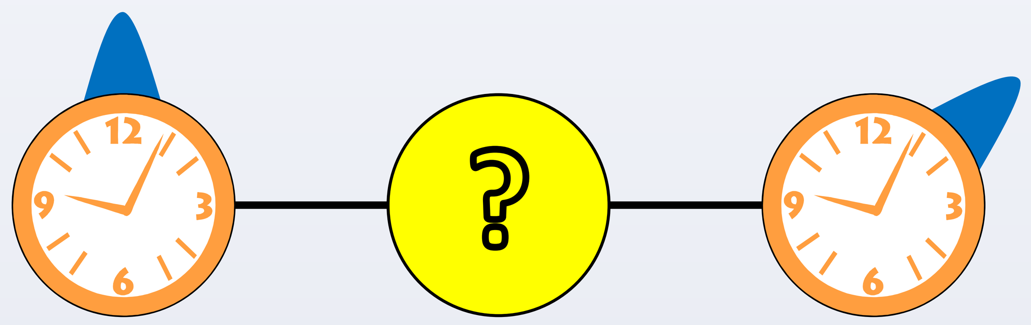
Wasserstein Propagation for Semi-Supervised Learning

Justin Solomon* • Raif Rustamov* • Leonidas Guibas* • Adrian Butscher†

*Stanford Department of Computer Science • †Autodesk Research

Motivating Example

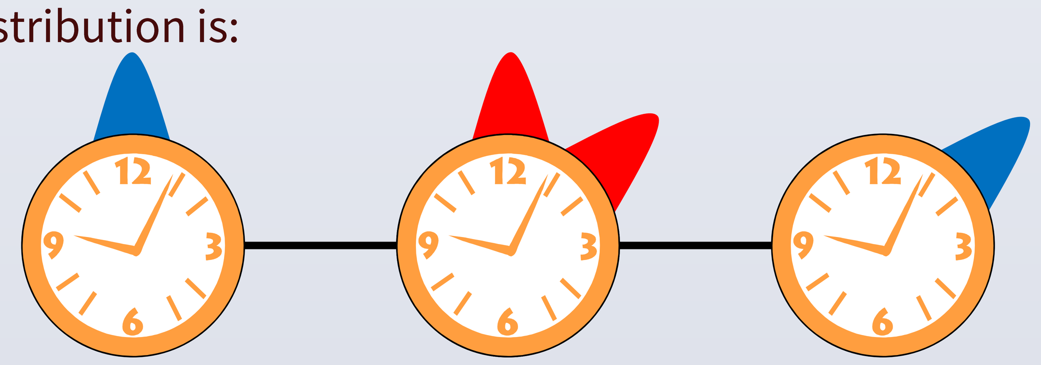
Suppose we have a website consisting of three pages connected by links. We collect traffic on two of three pages as **histograms over the clock**:



We wish to predict traffic statistics on the second page. If traffic flows along links, we might assume that adjacent histograms are similar and minimize:

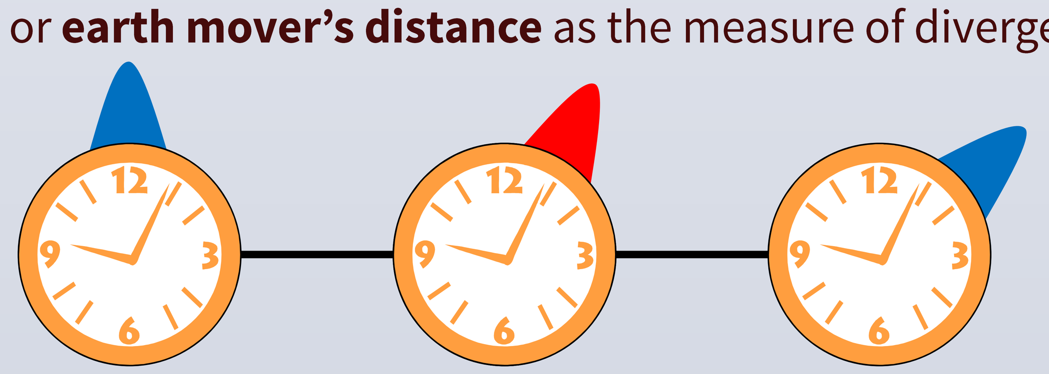
$$\min_h \sum_{(v,w)=e} d^2(h_v, h_w)$$

But, the **measure of divergence** $d(h_v, h_w)$ between histograms matters. For example, if d comes from the KL divergence [Subramanya & Bilmes 2011], the predicted distribution is:



This result is **bimodal** and does not slide along the clock as we might expect.

Instead, we propose **Wasserstein propagation**, which uses the quadratic **Wasserstein** or **earth mover's distance** as the measure of divergence:



Now, the predicted distribution of web traffic is **single-peaked** at the **intermediate time**.

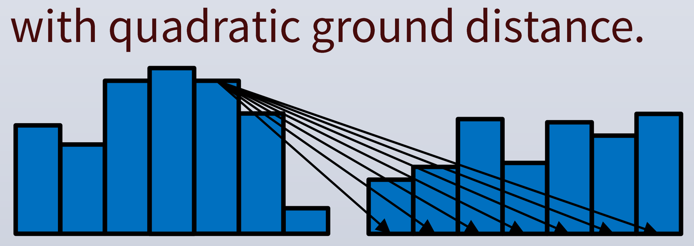
Our model respects the **geometry of the domain** and reduces to **Dirichlet label propagation** [Zhu et al. 2003] as the fixed boundary histograms become peaked about single values. We provide a **general linear programming formulation** and show that a common case can be solved using **positive definite linear machinery**.

Optimal Transportation Distances

Our technique is built using the **Wasserstein distance** between probability distributions. Take $\rho_v, \rho_w \in \mathrm{Prob}(\mathbb{R}^2)$. Then, this distance is given by:

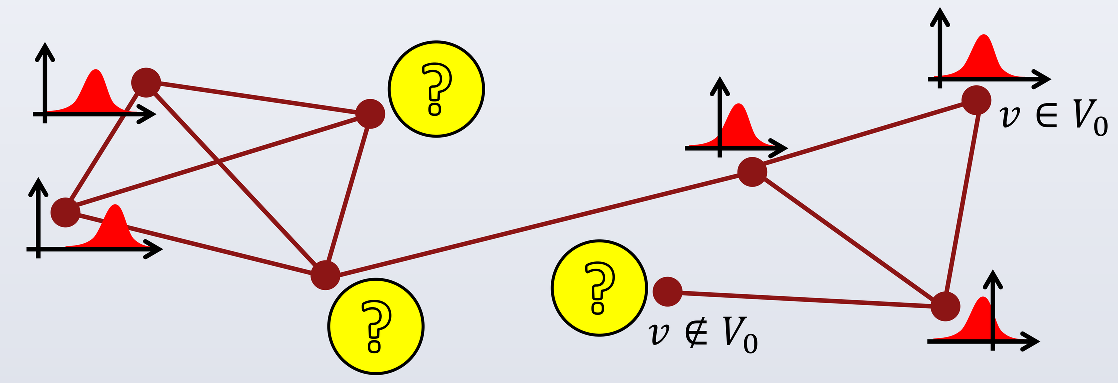
$$\mathcal{W}_2(\rho_v, \rho_w) := \inf_{\pi \in \Pi(\rho_v, \rho_w)} \left(\iint_{\mathbb{R}^2} |x-y|^2 \, d\pi(x,y) \right)^{1/2}$$

$\Pi(\rho_v, \rho_w)$ denotes the set of distributions over $\mathbb{R}^2 \times \mathbb{R}^2$ marginalizing to ρ_v and ρ_w, resp. Intuitively, this distance measures the **minimum work moving the mass of ρ_v to ρ_w** with quadratic ground distance.



Wasserstein Propagation

We study **semi-supervised propagation** of **probability distribution labels** associated with nodes of a graph $G = (V, E)$ given labels on a subset $V_0 \subseteq V$.



For a **distribution-valued map** $\rho: V \to \mathrm{Prob}(D)$ we define a **Dirichlet energy** measuring smoothness along edges:

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

Then, our technique for learning the missing histograms can be described as:

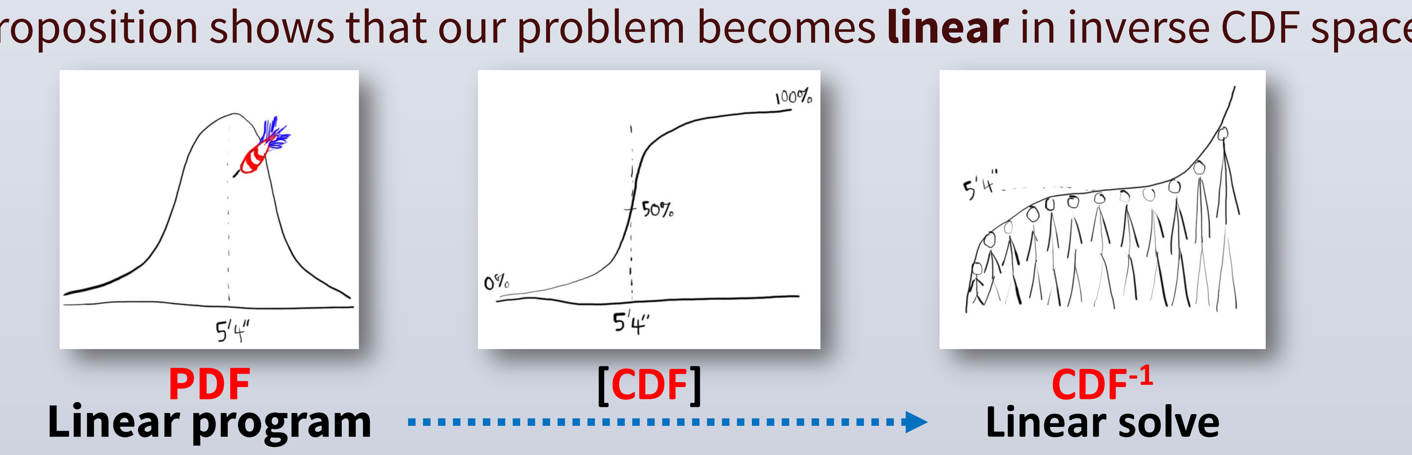
> WASSERSTEIN PROPAGATION
> Minimize $\mathcal{E}_D[\rho]$ in the space of distribution-valued maps with prescribed distributions at all $v \in V_0$.

Computation on Prob(ℝ)

Suppose $\rho_v, \rho_w \in \mathrm{Prob}(\mathbb{R})$ with **cumulative distribution functions (CDFs)** F_v, F_w. Then, $W_2(\rho_v, \rho_w) = \|F_v^{-1} - F_w^{-1}\|_2$, the Euclidean distance between inverse CDFs [Villani 2003]. Starting from this formula, we prove:

Proposition. *For each $v \in V_0$, let F_v be the CDF of ρ_v. For each $s \in [0,1]$ determine $g_s: V \to \mathbb{R}$ as the solution of the **classical Dirichlet problem** $\Delta g_s = 0 \; \forall v \in V \backslash V_0$ with $g_s(v) = F_v^{-1}(s) \; \forall v \in V_0$. Then, for each v, the function $s \mapsto g_s(v)$ is the inverse CDF of a probability distribution ρ_v, and the resulting map $v \mapsto \rho_v$ **minimizes the Dirichet energy**.*

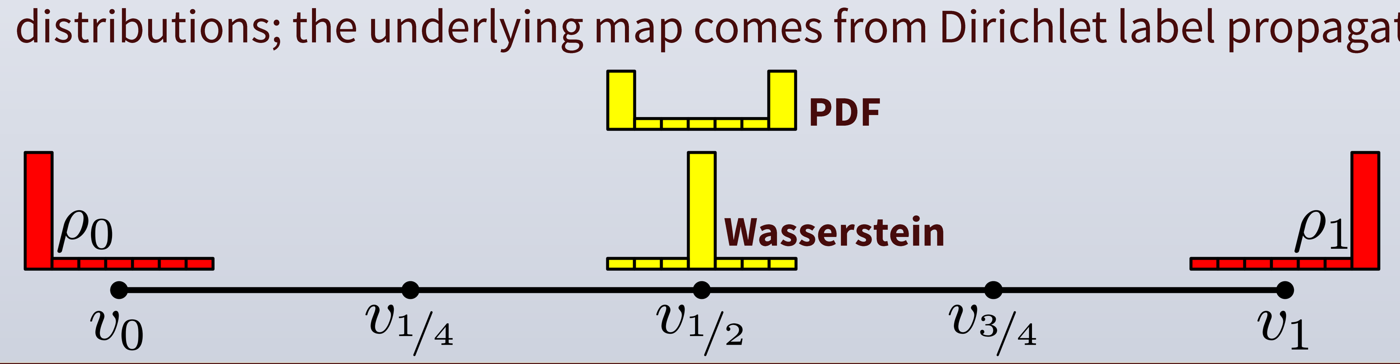
This proposition shows that our problem becomes **linear** in inverse CDF space:



Theoretical Properties

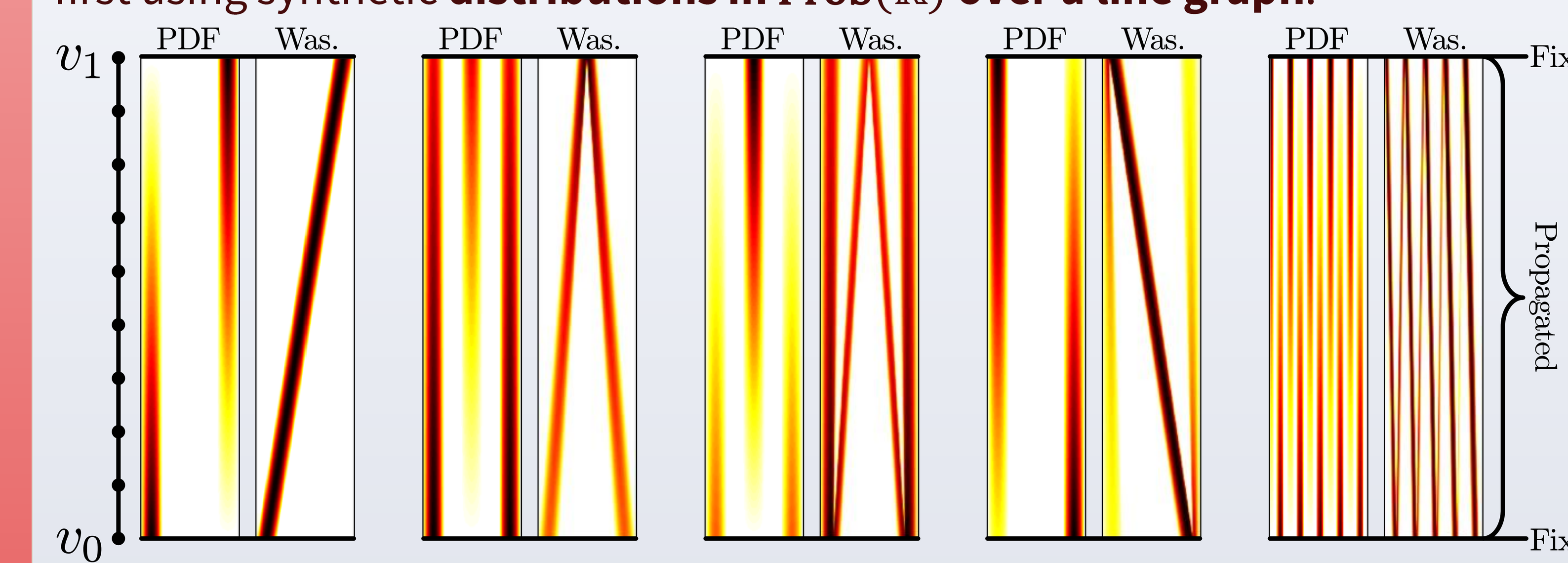
For Prob(ℝ), we can prove many theoretical properties that **may not hold for direct propagation** of bin values:

- **Means and variances** of propagated distributions are **bounded** by those on the boundary.
- If the boundary distributions are **delta functions**, so are the propagated distributions; the underlying map comes from Dirichlet label propagation.

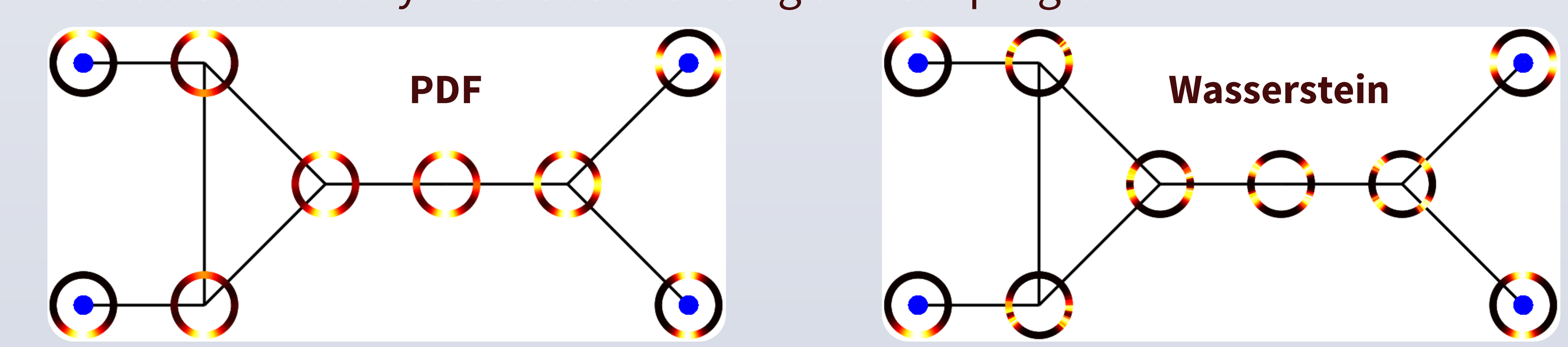


Experiments

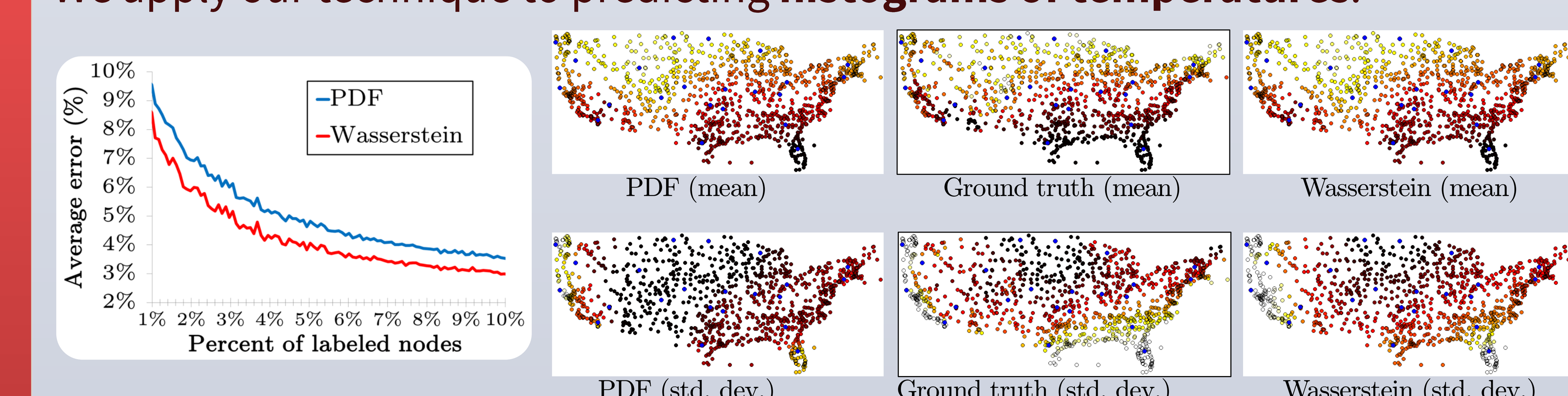
We compare to direct propagation of probability distribution functions (PDFs), first using synthetic **distributions in Prob(ℝ) over a line graph**:



Wasserstein propagation moves probability **across** the domain rather than "teleporting" it across. We carry out similar experiments in $\mathrm{Prob}(S^1)$ with fixed blue boundary distributions using a linear program:



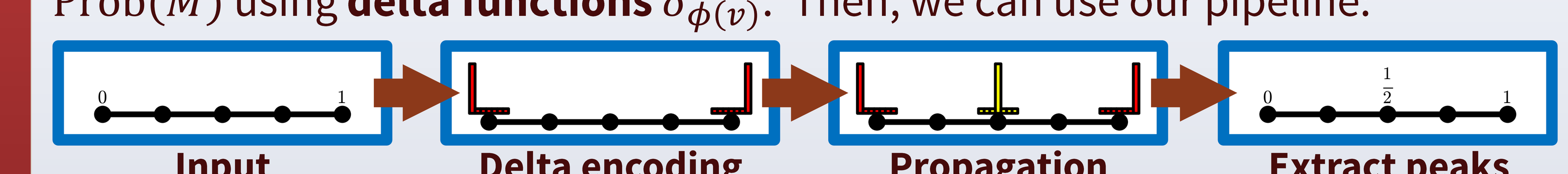
We apply our technique to predicting **histograms of temperatures**:



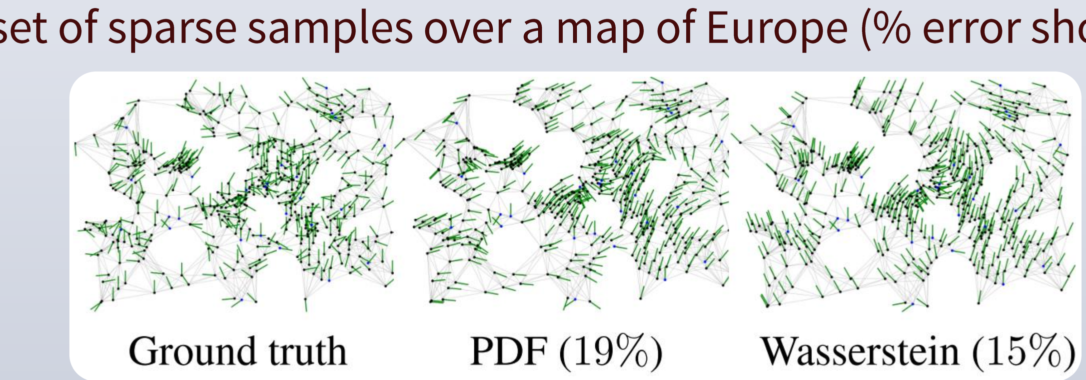
We similarly predict histograms of **wind directions**.

Application to Manifold-Valued Learning

For manifold M, we can **encode maps** $\phi: V \to M$ probabilistically as $\rho_\phi: V \to \mathrm{Prob}(M)$ using **delta functions** $\delta_{\phi(v)}$. Then, we can use our pipeline:



We test this method for predicting periodic **wind directions** on the unit circle S^1 from a set of sparse samples over a map of Europe (% error shown):



References

A. Subramanya & J. Bilmes. "Semi-supervised learning with measure propagation." JMLR 12, 2011.

X. Zhu et al. "Semi-supervised learning using Gaussian fields and harmonic functions." ICML, 2003.

C. Villani. *Topics in Optimal Transportation*. 2003.