Optimal Transport Mapping via Input Convex Neural Networks

International Conference on Machine Learning, Virtual, 2020

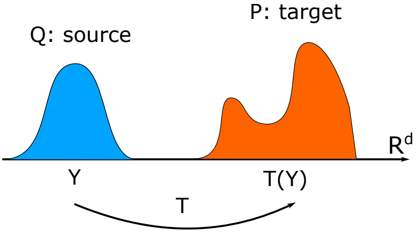
Ashok Vardhan Makkuva* and Amirhossein Taghvaei*
Joint work with Jason D. Lee and Sewoong Oh

University of Illinois at Urbana-Champaign
University of California, Irvine
Princeton University, University of Washington

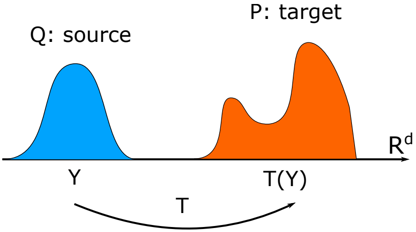
*equal contribution

July 12-18, 2020

Problem statement

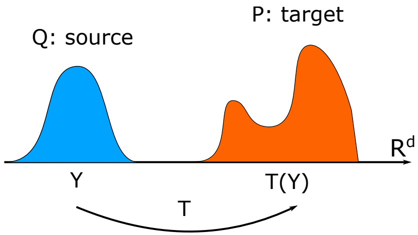


Problem statement



$$\text{OT map} = \operatorname*{argmin}_{T} \mathbb{E}_{Q} \| T(Y) - Y \|^2 \quad \text{s.t.} \quad T_{\#} Q = P$$

Problem statement



$$\text{OT map} = \underset{T}{\operatorname{argmin}}\ \mathbb{E}_Q \| T(Y) - Y \|^2 \quad \text{s.t.} \quad T_\# Q = P$$

Objective:

Given: $\{Y_i\}_{i=1}^n \sim Q, \quad \{X_i\}_{i=1}^n \sim P$

Goal: Approximate the OT map

Solution overview

- Min-max formulation:

$$\inf_{f \in \mathtt{CVX}(\mathbb{R}^d)} \sup_{g \in \mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla g(Y) \rangle - f(\nabla g(Y))]$$

Solution overview

- Min-max formulation:

$$\inf_{f \in \mathtt{CVX}(\mathbb{R}^d)} \sup_{g \in \mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla g(Y)\rangle - f(\nabla g(Y))]$$

- OT map = $\nabla g_{\text{optimal}}$

Solution overview

- Min-max formulation:

$$\inf_{f \in \mathtt{CVX}(\mathbb{R}^d)} \sup_{g \in \mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla g(Y) \rangle - f(\nabla g(Y))]$$

- OT map = $\nabla g_{\text{optimal}}$
- Parametrize $\mathtt{CVX}(\mathbb{R}^d)$ with ICNN

Solution overview

- Min-max formulation:

$$\inf_{f \in \texttt{CVX}(\mathbb{R}^d)} \sup_{g \in \texttt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla g(Y)\rangle - f(\nabla g(Y))]$$

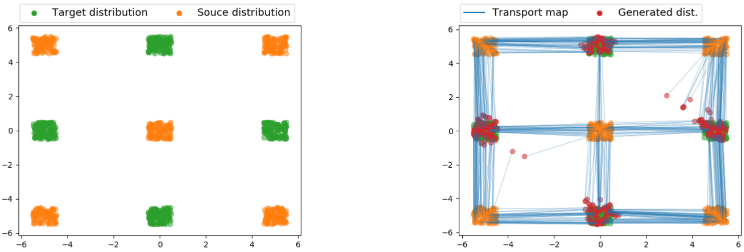
- OT map = $\nabla g_{\text{optimal}}$
- Parametrize $\texttt{CVX}(\mathbb{R}^d)$ with ICNN
- Solve using stochastic optimization algorithm

Solution overview

- Min-max formulation:

$$\inf_{f \in \mathtt{CVX}(\mathbb{R}^d)} \sup_{g \in \mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla g(Y)\rangle - f(\nabla g(Y))]$$

- OT map $= \nabla g_{\text{optimal}}$
- Parametrize $\mathtt{CVX}(\mathbb{R}^d)$ with ICNN
- Solve using stochastic optimization algorithm



Outline

- Motivation and related literature
- Proposed methodology and theoretical results
- Numerical algorithm and experiments

Why optimal transportation?

- Probability distributions appear in many machine learning models

Why optimal transportation?

- Probability distributions appear in many machine learning models
- OT gives a natural geometry for probability distributions

Why optimal transportation?

- Probability distributions appear in many machine learning models
- OT gives a natural geometry for probability distributions
- It has fascinating theory (one Nobel prize and two fields medal)

Why optimal transportation?

- Probability distributions appear in many machine learning models
- OT gives a natural geometry for probability distributions
- It has fascinating theory (one Nobel prize and two fields medal)

Applications:

- Generative models: (Arjovsky et al. 2017, Tolstikhin et al. 2018,...)
- Domain adaptation: (Courty et. al. 2017,...)
- Bayesian inference: (El Moselhy & Marzouk 2012, Reich 2013,...)
- Image processing: (Rabin et. al. 2011, Su et. al. 2015,...)
- Sensor fusion: (Staib et. al. 2017, Srivastav et. al. 2018,...)
- ...

Why optimal transportation?

- Probability distributions appear in many machine learning models
- OT gives a natural geometry for probability distributions
- It has fascinating theory (one Nobel prize and two fields medal)

Applications:

- Generative models: (Arjovsky et al. 2017, Tolstikhin et al. 2018,...)
- Domain adaptation: (Courty et. al. 2017,...)
- Bayesian inference: (El Moselhy & Marzouk 2012, Reich 2013,...)
- Image processing: (Rabin et. al. 2011, Su et. al. 2015,...)
- Sensor fusion: (Staib et. al. 2017, Srivastav et. al. 2018,...)
- ...

This work: Numerical approximation of optimal transport map

Related literature

Discrete OT: see (Peyré & Cuturi 2019) for complete list

- Linear programming
- Sinkhorn iterations (Cuturi, 2013)

Related literature

Discrete OT: see (Peyré & Cuturi 2019) for complete list

- Linear programming
- Sinkhorn iterations (Cuturi, 2013)

Semi-discrete OT:

- Computationoal geometry: (Mérigot 2011, Guo et. al. 2019)
- Stochastic optimization: (Genevay et. al. 2016)

Related literature

Discrete OT: see (Peyré & Cuturi 2019) for complete list

- Linear programming
- Sinkhorn iterations (Cuturi, 2013)

Semi-discrete OT:

- Computationoal geometry: (Mérigot 2011, Guo et. al. 2019)
- Stochastic optimization: (Genevay et. al. 2016)

Continuous approaches:

- with entropic/quadratic regularization (Seguy et. al. 2018)
- adversarial procedure (Leygonie et al. 2019)
- learn optimal coupling (Xie et al. 2019)

Related literature

Discrete OT: see (Peyré & Cuturi 2019) for complete list

- Linear programming
- Sinkhorn iterations (Cuturi, 2013)

Semi-discrete OT:

- Computationoal geometry: (Mérigot 2011, Guo et. al. 2019)
- Stochastic optimization: (Genevay et. al. 2016)

Continuous approaches:

- with entropic/quadratic regularization (Seguy et. al. 2018)
- adversarial procedure (Leygonie et al. 2019)
- learn optimal coupling (Xie et al. 2019)

This work:

- no regularization
- sample-based and scales to high-dimensions
- ICNN parametrization: built upon (T. & Jalali 2019)

Proposed methodology

Main steps:

1. Kantorovich dual formulation:

$$\inf_{(f,h)\in\tilde{\Phi}_c} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[h(Y)]$$

Proposed methodology

Main steps:

1. Kantorovich dual formulation:

$$\inf_{(f,h)\in\tilde{\Phi}_c} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[h(Y)]$$

   - Constraint $\tilde{\Phi}_c \triangleq \{f(x) + h(y) \geq \langle x, y\rangle,\ \forall x, y \in \mathbb{R}^d\}$ (challenging)

Proposed methodology

Main steps:

1. Kantorovich dual formulation:

$$\inf_{(f,h)\in\tilde{\Phi}_c} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[h(Y)]$$

   - Constraint $\tilde{\Phi}_c \triangleq \{f(x) + h(y) \geq \langle x, y\rangle,\ \forall x, y \in \mathbb{R}^d\}$ (challenging)
   - Add entropic/quadratic regularization
     (Cutury 2013, Genevay et. al. 2016, Seguy et. al. 2018,...)

Proposed methodology

Main steps:

1. Kantorovich dual formulation:

$$\inf_{(f,h)\in\tilde{\Phi}_c} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[h(Y)]$$

   - Constraint $\tilde{\Phi}_c \triangleq \{f(x) + h(y) \geq \langle x, y \rangle, \ \forall x, y \in \mathbb{R}^d\}$ (challenging)
   - Add entropic/quadratic regularization (Cutury 2013, Genevay et. al. 2016, Seguy et. al. 2018,...)
   - This work: <u>no regularization</u>

Proposed methodology

Main steps:

1. Kantorovich dual formulation:

$$\inf_{(f,h)\in\tilde{\Phi}_c} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[h(Y)]$$

Proposed methodology

Main steps:

1. Kantorovich dual formulation:

$$\inf_{(f,h)\in\tilde{\Phi}_c} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[h(Y)]$$

2. Semidual formulation: $(h_{\text{optimal}} = f^*)$

$$\inf_{f\in\mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[f^*(Y)]$$

Proposed methodology

Main steps:

1. Kantorovich dual formulation:

$$\inf_{(f,h)\in\tilde{\Phi}_c} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[h(Y)]$$

2. Semidual formulation: $(h_{\text{optimal}} = f^*)$

$$\inf_{f\in\mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] \ + \ \mathbb{E}_Q[f^*(Y)]$$

    - Hard to compute/estimate f^*

Proposed methodology

Main steps:

1. Kantorovich dual formulation:

$$\inf_{(f,h)\in\tilde{\Phi}_c} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[h(Y)]$$

2. Semidual formulation: $(h_{\text{optimal}} = f^*)$

$$\inf_{f\in\mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] \ + \ \mathbb{E}_Q[f^*(Y)]$$

Proposed methodology

Main steps:

1. Kantorovich dual formulation:

$$\inf_{(f,h)\in\tilde{\Phi}_c} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[h(Y)]$$

2. Semidual formulation: $(h_{\text{optimal}} = f^*)$

$$\inf_{f\in\mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[f^*(Y)]$$

3. **Min-max formulation**:

$$\inf_{f\in\mathtt{CVX}(\mathbb{R}^d)} \sup_{g\in\mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla g(Y)\rangle - f(\nabla g(Y))]$$

Theoretical results

Consistency: If source dist. Q admits density, then

$$\boxed{\exists \text{ optimal pair } (f_0, g_0) \text{ and } \nabla g_0 \text{ is the OT map}}$$

Theoretical results

Consistency: If source dist. Q admits density, then

$$\boxed{\exists \text{ optimal pair } (f_0, g_0) \text{ and } \nabla g_0 \text{ is the OT map}}$$

Stability: For any (f, g) such that f is α-strongly convex

$$\underbrace{\|\nabla g - \nabla g_0\|^2_{L^2(Q)}}_{\text{error in estimating OT}} \leq \frac{2}{\alpha} \underbrace{(\epsilon_1(f,g) + \epsilon_2(f))}_{\text{optimization gap}}$$

Theoretical results

Consistency: If source dist. Q admits density, then

$$\boxed{\exists \text{ optimal pair } (f_0, g_0) \text{ and } \nabla g_0 \text{ is the OT map}}$$

Stability: For any (f, g) such that f is $\underline{\alpha\text{-strongly convex}}$

$$\underbrace{\|\nabla g - \nabla g_0\|^2_{L^2(Q)}}_{\text{error in estimating OT}} \leq \frac{2}{\alpha} \underbrace{(\epsilon_1(f, g) + \epsilon_2(f))}_{\text{optimization gap}}$$

- ϵ_1 is the optimality gap for max
- ϵ_2 is the optimality gap for min

Consistency of the method

Main result

Consider the min-max formulation

$$\inf_{f \in \mathtt{CVX}(\mathbb{R}^d)} \sup_{g \in \mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla g(Y) \rangle - f(\nabla g(Y))]$$

If Q admits density, then

Consistency of the method

Main result

Consider the min-max formulation

$$\inf_{f \in \mathtt{CVX}(\mathbb{R}^d)} \sup_{g \in \mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla g(Y)\rangle - f(\nabla g(Y))]$$

If Q admits density, then

- There exists an optimal pair (f_0, g_0)

Consistency of the method

Main result

Consider the min-max formulation

$$\inf_{f \in \mathtt{CVX}(\mathbb{R}^d)} \sup_{g \in \mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla g(Y) \rangle - f(\nabla g(Y))]$$

If Q admits density, then

- There exists an optimal pair (f_0, g_0)
- ∇g_0 is the OT map from Q to P

Consistency of the method

Main result

Consider the min-max formulation

$$\inf_{f \in \mathtt{CVX}(\mathbb{R}^d)} \sup_{g \in \mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla g(Y)\rangle - f(\nabla g(Y))]$$

If Q admits density, then

- There exists an optimal pair (f_0, g_0)
- ∇g_0 is the OT map from Q to P

proof sketch: Using Fenchel's inequality

$$\langle y, \nabla g(y)\rangle - f(\nabla g(y)) \leq f^*(y), \quad \forall g$$

Stability analysis

Proposition

Consider the min-max formulation

$$\inf_{f \in \mathtt{CVX}(\mathbb{R}^d)} \sup_{g \in \mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla g(Y) \rangle - f(\nabla g(Y))]$$

Stability analysis

Proposition

Consider the min-max formulation

$$\inf_{f \in \mathtt{CVX}(\mathbb{R}^d)} \sup_{g \in \mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla g(Y)\rangle - f(\nabla g(Y))]$$

For any (f, g) such that f is α-strongly convex

$$\underbrace{\|\nabla g - \nabla g_0\|^2_{L^2(Q)}}_{\text{error in estimating OT}} \leq \frac{2}{\alpha} \underbrace{(\epsilon_1(f,g) + \epsilon_2(f))}_{\text{optimization gap}}$$

Stability analysis

Proposition

Consider the min-max formulation

$$\inf_{f \in \mathtt{CVX}(\mathbb{R}^d)} \sup_{g \in \mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla g(Y)\rangle - f(\nabla g(Y))]$$

For any (f, g) such that f is α-strongly convex

$$\underbrace{\|\nabla g - \nabla g_0\|^2_{L^2(Q)}}_{\text{error in estimating OT}} \leq \frac{2}{\alpha} \underbrace{(\epsilon_1(f, g) + \epsilon_2(f))}_{\text{optimization gap}}$$

- ϵ_1 is the optimality gap for max

Stability analysis

Proposition

Consider the min-max formulation

$$\inf_{f \in \mathtt{CVX}(\mathbb{R}^d)} \sup_{g \in \mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla g(Y)\rangle - f(\nabla g(Y))]$$

For any (f, g) such that f is α-strongly convex

$$\underbrace{\|\nabla g - \nabla g_0\|^2_{L^2(Q)}}_{\text{error in estimating OT}} \leq \frac{2}{\alpha} \underbrace{(\epsilon_1(f,g) + \epsilon_2(f))}_{\text{optimization gap}}$$

- ϵ_1 is the optimality gap for max
- ϵ_2 is the optimality gap for min

Stability analysis

Proposition

Consider the min-max formulation

$$\inf_{f \in \mathtt{CVX}(\mathbb{R}^d)} \sup_{g \in \mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla g(Y) \rangle - f(\nabla g(Y))]$$

For any (f, g) such that f is α-strongly convex

$$\underbrace{\|\nabla g - \nabla g_0\|^2_{L^2(Q)}}_{\text{error in estimating OT}} \leq \frac{2}{\alpha} \underbrace{(\epsilon_1(f, g) + \epsilon_2(f))}_{\text{optimization gap}}$$

- ϵ_1 is the optimality gap for max
- ϵ_2 is the optimality gap for min

proof: Extension of stability result for semi-dual formulation (Hütter & Rigollet 2019)

Proposed method

- Min-max formulation:

$$\inf_{f \in \mathtt{CVX}(\mathbb{R}^d)} \sup_{g \in \mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_{P}[f(X)] + \mathbb{E}_{Q}[\langle Y, \nabla g(Y) \rangle - f(\nabla g(Y))]$$

Proposed method

- Min-max formulation:

$$\inf_{f \in \mathtt{CVX}(\mathbb{R}^d)} \sup_{g \in \mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_{P}[f(X)] + \mathbb{E}_{Q}[\langle Y, \nabla g(Y) \rangle - f(\nabla g(Y))]$$

- Parametrization with ICNN

$$\inf_{f \in \mathtt{ICNN}(\mathbb{R}^d)} \sup_{g \in \mathtt{ICNN}(\mathbb{R}^d)} \mathbb{E}_{P}[f(X)] + \mathbb{E}_{Q}[\langle Y, \nabla g(Y) \rangle - f(\nabla g(Y))]$$

Proposed method

- Min-max formulation:

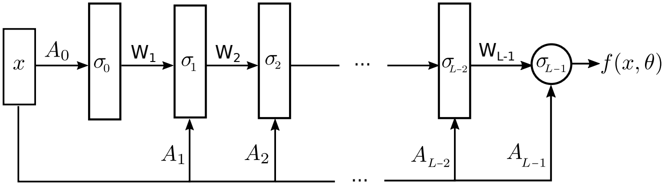
$$\inf_{f \in \mathtt{CVX}(\mathbb{R}^d)} \sup_{g \in \mathtt{CVX}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla g(Y) \rangle - f(\nabla g(Y))]$$

- Parametrization with ICNN

$$\inf_{f \in \mathtt{ICNN}(\mathbb{R}^d)} \sup_{g \in \mathtt{ICNN}(\mathbb{R}^d)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[\langle Y, \nabla g(Y) \rangle - f(\nabla g(Y))]$$

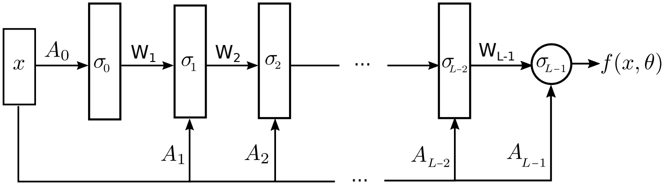
- Solve using stochastic optimization algorithm

Input Convex Neural Networks (ICNN)



(Amos et. al. 2016)

Input Convex Neural Networks (ICNN)

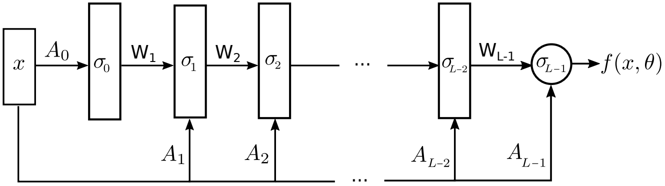


(Amos et. al. 2016)

$f(x,\theta)$ is convex in x if

- $W_l \geq 0$ element-wise
- σ_0 is convex
- σ_l is convex and non-decreasing for $l = 1, \dots, L-1$

Input Convex Neural Networks (ICNN)



(Amos et. al. 2016)

$f(x, \theta)$ is convex in x if

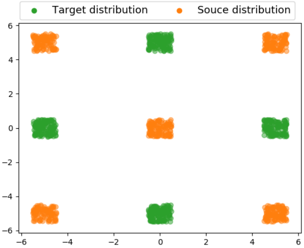
- $W_l \geq 0$ element-wise
- σ_0 is convex
- σ_l is convex and non-decreasing for $l = 1, \ldots, L-1$

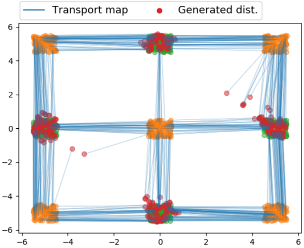
Representation power: (Chen et. al. 2018)

- ICNN can approximate any convex function over a compact domain

Proof-of-concept: Learning the OT map

- Example I: Checkerboard

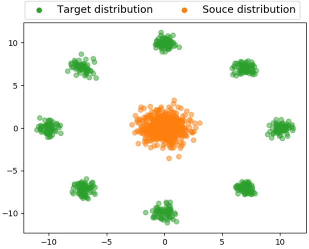


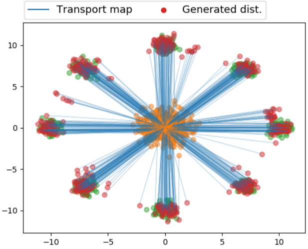


Proof-of-concept: Learning the OT map

- Example I: Checkerboard

- Example II: Mixture of Gaussians

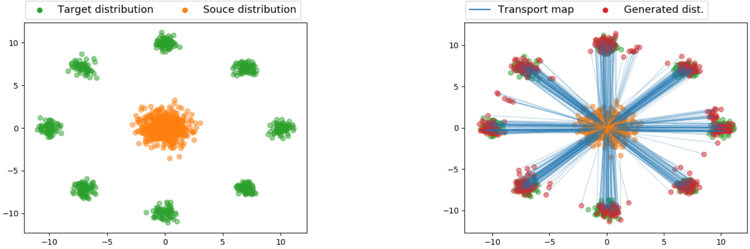




Proof-of-concept: Learning the OT map

- Example I: Checkerboard

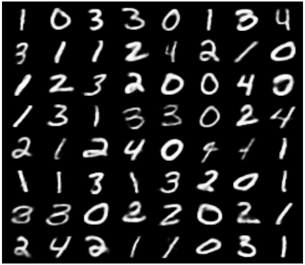
- Example II: Mixture of Gaussians



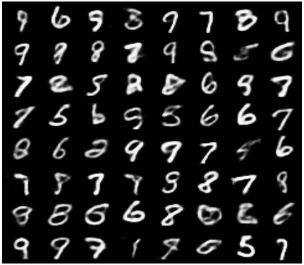
The algorithm learns the **OT** map

Results on high-dim real data

- MNIST $\{0, 1, 2, 3, 4\}$ to MNIST $\{5, 6, 7, 8, 9\}$ (in VAE latent space)

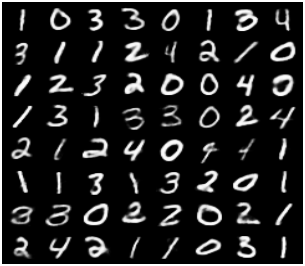


$\rightarrow$

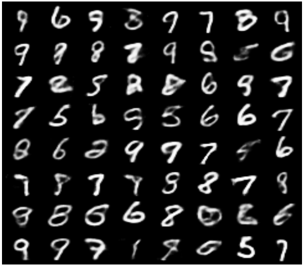


Results on high-dim real data

- MNIST $\{0, 1, 2, 3, 4\}$ to MNIST $\{5, 6, 7, 8, 9\}$ (in VAE latent space)



$\rightarrow$



For more information, please visit the poster

Thank you!