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

Tingran Gao[1], Shahab Asoodeh[2], Yi Huang[3], and James Evans[4]

Abstract

Inspired by recent interests in developing machine learning and data mining algorithms for hypergraphs, here we investigate 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 offer novel insight and deeper understanding about Wasserstein propagation on graphs.

I. INTRODUCTION

Recent decades have witnessed a growing interest in developing machine learning and data mining algorithms on *hypergraphs* [ZHS07], [JM18], [BP09], [LR15], [LM17], [HSJR13], [HZY15]. 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 the capture of higher order interactions among objects; applications have been found in many fields such as computer vision [Gov05], network clustering [DAC08], folksonomies [GZCN09], cellular networks [KHT09], and community detection [KBG18].

This paper develops a probably approximately correct (PAC) learning framework for *soft label propagation* or *Wasserstein propagation* [SRGB14], a recently proposed semi-supervised learning algorithm based on optimal transport [Vil03], [Vil08], on graphs and hypergraphs. Distinct from the prototypical semi-supervised learning algorithm of *label propagation* [BMN04], in which labels of interest are 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 and distributional information is crucial. For example, the traffic density at routers on the Internet network or topic distributions across the co-authorship network are more naturally modeled as probability distributions.

Semi-supervised learning is a paradigm that leverages unlabelled data to improve the generalization performance for supervised learning, under generic, unsupervised structural assumptions about the dataset (e.g. the manifold assumption); see [See01], [Zhu08], [CSZ06] for an overview. Given a graph $G = (V, E)$ and a subset of vertices $V_0 \subset V$, label propagation is a procedure for 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 from the classical heat equation, 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 distribution $\mathcal{P}(N)$ modeled on a complete, separable metric space (N, d_N).

Among the first works to address semi-supervised learning with soft labels are [CJ05], [Tsu05], [SB11]. In all of these works, the similarity between two soft labels is quantitatively measured using the Kullback-Leibler (KL) divergence, but the soft labels inferred from this process are often unstable and discontinuous. In [SRGB14] the authors proposed to replace 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 $f \upharpoonright V_0 = f_0$. Here $W_p(\mu, \nu)$ denotes the p-Wasserstein distance between probability distributions $\mu, \nu \in \mathcal{P}(N)$ defined as

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

[1]Department of Statistics, The University of Chicago, Chicago, IL 60637 `tingrangao@galton.uchicago.edu`
[2]Computation Institute and Institute of Genomics and System Biology, The University of Chicago, Chicago, IL 60637 `A@uchicago.edu`
[3]Department of Medicine, University of Chicago, Chicago IL, `yhuang10@uchicago.edu`
[4]Computation Institute and Department of Sociology, The University of Chicago, Chicago, IL 60637 `jevans@uchicago.edu`

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 $f \upharpoonright V_0 = f_0$ that takes value in a weak, metric-measure space sense [Ott01], [AGS05], [LV09], [Lav17]. Note that this is a nontrivial fact because harmonic maps (or minimizers of the Dirichlet energy) generally only exist when the target metric space $\mathcal{D}$ has negative Alexandrov curvature [Jos94], but $\mathcal{P}(N)$ equipped with the 2-Wasserstein distance has positive Alexandrov curvature [AGS05, §7.3]. When $\mathcal{D}$ is a one-dimensional distribution on the real line defined by 2-Wasserstein distance, [SRGB14] related (1) to a Dirichlet problem.

In this work, we first extend the framework of [SRGB14] to hypergraphs using the *Wasserstein barycenter* [AC11], [AGE18]. For 2-Wasserstein distances this is equivalent to solving a *multi-marginal optimal transport* [CE10] 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 [SRGB14] 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 applications for soft label propagation. The main technical ingredient is *algorithmic stability* [BE02]. To our knowledge, our generalization bound is the first of its type in the literature on Wasserstein distance-based soft label propagation; on graphs these results generalize the error bounds from [BMN04]. As no general semi-supervised learning algorithm is available for large datasets [PWZSK17], this new connection between the Wasserstein barycenter and semi-supervised learning might be of theoretical as well as computational interest.

In the last section, 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 UCI datasets including one on Congressional voting records and another on mushroom characteristics, which are naturally represented using a hypergraph representations.

A. Notation

We denote an undirected simple graph as $G = (V, E)$ where $V = [n] := \{1, \ldots, n\}$ is the vertex set and $E \in V \times V$ denote 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 $L := D - W$, 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

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

Fundamental properties of the minimizer in (3) are studied in [AC11]; similar results hold when the squared 2-Wasserstein distance are weighted differently. Given a hyperedge E of H, we use $\mathrm{bar}(E)$ to denote $\mathrm{bar}\left(\{\mu_i\}_{i=1}^{|E|}\right)$ where the probability measures $\mu_1, \ldots, \mu_{|E|}$ associated with each vertex i in E are clear from the context.

II. Message Passing and Label Propagation on Graphs and Hypergraphs

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 generalized version of Wasserstein label propagation [SRGB14] 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 then 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., $Y = \mathcal{P}(N)$. Because N is complete, the space Y equipped with Wasserstein distance is also a complete metric space [Vil03, Theorem 6.18].

A. 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 [BMN04] 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 $\mu_i := \mu(i)$ 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 remaining vertices. We 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 conceived of as an extension of the Dirichlet boundary problem studied in [SRGB14] 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 $\{(i, \mu_i) \mid 1 \le i \le n\}$ and the given graph structure $G = (V, E)$. We point out that the formulation in [SRGB14] is a special case (parameter-free "interpolated regularization") of (4) in the limit $\gamma \to 0$, for the same reason given in [BMN04, §2.2].

We now provide an algorithm for solving (4) based on belief propagation. Because this is only a motivating perspective, we assume for simplicity that the graph is unweighted, but all arguments below can be extended to weighted graphs with heavier notation. In this context, each vertex i updates its *belief* about the local minimizer of (4) f_i by exchanging messages to edges to which it is incident. The classical min-sum algorithm [MR09] 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

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

and

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

The first term in (5) is set to zero if $i \notin [m]$. The belief is then updated 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 mild conditions on initial beliefs if G is a tree (see e.g., [MR09]).

B. Wasserstein Label Propagation on Hypergraphs

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

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

and

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

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:

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

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.

III. BARYCENTER AND CLIQUE REPRESENTATION

In this section, we assume that 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 remainder of this paper 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 ([Vil03]). *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 ([BGKL17]). *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*

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

Because the inverse cdfs and 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, such as $\{1, 2, \ldots, k\}$ as

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

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

IV. GENERALIZATION BOUNDS FOR WASSERSTEIN PROPAGATION

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

A. Algorithmic Stability

The framework of algorithmic stability [DW79], [BE02], [MNPR06] was proposed in statistical learning as an alternative to the VC-dimension framework. The latter is often overly pessimistic because it attempts to bound the generalization performance uniformly over all 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—that the learning algorithm should return identical maps for two training sets with samples differing from each other only by permutation. We shall assume all maps considered here are measurable, and all measure spaces are separable. We are interested in the case where 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 a 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 [BE02], 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, [BE02]). *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' that differ by exactly 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 ([BE02]). *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 training samples m will be crucial here, otherwise any learning algorithm is uniformly stable for any bounded cost function. In [BE02] 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 [BE02] 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.

B. 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 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 come across in practice, in particular for propagating histograms but are not already satisfied with popular distribution classes such as the Gaussian distribution. It is therefore preferable to work with a model class for distributions with uniformly bounded pairwise Wasserstein distances under mild 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 typically difficult. Many existing bounds explore the multiscale structure of supports from the two distributions [Dav88], [Lei18], [SP18], but it is not clear how those technical conditions can be used as model class specifications. Here we 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

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

Because 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 notation, we 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 vertices from graph G. For simplicity, we 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 $s \in [0,1]$ where each sub-problem is 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 compatibility across 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 for each subproblem (15); see Proposition 2 below.

The solutions for Tikhonov regularization problems (15) were known back in [BMN04]. Let $\mathbf{1} = (1, \cdots, 1)^\top \in \mathbb{R}^n$ be a column vector of all ones, and

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

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 this notation, it becomes 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. Nevertheless, 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 [BMN04, 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, 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 $\Phi_s^* - \bar{y}_s\mathbf{1} = (T_\ell + m\gamma L)^{-1}(\mathbf{y}_s - \bar{y}_s T_\ell\mathbf{1})$. Therefore, the solution to (15) takes the form

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

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 meaning.

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 need 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 $\{\Phi_s^* \mid 0 \leq s \leq 1\}$ piece together and give rise to inverse c.d.f.'s at each vertex $i \in V$. This 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, because 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 this assertion follows from the closed-form expression (21). 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 $\mathbb{R}^n \ni \mathbf{y} \geq 0$ (entrywise) and $(T_\ell + m\gamma L) \Phi = \mathbf{y}$ then $\Phi \geq 0$ (entrywise). This is because we already have $\mathbf{y}_s - \mathbf{y}_t \geq 0$ for any $0 \leq t \leq s \leq 1$ by the monotonicity of the inverse c.d.f.'s, hence such a "maximum principle" would guarantee $\Phi_s - \Phi_t \geq 0$ (entrywise). Such maximum principles abound for graph Laplacians, see e.g., [HS97], [CCK07]. 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 $[(T_\ell + m\gamma L) \Phi](i) \geq 0$ for all $1 \leq i \leq \ell$ and $[(T_\ell + m\gamma L) \Phi](i) = 0$ for all $\ell + 1 \leq i \leq n$, then Φ attains both its maximum and minimum over $i = 1, \cdots, n$ within $\{1, \cdots, \ell\}$. In particular, $\Phi(i) \geq 0$ for all $1 \leq i \leq n$.*

Proof. The conditions on Φ can be written as

$$\left[\frac{t_i}{m\gamma} + \deg(i)\right] \Phi(i) - \sum_{j:j\sim i} \Phi(j) \geq 0 \qquad 1 \leq i \leq \ell \tag{22}$$

$$\deg(i) \Phi(i) - \sum_{j:j\sim i} \Phi(j) = 0 \qquad \ell + 1 \leq i \leq n \tag{23}$$

where $\deg(i) \geq 1$ is the degree of vertex i in graph G. First, we assert that the minimum of Φ must be attained among the vertices $1, \cdots, \ell$, for otherwise, if $\ell + 1 \leq i_* = \arg\min_{i\in V} \Phi(i) \leq n$, then by (23) we have

$$\deg(i_*) \Phi(i_*) = \sum_{j:j\sim i_*} \Phi(j)$$
$$\geq \sum_{j:j\sim i_*} \Phi(i_*) = \deg(i_*) \Phi(i_*)$$

which implies $\Phi(j) = \Phi(i_*)$ for all vertices $j \sim i_*$. This argument can be repeated until the constant value propagates into the vertices within $1, \cdots, \ell$, and the assertion follows from the connectivity of the graph. The assertion for the maximum can be established analogously. Next we argue that the minimum of Φ on the vertices of G must be non-negative. Assume the contracy, i.e. the minimum attained at $i_* \in [1, \ell]$ is strictly negative, then by (22) we have

$$0 \leq \left[\frac{t_{i_*}}{m\gamma} + \deg(i_*)\right] \Phi(i_*) - \sum_{j:j\sim i_*} \Phi(j)$$
$$= \frac{t_{i_*}}{m\gamma} \Phi(i_*) + \sum_{j:j\sim i_*} [\Phi(i_*) - \Phi(j)] < 0$$

where the strict inequalty follows from the counter-assumption $\Phi(i_*) < 0$. This contradiction completes our proof that $\Phi \geq 0$ on the entire graph G. ∎

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), solutions Φ_s satisfy the Euler-Lagrange equations for (15):

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

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

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

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 $\Phi_s^* \geq \Phi_t^*$. ∎

We can now rest assured that the solutions (21) constitute an inverse c.d.f. at each vertex $i \in V$. But there is more: it can be easily verified that (20) is equivalent to the Tikhonov regularization problem formulated in [BMN04] if we view $(\Phi_s - \bar{y}_s\mathbf{1})$ as variables. We can thus follow the idea of [BMN04, Theorem 5] to get algorithmic stability for each individual Φ_s, $s \in [0,1]$.

Theorem 4. *Assume $m \geq 4$ 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 $S = \{(v_i, \mu_i) \mid 1 \leq i \leq m, \ v_i \in V, \ \mu_i \in \mathcal{P}(\mathbb{R})\}$ and $S' = \{(v_i', \mu_i') \mid 1 \leq i \leq m, \ v_i \in V, \ \mu_i \in \mathcal{P}(\mathbb{R})\}$ 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*

$$\max\left\{\left|F_{\mu_i}^{-1}(s)\right|, \left|F_{\mu_i'}^{-1}(s)\right|, \ i = 1, \cdots, m\right\} \leq M_s < \infty. \tag{24}$$

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

$$\|\Phi_s^* - \Phi_s'^*\|_\infty \leq \frac{3M_s\sqrt{Tm}}{(m\gamma\lambda_1 - T)^2} + \frac{4M_s}{m\gamma\lambda_1 - T} + \frac{2M_s}{m}. \tag{25}$$

Proof. Following the same argument as in the proof of [BMN04, Theorem 5], we can assume without loss of generality that S, S' differ by a new point $(v_m, \mu_m) \leftrightarrow (v_m', \mu_m')$; the other case where only the multiplicities differ can be treated similarly. By our assumption (24), the two averages differ by at most an amount of

$$|\bar{y}_s - \bar{y}_s'| \leq \frac{2M_s}{m}.$$

For simplicity, introduce temporary notations

$$A := T_\ell + m\gamma L, \qquad B := T_\ell' + m\gamma L.$$

Using the simple fact that the 2-norm dominate the ∞-norm, we have

$$\|\Phi_s^* - \Phi_s'^*\|_\infty \leq \|\Phi_s^* - \Phi_s'^*\|_2$$
$$\leq \frac{2M_s}{m} + \left\|A^{-1}(\mathbf{y}_s - \bar{y}_s T_\ell \mathbf{1}) - B^{-1}(\mathbf{y}_s' - \bar{y}_s' T_\ell' \mathbf{1})\right\|_2$$
$$\leq \frac{2M_s}{m} + \left\|A^{-1}(\mathbf{y}_s - \bar{y}_s T_\ell \mathbf{1}) - A^{-1}(\mathbf{y}_s' - \bar{y}_s' T_\ell' \mathbf{1})\right\|_2$$
$$+ \left\|A^{-1}(\mathbf{y}_s' - \bar{y}_s' T_\ell' \mathbf{1}) - B^{-1}(\mathbf{y}_s' - \bar{y}_s' T_\ell' \mathbf{1})\right\|_2.$$

Standard functional analysis argument (the same perturbation reasoning we gave in (18)) tells us that $\|A^{-1}\|_2 \leq (m\gamma\lambda_1 - T)^{-1}$. Together with the observation that

$$\|(\mathbf{y}_s - \bar{y}_s T_\ell \mathbf{1}) - (\mathbf{y}_s' - \bar{y}_s' T_\ell' \mathbf{1})\|_2$$
$$\leq \|\mathbf{y}_s - \mathbf{y}_s'\|_2 + \|\bar{y}_s T_\ell \mathbf{1} - \bar{y}_s' T_\ell' \mathbf{1}\|_2$$
$$\leq 2M_s + \frac{2M_s}{m} < 4M_s$$

we have

$$\left\|A^{-1}(\mathbf{y}_s - \bar{y}_s T_\ell \mathbf{1}) - A^{-1}(\mathbf{y}_s' - \bar{y}_s' T_\ell' \mathbf{1})\right\|_2 \leq \frac{4M_s}{m\gamma\lambda_1 - T}.$$

In the meanwhile, noting that we also have $\|B^{-1}\|_2 \leq (m\gamma\lambda_1 - T)^{-1}$, and $\|A - B\|_2 = \|T_\ell' - T_\ell\|_2 \leq \sqrt{2} < 3/2$, we conclude that

$$\left\|A^{-1}(\mathbf{y}_s' - \bar{y}_s' T_\ell' \mathbf{1}) - B^{-1}(\mathbf{y}_s' - \bar{y}_s' T_\ell' \mathbf{1})\right\|_2$$
$$= \left\|B^{-1}(B - A)A^{-1}(\mathbf{y}_s' - \bar{y}_s' T_\ell' \mathbf{1})\right\|_2 \leq \frac{3M_s\sqrt{Tm}}{(m\gamma\lambda_1 - T)^2}.$$

Putting everything together completes the proof. $\blacksquare$

The boundedness assumption on Φ_s seems artificial, but is actually natural: an almost identical argument as the first part of the proof of Lemma 1, with minimum replaced with maximum and *mutatis mutandis*, establishes that the global maximum of Φ_s must be attained at the boundary $1 \leq i \leq \ell$. Hence, because there are only finitely many data in the training set, this boundedness is a mild requirement (e.g., satisfied if each $F_{\mu_i}^{-1}(s)$ is finite). We define a model class to reflect the requirement

that the inverse c.d.f.'s of 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 $\phi \geq 0$ 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 $\left|F_\mu^{-1}(s)\right| \leq \phi(s)$ for e.g., $s \in [0,1]$.*

Lemma 2 (*A Priori* Estimates)**.** *If in the training set $S = \{(v_i, \mu_i) \mid 1 \leq i \leq m, \ v_i \in V, \ \mu_i \in \mathcal{P}(\mathbb{R})\}$ all μ_i lie in a dominated quantile model class $\mathcal{M}_\phi^2$ for some $\phi \in L^2[0,1]$ with $\phi \geq 0$ 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 $\max\left\{\left|F_{\mu_i}^{-1}(s)\right|, \ i = 1, \cdots \right.$ $\phi(s)$ then $\|\Phi_s^*\|_\infty \leq \phi(s)$, 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 $m \geq 4$ 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 $\phi \geq 0$ 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*

$$\beta = 4\|\phi\|_2^2 \left[\frac{3\sqrt{Tm}}{(m\gamma\lambda_1 - T)^2} + \frac{4}{m\gamma\lambda_1 - T} + \frac{2}{m}\right]. \tag{26}$$

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

$$
\begin{aligned}
&\left|\Phi_s^*(j) - \Phi_s'^*(j)\right| \\
&\leq \left[\frac{3\sqrt{Tm}}{(m\gamma\lambda_1 - T)^2} + \frac{4}{m\gamma\lambda_1 - T} + \frac{2}{m}\right]\phi(s).
\end{aligned} \tag{27}
$$

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

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

where at $(*)$ we used (27) to bound the difference $|\Phi_s^*(j) - \Phi_s'^*(j)|$, and invoked Lemma 2 to conclude that

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

and hence

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

Note that the cost function is uniformly bounded by $M = 4\|\phi\|_2^2$ in our setting. Our main result 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 $m \geq 8M^2/\epsilon^2$*

$$\mathbb{P}_{S \sim D^m}\{|R_m(f_S) - R_D(f_S)| > \epsilon\} \leq \frac{64Mm\beta + 8M^2}{m\epsilon^2}, \tag{28}$$

and for any $m \geq 1$

$$\mathbb{P}_{S \sim D^m} \left\{ |R_m\left(f_S\right) - R_D\left(f_S\right)| > \epsilon + \beta \right\}$$

$$\leq 2 \exp\left(-\frac{m\epsilon^2}{2\left(m\beta + M\right)^2} \right), \tag{29}$$

where $M = 4\left\|\phi\right\|_2^2$ and β given by (26).

V. NUMERICAL EXPERIMENTS

A. Label Propagation Algorithm

Alg. 1 details the label propagation algorithm we use to obtain the results in the next two sections.

Algorithm 1: Alternating label propagation algorithm

Data: hypergraph $H = (V, \mathcal{E})$; a subset of vertices V_0 with known labels $\bar{l}(v)$, $\forall v \in V_0$; parameters $\alpha, \gamma > 0$, a condition EC for exiting the main loop on line 19.

Result: labels $l(v)$, $\forall v \in V$.

1   Randomly initialize labels $l(v)$, $\forall v \in V$
2   **for** *every $E \in \mathcal{E}$* **do**
3   **for** *every $v \in E$* **do**
4   **if** *$v \in V_0$* **then**
5         $W_E(v) = \alpha$
6   **else**
7         $W_E(v) = 1$
8   **end**
9   **end**
10   **end**
11   **for** *every $v \in V$* **do**
12   **for** *every $E \in \mathcal{E}$ incident to v* **do**
13       $w_v(E) = 1/|E|$
14   **end**
15   **if** *every $v \in V_0$* **then**
16   append vector W_v with γ
17   **end**
18   **end**
19   **while** EC *is not met* **do**
20   Initialize $loss = 0$
21   **for** *every $E \in \mathcal{E}$* **do**
22       $L_E = (l(v))_{v \in E}$
23       $l(E) = \mathsf{Barycenter}\left(W_E, L_E\right)$
24   **end**
25   **for** *every $v \in V$* **do**
26       $L_v = (l(E))_{E \text{ incident to } v}$
27   **if** *$v \in V_0$* **then**
28   append L_v with $\bar{l}(v)$
29   **end**
30       $l(v) = \mathsf{Barycenter}\left(W_v, L_v\right)$
31   **for** *every $E \in \mathcal{E}$ incident to v* **do**
32         $loss = loss + W_v(E) \cdot \mathsf{WassDist}(l(v), l(E))$
33   **end**
34   **if** *$v \in V_0$* **then**
35         $loss = loss + \gamma \cdot \mathsf{WassDist}\left(l(v), \bar{l}(v)\right)$
36   **end**
37   **end**
38   **end**

The functions Barycenter and WassDist can be any algorithms that calculate the weighted Wasserstein barycenter of a vector of labels L with weights W, and the Wasserstein distance between two input labels, respectively. Note that we introduce

another parameter $\alpha > 1$ to adjust the weights of vertices with known labels (in line 5) in order to increase their influences to hyperedge barycenters. Similar techniques are explored in [SOZ17], [SOZ18].

The algorithm relies on the alternating technique in minimizing (9) in each iteration. This technique consists of two steps: (i) first calculates the barycenters $bar(E)$ of all hyperedges E using the current labels of vertices they contain and treats the derived barycenters as the labels of the hyperedges (lines 21 to 24), and (ii) then calculates the barycenters, i.e. the new labels, of all vertices using labels of the hyperedges incident to them, together with their targeted labels if the latter are known (line 25 to 37). Due to the alternating nature of the algorithm, we call it *alternating label propagation*.

B. Stochastic Block Model

In the first two experiments, we run label propagation on 3-uniform hypergraphs generated using the stochastic block model (SBM) over 100 vertices that are grouped into either 2 or 3 blocks. More specifically, the probability that a hyperedge $\{v_i, v_j, v_r\}$ exists is $p = 0.01$ if all v_i, v_j, and v_r belong to the same block and is $q = 0.002$ otherwise.

We set the soft labels to be b-dimensional Gaussian distributions, where b is the number of blocks. For any vertex from block i, $i = 1, \ldots, b$, whose label is known, we set the mean of its label to be e_i, where e_i is the base vector with the i-th coordinate being 1 and the rest being 0. The covariance matrix of each known label is set to be $0.05I_b$, where I_b is the b-dimensional identity matrix. The predicted block assignment of a vertex is the $\arg\max$ of its predicted mean. In both of the experiments, we use $\alpha = 20$ and $\gamma = 10$. We run the experiments with 5 to 15 vertices of known block assignment from each block, and the error bars are obtained by averaging over 20 random selections of vertices with known labels.

We compare the performance of our label propagation approach with with AdaBoost, random forest, and SVM in Fig. 1. We use incidence matrix as the feature matrix in AdaBoost, Random forest, and SVM to solve the classification problem.
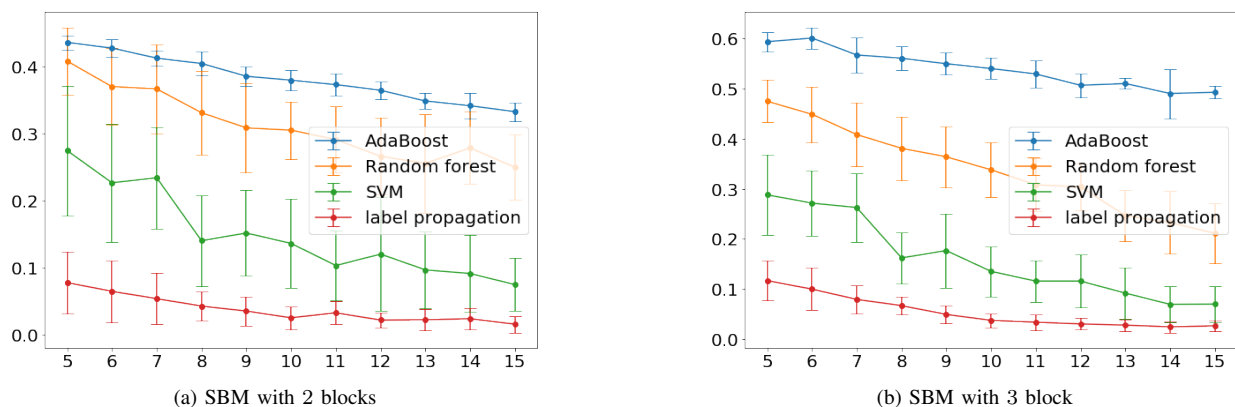

(a) SBM with 2 blocks   (b) SBM with 3 block

Fig. 1. Comparison of traditional classification algorithms with hypergraph label propagation on SBM.

SBM with two blocks: The hypergraph generated for this experiment has two blocks of sizes 50 and 50, and 629 hyperedges with 388 of them containing vertices from one block.

SBM with three blocks: The hypergraph generated for this experiment has three blocks of sizes 33, 33, and 34, and 384 hyperedges with 182 of them has vertices from one block.

C. UCI datasets

In the next two experiments, we apply our label propagation as a classification algorithm to the following two datasets with categorical features from the UCI machine learning repository:

Congressional Voting Records: This dataset contains voting records on 16 issues of the 2nd session of the 98th Congress. We form a pair of hyperedges for each issue each of which contains voters who voted "Yay" and "Nay", respectively. For voters whose votes were missing, we don't include them in any of the hyperedges constructed for the corresponding issue. This resilience to the missing data samples illustrates another advantage of applying hypergraph label propagation to classification problems. We test label propagation algorithm with 5, 10, 15, 20, 25, and 30 congressmen and women from each party whose affiliation are given.

Mushrooms: This dataset contains 22 features (e.g., shapes, colors, and habitats, etc) of 8124 mushrooms. We form 97 hyperedges each of which contains mushrooms sharing identical features. We choose 1000 edible and 1000 poisonous mushrooms to run the experiment. We run the algorithm in 6 cases where 10, 20, 30, 40, 50, and 60 mushrooms are given labels from each category.

In both datasets, the soft labels are either 1-dimensional Gaussian distributions $N(+1, 0.01)$ and $N(-1, 0.01)$ or 2-dimensional Gaussian distributions $N((1, 0), 0.01I_2)$ and $N((0, 1), 0.01I_2)$ depending on which class the labelled sample belongs to. The

predicted class of a vertex is obtained as follows: For the 1-dimensional case, it is the sign of the mean of its label and for the 2-dimensional case, it is $+1$ if the first coordinate of the mean vector of its label is larger than the second coordinate and -1 otherwise. For both experiments, we set $\alpha = 10$ and $\gamma = 1$. The error bars are obtained by averaging 20 random selections of vertices with known labels. We compare the performance of hypergraph label propagation (as a classification algorithm) with SVM in Fig. 2.
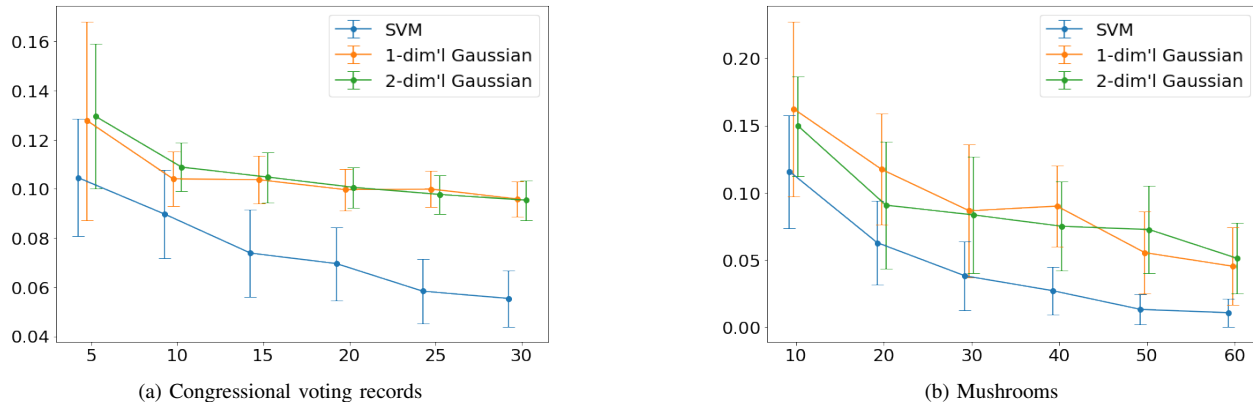

(a) Congressional voting records   (b) Mushrooms

Fig. 2. Comparison of SVM with hypergraph label propagation as a classification algorithm.

D. Discussion of numerical experiments

The above experiments demonstrate that the hypergraph label propagation can serve as a powerful alternative classification algorithm especially when the dataset is structured as a network (for example as in SBM). The reason as to why the traditional classification algorithms may fail on network-like datasets (as illustrated in Fig. 1) is because for these datasets almost all coordinates of a feature vector tend to be identical except for few of them. We can understand these features as describing only local properties of the dataset. Therefore, they can give rise to global characterizations of the datasets, in a substantial way, only when properly "patched" together. Label propagation algorithm provides a novel way of combining features which is shown in Fig. 1 to outperform the classical algorithms.

VI. 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 graphs and simplicial complexes. 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 underlying Wasserstein spaces will be indispensable for those purposes. Future work includes (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

[AC11] M. Agueh and G. Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[AGE18] S. Asoodeh, T. Gao, and J. Evans. Curvature of hypergraphs via multi-marginal optimal transport. In *The 57th IEEE Conference on Decision and Control (CDC 2018)*, 2018.

[AGS05] L. Ambrosio, N. Gigli, and G. Savare. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2005.

[BE02] O. Bousquet and A. Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, March 2002.

[BGKL17] J. Bigot, R. Gouet, T. Klein, and A. López. Geodesic PCA in the wasserstein space by convex PCA. *Ann. Inst. H. Poincaré Probab. Statist.*, 53(1):1–26, 02 2017.

[BMN04] Mikhail Belkin, Irina Matveeva, and Partha Niyogi. Regularization and Semi-Supervised Learning on Large Graphs. In *International Conference on Computational Learning Theory*, pages 624–638. Springer, 2004.

[BP09] S. R. Bulò and M. Pelillo. A game-theoretic approach to hypergraph clustering. In *Advances in Neural Information Processing Systems 22*, pages 1571–1579, 2009.

[CCK07] Soon-Yeong Chung, Yun-Sung Chung, and Jong-Ho Kim. Diffusion and Elastic Equations on Networks. *Publications of the Research Institute for Mathematical Sciences*, 43(3):699–725, sep 2007.

[CE10]   G. Carlier and I. Ekeland. Matching for Teams. *Economic Theory*, 42(2):397–418, 2010.

[CJ05]   A. Corduneanu and T. S. Jaakkola. Distributed information regularization on graphs. In *Advances in Neural Information Processing Systems*, pages 297–304. MIT Press, 2005.

[CSZ06]   O. Chapelle, B. Schölkopf, and A. Zien. *Semi-supervised Learning*. Adaptive computation and machine learning. MIT Press, 2006.

[DAC08]   E. Demir, C. Aykanat, and B. B. Cambazoglu. Clustering spatial networks for aggregate query processing: A hypergraph approach. *Information Systems*, 33(1):1–17, 2008.

[Dav88]   Guy David. Morceaux de Graphes Lipschitziens et Intégrales Singulières sur une Surface. *Revista Matemática Iberoamericana*, 4(1):73–114, apr 1988.

[DW79]   L. Devroye and T. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, sep 1979.

[Gov05]   V. M. Govindu. A tensor decomposition for geometric grouping and segmentation. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 1150–1157 vol. 1, June 2005.

[GZCN09]   Gourab Ghoshal, Vinko Zlatić, Guido Caldarelli, and MEJ Newman. Random hypergraphs and their applications. *Physical Review E*, 79(6):066118, 2009.

[HS97]   Ilkka Holopainen and Paolo M. Soardi. p-Harmonic Functions on Graphs and Manifolds. *Manuscripta Mathematica*, 94(1):95–110, dec 1997.

[HSJR13]   M. Hein, S. Setzer, L. Jost, and S. S. Rangapuram. The total variation on hypergraphs - learning on hypergraphs revisited. In *Advances in Neural Information Processing Systems*, pages 2427–2435, 2013.

[HZY15]   Jin Huang, Rui Zhang, and Jeffrey Xu Yu. Scalable hypergraph learning and processing. In *Proc. of IEEE Int. Conf. on Data Mining (ICDM)*, pages 775–780, 2015.

[JM18]   J. Jost and R. Mulas. Hypergraph laplace operators for chemical reaction networks, 2018.

[Jos94]   J. Jost. Equilibrium maps between metric spaces. *Calculus of Variations and Partial Differential Equations*, 2(2):173–204, May 1994.

[KBG18]   Chiheon Kim, Afonso S Bandeira, and Michel X Goemans. Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach. *arXiv preprint arXiv:1807.02884*, 2018.

[KHT09]   Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. Hypergraphs and cellular networks. *PLoS computational biology*, 5(5):e1000385, 2009.

[Lav17]   H. Lavenant. Harmonic mappings valued in the wasserstein space, 2017.

[Lei18]   Jing Lei. Convergence and Concentration of Empirical Measures under Wasserstein Distance in Unbounded Functional Spaces. *arxiv preprint*, apr 2018.

[LM17]   P. Li and O. Milenkovic. Inhomogeneous hypergraph clustering with applications. In *Advances in Neural Information Processing Systems 30*, pages 2308–2318, 2017.

[LR15]   X. Li and K. Ramchandran. An active learning framework using sparse-graph codes for sparse polynomials and graph sketching. In *Advances in Neural Information Processing Systems 28*, pages 2170–2178, 2015.

[LV09]   John Lott and Cédric Villani. Ricci Curvature for Metric-Measure Spaces via Optimal Transport. *Annals of Mathematics*, pages 903–991, 2009.

[MNPR06]   Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. 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, Jul 2006.

[MR09]   C. C. Moallemi and B. Van Roy. Convergence of min-sum message passing for quadratic optimization. *IEEE Trans. Inf. Theory*, 55(5):2413–2423, May 2009.

[Ott01]   Felix Otto. The Geometry of Dissipative Evolution Equations: the Porous Medium Equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.

[PWZSK17] L. Petegrosso, Z. Li W. Zhang, Y. Saad, and R. Kuang. Low rank label propagation for semi-supervised learning with 1000 millions samples. *arxiv preprint*, Feb. 2017.

[SB11]   A. Subramanya and J. Bilmes. Semi-supervised learning with measure propagation. *Journal of Machine Learning Research*, 12:3311–3370, Nov. 2011.

[See01]   Matthias Seeger. Learning with labeled and unlabeled data. Technical report, University of Edinburgh, 2001.

[SOZ17]   Zuoqiang Shi, Stanley Osher, and Wei Zhu. Weighted nonlocal laplacian on interpolation from sparse data. *Journal of Scientific Computing*, 73(2-3):1164–1177, 2017.

[SOZ18]   Zuoqiang Shi, Stanley Osher, and Wei Zhu. Generalization of the weighted nonlocal laplacian in low dimensional manifold model. *Journal of Scientific Computing*, 75(2):638–656, 2018.

[SP18]   Shashank Singh and Barnabás Póczos. Minimax Distribution Estimation in Wasserstein Distance. *arxiv preprint*, feb 2018.

[SRGB14]   J. Solomon, R. M. Rustamov, L. Guibas, and A. Butscher. Wasserstein propagation for semi-supervised learning. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages 306–314, 2014.

[Tsu05]   Koji Tsuda. Propagating distributions on a hypergraph by dual information regularization. In *Proceedings of the 22Nd International Conference on Machine Learning*, pages 920–927, 2005.

[Vil03]   C. Villani. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003.

[Vil08]   Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[ZHS07]   D. Zhou, Jiayuan H., and B. Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems 19*, pages 1601–1608. MIT Press, 2007.

[Zhu08]   Xiaojin Zhu. Semi-Supervised Learning Literature Survey. Technical report, University of Wisconsin-Madison, 2008.