# Wasserstein Propagation for Semi-Supervised Learning

**Justin Solomon** JUSTIN.SOLOMON@STANFORD.EDU
**Raif M. Rustamov** RUSTAMOV@STANFORD.EDU
**Leonidas Guibas** GUIBAS@CS.STANFORD.EDU
Department of Computer Science, Stanford University, 353 Serra Mall, Stanford, California 94305 USA

**Adrian Butscher** ADRIAN.BUTSCHER@GMAIL.COM
Max Planck Center for Visual Computing and Communication, Campus E1 4, 66123 Saarbrücken, Germany

## Abstract

Probability distributions and histograms are natural representations for product ratings, traffic measurements, and other data considered in many machine learning applications. Thus, this paper introduces a technique for graph-based semi-supervised learning of histograms, derived from the theory of optimal transportation. Our method has several properties making it suitable for this application; in particular, its behavior can be characterized by the moments and shapes of the histograms at the labeled nodes. In addition, it can be used for histograms on non-standard domains like circles, revealing a strategy for manifold-valued semi-supervised learning. We also extend this technique to related problems such as smoothing distributions on graph nodes.

## 1. Introduction

Graph-based semi-supervised learning is an effective approach for learning problems involving a limited amount of labeled data (Singh et al., 2008). Methods in this class typically propagate labels from a subset of nodes of a graph to the rest of the nodes. Usually each node is associated with a real number, but in many applications labels are more naturally expressed as histograms or probability distributions. For instance, the traffic density at a given location can be seen as a histogram over the 24-hour cycle; these densities may be known only where a service has cameras installed but need to be propagated to the entire map. Product ratings, climatic measurements, and other data sources exhibit similar structure.

While methods for numerical labels, such as Belkin & Niyogi (2001); Zhu et al. (2003); Belkin et al. (2006); Zhou & Belkin (2011); Ji et al. (2012) (also see the survey by Zhu (2008) and references therein), can be applied bin-by-bin to propagate normalized frequency counts, this strategy does not model interactions between histogram bins. As a result, a fundamental aspect of this type of data is ignored, leading to artifacts even when propagating Gaussian distributions.

Among first works directly addressing semi-supervised learning of probability distributions is Subramanya & Bilmes (2011), which propagates distributions representing class memberships. Their loss function, however, is based on Kullback-Leibler divergence, which cannot capture interactions between histogram bins. Talukdar & Crammer (2009) allow interactions between bins by essentially modifying the underlying graph to its tensor product with a prescribed bin interaction graph; this approach loses probabilistic structure and tends to oversmooth. Similar issues have been encountered in the mathematical literature (McCann, 1997; Agueh & Carlier, 2011) and in vision/graphics applications (Bonneel et al., 2011; Rabin et al., 2012) involving interpolating probability distributions. Their solutions attempt to find weighted barycenters of distributions, which is insufficient for propagating distributions along graphs.

The goal of our work is to provide an efficient and theoretically sound approach to graph-based semi-supervised learning of probability distributions. Our strategy uses the machinery of *optimal transportation* (Villani, 2003). Inspired by (Solomon et al., 2013), we employ the two-Wasserstein distance between distributions to construct a regularizer measuring the "smoothness" of an assignment of a probability distribution to each graph node. The final assignment is produced by optimizing this energy while fitting the histogram predictions at labeled nodes.

Our technique has many notable properties. As certainty in the known distributions increases, it reduces to the method of label propagation via harmonic functions (Zhu et al., 2003). Also, the moments and other characteristics of the

*Proceedings of the $31^{st}$ International Conference on Machine Learning*, Beijing, China, 2014. JMLR: W&CP volume 32. 

propagated distributions are well-characterized by those of the labeled nodes at minima of our smoothness energy. Our approach does not restrict the class of the distributions provided at labeled nodes, allowing for bi-modality and other non-Gaussian properties. Finally, we prove that under an appropriate change of variables our objective can be minimized using a fast linear solve.

**Overview** We first motivate the problem of propagating distributions along graphs and show why naïve techniques are ineffective (§2). Given this setup, we develop the Wasserstein propagation technique (§3) and discuss its theoretical properties (§3.1). We also show how it can be used to smooth distribution-valued maps from graphs (§3.2) and extend it to more general domains (§4). Finally, after providing algorithmic details (§5) we demonstrate our techniques on both synthetic (§6.1) and real-world (§6.2) data.

## 2. Preliminaries and Motivation

### 2.1. Label Propagation on Graphs

We consider generalization of the problem of label propagation on a graph $G = (V, E)$. Suppose a label function $f$ is known on a subset of vertices $V_0 \subseteq V$, and we wish to extend $f$ to the remainder $V \backslash V_0$. The classical approach of Zhu et al. (2003) minimizes the Dirichlet energy $\mathcal{E}_D[f] := \sum_{(v,w)\in E} \omega_e (f_v - f_w)^2$ over the space of functions taking the prescribed values on $V_0$. Here $\omega_e$ is the weight associated to the edge $e = (v, w)$. $\mathcal{E}_D$ is a measure of smoothness; therefore the minimizer matches the prescribed labels with minimal variation in between. Minimizing this quadratic objective is equivalent to solving $\Delta f = 0$ on $V \backslash V_0$ for an appropriate positive definite Laplacian matrix $\Delta$ (Chung & Yau, 2000). Solutions of this system are well-known to enjoy many regularity properties, making it a sound choice for smooth label propagation.

### 2.2. Propagating Probability Distributions

Suppose, however, that each vertex in $V_0$ is decorated with a probability distribution rather than a real number. That is, for each $v \in V_0$, we are given a probability distribution $\rho_v \in \text{Prob}(\mathbb{R})$. Our goal now is to propagate these distributions to the remaining vertices, generating a *distribution-valued map* $\rho : v \in V \mapsto \rho_v \in \text{Prob}(\mathbb{R})$ associating a probability distribution with every vertex $v \in V$. It must satisfy $\rho_v(x) \geq 0$ for all $x \in \mathbb{R}$ and $\int_{\mathbb{R}} \rho_v(x)\, dx = 1$. In §4 we consider the generalized case $\rho : V \to \text{Prob}(\Gamma)$ for alternative domains $\Gamma$ including subsets of $\mathbb{R}^n$; most of the statements we prove about maps into $\text{Prob}(\mathbb{R})$ extend naturally to this setting with suitable technical adjustments.

In the applications we consider, such a propagation process should satisfy a number of properties:

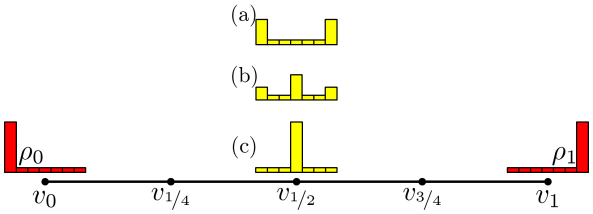



*Figure 1.* Propagating prescribed probability distributions (in red) to interior nodes of path graph identified with the interval $[0, 1]$: (a) naive approach; (b) statistical approach; (c) desirable output.

- The spread of the propagated distributions should be related to the spread of the prescribed distributions.
- As the prescribed distributions in $V_0$ become peaked (concentrated around the mean), the propagated distributions should become peaked around the values obtained by propagating means of prescribed distributions via label propagation (e.g. Zhu et al. (2003)).
- The computational complexity of distribution propagation should be similar to that of scalar propagation.

The simplest method for propagating probability distributions is to extend Zhu et al. (2003) naïvely. For each $x \in \mathbb{R}$, we can view $\rho_v(x)$ as a label at $v \in V$ and solve the Dirichlet problem $\Delta \rho_v(x) = 0$ with $\rho_{v_0}(x)$ prescribed for all $v \in V_0$. The resulting functions $\rho_v(x)$ are distributions because the maximum principle guarantees $\rho_v(x) \geq 0$ for all $x$ and $\int_{\mathbb{R}} \rho_v(x)\, dx = 1$ for all $v \in V$ since these properties hold at the boundary (Chung et al., 2007).

It is easy to see, however, that this method has shortcomings. For instance, consider the case where $G$ is a path graph representing the segment $[0, 1]$ and the labeled vertices are the endpoints, $V_0 = \{0, 1\}$. In this case, the naïve approach results in the linear interpolation $\rho_t(x) := (1 - t)\rho_0(x) + t\rho_1(x)$ at all intermediate graph vertices for $t \in (0, 1)$. The propagated distributions are thus *bimodal* as in Figure 1a. Given our criteria, however, we would prefer an interpolation result closer to Figure 1c, which causes the peak in the boundary data simply to slide from left to right without introducing variance as $t$ changes.

An alternative strategy for propagating probability distributions over $V$ given boundary data on $V_0$ is to use a statistical approach. We could repeatedly draw an independent sample from each distribution in $\{\rho_v : v \in V_0\}$ and propagate the resulting scalars using a classical approach; binning the results of these repeated experiments provides a histogram-style distribution at each vertex in $V$. This strategy has a similar shortcomings to the naïve approach above. For instance, in the path graph example, the interpolated distribution is *trimodal* as in Figure 1b, with nonzero probability at both endpoints and for some $v$ in the interior of $V$.

Of course, the desiderata above are application-specific. One key assumption is that the spread of the distributions is preserved, which differs from existing approaches which tend to blur the distributions. While this property is not intrinsically superior, in a way the experiments in §6 validate not only the algorithmic effectiveness of our technique but also this assumption about probabilistic data on graphs.

## 3. Wasserstein Propagation

*Ad hoc* methods for propagating distributions based on methods for scalar functions tend to have a number of drawbacks. Therefore, we tackle this problem using a technique designed explicitly for the probabilistic setting. To this end, we formulate the semi-supervised problem at hand as the optimization of a Dirichlet energy for distribution-valued maps generalizing the classical Dirichlet energy.

Similar to the construction in (Subramanya & Bilmes, 2011), we replace the square distance between scalar function values appearing in the classical Dirichlet energy (namely the quantity $|f_v - f_w|^2$) with an appropriate distance between the distributions $\rho_v$ and $\rho_w$. Rather than using the bin-by-bin KL divergence, however, we use the *Wasserstein distance* with quadratic cost between probability distributions with finite second moment on $\mathbb{R}$. This distance is defined as

$$\mathcal{W}_2(\rho_v, \rho_w) := \inf_{\pi \in \Pi(\rho_v, \rho_w)} \left( \iint_{\mathbb{R}^2} |x-y|^2 \, d\pi(x,y) \right)^{1/2}$$

where $\Pi(\rho_0, \rho_1) \subseteq \mathrm{Prob}(\mathbb{R}^2)$ is the set of probability distributions $\pi$ on $\mathbb{R}^2$ satisfying the marginal constraints

$$\int_0^1 \pi(x,y)\, dx = \rho_w(y) \quad \text{and} \quad \int_0^1 \pi(x,y)\, dy = \rho_v(x)\,.$$

The Wasserstein distance is a well-known distance metric for probability distributions, sometimes called the quadratic Earth Mover's Distance, and is studied in the field of optimal transportation. It measures the optimal cost of *transporting* one distribution to another, given that the cost of transporting a unit amount of mass from $x$ to $y$ is $|x-y|^2$. $\mathcal{W}_2(\rho_v, \rho_w)$ takes into account not only the values of $\rho_v$ and $\rho_w$ but also the ground distance in the sample space $\mathbb{R}$. It already has shown promise for search and clustering techniques (Irpino et al., 2011; Applegate et al., 2011) and interpolation problems in graphics and vision (Bonneel et al., 2011).

With these ideas in place, we define a Dirichlet energy for a distribution-valued map from a graph into $\mathrm{Prob}(\mathbb{R})$ by

$$\mathcal{E}_D[\rho] := \sum_{(v,w)\in E} \mathcal{W}_2^2(\rho_v, \rho_w)\,, \tag{1}$$

along with the notion of *Wasserstein propagation* of distribution-valued maps given prescribed boundary data.

> WASSERSTEIN PROPAGATION
>
> Minimize $\mathcal{E}_D[\rho]$ in the space of distribution-valued maps with prescribed distributions at all $v \in V_0$.

### 3.1. Theoretical Properties

Solutions of the Wasserstein propagation problem satisfy many desirable properties that we will establish below. Before proceeding, however, we recall a fact about the Wasserstein distance. Let $\rho \in \mathrm{Prob}(\mathbb{R})$ be a probability distribution. Then its cumulative distribution function (CDF) is given by $F(x) := \int_{-\infty}^{x} \rho(y)\, dy$, and the *generalized inverse* of the its CDF is given by $F^{-1}(s) := \inf\{x \in \mathbb{R} : F(x) > s\}$. Then the following result holds.

**Proposition 1.** *[Villani (2003), Theorem 2.18] Let $\rho_0, \rho_1 \in \mathrm{Prob}(\mathbb{R})$ with CDFs $F_0, F_1$. Then*

$$\mathcal{W}_2^2(\rho_0, \rho_1) = \int_0^1 (F_1^{-1}(s) - F_0^{-1}(s))^2\, ds\,. \tag{2}$$

By applying (2) to the minimization problem (1), we obtain a *linear* strategy for our propagation problem.

**Proposition 2.** *Wasserstein propagation can be characterized in the following way. For each $v \in V_0$ let $F_v$ be the CDF of the distribution $\rho_v$. Now suppose that for each $s \in [0,1]$ we determine $g_s : V \to \mathbb{R}$ as the solution of the classical Dirichlet problem*

$$\begin{aligned} \Delta g_s &= 0 \quad \forall\, v \in V \setminus V_0 \\ g_s(v) &= F_v^{-1}(s) \quad \forall\, v \in V_0\,. \end{aligned} \tag{3}$$

*Then for each $v$, the function $s \mapsto g_s(v)$ is the inverse CDF of a probability distribution $\rho_v$. Moreover, the distribution-valued map $v \mapsto \rho_v$ minimizes the Dirichet energy (1).*

*Proof.* Let $\mathcal{X}$ be the set of functions $g : V \times [0,1] \to \mathbb{R}$ satisfying the constraints $g_s(v) = F_v^{-1}(s)$ for all $s \in [0,1]$ and all $v \in V_0$. Consider the minimization problem

$$\min_{g \in \mathcal{X}} \hat{\mathcal{E}}_D(g) := \sum_{(u,v)\in E} \int_0^1 (g_s(u) - g_s(v))^2\, ds\,.$$

The solution of this optimization for each $s$ is exactly a solution of the classical Dirichlet problem (3) on $G$. Moreover, the maximum principle implies that $g_s(v) \le g_{s'}(v)$ whenever $s < s'$, which holds by definition for all $v \in V_0$, can be extended to all $v \in V$ (Chung et al., 2007). Hence $g_s(v)$ can be interpreted as an inverse CDF for each $v \in V$ form which we can define a distribution-valued map $\rho : v \mapsto \rho_v$. Since $\hat{\mathcal{E}}_D$ takes on its minimum value in the subset of $\mathcal{X}$ consisting of inverse CDFs, and $\hat{\mathcal{E}}_D$ coincides with $\mathcal{E}_D$ on this set, $\rho$ is a solution of the Wasserstein propagation problem. □

Distribution-valued maps $\rho : V \to \text{Prob}(\mathbb{R})$ propagated by optimizing (1) satisfy many analogs of functions extended using the classical Dirichlet problem. Two results of this kind concern the *mean* $m(v)$ and the *variance* $\sigma(v)$ of the distributions $\rho_v$ as functions of $V$. These are defined as

$$m(v) := \int_{-\infty}^{\infty} x\rho_v(x)\,dx$$
$$\sigma^2(v) := \int_{-\infty}^{\infty} (x - m(v))^2\rho_v(x)\,dx\,.$$

**Proposition 3.** *Suppose the distribution-valued map $\rho : V \to Prob(\mathbb{R})$ is obtained using Wasserstein propagation. Then for all $v \in V$ the following estimates hold.*

- $\inf_{v_0 \in V_0} m(v_0) \le m(v) \le \sup_{v_0 \in V_0} m(v_0)$.
- $0 \le \sigma(v) \le \sup_{v_0 \in V_0} \sigma(v_0)$.

*Proof.* Both estimates can be derived from the following formula. Let $\rho \in \text{Prob}(\mathbb{R})$ and let $\phi : \mathbb{R} \to \mathbb{R}$ be any integrable function. If we apply the change of variables $s = F(x)$ where $F$ is the CDF of $\rho$ in the integral defining the expectation value of $\phi$ with respect to $\rho$, we get

$$\int_{-\infty}^{\infty} \phi(x)\rho(x)\,dx = \int_0^1 \phi(F^{-1}(s))\,ds\,.$$

Thus $m(v) = \int_0^1 F_v^{-1}(s)\,ds$ and $\sigma^2(v) = \int_0^1 (F_v^{-1}(s) - m(v))^2\,ds$ where $F_v$ is the CDF of $\rho_v$ for each $v \in V$.

Assume $\rho$ minimizes (1) with fixed boundary constraints on $V_0$. By Proposition 2, we then have $\Delta F_v^{-1} = 0$ for all $v \in V$. Therefore $\Delta m(v) = \int_0^1 \Delta F_v^{-1}(s)\,ds = 0$, so $m$ is a harmonic function on $V$. The estimates for $m$ follow by the maximum principle for harmonic functions. Also,

$$\begin{aligned}\Delta[\sigma^2(v)] &= \int_0^1 \Delta(F_v^{-1}(s) - m(v))^2\,ds \\ &= \sum_{(v,v')\in E} \int_0^1 \big(a(v,s) - a(v',s)\big)^2\,ds \\ &\ge 0 \qquad \text{— where } a(v,s) := F_v^{-1}(s) - m(v),\end{aligned}$$

since $\Delta F_v^{-1}(s) = \Delta m(v) = 0$. Thus $\sigma^2$ is a subharmonic function and the upper bound for $\sigma^2$ follows by the maximum principle for subharmonic functions. □

Finally, we check that if we encode a classical interpolation problem using Dirac delta distributions, we recover the classical solution. The essence of this result is that if the boundary data for Wasserstein propagation has zero variance, then the solution must also have zero variance.

**Proposition 4.** *Suppose that there exists $u : V_0 \to \mathbb{R}$ such that $\rho_v(x) = \delta(x - u(v))$ for all $v \in V_0$. Then, the solutions of the classical Dirichlet problem and the Wasserstein propagation problem coincide in the following way. Suppose that $f : V \to \mathbb{R}$ satisfies the classical Dirichlet problem with boundary data $u$. Then $\rho_v(x) := \delta(x - f(v))$ minimizes (1) subject to the fixed boundary constraints.*

*Proof.* The boundary data for $\rho$ given here yields the boundary data $g_s(v) = u(v)$ for all $v \in V_0$ and $s \in [0, 1)$ in the Dirichlet problem (3). The solution of this Dirichlet problem is thus also constant in $s$, let us say $g_s(v) = f(v)$ for all $s \in [0, 1)$ and $v \in V$. The only distributions whose inverse CDFs are of this form are $\delta$-distributions; hence $\rho_v(x) = \delta(x - f(v))$ as desired. □

### 3.2. Application to Smoothing

Using the connection to the classical Dirichlet problem in Proposition 2 we can extend our treatment to other differential equations. There is a large space of differential equations that have been adapted to graphs via the discrete Laplacian $\Delta$; here we focus on the heat equation, considered e.g. in Chung et al. (2007).

The heat equation for scalar functions is applied to smoothing problems; for example, in $\mathbb{R}^n$ solving the heat equation is equivalent to Gaussian convolution. Just as the Dirichlet equation on $F^{-1}$ is equivalent to Wasserstein propagation, heat diffusion on $F^{-1}$ is equivalent to gradient flows of the energy $\mathcal{E}_D$ in (1), providing a straightforward way to understand and implement such a diffusive process.

**Proposition 5.** *Let $\rho : V \to Prob(\mathbb{R})$ be a distribution-valued map and let $F_v : [0, 1] \to \mathbb{R}$ be the CDF of $\rho_v$ for each $v \in V$. Then these two procedures are equivalent:*

- *Mass-preserving flow of $\rho$ in the direction of steepest descent of the Dirichlet energy.*
- *Heat flow of the inverse CDFs.*

*Proof.* A mass-preserving flow of $\rho$ is a family of distribution-valued maps $\rho_\varepsilon : V \to \text{Prob}(\mathbb{R})$ with $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$ that satisfies the equations

$$\left.\begin{aligned}\frac{\partial \rho_{v,\varepsilon}(t)}{\partial \varepsilon} + \frac{\partial}{\partial t}\big(Y_v(\varepsilon, t)\rho_{v,\varepsilon}(t)\big) = 0 \\ \rho_{v,0}(t) = \rho_v(t)\end{aligned}\right\} \quad \forall\, v \in V$$

where $Y_v : (-\varepsilon_0, \varepsilon_0) \times \mathbb{R} \to \mathbb{R}$ is an arbitrary function that governs the flow. By applying the change of variables $t = F_{v,\varepsilon}^{-1}(s)$ using the inverse CDFs of the $\rho_{v,\varepsilon}$, we find that this flow is equivalent to the equations

$$\left.\begin{aligned}\frac{\partial F_{v,\varepsilon}^{-1}(s)}{\partial \varepsilon} = Y_v(\varepsilon, F_{v,\varepsilon}^{-1}(s)) \\ F_{v,0}^{-1}(s) = F_v^{-1}(s)\end{aligned}\right\} \quad \forall\, v \in V.$$

A short calculation starting from (1) now leads to the derivative of the Dirichlet energy under such a flow, namely

$$\frac{d\mathcal{E}_D(\rho_\varepsilon)}{d\varepsilon} = -2\sum_{v\in V}\int_0^1 \Delta(F_{v,\varepsilon}^{-1})\cdot Y_v(\varepsilon, F_{v,\varepsilon}^{-1}(s))\,ds\,.$$

Thus, steepest descent for the Dirichlet energy is achieved by choosing $Y_v(\varepsilon, F_{v,\varepsilon}^{-1}(s)) := \Delta(F_{v,\varepsilon}(s))$ for each $v, \varepsilon, s$. As a result, the equation for the evolution of $F_{v,\varepsilon}^{-1}$ becomes

$$\left.\begin{aligned}\frac{\partial F_{v,\varepsilon}^{-1}(s)}{\partial\varepsilon} &= \Delta(F_{v,\varepsilon}^{-1}(s))\\ F_{v,0}^{-1}(s) &= F_v^{-1}(s)\end{aligned}\right\}\ \forall\, v\in V$$

which is exactly heat flow of $F_{v,\varepsilon}^{-1}$. □

## 4. Generalization

Our preceding discussion involves distribution-valued maps into Prob($\mathbb{R}$), but in a more general setting we might wish to replace Prob($\mathbb{R}$) with Prob($\Gamma$) for an alternative domain $\Gamma$ carrying a distance metric $d$. Our original formulation of Wasserstein propagation easily handles such an extension by replacing $|x-y|^2$ with $d(x,y)^2$ in the definition of $\mathcal{W}_2$. Furthermore, although proofs in this case are considerably more involved, some key properties proved above for Prob($\mathbb{R}$) extend naturally.

In this case, we no longer can rely on the computational benefits of Propositions 2 and 5 but can solve the propagation problem directly. If $\Gamma$ is discrete, then Wasserstein distances between $\rho_v$'s can be computed using a linear program. Suppose we represent two histograms as $\{a_1,\dots,a_m\}$ and $\{b_1,\dots,b_m\}$ with $a_i, b_i \geq 0\ \forall i$ and $\sum_i a_i = \sum_i b_i = 1$. Then, the definition of $\mathcal{W}_2$ yields the optimization:

$$\begin{aligned}\mathcal{W}_2^2(\{a_i\},\{b_j\}) &= \min \sum_{ij} d_{ij}^2 x_{ij} \qquad (4)\\ \text{s.t. } \sum_j x_{ij} = a_i\ \forall i &\quad \sum_i x_{ij} = b_j\ \forall j \quad x_{ij}\geq 0\ \forall i,j\end{aligned}$$

Here $d_{ij}$ is the distance from bin $i$ to bin $j$, which need not be proportional to $|i-j|$.

From this viewpoint, the energy $\mathcal{E}_D$ from (1) remains convex in $\rho$ and can be optimized using a linear program simply by summing terms of the form (4) above:

$$\begin{aligned}\min_{\rho,x} &\sum_{e\in E}\sum_{ij} d_{ij}^2 x_{ij}^{(e)}\\ \text{s.t. } &\sum_j x_{ij}^{(e)} = \rho_{vi}\ \forall e=(v,w)\in E, i\in S\\ &\sum_i x_{ij}^{(e)} = \rho_{wj}\ \forall e=(v,w)\in E, j\in S\\ &\sum_i \rho_{vi} = 1\ \forall v\in V \qquad \rho_{vi}\text{ fixed }\forall v\in V_0\\ &\rho_{vi}\geq 0\ \forall v\in V, i\in S \qquad x_{ij}\geq 0\ \forall i,j\in S\end{aligned}$$

where $S=\{1,\dots,m\}$.

## 5. Algorithm Details

We handle the general case from §4 by optimizing the linear programming formulation directly. Given the size of these linear programs, we use large-scale barrier method solvers.

The characterizations in Propositions 2 and 5, however, suggest a straightforward discretization and accompanying set of optimization algorithms in the linear case. In fact, we can recover propagated distributions by inverting the graph Laplacian $\Delta$ via a sparse linear solve, leading to near-real-time results for moderately-sized graphs $G$.

For a given graph $G=(V,E)$ and subset $V_0\subseteq V$, we discretize the domain $[0,1]$ of $F_v^{-1}$ for each $v$ using a set of evenly-spaced samples $s_0 = 0, s_1, \dots, s_m = 1$. This representation supports any $\rho_v$ provided it is possible to sample the inverse CDF from Proposition 1 at each $s_i$. In particular, when the underlying distributions are histograms, we model $\rho_v$ using $\delta$ functions at evenly-spaced bin centers, which have piecewise constant CDFs; we model continuous $\rho_v$ using piecewise linear interpolation. Regardless, in the end we obtain a non-decreasing set of samples $(F^{-1})_v^1,\dots,(F^{-1})_v^m$ with $(F^{-1})_v^1 = 0$ and $(F^{-1})_v^m = 1$.

Now that we have sampled $F_v^{-1}$ for each $v\in V_0$, we can propagate to the remainder $V\backslash V_0$. For each $i\in\{1,\dots,m\}$, we solve the system from (3):

$$\begin{aligned}\Delta g &= 0 \quad \forall\, v\in V\setminus V_0\\ g(v) &= (F^{-1})_v^i \quad \forall\, v\in V_0\,.\end{aligned} \qquad (5)$$

In the diffusion case, we replace this system with implicit time stepping for the heat equation, iteratively applying $(I-t\Delta)^{-1}$ to $g$ for diffusion time step $t$. In either case, the linear solve is sparse, symmetric, and positive definite; we apply Cholesky factorization to solve the systems directly.

This process propagates $F^{-1}$ to the entire graph, yielding samples $(F^{-1})_v^i$ for all $v\in V$. We invert once again to yield samples $\rho_v^i$ for all $v\in V$. Of course, each inversion incurs some potential for sampling and discretization error, but in practice we are able to oversample sufficiently to overcome most potential issues. When the inputs $\rho_v$ are discrete histograms, we return to this discrete representation by integrating the resulting $\rho_v\in$ Prob($[0,1]$) over the width of the bin about the center defined above.

This algorithm is efficient even on large graphs and is easily parallelizable. For instance, the initial sampling steps for obtaining $F^{-1}$ from $\rho$ are parallelizable over $v\in V_0$, and the linear solve (5) can be parallelized over samples $i$. Direct solvers can be replaced with iterative solvers for particularly

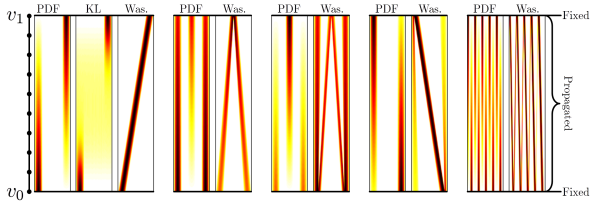



*Figure 2.* Comparison of propagation strategies on a linear graph (coarse version on left); each horizontal slice represents a vertex $v \in V$, and the colors from left to right in a slice show $\rho_v$. (Subramanya & Bilmes, 2011) (KL) is shown only in one example because it has qualitatively similar behavior to the PDF strategy.

large graphs $G$; regardless, the structure of such a solve is well-understood and studied, e.g. in Krishnan et al. (2013).

## 6. Experiments

We run our scheme through a number of tests demonstrating its strengths and weaknesses compared to other potential methods for propagation. We compare Wasserstein propagation with the strategy of propagating probability distribution functions (PDFs) directly, as described in §2.2.

### 6.1. Synthetic Tests

We begin by considering the behavior of our technique on synthetic data designed to illustrate its various properties.

**One-Dimensional Examples** Figure 2 shows "displacement interpolation" properties inherited by our propagation technique from the theory of optimal transportation. The underlying graph is a line as in Figure 1, along the vertical axis. Horizontally, each image is colored by values in $\rho_v$.

The bottom and top vertices $v_0$ and $v_1$ have fixed distributions $\rho_{v_0}$ and $\rho_{v_1}$, and the remaining vertices receive $\rho_v$ via one of two propagation techniques. The left of each pair propagates distributions by solving a classical Dirichlet problem independently for each bin of the probability distribution function (PDF) $\rho_v$, whereas the right of each pair propagates inverse CDFs using our method in §5.

By examining the propagation behavior from the bottom to the top of this figure, it is easy to see how the naïve PDF method varies from Wasserstein propagation. For instance, in the leftmost example both $\rho_{v_0}$ and $\rho_{v_1}$ are unimodal, yet when propagating PDFs all the intermediate vertices have bimodal distributions; furthermore, no relationship is determined between the two peaks. Contrastingly, our technique identifies the modes of $\rho_{v_0}$ and $\rho_{v_1}$, linearly moving the peak from one side to the other.

**Boundary Value Problems** Figure 3 illustrates our algorithm on a less trivial graph $G$. To mimic a typical test case for classical Dirichlet problems, our graph is a mesh of the unit circle, and we propagate $\rho_v$ from fixed distributions on the boundary. Unlike the classical case, however, our prescribed boundary distributions $\rho_v$ are multimodal. Once again, Wasserstein propagation recovers a smoothly-varying set of distributions whose peaks behave like solutions to the classical Dirichlet problem. Propagating probability directions rather than inverse CDFs yields somewhat similar modes, but with much higher entropy and variance especially at the center of the circle.

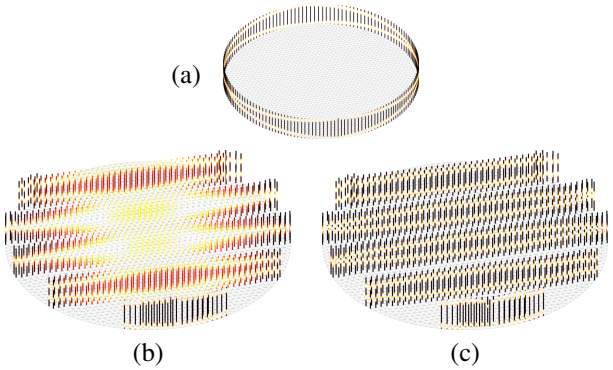



*Figure 3.* PDF (b) and Wasserstein (c) propagation on a meshed circle with prescribed boundary distributions (a). The underlying graph is shown in grey, and probability distributions at vertices $v \in V$ are shown as vertical bars colored by the density $\rho_v$; we invert the color scheme of Figures 2 and 4 to improve contrast. Propagated distributions in (b) and (c) are computed for *all* vertices but for clarity are shown at representative slices of the circle.

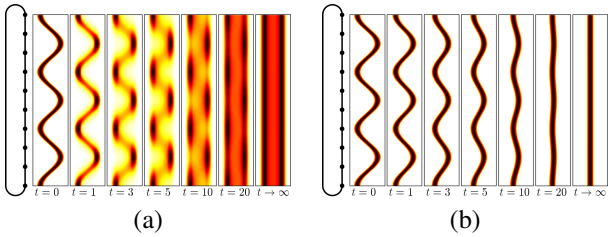
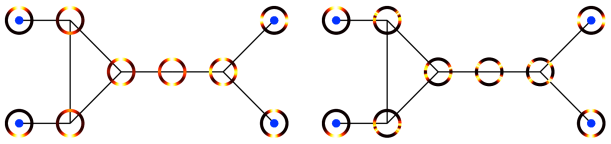



*Figure 4.* Comparison of PDF diffusion (a) and Wasserstein diffusion (b); in both cases the leftmost distribution comprises the initial conditions, and several time steps of diffusion are shown left-to-right. The underlying graph $G$ is the circle on the left.

**Diffusion** Figure 4 illustrates the behavior of Wasserstein diffusion compared with simply diffusing distribution values directly. When PDF values are diffused directly, as time $t$ increases the distributions simply become more and more smooth until they are uniform not only along $G$ but also as distributions on $\text{Prob}([0, 1])$. Contrastingly, Wasserstein diffusion preserves the uncertainty from the initial distributions but does not increase it as time progresses.

**Alternative Target Domain** Figure 5 shows an example in which the target is $\text{Prob}(\mathbb{S}^1)$, where $\mathbb{S}^1$ is the unit circle, rather than $\text{Prob}([0, 1])$. We optimize the $\mathcal{E}_D$ using the



(a) (b)

*Figure 5.* Interpolation of distributions on $\mathbb{S}^1$ via (a) PDF propagation and (b) Wasserstein propagation; in these figures the vertices with valence 1 have prescribed distributions $\rho_v$ and the remaining vertices have distributions from propagation.

linear program in §4 rather than the linear algorithm for $\text{Prob}([0,1])$. Conclusions from this example are similar to those from Figure 3: Wasserstein propagation identifies peaks from different prescribed boundary distributions without introducing variance, while PDF propagation exhibits much higher variance in the interpolated distributions and does not "move" peaks from one location to another.

### 6.2. Real-World Data

We now evaluate our techniques on real-world input. To evaluate the quality of our approach relative to ground truth, we will use the *one*-Wasserstein distance, or Earth Mover's Distance (Rubner et al., 2000), formulated by removing the square in the formula for $\mathcal{W}_2^2$. We use this distance, given on $\text{Prob}(\mathbb{R})$ by the $L^1$ distance between (non-inverted) CDFs, because it does not favor the $\mathcal{W}_2$ distance used in Wasserstein propagation while taking into account the ground distances. We consider weather station coordinates as defining a point cloud on the plane and compute the point cloud Laplacian using the approach of (Coifman & Lafon, 2006).

**Temperature Data** Figure 6 illustrates the results of a series of experiments on weather data on a map of the United States.[1] Here, we have $|V| = 1113$ sites each collecting daily temperature measurements, which we classify into 100 bins at each vertex. In each experiment, we choose a subset $V_0 \subseteq V$ of vertices, propagate the histograms from these vertices to the remainder of $V$, and measure the error between the propagated and ground-truth histograms.

Figure 6a shows quantitative results of this experiment. Here we show the average histogram error per vertex as a function of the percent of nodes in $V$ with fixed labels; the fixed vertices are chosen randomly, and errors are averaged over 20 trials for each percentage. The Wasserstein strategy consistently outperforms naïve PDF interpolation with respect to our error metric and approaches relatively small error with as few as $5\%$ of the labels fixed.

Figures 6b and 6c show results for a single trial. We color the vertices $v \in V$ by the mean (b) and standard deviation (c) of $\rho_v$ from PDF and Wasserstein propagation. Both yield similar mean temperatures on $V \backslash V_0$, which agree with the means of the ground truth data. The standard deviations, however, better illustrate differences between the approaches. In particular, the standard deviations of the Wasserstein-propagated distributions approximately follow those of the ground truth histograms, whereas the PDF strategy yields high standard deviations nearly everywhere on the map due to undesirable smoothing effects.

**Wind Directions** We apply the general formulation in §4 to propagating distributions on the unit circle $\mathbb{S}^1$ by considering histograms of wind *directions* collected over time by nodes on the ocean outside of Australia.[2]

In this experiment, we keep approximately $4\%$ of the data points and propagate to the remaining vertices. Both the PDF and Wasserstein propagation strategies score similarly with respect to our error metric; in the experiment shown, Wasserstein propagation exhibits $6.6\%$ average error per node and PDF propagation exhibits $6.1\%$ average error per node. Propagation results are illustrated in Figure 7a.

The nature of the error from the two strategies, however, is quite different. In particular, Figure 7b shows the same map colored by the entropy of the propagated distributions. PDF propagation exhibits high entropy away from the prescribed vertices, reflecting the fact that the propagated distributions at these points approach uniformity. Wasserstein propagation, on the other hand, has a more similar pattern of entropy to that of the ground truth data, reflecting structure like that demonstrated in Proposition 3.

**Non-Euclidean Interpolation** Proposition 4 suggests an application outside histogram propagation. In particular, if the vertices of $V_0$ have prescribed distributions that are $\delta$ functions encoding individual points as mapping targets, all propagated distributions also will be $\delta$ functions. Thus, one strategy for interpolation is to encode the problem probabilistically using $\delta$ distributions, interpolate using Wasserstein propagation, and then extract peaks of the propagated distributions. Experimentally we find that optima of the linear program in §4 with peaked prescribed distributions yield peaked distributions $\rho_v$ for all $v \in V$ even when the target is not $\text{Prob}(\mathbb{R})$; we leave a proof for future work.

In Figure 8, we apply this strategy to interpolating angles on $\mathbb{S}^1$ from a single day of wind data on a map of Europe.[3] Classical Dirichlet interpolation fails to capture the identification of angles $0$ and $2\pi$. Contrastingly, if we encode the boundary conditions as peaked distributions on $\text{Prob}(\mathbb{S}^1)$, we can interpolate using Wasserstein propagation without losing structure. The resulting distributions are peaked about a sin-

[1] National Climatic Data Center

[2] WindSat Remote Sensing Systems

[3] Carbon Dioxide Information Analysis Center

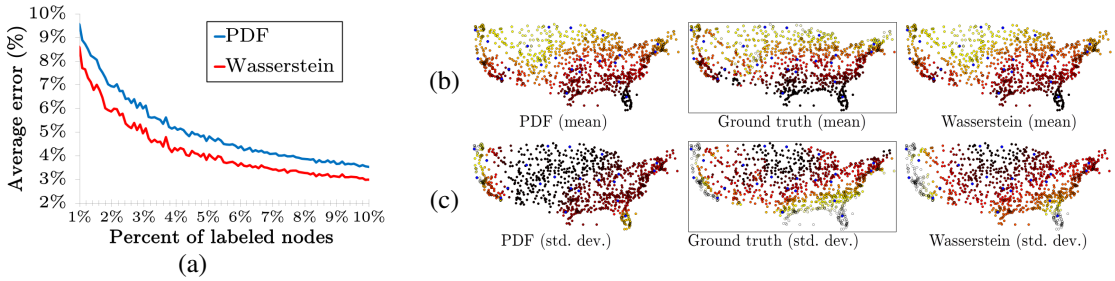



*Figure 6.* We propagate histograms of temperatures collected over time to a map of the United States: (a) Average error at propagated sites as a function of the number of nodes with labeled distributions; (b) means of the histograms at the propagated sites from a typical trial in (a); (c) standard deviations at the propagated sites. Vertices with prescribed distributions are shown in blue and comprise $\sim 2\%$ of $V$.


*Figure 7.* (a) Interpolating histograms of wind directions using the PDF and Wasserstein propagation methods, illustrated using the same scheme as Figure 5; (b) entropy values from the same distributions.

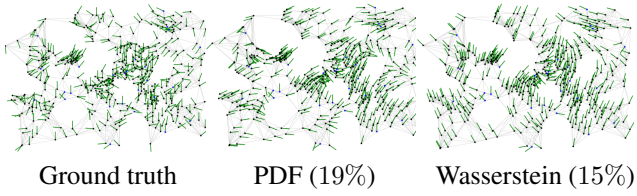



*Figure 8.* Learning wind directions on the unit circle $\mathbb{S}^1$.

gle maximum, so we extract a direction field as the mode of each $\rho_v$. Despite noise in the dataset we achieve $15\%$ error rather than the $19\%$ error obtained by classical Dirichlet interpolation of angles disregarding periodicity.

## 7. Conclusion

It is easy to formulate strategies for histogram propagation by applying methods for propagating scalar functions bin-by-bin. Here, however, we have shown that propagating instead inverse CDFs has a deep connections to the theory of optimal transportation and provides superior results, making it a strong yet still efficient choice. This basic connection gives our method theoretical and practical soundness that is difficult to guarantee otherwise.

While our algorithms show promise as practical techniques, we leave many avenues for future study. Most prominently, the generalization in §4 can be applied to many problems, such as the surface mapping problem in Solomon et al. (2013). Such an optimization, however, has $O(m^2|E|)$ variables, which is intractable for dense or large graphs. An open theoretical problem might be to reduce the number of variables asymptotically. Some simplifications may also be afforded using approximations like (Pele & Werman, 2009), which simplify the form of $d_{ij}$ at the cost of complicating theoretical analysis and understanding of optimal distributions $\rho_v$. Alternatively, work such as (Rabin et al., 2011) suggests the potential to formulate efficient algorithms when replacing $\text{Prob}([0,1])$ with $\text{Prob}(\mathbb{S}^1)$ or other domains with special structure.

In the end, our proposed algorithms are equally as lightweight as less principled alternatives, while exhibiting practical performance, theoretical soundness, and the possibility of extension into several alternative domains.

**Acknowledgments** The authors gratefully acknowledge the support of NSF grants CCF 1161480 and DMS 1228304, AFOSR grant FA9550-12-1-0372, a Google research award, the Max Planck Center for Visual Computing and Communications, the National Defense Science and Engineering Graduate Fellowship, the Hertz Foundation Fellowship, and the NSF GRF program.

## References


Agueh, M. and Carlier, G. Barycenters in the Wasserstein space. *J. Math. Anal.*, 43(2):904–924, 2011. 1

Applegate, David, Dasu, Tamraparni, Krishnan, Shankar, and Urbanek, Simon. Unsupervised clustering of multidimensional distributions using earth mover distance. In *KDD*, pp. 636–644, 2011. 3

Belkin, Mikhail and Niyogi, Partha. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, pp. 585–591, 2001. 1

Belkin, Mikhail, Niyogi, Partha, and Sindhwani, Vikas. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7: 2399–2434, December 2006. 1

Bonneel, Nicolas, van de Panne, Michiel, Paris, Sylvain, and Heidrich, Wolfgang. Displacement interpolation using Lagrangian mass transport. *Trans. Graph.*, 30(6):158:1–158:12, December 2011. 1, 3

Chung, Fan and Yau, S.-T. Discrete Green's functions. *J. Combinatorial Theory*, 91(1–2):191–214, 2000. 2.1

Chung, Soon-Yeong, Chung, Yun-Sung, and Kim, Jong-Ho. Diffusion and elastic equations on networks. *Pub. RIMS*, 43(3):699–726, 2007. 2.2, 3.1, 3.2

Coifman, Ronald R. and Lafon, Stéphane. Diffusion maps. *Applied and Computational Harmonic Anal.*, 21(1):5–30, 2006. 6.2

Irpino, Antonio, Verde, Rosanna, and de A.T. de Carvalho, Francisco. Dynamic clustering of histogram data based on adaptive squared Wasserstein distances. *CoRR*, abs/1110.1462, 2011. 3

Ji, Ming, Yang, Tianbao, Lin, Binbin, Jin, Rong, and Han, Jiawei. A simple algorithm for semi-supervised learning with improved generalization error bound. In *ICML*, 2012. 1

Krishnan, Dilip, Fattal, Raanan, and Szeliski, Richard. Efficient preconditioning of Laplacian matrices for computer graphics. *Trans. Graph.*, 32(4):142:1–142:15, July 2013. 5

McCann, Robert J. A convexity principle for interacting gases. *Advances in Math.*, 128(1):153–179, 1997. 1

Pele, O. and Werman, M. Fast and robust earth mover's distances. In *ICCV*, pp. 460–467, 2009. 7

Rabin, Julien, Delon, Julie, and Gousseau, Yann. Transportation distances on the circle. *J. Math. Imaging Vis.*, 41(1–2):147–167, September 2011. 7

Rabin, Julien, Peyre, Gabriel, Delon, Julie, and Bernot, Marc. Wasserstein barycenter and its application to texture mixing. volume 6667 of *LNCS*, pp. 435–446. Springer, 2012. 1

Rubner, Yossi, Tomasi, Carlo, and Guibas, Leonidas. The earth mover's distance as a metric for image retrieval. *IJCV*, 40(2):99–121, November 2000. 6.2

Singh, Aarti, Nowak, Robert D., and Zhu, Xiaojin. Unlabeled data: Now it helps, now it doesn't. In *NIPS*, pp. 1513–1520, 2008. 1

Solomon, Justin, Guibas, Leonidas, and Butscher, Adrian. Dirichlet energy for analysis and synthesis of soft maps. *Comp. Graph. Forum*, 32(5):197–206, 2013. 1, 7

Subramanya, Amarnag and Bilmes, Jeff. Semi-supervised learning with measure propagation. *JMLR*, 12:3311–3370, 2011. 1, 3, 2

Talukdar, Partha Pratim and Crammer, Koby. New regularized algorithms for transductive learning. *ECML-PKDD*, 5782:442–457, 2009. 1

Villani, Cédric. *Topics in Optimal Transportation*. Graduate Studies in Mathematics. AMS, 2003. 1, 1

Zhou, Xueyuan and Belkin, Mikhail. Semi-supervised learning by higher order regularization. *ICML*, 15:892–900, 2011. 1

Zhu, Xiaojin. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2008. 1

Zhu, Xiaojin, Ghahramani, Zoubin, and Lafferty, John D. Semi-supervised learning using Gaussian fields and harmonic functions. pp. 912–919, 2003. 1, 2.1, 2.2