ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Wasserstein Training of Boltzmann Machines

Grégoire Montavon, Klaus-Rober Muller, Marco Cuturi

Presenter: Shiyu Liang

December 1, 2016

Coordinated Science Laboratory
Department of Electrical and Computer Engineering
University of Illinois at Urbana Champaign

Outline

1. Introduction
2. Wasserstein Distance
3. Wasserstein training
4. Experiments
5. Conclusions

Introduction

Boltzmann Machines

Definition: Boltzmann machines

Boltzmann machine are a class of generative models used to approximate real-world data distributions.

Boltzmann machines are largely used in

- Approximation of real-world data distribution: handwritten characters, speech segments or multimodal data
- Extraction of features
- Building hierarchical data representations

Restrict Boltzmann Machines

Definition: Restricted Boltzmann machines (RBM)

Restricted Boltzmann machine is a special type of Boltzmann machine with d binary observed variables $\boldsymbol{x} \in \{0,1\}^d$ and h binary explanatory variables $\boldsymbol{y} \in \{0,1\}^h$.

Probability distribution denoted by RBM:

$$p_{\boldsymbol{a},\boldsymbol{w},\boldsymbol{b}}(\boldsymbol{x}) = \frac{1}{Z_{\boldsymbol{a},\boldsymbol{w},\boldsymbol{b}}} \sum_{y \in \{0,1\}^h} \exp\left(-\boldsymbol{a}^T x - \sum_{j=1}^{h} y_j(\boldsymbol{w}_j^T \boldsymbol{x} + b_j)\right)$$

Learning task: given an empirical distribution $\hat{p}$, solve

$$\min_{\boldsymbol{a},\boldsymbol{w},\boldsymbol{b}} \mathrm{KL}(\hat{p} || p_{\boldsymbol{a},\boldsymbol{w},\boldsymbol{b}})$$

Q: Is KL divergence the only way? (No)

Wasserstein Distance

Wasserstein Distance

Definition: Wasserstein distance

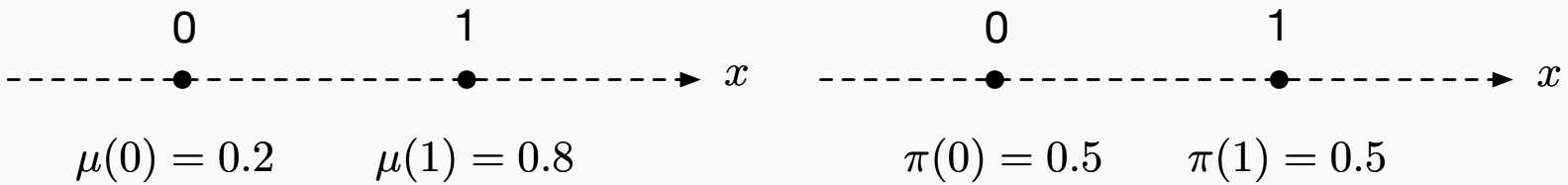
The p-th Wasserstein distance between two probability distribution μ and π is defined as

$$W_p(\mu, \pi) = \left(\inf_{\beta \in \Gamma(\mu, \pi)} \mathbb{E}_{X,Y \sim \beta}[d(x, y)^p] \right)^{1/p}$$

- $d(x, y)$: distance between x and y
- $\beta \in \Gamma(\mu, \pi)$: joint distribution $\beta(x, y)$ has marginal distribution $\mu(x)$ and $\pi(y)$.

Intuition: earth-mover's distance

The minimum cost of moving probability mass μ to match π



Wasserstein Distance

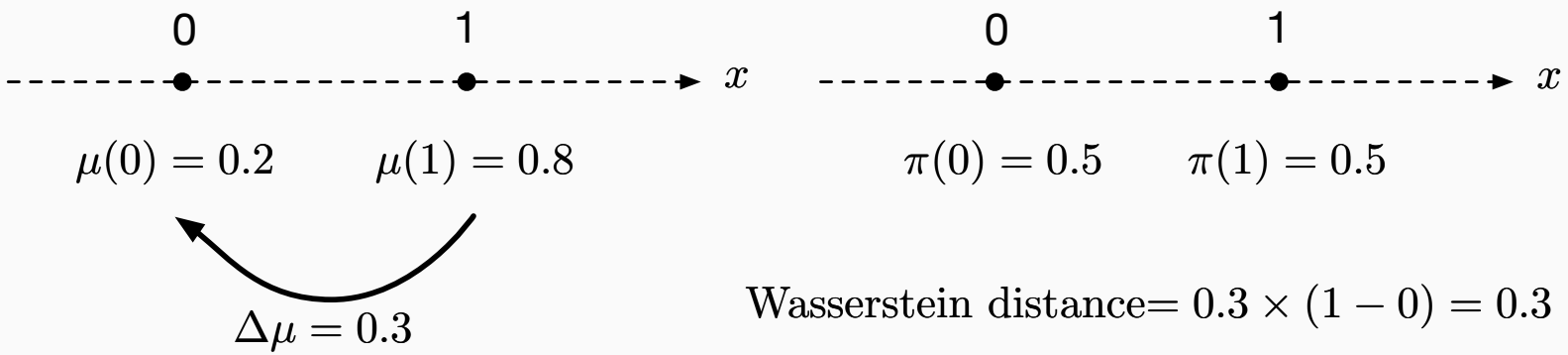
Definition: Wasserstein distance

The p-th Wasserstein distance between two probability distribution μ and π is defined as

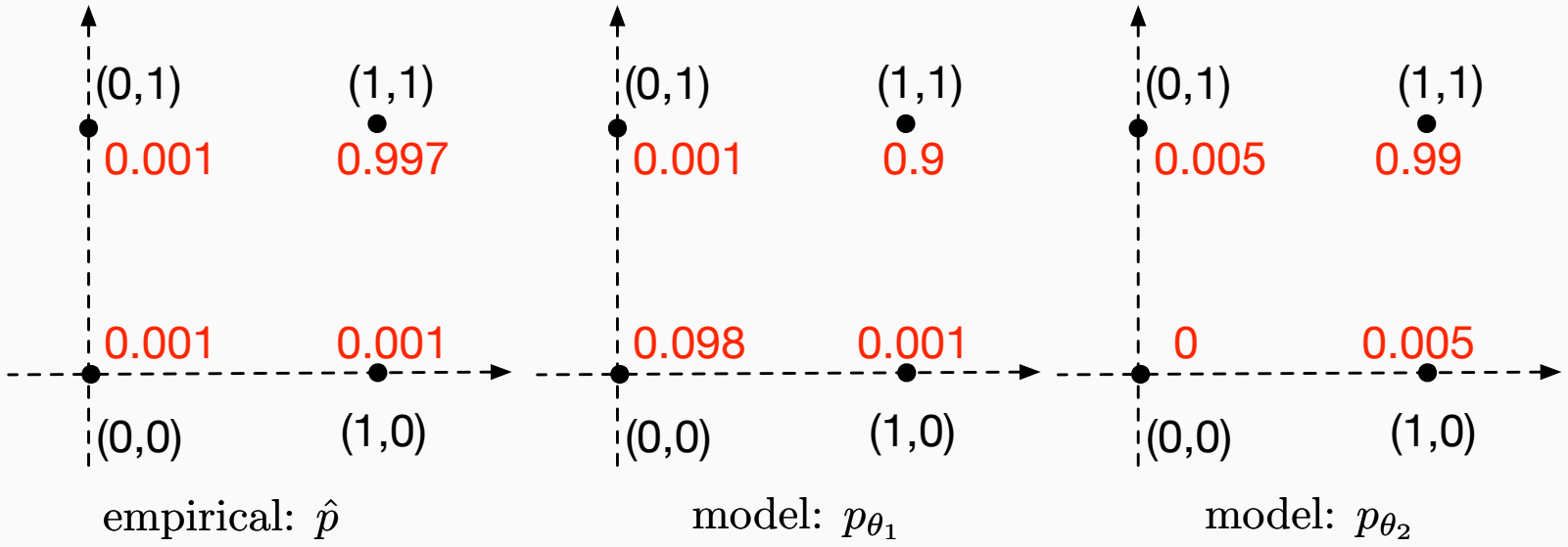
$$W_p(\mu, \pi) = \left(\inf_{\beta \in \Gamma(\mu, \pi)} \mathbb{E}_{X,Y \sim \beta}[d(x, y)^p] \right)^{1/p}$$

- $d(x, y)$: distance between x and y
- $\beta \in \Gamma(\mu, \pi)$: joint distribution $\beta(x, y)$ has marginal distribution $\mu(x)$ and $\pi(y)$.

Intuition: earth-mover's distance



Benefits of Wasserstein Distance



Minimizing KL divergence:

$\mathrm{KL}(\hat{p}||p_{\theta_1}) < \infty = \mathrm{KL}(\hat{p}||p_{\theta_2})$, $\min_\theta \mathrm{KL}(\hat{p}||p_\theta) \Longrightarrow p_{\theta_1}$

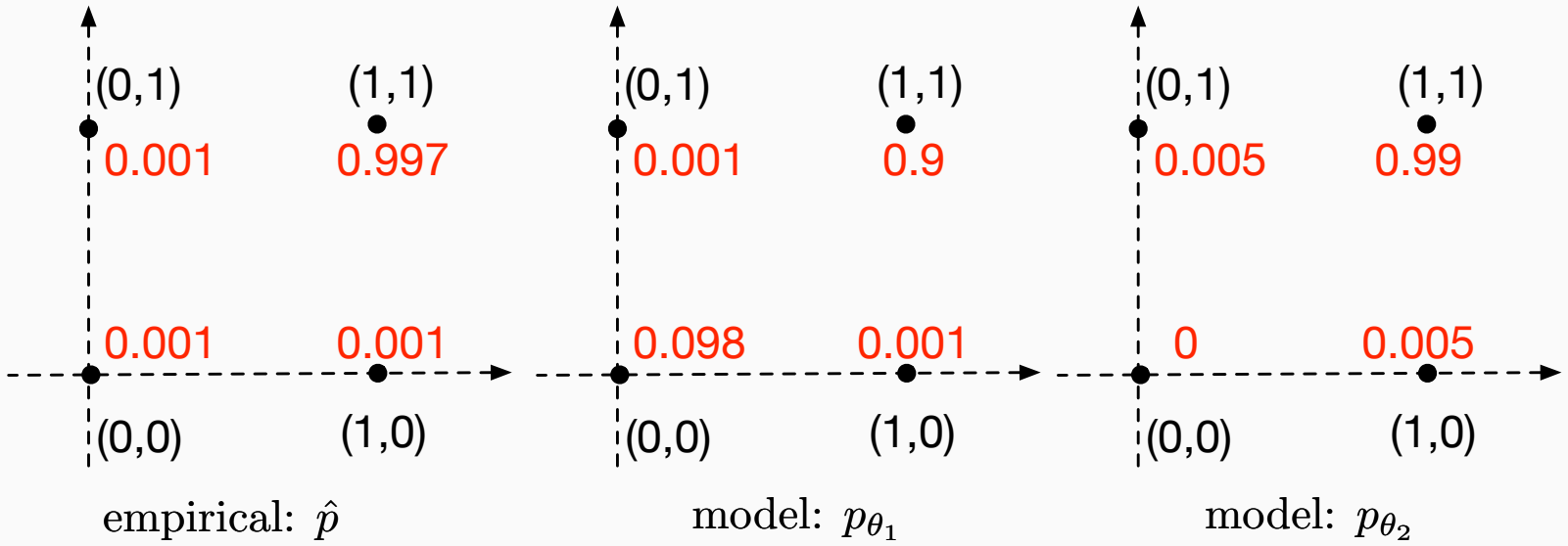
Minimizing Wasserstein distance:

$W(\hat{p}, p_{\theta_1}) = 0.097 \times 2 = 0.194 > 0.008 = W(\hat{p}, p_{\theta_2})$

$\min_\theta \mathrm{W}(\hat{p}, p_\theta) \Longrightarrow p_{\theta_2}$

Some good models can be found by minimizing Wasserstein distance but cannot be found by minimizing KL divergence.

Benefits of Wasserstein Distance



Measuring the goodness-of-fit:

- Wasserstein distance provides a different view than that provided by KL-divergence approach.
- The distance between p_θ and empirical $\hat{p}$ can be small in KL sense but large in Wasserstein sense.
- The distance between p_θ and empirical $\hat{p}$ can be large in KL sense but small in Wasserstein sense.

Wasserstein training

Wasserstein Training

Equivalent dual definition of Wasserstein distance:

- Two probabilities p, q on $\mathcal{X} = \{0,1\}^d$
- A distance function $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$
- Two real-valued functions α, β on $\mathcal{X}$
- γ-smoothed Wasserstein distance is

$$W_\gamma(p,q) = \max_{\alpha,\beta} \Bigg[\mathbb{E}_{\boldsymbol{X}\sim p}[\alpha(\boldsymbol{X})] + \mathbb{E}_{\boldsymbol{X'}\sim q}[\beta(\boldsymbol{X'})] - \gamma \sum_{\boldsymbol{x},\boldsymbol{x'}\in\{0,1\}^d} \exp\left(\frac{1}{\gamma}(\alpha(\boldsymbol{x}) + \beta(\boldsymbol{x'}) - D(\boldsymbol{x},\boldsymbol{x'})) - 1\right) \Bigg]$$

Learning task: given the empirical distribution $\hat{p}$ and $\gamma > 0$, solving

$$\min_\theta W_\gamma(\hat{p}, p_\theta)$$

Gradient Descent

γ-smoothed Wasserstein distance is

$$W_\gamma(p,q) = \max_{\alpha,\beta} \Bigg[\mathbb{E}_{\boldsymbol{X}\sim p}[\alpha(\boldsymbol{X})] + \mathbb{E}_{\boldsymbol{X}'\sim q}[\beta(\boldsymbol{X}')] - \gamma \sum_{\boldsymbol{x},\boldsymbol{x}'\in\{0,1\}^d} \exp\left(\frac{1}{\gamma}(\alpha(\boldsymbol{x}) + \beta(\boldsymbol{x}') - D(\boldsymbol{x},\boldsymbol{x}')) - 1\right) \Bigg] \quad (1)$$

Proposition: gradient descent

For probability distribution $p_\theta(\boldsymbol{x}) = \frac{1}{Z} e^{-F_\theta(\boldsymbol{x})}$. Let α^* be the optimal dual solution of $W_\gamma(p_\theta, p)$. Then

$$\nabla_\theta W_\gamma(p_\theta, p) = \mathbb{E}_{p_\theta}[\nabla_\theta F_\theta(\boldsymbol{x})] \cdot \mathbb{E}_{p_\theta}[\alpha^*(\boldsymbol{x})] - \mathbb{E}_{p_\theta}[\alpha^*(\boldsymbol{x}) \nabla_\theta F(\theta(\boldsymbol{x}))]$$

Notice: $\gamma > 0$, the optimal variable α^* can be recovered by Sinkhorn algorithm.

KL Regularization

Probability distribution denoted by RBM:

$$p_{\boldsymbol{a},\boldsymbol{w},\boldsymbol{b}}(\boldsymbol{x}) = \frac{1}{Z_{\boldsymbol{a},\boldsymbol{w},\boldsymbol{b}}} \sum_{y \in \{0,1\}^h} \exp\left(-\boldsymbol{a}^T x - \sum_{j=1}^{h} y_j(\boldsymbol{w}_j^T \boldsymbol{x} + b_j)\right)$$

Regularization term:

$$\Omega(\boldsymbol{a}, \boldsymbol{w}, \boldsymbol{b}) = \mathrm{KL}(\hat{p} \| p_{\boldsymbol{a},\boldsymbol{w},\boldsymbol{b}}) + \eta \cdot \left(\|\boldsymbol{a}\|^2 + \sum_j \|\boldsymbol{w}_j\|^2\right)$$

Learning task:

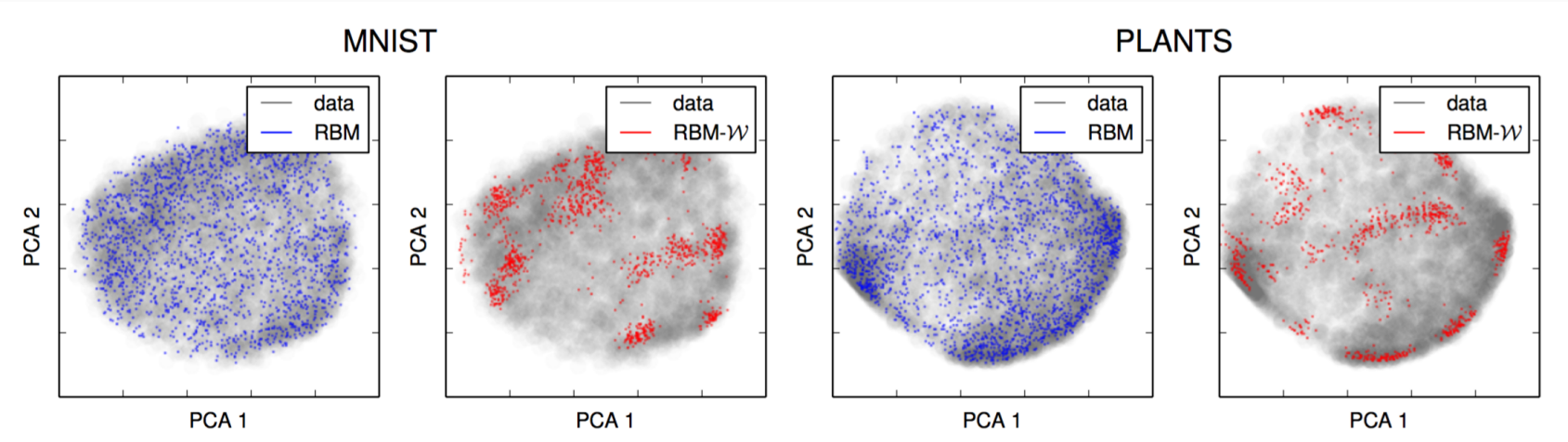
$$\min_{\boldsymbol{a},\boldsymbol{w},\boldsymbol{b}} W_\gamma(p_{\boldsymbol{a},\boldsymbol{w},\boldsymbol{b}}, \hat{p}) + \lambda \cdot \Omega(\boldsymbol{a}, \boldsymbol{w}, \boldsymbol{b})$$

Experiments

Experiments

Results:

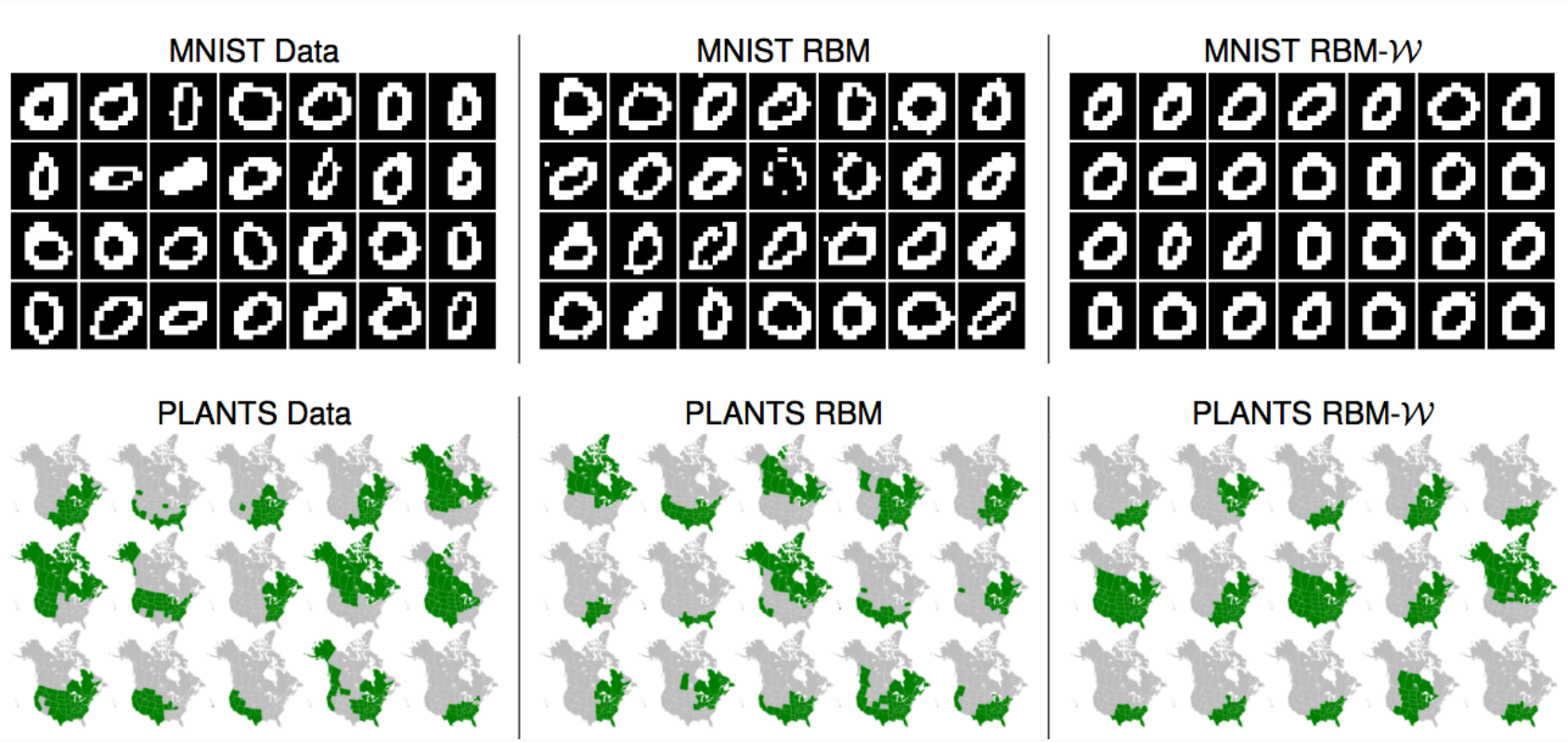
- MNIST dataset: 60000 images of handwritten digits.
- PLANTS dataset: binary vectors indicating the presence or absence of each plant in each US state or Canadian province.
- Standard RBM distribution uniformly covers the data (looks better)
- RBM-W consists small dense clusters.



Experiments

Generating examples:

- Examples produced by RBM-W is less noise than those produced by standard RBM but less various.



Conclusions

Conclusions

- A new objective function for Boltzmann machines based on Wasserstein distance is introduced.
- The objective function has a simple gradient and thus can be optimized by gradient descent algorithm.
- Experiments show that the distribution learned by Wasserstein RBM strongly departed from the one learned by KL model.

Thanks