0$ is a regularization parameter. This minimization problem can be thought of as an extension of the Dirichlet boundary problem studied in (Solomon et al. 2014) as here we do not impose $f_i = \mu_i$ for $i \in [m]$. The minimizer of (4) is the measure-valued map "learned" from the training data $\{(i, \mu_i) \mid 1 \leq i \leq n\}$ and the given graph structure

$G = (V, E)$. We point out that the formulation in (Solomon et al. 2014) is a special case (parameter-free "interpolated regularization") of (4) in the limit $\gamma \to 0$, for the same reason as given in (Belkin, Matveeva, and Niyogi 2004, §2.2).

We now provide an algorithm for solving (4) based on belief propagation. Since this is only a motivating perspective, we assume for simplicity that the graph is unweighted; all arguments below can be extended to weighted graphs with heavier notations. In this context, each vertex i updates its *belief* about the local minimizer of (4) f_i by exchanging messages to edges it is incident to. The classical min-sum algorithm (Moallemi and Roy 2009) describes this process as follows. At time t, vertex $i \in [m]$ has belief $b_i^{(t)}$ about the minimizer f_i of (4); then, at time $t+1$, i sends message $J_{i\to e}^{(t)}$ to edge $e = (i, j)$ and receives message $J_{e\to i}^{(t)}$ from e, then updates the message for the next iteration according to

$$J_{i\to e}^{(t)}\left(b_i^{(t)}\right) = W_2^2\left(\mu_i, b_i^{(t)}\right) + \sum_{k\in N(i)\setminus\{j\}} J_{(i,k)\to i}^{(t-1)}\left(b_i^{(t-1)}\right) \quad (5)$$

and

$$J_{e\to i}^{(t)}\left(b_i^{(t)}\right) = \min_{f_j\in\mathcal{P}(N)}\left[W_2^2\left(b_i^{(t)}, f_j\right) + J_{j\to e}^{(t-1)}(f_j)\right]. \quad (6)$$

The first term in (5) is set to be zero if $i \notin [m]$. The belief is then updated at at time $t+1$ according to evolution

$$b_i^{(t+1)} := \arg\min_{f_i}\left[W_2(\mu_i, f_i) + \sum_{k\in V:(i,k)\in E} J_{(i,k)\to i}^{(t)}(f_i)\right].$$

Convergence of $b_i^{(t)}$ to the true minimizer f_i^* can be guaranteed under some (mild) conditions on initial beliefs if G is a tree (see e.g., (Moallemi and Roy 2009)).

Wasserstein Label Propagation on Hypergraphs

Let now X be represented by a hypergraph $H = (V, \mathcal{E})$. Since each hyperedge may contain arbitrary number of vertices, the minimization (4) fails to formulate our learning objective. Nevertheless, the belief propagation updates (5) and (6) can naturally be extended to the message passing between vertex i and hyperedge E containing i as

$$J_{i\to E}^{(t)}\left(b_i^{(t)}\right) = W_2^2\left(\mu_i, b_i^{(t)}\right) + \sum_{E'\in\mathcal{E}\setminus\{E\}:i\in E'} J_{E'\to i}^{(t-1)}\left(b_i^{(t-1)}\right) \quad (7)$$

and

$$J_{E\to i}^{(t)}\left(b_i^{(t)}\right) = \min_{f_{E\setminus\{i\}}}\left[\mathsf{bar}(E) + \sum_{k\in E\setminus\{i\}} J_{k\to E}^{(t-1)}(f_k)\right]. \quad (8)$$

where $f_{E\setminus\{i\}} = \{f_k \in \mathcal{P}(N) : k \in E\setminus\{i\}\}$. The belief of vertex $i \in [m]$ is then obtained according to the following rule:

$$b_i^{(t+1)} = \arg\min_{f_i\in\mathcal{P}(N)}\left[W_2^2(\mu_i, f_i) + \sum_{E\in\mathcal{E}:i\in E} J_{E\to i}^{(t)}(f_i)\right].$$

These belief propagation update rules justify the following formulation of label propagation for hypergraphs:

$$\min_{f:V\to\mathcal{P}(N)} \frac{1}{m}\sum_{i=1}^{m} W_2^2\left(\mu_i, f_i\right) + \gamma\sum_{E\in\mathcal{E}} \mathsf{bar}(E) \quad (9)$$

which is a natural generalization of (4) when the graph is unweighted. For weighted graphs, (9) still holds with properly adjusted $\mathsf{bar}(E)$ with weights.

Barycenter and Clique Representation

In this section, we assume that the labels are one-dimensional probability distributions, i.e., $N \subset \mathbb{R}$, and work solely with the 2-Wasserstein distance. We will see that in this case hypergraph label propagation can be cast into a Wasserstein propagation on a weighted graph arising from the clique representation of the hypergraph. The rest of this paper thus focuses on establishing generalization error bounds for graphs. The main advantage of one-dimensional soft labels is illustrated by the following classical result in optimal transportation theory.

Theorem 1 ((Villani 2003)). *Let $\mu, \nu \in \mathcal{P}(N)$ with $N \subset \mathbb{R}$ with cumulative density functions (c.d.f.) F_μ and F_ν, respectively. Then*

$$W_2^2(\mu, \nu) = \int_0^1 \left(F_\mu^{-1}(s) - F_\nu^{-1}(s)\right)^2 ds,$$

where F_μ^{-1} and F_ν^{-1} are the generalized inverses of F_μ and F_ν, respectively, i.e., $F_\mu^{-1}(s) := \inf\{x \in N : F_\mu(x) > s\}$.

The explicit expression for Wasserstein distance enables us to derive the barycenter of any number of one-dimensional distributions in a closed form.

Theorem 2 ((Bigot et al. 2017)). *Let $\rho_1, \ldots, \rho_k \in \mathcal{P}(N)$ be m probability distributions on $N \subset \mathbb{R}$ with cumulative density functions F_{ρ_i}, $i \in [k]$. Let ρ_b be the (unique) Wasserstein barycenter of $\{\rho_i\}_{i=1}^k$. Then the generalized inverse c.d.f. F_b^{-1} of ρ_b is given by*

$$F_\mathsf{b}^{-1}(s) = \frac{1}{k}\sum_{i=1}^{k} F_{\rho_i}^{-1}(s).$$

Since the inverse cdf's and the distributions are in one-to-one correspondence, this theorem characterizes the 2-Wasserstein barycenter of $\{\rho_i\}_{i=1}^m$. In light of Theorem 2, one can simplify the barycenter of hyperedge E that contains vertices, say, $\{1, 2, \ldots, k\}$ as

$$\begin{aligned}
\mathsf{bar}(E) &= \frac{1}{k}\sum_{i=1}^{k} W_2^2(\mu_i, \mu_\mathsf{b}) \\
&= \frac{1}{k}\sum_{i=1}^{k}\int_0^1\left(F_{\mu_i}^{-1}(s) - \frac{1}{k}\sum_{i=1}^{k} F_{\mu_i}^{-1}(s)\right)^2 \mathrm{d}s \\
&= \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=i+1}^{n}\int_0^1\left(F_{\mu_i}^{-1}(s) - F_{\mu_j}^{-1}(s)\right)^2 \mathrm{d}s \\
&= \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=i+1}^{k} W_2^2(\mu_i, \mu_j) \qquad (10)
\end{aligned}$$

where the first and second equalities follow from Theorems 1 and 2, respectively. Comparing (10) with (9), we have

Proposition 1. *Soft label propagation with 2-Wasserstein distance for one-dimensional distributions on hypergraphs H using (9) is equivalent to Wasserstein propagation on a weighted graph arising from the clique representation G_H of H. The weight of each edge e in G_H depends only on the degrees of the hyperedges containing e.*

Proof. Recall that the *clique representation* of a hypergraph $H = (V, \mathcal{E})$ is a graph $G_H = (V, E_H)$, where $E_H = \{(i,j) : \exists E \in \mathcal{E}, \{i,j\} \subset E\}$. The rest of the proof follows from checking definitions. ∎

Generalization Bounds for Wasserstein Propagation

In this section we derive generalization bounds for label propagation (4) on graphs. The same results apply to hypergraphs as well, by Proposition 1. We begin with briefly reviewing empirical risk, generalization error, and algorithmic stability in the passing.

Algorithmic Stability

The framework of algorithmic stability (Devroye and Wagner 1979; Bousquet and Elisseeff 2002; Mukherjee et al. 2006) was proposed in statistical learning as an alternative to the VC-dimension framework; the latter is often overpessimistic since it attempts to bound the generalization performance uniformly over possible algorithms. We briefly recapture the essence of algorithmic stability here. Let X and Y be two measurable spaces, and a set of training samples $S = \{z_i = (x_i, y_i), i = 1, \cdots, m\}$ of size m sampled i.i.d. with respect to an unknown joint distribution D on the product space $Z = X \times Y$. A learning algorithm is a mechanism that maps S to a global map $f_S : X \to Y$ defined on the entire X. It is often assumed for simplicity that the algorithm is symmetric with respect to training sets, i.e., the learning algorithm should return identical maps for two training sets with samples differing from each other only by a permutation. We shall assume all maps considered here are measurable, and all the measure spaces are separable. We are interested in the case in which X is a simple finite graph and Y is the probability space $\mathcal{P}(N)$. The *empirical risk* or *empirical error* of a mapping $f_S : X \to Y$ learned from the training set S of size $m > 0$ is defined as

$$R_m(f_S) := \frac{1}{m} \sum_{i=1}^{m} c(f_S, z_i)$$

where $c(\cdot,\cdot) : Y^X \times (X \times Y) \to \mathbb{R}_{\geq 0}$ is a cost function evaluating the predictive error of $f_S : X \to Y$ at a point sampled from the joint distribution D on $X \times Y$. The *generalization error* of the learned map is

$$R_D(f_S) = \mathbb{E}_{z \sim D}[c(f_S, z)]$$

which measures the average prediction error for a map learned from training data. The central problem in the PAC learning framework is bounding the discrepancy between R_m and R_D. In (Bousquet and Elisseeff 2002), the authors proved that such a bound exists if the algorithm satisfies a *uniform stability* property, essentially meaning that the learned mapping changes very little in terms of predictive power if the training sample undergoes a small change.

Definition 1 (Uniform Stability, (Bousquet and Elisseeff 2002))**.** *Fix a positive integer $m \in \mathbb{Z}_+$. Let $S = \{z_1, \cdots, z_m\} \subset X \times Y$ be a training set, and S' be another training set that contains the same elements as S with the only exception that the sample z_i is replaced with a different sample $z_i' \neq z_i$. A learning algorithm $A : (X \times Y)^m \to Y^X$ that sends any training set S to a mapping $f_S : X \times Y$ is said to be (uniform) β-stable for some positive constant $\beta > 0$ if for any pair of training sets S, S' differing by exact one element the following inequality holds:*

$$|c(f_S, z) - c(f_{S'}, z)| \leq \beta \qquad \forall z \in X \times Y.$$

Theorem 3 ((Bousquet and Elisseeff 2002))**.** *Let $S \mapsto f_S$ be a β-stable learning algorithm, such that $0 \leq c(f_S, z) \leq M$ for all $z \in X \times Y$ and all learning set S. For any arbitrary $\epsilon > 0$ we have for all $m \geq 8M^2/\epsilon^2$*

$$\mathbb{P}_{S \sim D^m}\{|R_m(f_S) - R_D(f_S)| > \epsilon\} \leq \frac{64Mm\beta + 8M^2}{m\epsilon^2}, \tag{11}$$

and for any $m \geq 1$

$$\mathbb{P}_{S \sim D^m}\{|R_m(f_S) - R_D(f_S)| > \epsilon + \beta\} \leq 2\exp\left(-\frac{m\epsilon^2}{2(m\beta + M)^2}\right). \tag{12}$$

Of course, the order of β in terms of the number m of training samples will be crucial here, otherwise any learning algorithm is uniformly stable for any bounded cost function. In (Bousquet and Elisseeff 2002) it was pointed out that a sufficient condition for these bounds to be tight is $\beta = O(1/m)$ as $m \to \infty$. It was verified in (Bousquet and Elisseeff 2002) that the Tikhonov regularization framework for scalar-valued functions with quadratic cost function satisfies this requirement; but Theorem 3 is indeed much more general and applicable to any measurable spaces X and Y. The rest of this paper is devoted to establishing algorithmic stability for(hyper)graph soft label propagation.

Generalization bounds for Soft Label Propagation

The goal of this subsection is to verify that the conditions of Theorem 3 are satisfied for the Tikhonov regularization framework (4). The first task is to find an appropriate model class for the distributions in $\mathcal{P}(N)$ that ensures the uniform boundedness of the cost function

$$c(f, (j, \mu_j)) = W_2^2(f_j, \mu_j). \tag{13}$$

This can be fulfilled trivially, for instance, if the metric space (N, d_N) is of bounded diameter. This includes many generic applications we run into in practice, in particular for propagating histograms but is already not satisfied by popular distribution classes such as the Gaussian distributions. It is therefore preferable to work with a model class for distributions with uniformly bounded pairwise Wasserstein distances under milder assumptions. By definition (2), bounding the Wassertein distance from above can be achieved by plugging an arbitrary coupling into the variational energy functional defining (2). However, explicitly constructing meaningful couplings is difficult in general. Many existing bounds explore the multiscale structure of the supports of the two distributions (David 1988; Lei 2018; Singh and Póczos 2018), but it is not clear how those technical conditions can be used as model class specifications. We shall bypass this difficulty by leveraging the simple characterization

of Wasserstein distances between one-dimensional distributions using quantile functions.

According to Theorem 1, one can simplify (4) as

$$\min_{f:V\to\mathcal{P}(N)} \int_0^1 \Big[\frac{1}{m}\sum_{i=1}^m \left(F_{\mu_i}^{-1}(s) - F_{f_i}^{-1}(s)\right)^2 + \gamma \sum_{(i,j)\in E} \left(F_{f_i}^{-1}(s) - F_{f_j}^{-1}(s)\right)^2\Big]\mathrm{d}s.$$

Since the inverse c.d.f.'s and the distributions are in one-to-one correspondences, and all $F_{\mu_i}^{-1}$ are given, it suffices to solve for the $F_{f_i}^{-1}$'s in their entirety and then recover each probability distribution at vertex i from $F_{f_i}^{-1}:[0,1]\to\mathbb{R}$. To simplify notations, define $\Phi: V\times[0,1]\to\mathbb{R}$ as $\Phi(i,s) := F_{f_i}^{-1}(s)$ and denote $\Phi_s(i) := \Phi(i,s)$ for all $s\in[0,1]$ and $i\in V$. For each fixed $s\in[0,1]$, Φ_s can be viewed as a function defined on the vertices of the graph G. For simplicity, we will identify each Φ_s with a real column vector of length $n=|V|$. Then the regularization term in (4) can be written in terms of L, the weighted graph Laplacian of G. Thus (4) transforms into

$$\min_{\Phi:V\times[0,1]\to\mathbb{R}} \frac{1}{m}\sum_{i=1}^m \int_0^1 \left|F_{\mu_i}^{-1}(s) - \Phi_s(i)\right|^2 \mathrm{d}s + \gamma\int_0^1 \Phi_s^\top L\Phi_s\,\mathrm{d}s. \tag{14}$$

The optimization problem (14) can be viewed as a linear combination of infinitely many Tikhonov regularization problems, one for each $s\in[0,1]$; each sub-problems is completely decoupled from others. Indeed, standard variational analysis shows that it suffices to solve each subproblem individually, i.e., solve for each fixed $s\in[0,1]$

$$\min_{\Phi_s\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^m \left(F_{\mu_i}^{-1}(s) - \Phi_s(i)\right)^2 + \gamma\Phi_s^\top L\Phi_s. \tag{15}$$

Once all subproblems are solved, it is necessary to check the compatibility across the solutions $\{\Phi_s : s\in[0,1]\}$, i.e., for any fixed $i\in V$, the map $s\mapsto\Phi_s(i)$ is indeed the inverse c.d.f. of a probability distribution. This compatibility will become straightforward after we derive the closed-form solution of each subproblem (15); see Proposition 2 below.

The solutions to Tikhonov regularization problems (15) is known back in (Belkin, Matveeva, and Niyogi 2004). Let $\mathbf{1}=(1,\cdots,1)^\top\in\mathbb{R}^n$ be the column vector of all ones, and

$$T_\ell = \mathrm{diag}(t_1,\cdots,t_\ell,0,\cdots,0)^\top\in\mathbb{R}^n$$

where t_i is the multiplicity of vertex $i\in V$ in the training set S (we assumed without loss of generality that the training samples are the first ℓ vertices, for notational convenience), and

$$\mathbf{y}_s = \left(\sum_{v_i=1} F_{\mu_i}^{-1}(s),\cdots,\sum_{v_i=\ell} F_{\mu_i}^{-1}(s),0,\cdots,0\right)^\top\in\mathbb{R}^n \tag{16}$$

i.e., for $1\le i\le\ell$, the i-th entry of $\mathbf{y}_s$ is the sum of the t_i values of the inverse c.d.f.'s of $i\in V$. With these notations, it is easy to write down the Euler-Lagrange equation of the optimization problem (15) as

$$(T_\ell + m\gamma L)\,\Phi_s^* = \mathbf{y}_s. \tag{17}$$

To solve this equation, note that the operator $T_\ell+m\gamma L$ may not be invertible — in fact, neither T_ℓ nor L is invertible. Nonetheless, assuming the graph is connected, the nullspace of L is one-dimensional and spanned precisely by the all-one vector $\mathbf{1}$. This means that L will be invertible on the orthogonal complement of the one-dimensional subspace spanned by $\mathbf{1}$. Furthermore, noting that

$$T_\ell + m\gamma L = m\gamma\left(\frac{1}{m\gamma}T_\ell + L\right), \tag{18}$$

by standard functional analysis (or (Belkin, Matveeva, and Niyogi 2004, Proof of Theorem 5)) we know that the perturbed operator $L+(m\gamma)^{-1}T_\ell$ is invertible on the orthogonal complement as well provided that $m\gamma$ is sufficiently large. More precisely, the invertibility holds for

$$\gamma \ge \frac{\max\{t_1,\cdots,t_\ell\}}{m\lambda_1}$$

where λ_1 is the smallest non-zero eigenvalue of L, or the *spectral gap* of the (possibly weighted) connected graph G. This observation, together with the invariance of the quadratic cost in (15) under global translations, allow us to preprocess the input data by subtracting scalar

$$\bar{y}_s := \frac{1}{m}\mathbf{1}^\top\mathbf{y}_s = \frac{1}{m}\sum_{i=1}^m F_{\mu_i}^{-1}(s) \tag{19}$$

from each $F_{\mu_i}^{-1}(s)$, applying the inverse of $T_\ell+m\gamma L$, and finally adding $\bar{y}_s$ back to the obtained solution. More specifically, we would like to solve the equivalent optimization problem

$$\Phi_s^* = \arg\min_{\Phi_s\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^m \left[\left(F_{\mu_i}^{-1}(s) - \bar{y}_s\right) - \left(\Phi_s(i) - \bar{y}_s\right)\right]^2 + \gamma\left(\Phi_s - \bar{y}_s\mathbf{1}\right)^\top L\left(\Phi_s - \bar{y}_s\mathbf{1}\right), \tag{20}$$

which gives

$$\Phi_s^* - \bar{y}_s\mathbf{1} = (T_\ell + m\gamma L)^{-1}\left(\mathbf{y}_s - \bar{y}_s T_\ell\mathbf{1}\right).$$

Therefore, the solution to (15) takes the form

$$\Phi_s^* = (T_\ell + m\gamma L)^{-1}\left(\mathbf{y}_s - \bar{y}_s T_\ell\mathbf{1}\right) + \bar{y}_s\mathbf{1}. \tag{21}$$

We emphasize here that the notation $(T_\ell+m\gamma L)^{-1}$ alone does not make sense because the matrix $T_\ell+m\gamma L$ may well be non-invertible; only the notation $(T_\ell+m\gamma L)^{-1}u$ for $u\in\mathbb{R}^n$ satisfying $\mathbf{1}^\top u = 0$ bears actual meanings.

Remark 1. Alternatively, one can derive a solution to (15) by directly applying the pseudo-inverse of $T_\ell+m\gamma L$ to $\mathbf{y}_s$, i.e. setting $\Phi_s^* := (T_\ell+m\gamma L)^\dagger\mathbf{y}_s$; this avoids the requirement that γ needs not be too small, but leaves the algorithmic stability of the resulting solution Φ_s^* in question.

Now that we have obtained closed-form solutions (21) to subproblems (15) for each $s \in [0,1]$, it is imperative to guarantee that the closed-form solutions $\{\Phi_s^* \mid 0 \le s \le 1\}$ do piece together and give rise to inverse c.d.f.'s at each vertex $i \in V$. This basically requires that, for each $i \in V$, the map $[0,1] \ni s \mapsto \Phi_s^*(i) \in \mathbb{R}$ should be non-decreasing and right continuous. The right continuity is obvious, since for each $i \in V$ the map $[0,1] \ni s \mapsto \mathbf{y}_s(i)$ is right continuous, and the linear combination of right continuous functions is still right continuous, thus the assertion follows from the closed-form expression (21). The monotonicity would be guaranteed if there is a "maximum principle" for the operator $T_\ell + m\gamma L$, or equivalently $L + (m\gamma)^{-1} T$, on the graph G, i.e. if $\mathbb{R}^n \ni \mathbf{y} \ge 0$ (entrywise) and $(T_\ell + m\gamma L)\Phi = \mathbf{y}$ then $\Phi \ge 0$ (entrywise). This is because: we already have $\mathbf{y}_s - \mathbf{y}_t \ge 0$ for any $0 \le t \le s \le 1$ by the monotonicity of the inverse c.d.f.'s, hence such a "maximum principle" would then guarantee $\Phi_s - \Phi_t \ge 0$ (entrywise). Such maximum principles abound for graph Laplacians, see e.g. (Holopainen and Soardi 1997; Chung, Chung, and Kim 2007). It is natural to expect such a maximum principle to hold for $L + (m\gamma)^{-1} T$ as well, since T is a non-negative.

Lemma 1 (Maximum Principle)**.** *If $\Phi \in \mathbb{R}^n$ is such that $[(T_\ell + m\gamma L)\Phi](i) \ge 0$ for all $1 \le i \le \ell$ and $[(T_\ell + m\gamma L)\Phi](i) = 0$ for all $\ell + 1 \le i \le n$, then Φ attains both its maximum and minimum over $i = 1, \cdots, n$ within $\{1, \cdots, \ell\}$. In particular, $\Phi(i) \ge 0$ for all $1 \le i \le n$.*

This lemma then implies the promised monotonicity.

Proposition 2. *For any vertex $i \in V$, the closed-form solutions* (21) *is non-decreasing with respect to $s \in [0,1]$.*

Proof. By the equivalence of (20) and (15), the solutions Φ_s satisfies the Euler-Lagrange equations for (15):

$$(T_\ell + m\gamma L)\Phi_s^* = \mathbf{y}_s.$$

For any $0 \le t \le s \le 1$, subtracting two Euler-Lagrange equations yields

$$(T_\ell + m\gamma L)(\Phi_s^* - \Phi_t^*) = \mathbf{y}_s - \mathbf{y}_t \ge 0$$

where the inequality follows from the definition of $\mathbf{y}_s$ in (16). Furthermore, it is straightforward to see that $\mathbf{y}_s - \mathbf{y}_t$ satisfies the assumption in Lemma 1, which then implies $\Phi_s^* \ge \Phi_t^*$. ∎

We can now rest assured that the solutions (21) indeed constitute an inverse c.d.f. at each vertex $i \in V$. But there is more to this: it can actually be easily verified that (20) is equivalent to the Tikhonov regularization problem formulated in (Belkin, Matveeva, and Niyogi 2004) if we view $(\Phi_s - \bar{y}_s \mathbf{1})$ as variables. We can thus follow the idea of (Belkin, Matveeva, and Niyogi 2004, Theorem 5) to get algorithmic stability of each individual Φ_s, $s \in [0,1]$.

Theorem 4. *Assume $m \ge 4$ and $0 < T := \max\{t_1, \cdots, t_\ell\} < \infty$ satisfies $m\gamma\lambda_1 - T > 0$, where λ is the regularization parameter in* (15) *and λ_1 is the spectral gap of the connected graph G. Let $S = \{(v_i, \mu_i) \mid 1 \le i \le m,\ v_i \in V,\ \mu_i \in \mathcal{P}(\mathbb{R})\}$ and $S' = \{(v_i', \mu_i') \mid 1 \le i \le m,\ v_i \in V,\ \mu_i \in \mathcal{P}(\mathbb{R})\}$ be two training sets that differ from each other by exactly one data sample. Assume further that, for a fixed $s \in [0,1]$ there holds*

$$\max\left\{\left|F_{\mu_i}^{-1}(s)\right|, \left|F_{\mu_i'}^{-1}(s)\right|,\ i = 1, \cdots, m\right\} \le M_s < \infty. \tag{22}$$

Let $\Phi_s^*, \Phi_s'^*$ be solutions of (15) *for S and S', respectively,*

$$\Phi_s^* = (T_\ell + m\gamma L)^{-1}(\mathbf{y}_s - \bar{y}_s T_\ell \mathbf{1}) + \bar{y}_s \mathbf{1}$$
$$\Phi_s'^* = (T_\ell' + m\gamma L)^{-1}(\mathbf{y}_s' - \bar{y}_s' T_\ell' \mathbf{1}) + \bar{y}_s' \mathbf{1}$$

where T_ℓ', $\mathbf{y}_s'$, $\bar{y}_s'$ are defined analogously to T_ℓ, $\mathbf{y}_s$, $\bar{y}_s$ but with respect to S' instead of S. Then

$$\|\Phi_s^* - \Phi_s'^*\|_\infty \le \frac{3M_s\sqrt{Tm}}{(m\gamma\lambda_1 - T)^2} + \frac{4M_s}{m\gamma\lambda_1 - T} + \frac{2M_s}{m}. \tag{23}$$

The boundedness assumption on Φ_s seems artificial but is indeed very natural: an almost identical argument as the first part of the proof of Lemma 1, with minimum replaced with maximum and *mutatis mutandis*, establishes the fact that the global maximum of Φ_s must be attained at the boundary $1 \le i \le \ell$. Hence, since there are only finitely many data in the training set, this boundedness is a very mild requirement (e.g. satisfied if each $F_{\mu_i}^{-1}(s)$ is finite). We define the model class to reflect the requirement that the inverse c.d.f.'s of the one-dimensional probability distributions in the training set should be controlled. We define the model class in Definition 2 and summarize the maximum principle argument as a lemma on *a priori* estimates for future convenience.

Definition 2 (Dominated Quantile Class)**.** *Let $\phi \in L^2[0,1]$ and $\phi \ge 0$ on $[0,1]$. A probability distribution $\mu \in \mathcal{P}(\mathbb{R})$ is said to belong to* dominated quantile class $\mathcal{M}_\phi^2$ *if $\left|F_\mu^{-1}(s)\right| \le \phi(s)$ for a.e. $s \in [0,1]$.*

Lemma 2 (*A Priori* Estimates)**.** *If in the training set $S = \{(v_i, \mu_i) \mid 1 \le i \le m,\ v_i \in V,\ \mu_i \in \mathcal{P}(\mathbb{R})\}$ all μ_i lie in a dominated quantile model class $\mathcal{M}_\phi^2$ for some $\phi \in L^2[0,1]$ with $\phi \ge 0$ on $[0,1]$, then any map $f : V \to \mathcal{P}(\mathbb{R})$ minimizing* (4) *takes values in $\mathcal{M}_\phi^2$ as well.*

Proof. By the equivalence between (4) and (14), it suffices to show the following fact: for each fixed $s \in [0,1]$, if $\max\left\{\left|F_{\mu_i}^{-1}(s)\right|,\ i = 1, \cdots, m\right\} \le \phi(s)$ then $\|\Phi_s^*\|_\infty \le \phi(s)$, where Φ_s^* is defined in (20). But this follows straightforwardly from the maximum principle. ∎

We now present the main theoretical result of this paper. In our setting these results apply to graphs as well as hypergraphs by Proposition 1.

Proposition 3 (Algorithmic Stability for Soft Label Propagation of One-Dimensional Distributions)**.** *Assume $m \ge 4$ and $0 < T := \max\{t_1, \cdots, t_\ell\} < \infty$ satisfying $m\gamma\lambda_1 - T > 0$, where γ is the regularization parameter in* (15) *and λ_1 is the spectral gap of the weighted, connected graph G. If the joint distribution $D \in \mathcal{P}(V \times \mathcal{P}(\mathbb{R}))$ is supported on $V \times \mathcal{M}_\phi^2$ for a quantile model class $\mathcal{M}_\phi^2 \subset \mathcal{P}(\mathbb{R})$ for some $\phi \in L^2[0,1]$ with $\phi \ge 0$ on $[0,1]$, then the solutions of* (4)

or (9) *are* β*-stable in the sense of Definition 1 with respect to cost function* (13)*, where*

$$\beta = 4\left\|\phi\right\|_2^2 \left[\frac{3\sqrt{Tm}}{\left(m\gamma\lambda_1 - T\right)^2} + \frac{4}{m\gamma\lambda_1 - T} + \frac{2}{m}\right]. \quad (24)$$

Proof. Let (j, θ_j) be a new sample drawn from the joint distribution D. Then $\theta_j \in \mathcal{M}_\phi^2$ with probability 1. Let S, S' be two training samples with values in $\mathcal{M}_\phi^2$ and differ by exactly one data point. By Theorem 4 we have

$$\begin{aligned}&\left|\Phi_s^*\left(j\right) - \Phi_s'^*\left(j\right)\right| \\ &\leq \left[\frac{3\sqrt{Tm}}{\left(m\gamma\lambda_1 - T\right)^2} + \frac{4}{m\gamma\lambda_1 - T} + \frac{2}{m}\right]\phi\left(s\right). \end{aligned} \quad (25)$$

By (10), the difference between the squared Wasserstein losses satisfy

$$\begin{aligned}&\left|c\left(f_S, (j,\theta_j)\right) - c\left(f_{S'}, (j,\theta_j)\right)\right| \\ &= \left|W_2^2\left(f_S\left(j\right), \theta_j\right) - W_2^2\left(f_{S'}\left(j\right), \theta_j\right)\right| \\ &= \left|\int_0^1 \left|\Phi_s^*\left(j\right) - F_{\theta_j}^{-1}\left(s\right)\right|^2 \mathrm{d}s - \int_0^1 \left|\Phi_s'^*\left(j\right) - F_{\theta_j}^{-1}\left(s\right)\right|^2 \mathrm{d}s\right| \\ &\leq \int_0^1 \left|\left(\Phi_s^*\left(j\right) + \Phi_s'^*\left(j\right) - 2F_{\theta_j}^{-1}\left(s\right)\right)\left(\Phi_s^*\left(j\right) - \Phi_s'^*\left(j\right)\right)\right| \mathrm{d}s \\ &\overset{(*)}{\leq} \left[\frac{3\sqrt{Tm}}{\left(m\gamma\lambda_1 - T\right)^2} + \frac{4}{m\gamma\lambda_1 - T} + \frac{2}{m}\right] \cdot \int_0^1 4\phi\left(s\right)\cdot\phi\left(s\right)\mathrm{d}s \\ &= 4\left\|\phi\right\|_2^2 \left[\frac{3\sqrt{Tm}}{\left(m\gamma\lambda_1 - T\right)^2} + \frac{4}{m\gamma\lambda_1 - T} + \frac{2}{m}\right] = \beta, \end{aligned}$$

where at $(*)$ we used (25) to bound the difference $\left|\Phi_s^*\left(j\right) - \Phi_s'^*\left(j\right)\right|$, and invoked Lemma 2 to conclude that

$$\Phi_s^*\left(j\right), \Phi_s'^*\left(j\right) \leq \phi\left(s\right)$$

and hence

$$\left|\Phi_s^*\left(j\right) + \Phi_s'^*\left(j\right) - 2F_{\theta_j}^{-1}\left(s\right)\right| \leq 4\phi\left(s\right). \qquad \blacksquare$$

Note that the cost function is uniformly bounded by $M = 4\left\|\phi\right\|_2^2$ in our setting. Our main result this follows from combining Proposition 3 and Theorem 3.

Theorem 5 (Generalization Error for Soft Label Propagation for One-Dimensional Distributions)**.** *Under the same assumptions as Proposition 3, for any* $\epsilon > 0$ *we have for all* $m \geq 8M^2/\epsilon^2$

$$\mathbb{P}_{S\sim D^m}\left\{\left|R_m\left(f_S\right) - R_D\left(f_S\right)\right| > \epsilon\right\} \leq \frac{64Mm\beta + 8M^2}{m\epsilon^2}, \quad (26)$$

and for any $m \geq 1$

$$\begin{aligned}&\mathbb{P}_{S\sim D^m}\left\{\left|R_m\left(f_S\right) - R_D\left(f_S\right)\right| > \epsilon + \beta\right\} \\ &\qquad \leq 2\exp\left(-\frac{m\epsilon^2}{2\left(m\beta + M\right)^2}\right), \end{aligned} \quad (27)$$

where $M = 4\left\|\phi\right\|_2^2$ *and* β *given by* (24)*.*

Conclusion

In this paper, we proposed a novel framework for a semi-supervised learning problem where (i) the labels are given by probability measures on a metric space ("soft labels") and (ii) the underlying similarity structure is given by a hypergraph (which subsumes graph and simplicial complex). Our framework was inspired by a re-formulation of graph-based label propagation in terms of message passing and borrowed ideas from the theory of multi-marginal optimal transport. We then established generalization error bounds for propagating one-dimensional distributions using 2-Wasserstein distances. To the best of our knowledge, this constitutes the first generalization error bounds for Wasserstein distance based soft label propagation, even on graphs. We expect similar generalization bounds to hold for propagating higher-dimensional probability distributions as well as using other Wasserstein distances, but a deeper understanding of the geometry of Wasserstein spaces will be indispensable for those purposes. Future work include (i) generalization of our results to higher-dimensional probability measures, (ii) investigating the scalability and efficiency of our message-passing algorithm, and (iii) experimental study of our framework on real-work networks that can be naturally represented by hypergraphs.

References

Agueh, M., and Carlier, G. 2011. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis* 43(2):904–924.

Ambrosio, L.; Gigli, N.; and Savare, G. 2005. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel.

Asoodeh, S.; Gao, T.; and Evans, J. 2018. Curvature of hypergraphs via multi-marginal optimal transport. In *The 57th IEEE Conference on Decision and Control (CDC 2018)*.

Belkin, M.; Matveeva, I.; and Niyogi, P. 2004. Regularization and Semi-Supervised Learning on Large Graphs. In *International Conference on Computational Learning Theory*, 624–638. Springer.

Bigot, J.; Gouet, R.; Klein, T.; and López, A. 2017. Geodesic PCA in the wasserstein space by convex PCA. *Ann. Inst. H. Poincaré Probab. Statist.* 53(1):1–26.

Bousquet, O., and Elisseeff, A. 2002. Stability and generalization. *J. Mach. Learn. Res.* 2:499–526.

Bulò, S. R., and Pelillo, M. 2009. A game-theoretic approach to hypergraph clustering. In *Advances in Neural Information Processing Systems 22*, 1571–1579.

Carlier, G., and Ekeland, I. 2010. Matching for Teams. *Economic Theory* 42(2):397–418.

Chapelle, O.; Schölkopf, B.; and Zien, A. 2006. *Semi-supervised Learning*. Adaptive computation and machine learning. MIT Press.

Chung, S.-Y.; Chung, Y.-S.; and Kim, J.-H. 2007. Diffusion and Elastic Equations on Networks. *Publications of the Research Institute for Mathematical Sciences* 43(3):699–725.

Corduneanu, A., and Jaakkola, T. S. 2005. Distributed information regularization on graphs. In *Advances in Neural Information Processing Systems*. MIT Press. 297–304.

David, G. 1988. Morceaux de Graphes Lipschitziens et Intégrales Singulières sur une Surface. *Revista Matemática Iberoamericana* 4(1):73–114.

Demir, E.; Aykanat, C.; and Cambazoglu, B. B. 2008. Clustering spatial networks for aggregate query processing: A hypergraph approach. *Information Systems* 33(1):1–17.

Devroye, L., and Wagner, T. 1979. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory* 25(5):601–604.

Ghoshal, G.; Zlatić, V.; Caldarelli, G.; and Newman, M. 2009. Random hypergraphs and their applications. *Physical Review E* 79(6):066118.

Govindu, V. M. 2005. A tensor decomposition for geometric grouping and segmentation. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, 1150–1157 vol. 1.

Hein, M.; Setzer, S.; Jost, L.; and Rangapuram, S. S. 2013. The total variation on hypergraphs - learning on hypergraphs revisited. In *Advances in Neural Information Processing Systems*, 2427–2435.

Holopainen, I., and Soardi, P. M. 1997. p-Harmonic Functions on Graphs and Manifolds. *Manuscripta Mathematica* 94(1):95–110.

Huang, J.; Zhang, R.; and Yu, J. X. 2015. Scalable hypergraph learning and processing. In *Proc. of IEEE Int. Conf. on Data Mining (ICDM)*, 775–780.

Jost, J., and Mulas, R. 2018. Hypergraph laplace operators for chemical reaction networks.

Jost, J. 1994. Equilibrium maps between metric spaces. *Calculus of Variations and Partial Differential Equations* 2(2):173–204.

Kim, C.; Bandeira, A. S.; and Goemans, M. X. 2018. Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach. *arXiv preprint arXiv:1807.02884*.

Klamt, S.; Haus, U.-U.; and Theis, F. 2009. Hypergraphs and cellular networks. *PLoS computational biology* 5(5):e1000385.

Lavenant, H. 2017. Harmonic mappings valued in the wasserstein space.

Lei, J. 2018. Convergence and Concentration of Empirical Measures under Wasserstein Distance in Unbounded Functional Spaces. *arxiv preprint*.

Li, P., and Milenkovic, O. 2017. Inhomogeneous hypergraph clustering with applications. In *Advances in Neural Information Processing Systems 30*, 2308–2318.

Li, X., and Ramchandran, K. 2015. An active learning framework using sparse-graph codes for sparse polynomials and graph sketching. In *Advances in Neural Information Processing Systems 28*, 2170–2178.

Lott, J., and Villani, C. 2009. Ricci Curvature for Metric-Measure Spaces via Optimal Transport. *Annals of Mathematics* 903–991.

Moallemi, C. C., and Roy, B. V. 2009. Convergence of min-sum message passing for quadratic optimization. *IEEE Trans. Inf. Theory* 55(5):2413–2423.

Mukherjee, S.; Niyogi, P.; Poggio, T.; and Rifkin, R. 2006. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics* 25(1):161–193.

Otto, F. 2001. The Geometry of Dissipative Evolution Equations: the Porous Medium Equation. *Communications in Partial Differential Equations* 26(1-2):101–174.

Petegrosso, L.; W. Zhang, Z. L.; Saad, Y.; and Kuang, R. 2017. Low rank label propagation for semi-supervised learning with 1000 millions samples. *arxiv preprint*.

Seeger, M. 2001. Learning with labeled and unlabeled data. Technical report, University of Edinburgh.

Singh, S., and Póczos, B. 2018. Minimax Distribution Estimation in Wasserstein Distance. *arxiv preprint*.

Solomon, J.; Rustamov, R. M.; Guibas, L.; and Butscher, A. 2014. Wasserstein propagation for semi-supervised learning. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, 306–314.

Subramanya, A., and Bilmes, J. 2011. Semi-supervised learning with measure propagation. *Journal of Machine Learning Research* 12:3311–3370.

Tsuda, K. 2005. Propagating distributions on a hypergraph by dual information regularization. In *Proceedings of the 22Nd International Conference on Machine Learning*, 920–927.

Villani, C. 2003. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society.

Villani, C. 2008. *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Zhou, D.; H., J.; and Schölkopf, B. 2007. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems 19*. MIT Press. 1601–1608.

Zhu, X. 2008. Semi-Supervised Learning Literature Survey. Technical report, University of Wisconsin-Madison.