MATHÉMATIQUES
VISION
APPRENTISSAGE

Presentation: Neural Optimal Transport

Mathis Hardion, Antoine Olivier

Plan

Introduction

- OT as loss faithful to the ground metric, approximated with NNs: widely used in generative modeling since WGAN and extensions (rich literature: [3, 30, 14, 22, 9, 12, 25, 10]...)

- More recently, OT map/plan itself as generator

- Previous works still have limitations:
  - Restrictive assumptions on existence of Monge/Brenier maps
  - High dimensional sampling with diffusion models
  - Difficulties to scale
  - Often difficult to train

- Studied paper: generalization to WOT, try to obtain scaleable procedure

MATHÉMATIQUES VISION APPRENTISSAGE

Plan

Monge/Kantorovich OT

Definition

$$\mathbb{M}(\alpha, \beta) := \inf_{T_\sharp \alpha = \beta} \int_{\mathcal{X}} c(x, T(x)) d\alpha(x)$$

Monge/Kantorovich OT

Definition

$$\mathbb{M}(\alpha, \beta) := \inf_{T_\sharp \alpha = \beta} \int_{\mathcal{X}} c(x, T(x)) d\alpha(x)$$

- OT map T: unpaired translation/style transfer, inpainting, faithful to the original image.

MATHÉMATIQUES VISION APPRENTISSAGE

Monge/Kantorovich OT

Definition

$$\mathbb{M}(\alpha, \beta) := \inf_{T_\sharp \alpha = \beta} \int_{\mathcal{X}} c(x, T(x)) d\alpha(x)$$

- OT map T: unpaired translation/style transfer, inpainting, faithful to the original image.
- Problem: T may not exist.

MATHÉMATIQUES VISION APPRENTISSAGE

Monge/Kantorovich OT

Definition

$$\mathbb{M}(\alpha, \beta) := \inf_{T_\sharp \alpha = \beta} \int_{\mathcal{X}} c(x, T(x)) d\alpha(x)$$

- OT map T: unpaired translation/style transfer, inpainting, faithful to the original image.
- Problem: T may not exist.

Definition

$$\mathbb{K}(\alpha, \beta) := \inf_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y).$$

Monge/Kantorovich OT

Definition

$$\mathbb{M}(\alpha, \beta) := \inf_{T_\sharp \alpha = \beta} \int_{\mathcal{X}} c(x, T(x)) d\alpha(x)$$

- OT map T: unpaired translation/style transfer, inpainting, faithful to the original image.
- Problem: T may not exist.

Definition

$$\mathbb{K}(\alpha, \beta) := \inf_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y).$$

- Instead of using the deterministic $T(x)$, sample OT plan $\pi(\cdot|x)$.

Monge/Kantorovich OT

Definition

$$\mathbb{M}(\alpha, \beta) := \inf_{T_\sharp \alpha = \beta} \int_{\mathcal{X}} c(x, T(x)) d\alpha(x)$$

- OT map T: unpaired translation/style transfer, inpainting, faithful to the original image.
- Problem: T may not exist.

Definition

$$\mathbb{K}(\alpha, \beta) := \inf_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y).$$

- Instead of using the deterministic $T(x)$, sample OT plan $\pi(\cdot|x)$.
- use the average $\int_{\mathcal{Y}} y d\pi(y|x)$ if a deterministic map is needed.

Weak OT

Definition

For a "weak" cost $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$,

$$\mathbb{T}(\alpha, \beta) := \inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}} C(x, \pi(\cdot|x)) d\alpha(x).$$

Weak OT

Definition

For a "weak" cost $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$,

$$\mathbb{T}(\alpha, \beta) := \inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}} C(x, \pi(\cdot|x)) d\alpha(x).$$

- Recovers Kantorovitch for $C(x, \mu) = \int_{\mathcal{Y}} c(x, y) d\mu(y)$.

MATHÉMATIQUES VISION APPRENTISSAGE

Weak OT

Definition

For a "weak" cost $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$,

$$\mathbb{T}(\alpha, \beta) := \inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}} C(x, \pi(\cdot|x)) d\alpha(x).$$

- Recovers Kantorovitch for $C(x, \mu) = \int_{\mathcal{Y}} c(x, y) d\mu(y)$.
- Added flexibility: directly regularize the generator $\pi(\cdot|x)$ through C.

Weak OT

Definition

For a "weak" cost $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$,

$$\mathbb{T}(\alpha, \beta) := \inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}} C(x, \pi(\cdot|x)) d\alpha(x).$$

- Recovers Kantorovitch for $C(x, \mu) = \int_{\mathcal{Y}} c(x, y) d\mu(y)$.
- Added flexibility: directly regularize the generator $\pi(\cdot|x)$ through C.
- e.g. γ-weak cost:

$$C_\gamma(x, \mu) := \frac{1}{2} \int_{\mathcal{Y}} \|x - y\|^2 \, d\mu(y) - \frac{\gamma}{2} \mathrm{Var}\left(\mu\right).$$

Dual WOT

- Dual OT problems usually enjoy simpler optimization procedures

Dual WOT

- Dual OT problems usually enjoy simpler optimization procedures

Proposition

Under appropriate assumptions,

$$\mathbb{T}(\alpha,\beta) = \sup_{f\in\mathscr{C}_{b,s}(\mathcal{Y})} \left(\int_{\mathcal{X}} f^C(x)d\alpha(x) + \int_{\mathcal{Y}} f(y)d\beta(y) \right),$$

where

$$\forall x \in \mathcal{X}, f^C(x) = \inf_{\mu\in\mathcal{P}(\mathcal{Y})} C(x,\mu) - \int_{\mathcal{Y}} f(y)d\mu(y).$$

MATHÉMATIQUES
VISION
APPRENTISSAGE

Plan

Contribution of the paper

Proposition

We consider atomless distribution ζ on $\mathcal{Z}$ "the outsourcing distribution".

The dual WOT problem can be rewritten as

$$\mathbb{T}(\alpha, \beta) = \sup_{f \in \mathscr{C}_{b,s}(\mathcal{Y})} \inf_{T : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}} \mathcal{L}_{\alpha,\beta}(f, T)$$

where

$$\mathcal{L}_{\alpha,\beta}(f, T) = \int_{\mathcal{Y}} f(y) d\beta(y) + \int_{\mathcal{X}} \left(C\left(x, T(x, \cdot)_\sharp \zeta\right) - \int_{\mathcal{Z}} f(T(x, z)) d\zeta(z) \right) d\alpha(x).$$

Contribution of the paper

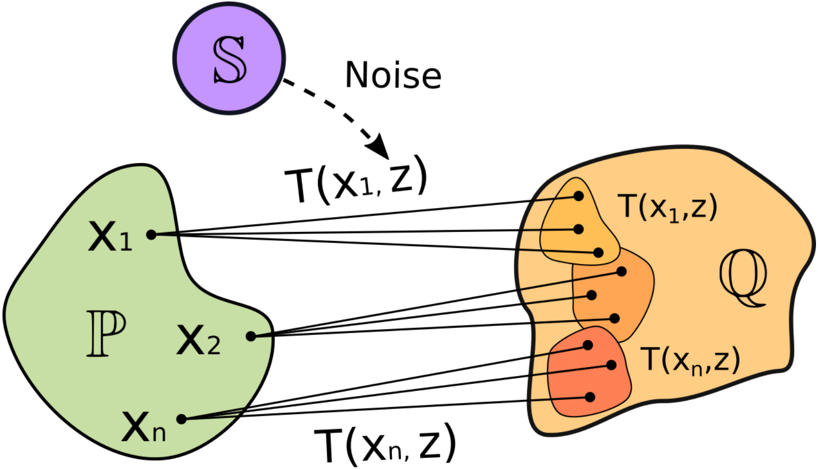


FIGURE 1: Extracted from the paper: Stochastic map illustration

Intuition

Proposition

Dual WOT reformulation

$$\mathbb{T}(\alpha,\beta) = \sup_{f\in\mathscr{C}_{b,s}(\mathcal{Y})} \inf_{T:\mathcal{X}\times\mathcal{Z}\to\mathcal{Y}} \mathcal{L}_{\alpha,\beta}(f,T)$$

where

$$\mathcal{L}_{\alpha,\beta}(f,T) = \int_{\mathcal{Y}} f(y)d\beta(y) + \int_{\mathcal{X}} \left(C\left(x, T(x,\cdot)_\sharp\zeta\right) - \int_{\mathcal{Z}} f(T(x,z))d\zeta(z) \right) d\alpha(x).$$

- $T(x,\cdot)_\sharp\zeta$ corresponds to the conditional $\pi(\cdot|x)$
- The paper shows that neural networks are universal approximators of stochastic transport maps

Strenghts/Limitations

Proposition

Dual WOT reformulation

$$\mathbb{T}(\alpha, \beta) = \sup_{f \in \mathscr{C}_{b,s}(\mathcal{Y})} \inf_{T : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}} \mathcal{L}_{\alpha,\beta}(f, T)$$

where

$$\mathcal{L}_{\alpha,\beta}(f, T) = \int_{\mathcal{Y}} f(y) d\beta(y) + \int_{\mathcal{X}} \left(C\left(x, T(x, \cdot)_\sharp \zeta\right) - \int_{\mathcal{Z}} f(T(x, z)) d\zeta(z) \right) d\alpha(x).$$

- T is easily interpretable.
- Freedom in the choice of ζ (and $\mathcal{Z}$).
- Easy to sample
- Few hyperparameters
- Maximin problem: slow to optimize and potentially unstable
- Added biais with γ-weak cost.

Plan

Toy 2D Dataset

Definition

- The distribution on $\mathcal{X}$ is $\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$
- The distribution on $\mathcal{Y}$ is made of m uniform circles of center y_i and radius r.
- $\zeta \sim \mathcal{N}(0, 0.01\, I_4)$
- $C(x,\mu) \coloneqq \left\|x - \int_{\mathcal{X}} y d\mu(y)\right\|^2$

MATHÉMATIQUES VISION APPRENTISSAGE

Toy 2D Dataset

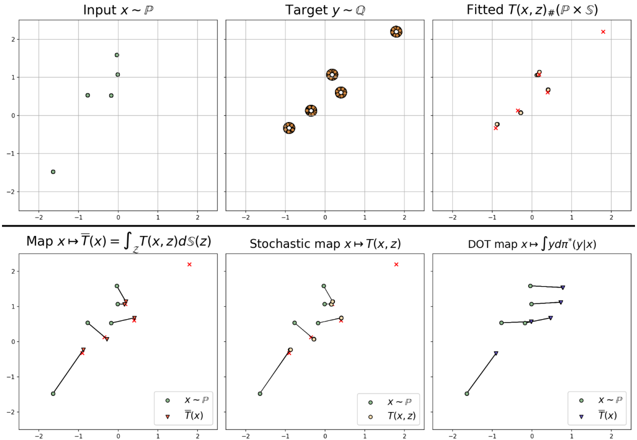
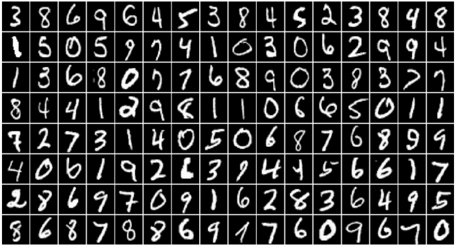


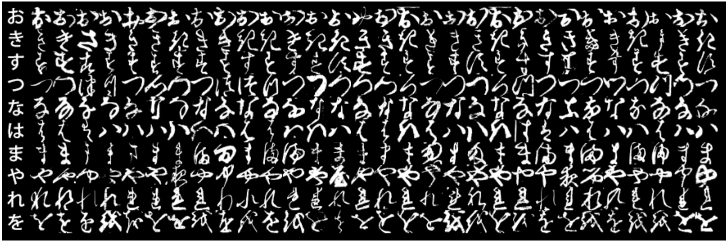
FIGURE 2: Toy 2D example with discrete input and multimodal target. Note how the algorithm fails to recover the upper right mode.

MATHÉMATIQUES VISION APPRENTISSAGE

MNIST to KMNIST



(A) Sample from the MNIST Dataset



(B) Sample from the KMNIST Dataset

MNIST to KMNIST: Training

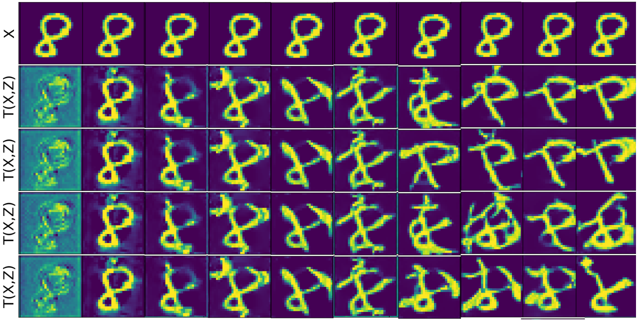
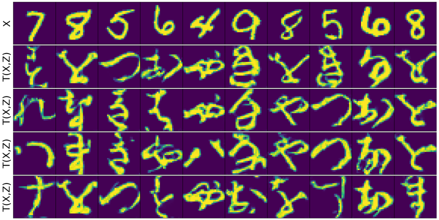


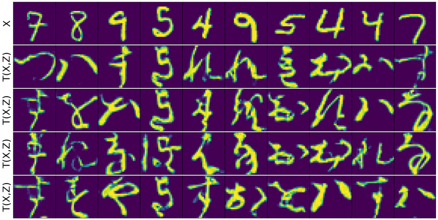
FIGURE 4: Evolution of the output during training (the intervals are irregular)

MATHÉMATIQUES VISION APPRENTISSAGE

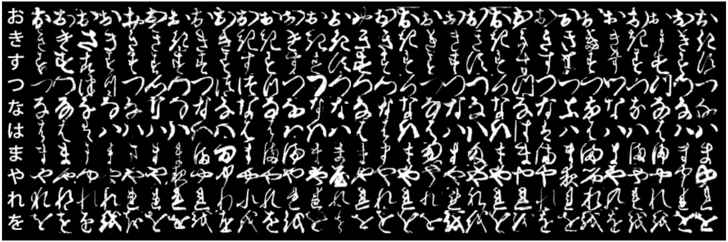
MNIST to KMNIST



(A) Random Images



(B) Random Test Images



(C) Sample from the KMNIST Dataset

Plan

MATHÉMATIQUES VISION APPRENTISSAGE

Entropic OT as WOT

Definition

Entropic OT:

$$\mathbb{S}_\varepsilon(\alpha,\beta) = \inf_{\pi\in\Pi(\alpha,\beta)} \left(\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y) + \varepsilon \mathrm{KL}\left(\pi || \alpha\otimes\beta\right) \right)$$

Proposition

$$\mathbb{S}_\varepsilon(\alpha,\beta) = \mathbb{T}_\varepsilon(\alpha,\beta),$$

with $\mathbb{T}_\varepsilon$ the WOT for the cost

$$C_\varepsilon(x,\mu) := \int_{\mathcal{Y}} c(x,y)d\mu(y) + \varepsilon \mathrm{KL}\left(\mu || \beta\right).$$

Entropic OT as WOT

$$C_\varepsilon(x,\mu) := \int_{\mathcal{Y}} c(x,y)d\mu(y) + \varepsilon \mathrm{KL}\left(\mu || \beta\right).$$

- With regards to the optimization procedure of [19], no computational advantage...

Entropic OT as WOT

$$C_\varepsilon(x, \mu) := \int_{\mathcal{Y}} c(x, y) d\mu(y) + \varepsilon \mathrm{KL}\left(\mu || \beta\right).$$

- With regards to the optimization procedure of [19], no computational advantage...
- However, regularization may enforce samples to appear 'likely' from β

Entropic OT as WOT

$$C_\varepsilon(x,\mu) := \int_{\mathcal{Y}} c(x,y)d\mu(y) + \varepsilon \mathrm{KL}\left(\mu || \beta\right).$$

- With regards to the optimization procedure of [19], no computational advantage...
- However, regularization may enforce samples to appear 'likely' from β
- The challenge to avoid divergence of the KL to $+\infty$ could be solved by adding the regularization mid-way

Approximate Monge with WOT

Proposition

Define

$$\tilde{C}_\gamma : x, \mu \mapsto \int_{\mathcal{Y}} c(x,y) d\mu(y) + \gamma \mathrm{Var}(\mu),$$

Then, it holds that

$$\inf_{\pi \in \Pi(\alpha,\beta)} \lim_{\gamma \to +\infty} \int_{\mathcal{X}} \tilde{C}_\gamma(x, \pi(\cdot|x)) d\alpha(x) = \inf_{T_\sharp \alpha = \beta} \int_{\mathcal{X}} c(x, T(x)) d\alpha(x). \quad (1)$$

MATHÉMATIQUES VISION APPRENTISSAGE

Approximate Monge with WOT

Proposition

Define

$$\tilde{C}_\gamma : x, \mu \mapsto \int_{\mathcal{Y}} c(x,y) d\mu(y) + \gamma \mathrm{Var}(\mu),$$

Then, it holds that

$$\inf_{\pi \in \Pi(\alpha,\beta)} \lim_{\gamma \to +\infty} \int_{\mathcal{X}} \tilde{C}_\gamma(x, \pi(\cdot|x)) d\alpha(x) = \inf_{T_\sharp \alpha = \beta} \int_{\mathcal{X}} c(x, T(x)) d\alpha(x). \tag{1}$$

- Future work could:
  - Swap inf and lim under appropriate conditions

Approximate Monge with WOT

Proposition

Define

$$\tilde{C}_\gamma : x, \mu \mapsto \int_{\mathcal{Y}} c(x,y) d\mu(y) + \gamma \mathrm{Var}\left(\mu\right),$$

Then, it holds that

$$\inf_{\pi \in \Pi(\alpha,\beta)} \lim_{\gamma \to +\infty} \int_{\mathcal{X}} \tilde{C}_\gamma(x, \pi(\cdot|x)) d\alpha(x) = \inf_{T_\sharp \alpha = \beta} \int_{\mathcal{X}} c(x, T(x)) d\alpha(x). \tag{1}$$

- Future work could:
  - Swap inf and lim under appropriate conditions
  - Study convergence of arginf if possible ($\tilde{C}_\gamma$ is concave)

Approximate Monge with WOT

Proposition

Define

$$\tilde{C}_\gamma : x, \mu \mapsto \int_{\mathcal{Y}} c(x,y) d\mu(y) + \gamma \mathrm{Var}\left(\mu\right),$$

Then, it holds that

$$\inf_{\pi \in \Pi(\alpha,\beta)} \lim_{\gamma \to +\infty} \int_{\mathcal{X}} \tilde{C}_\gamma(x, \pi(\cdot|x)) d\alpha(x) = \inf_{T_\sharp \alpha = \beta} \int_{\mathcal{X}} c(x, T(x)) d\alpha(x). \tag{1}$$

- Future work could:
  - Swap inf and lim under appropriate conditions
  - Study convergence of arginf if possible ($\tilde{C}_\gamma$ is concave)
- Convergence is observed on a toy dataset (appendix)

WOT as "stochastic Monge"

$$\mathbb{M}(\alpha, \beta) := \inf_{T_\sharp \alpha = \beta} \int_{\mathcal{X}} c(x, T(x)) d\alpha(x), \tag{2}$$

$$\mathbb{T}(\alpha, \beta) := \inf_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X}} C(x, \pi(\cdot|x)) d\alpha(x). \tag{3}$$

WOT as "stochastic Monge"

$$\mathbb{M}(\alpha,\beta) := \inf_{T_\sharp \alpha = \beta} \int_{\mathcal{X}} c(x, T(x)) d\alpha(x), \tag{2}$$

$$\mathbb{T}(\alpha,\beta) := \inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}} C(x, \pi(\cdot|x)) d\alpha(x). \tag{3}$$

- T replaced by $\mathscr{T} : \left| \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathcal{P}(\mathcal{Y}) \\ x & \longmapsto & \mathscr{T}_x := \pi(\cdot|x) \end{array} \right.$

WOT as "stochastic Monge"

$$\mathbb{M}(\alpha,\beta) := \inf_{T_\sharp \alpha = \beta} \int_{\mathcal{X}} c(x, T(x))d\alpha(x), \tag{2}$$

$$\mathbb{T}(\alpha,\beta) := \inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}} C(x, \pi(\cdot|x))d\alpha(x). \tag{3}$$

- T replaced by $\mathscr{T} : \left| \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathcal{P}(\mathcal{Y}) \\ x & \longmapsto & \mathscr{T}_x := \pi(\cdot|x) \end{array} \right.$
- $T_\sharp \alpha = \beta$ replaced by

$$\pi \in \Pi(\alpha,\beta) \iff \begin{cases} \pi_1 = \alpha \\ \int_{\mathcal{X}} \mathscr{T}_x(A) d\alpha(x) = \beta(A) \end{cases}$$

WOT as "stochastic Monge"

$$\mathbb{M}(\alpha, \beta) := \inf_{T_\sharp \alpha = \beta} \int_{\mathcal{X}} c(x, T(x)) d\alpha(x), \tag{2}$$

$$\mathbb{T}(\alpha, \beta) := \inf_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X}} C(x, \pi(\cdot|x)) d\alpha(x). \tag{3}$$

- T replaced by $\mathscr{T} : \left| \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathcal{P}(\mathcal{Y}) \\ x & \longmapsto & \mathscr{T}_x := \pi(\cdot|x) \end{array} \right.$
- $T_\sharp \alpha = \beta$ replaced by

$$\begin{aligned} \pi \in \Pi(\alpha, \beta) &\iff \left\{ \begin{array}{l} \pi_1 = \alpha \\ \int_{\mathcal{X}} \mathscr{T}_x(A) d\alpha(x) = \beta(A) \end{array} \right. \\ &\iff \left\{ \begin{array}{l} \pi_1 = \alpha \\ \int_{\mathcal{P}(\mathcal{Y})} \mu(A) d(\mathscr{T}_\sharp \alpha)(\mu) = \beta(A) \end{array} \right. \end{aligned}$$

WOT as "stochastic Monge"

$$\mathbb{M}(\alpha, \beta) := \inf_{T_\sharp \alpha = \beta} \int_{\mathcal{X}} c(x, T(x)) d\alpha(x), \tag{2}$$

$$\mathbb{T}(\alpha, \beta) := \inf_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X}} C(x, \pi(\cdot|x)) d\alpha(x). \tag{3}$$

- T replaced by $\mathscr{T} : \left| \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathcal{P}(\mathcal{Y}) \\ x & \longmapsto & \mathscr{T}_x := \pi(\cdot|x) \end{array} \right.$
- $T_\sharp \alpha = \beta$ replaced by

$$\begin{aligned} \pi \in \Pi(\alpha, \beta) &\iff \begin{cases} \pi_1 = \alpha \\ \int_{\mathcal{X}} \mathscr{T}_x(A) d\alpha(x) = \beta(A) \end{cases} \\ &\iff \begin{cases} \pi_1 = \alpha \\ \int_{\mathcal{P}(\mathcal{Y})} \mu(A) d(\mathscr{T}_\sharp \alpha)(\mu) = \beta(A) \end{cases} \\ &\iff \begin{cases} \pi_1 = \alpha \\ \mathbb{E}\left[\mathscr{T}_\sharp \alpha\right] = \beta \end{cases} \end{aligned}$$

WOT as "stochastic Monge"

Proposition

The following equality holds:

$$\mathbb{T}(\alpha, \beta) = \inf_{\mathbb{E}\left[\mathscr{T}_\sharp \alpha\right] = \beta} \int_{\mathcal{X}} C(x, \mathscr{T}_x) d\alpha(x) =: \mathbb{SM}(\alpha, \beta). \tag{4}$$

WOT as "stochastic Monge"

Proposition

The following equality holds:

$$\mathbb{T}(\alpha, \beta) = \inf_{\mathbb{E}[\mathscr{T}_\sharp \alpha] = \beta} \int_{\mathcal{X}} C(x, \mathscr{T}_x) d\alpha(x) =: \mathbb{SM}(\alpha, \beta). \tag{4}$$

- Since WOT may not be convex for nonconvex costs like the previous case, this motivates the Kantorovitch relaxation

MATHÉMATIQUES VISION APPRENTISSAGE

Kantorovitch relaxation

Definition

The stochastic Kantorovitch cost between two probability measures α, β is defined as

$$\mathbb{SK}(\alpha,\beta) := \inf_{\substack{\mathbb{E}[\Gamma_1]=\alpha \\ \mathbb{E}[\Gamma_2]=\beta}} \int_{\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y})} C(\mu,\nu) d\Gamma(\mu,\nu) \tag{5}$$

where Γ denotes an element of $\mathcal{P}(\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}))$ and Γ_1, Γ_2 its respective marginals.

Kantorovitch relaxation

Definition

The stochastic Kantorovitch cost between two probability measures α, β is defined as

$$\mathbb{SK}(\alpha, \beta) := \inf_{\substack{\mathbb{E}[\Gamma_1]=\alpha \\ \mathbb{E}[\Gamma_2]=\beta}} \int_{\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y})} C(\mu, \nu) d\Gamma(\mu, \nu) \tag{5}$$

where Γ denotes an element of $\mathcal{P}(\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}))$ and Γ_1, Γ_2 its respective marginals.

- $\mathbb{SK}$ is a convex and symmetric optimization problem, recovers $\mathbb{SM}$ when $\Gamma = (\delta, \mathscr{T})_\sharp \alpha$

Kantorovitch relaxation

Definition

The stochastic Kantorovitch cost between two probability measures α, β is defined as

$$\mathbb{SK}(\alpha, \beta) := \inf_{\substack{\mathbb{E}[\Gamma_1]=\alpha \\ \mathbb{E}[\Gamma_2]=\beta}} \int_{\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})} C(\mu, \nu) d\Gamma(\mu, \nu) \tag{5}$$

where Γ denotes an element of $\mathcal{P}(\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}))$ and Γ_1, Γ_2 its respective marginals.

- $\mathbb{SK}$ is a convex and symmetric optimization problem, recovers $\mathbb{SM}$ when $\Gamma = (\delta, \mathscr{T})_\sharp \alpha$
- A dual can be derived similary to the Kantorovitch duality

Kantorovitch relaxation

Proposition

Assuming duality holds, the dual Stochastic Kantorovitch reads

$$\mathbb{SK}(\alpha, \beta) = \inf_{\langle f, \cdot \rangle \oplus \langle g, \cdot \rangle \leq C} \langle f, \alpha \rangle + \langle g, \beta \rangle \,, \tag{6}$$

where $\langle f, \mu \rangle \coloneqq \int f d\mu$, $\langle f, \cdot \rangle \oplus \langle g, \cdot \rangle : (\mu, \nu) \mapsto \langle f, \mu \rangle + \langle g, \nu \rangle$.

MATHÉMATIQUES VISION APPRENTISSAGE

Kantorovitch relaxation

Proposition

Assuming duality holds, the dual Stochastic Kantorovitch reads

$$\mathbb{SK}(\alpha, \beta) = \inf_{\langle f, \cdot \rangle \oplus \langle g, \cdot \rangle \leq C} \langle f, \alpha \rangle + \langle g, \beta \rangle \,, \tag{6}$$

where $\langle f, \mu \rangle := \int f d\mu$, $\langle f, \cdot \rangle \oplus \langle g, \cdot \rangle : (\mu, \nu) \mapsto \langle f, \mu \rangle + \langle g, \nu \rangle$.

- Further work could:
  - Study the tightness of $\mathbb{SK}$ vs $\mathbb{SM}$

MATHÉMATIQUES VISION APPRENTISSAGE

Kantorovitch relaxation

Proposition

Assuming duality holds, the dual Stochastic Kantorovitch reads

$$\mathbb{SK}(\alpha, \beta) = \inf_{\langle f, \cdot \rangle \oplus \langle g, \cdot \rangle \leq C} \langle f, \alpha \rangle + \langle g, \beta \rangle \,, \tag{6}$$

where $\langle f, \mu \rangle := \int f d\mu$, $\langle f, \cdot \rangle \oplus \langle g, \cdot \rangle : (\mu, \nu) \mapsto \langle f, \mu \rangle + \langle g, \nu \rangle$.

- Further work could:
  - Study the tightness of $\mathbb{SK}$ vs $\mathbb{SM}$
  - Potentially deduce existence/unicity of stochastic Monge maps (or equivalently WOT plans) in a more general setting than previous proofs (limited to convex weak cost [31])

Kantorovitch relaxation

Proposition

Assuming duality holds, the dual Stochastic Kantorovitch reads

$$\mathbb{SK}(\alpha, \beta) = \inf_{\langle f, \cdot \rangle \oplus \langle g, \cdot \rangle \leq C} \langle f, \alpha \rangle + \langle g, \beta \rangle \,, \tag{6}$$

where $\langle f, \mu \rangle := \int f d\mu$, $\langle f, \cdot \rangle \oplus \langle g, \cdot \rangle : (\mu, \nu) \mapsto \langle f, \mu \rangle + \langle g, \nu \rangle$.

- Further work could:
  - Study the tightness of $\mathbb{SK}$ vs $\mathbb{SM}$
  - Potentially deduce existence/unicity of stochastic Monge maps (or equivalently WOT plans) in a more general setting than previous proofs (limited to convex weak cost [31])
  - Find under which conditions duality holds

Kantorovitch relaxation

Proposition

Assuming duality holds, the dual Stochastic Kantorovitch reads

$$\mathbb{SK}(\alpha, \beta) = \inf_{\langle f, \cdot \rangle \oplus \langle g, \cdot \rangle \leq C} \langle f, \alpha \rangle + \langle g, \beta \rangle \,, \tag{6}$$

where $\langle f, \mu \rangle := \int f d\mu$, $\langle f, \cdot \rangle \oplus \langle g, \cdot \rangle : (\mu, \nu) \mapsto \langle f, \mu \rangle + \langle g, \nu \rangle$.

- Further work could:
  - Study the tightness of $\mathbb{SK}$ vs $\mathbb{SM}$
  - Potentially deduce existence/unicity of stochastic Monge maps (or equivalently WOT plans) in a more general setting than previous proofs (limited to convex weak cost [31])
  - Find under which conditions duality holds
  - Possibly reformulate the dual $\mathbb{SK}$ with some generalization of C-transform and derive an approach similar to [19]

Plan

Summary

- WOT can prove useful when using an OT plan as generator since it allows direct regularization
- [19] suggests a noise outsourcing formulation and derive a neural approximation scheme
  - More general than previous works, can function well on some decently large datasets
  - Slow training, can fail to recover all modes
- EOT could be used as regularzied cost in this WOT framework
- Monge maps could be approached with a variance regularization
- A Kantorovitch relaxation of WOT could be explored to try and prove more general results, dual could be extended for practical use

Appendix

Proof of (1)

First, observe that for $x \in \mathcal{X}, \pi \in \Pi(\alpha, \beta)$,

$$\tilde{C}_\gamma(x, \pi(\cdot|x)) \xrightarrow[\gamma \to +\infty]{} \begin{cases} c(x, T(x)) & \text{if } \pi(\cdot|x) = \delta_{T(x)} \text{ for some } T(x) \in \mathcal{Y} \\ +\infty & \text{otherwise} \end{cases}$$

since for a probability measure μ, $\mathrm{Var}\,(\mu) = 0 \iff \exists y, \mu = \delta_y$. Therefore, by Beppo-Levi's lemma (the sequence is increasing in γ and nonnegative), one has

$$\int_{\mathcal{X}} \tilde{C}_\gamma(x, \pi(\cdot|x)) \xrightarrow[\gamma \to +\infty]{} \begin{cases} \int_{\mathcal{X}} c(x, T(x)) d\alpha(x) & \text{if } \pi(\cdot|x) = \delta_{T(x)} \ \alpha\text{-a.e. for } T : \mathcal{X} \to \mathcal{Y} \\ +\infty & \text{otherwise.} \end{cases} \tag{7}$$

Notice that if $\pi(\cdot|x) = \delta_{T(x)}$ α-a.e., then $T(x) = \int_{\mathcal{Y}} y d\pi(y|x)$ and therefore T is measurable (at least when restricted to a set of probability 1 under α). Additionally in that case, for any continuous bounded $f : \mathcal{Y} \longrightarrow \mathbb{R}$, one has

$$\begin{aligned} \int_{\mathcal{Y}} f(y) d\beta(y) &= \int_{\mathcal{Y} \times \mathcal{X}} f(y) d\pi(y|x) d\alpha(x) \\ &= \int_{\mathcal{X}} f(T(x)) d\alpha(x) \end{aligned}$$

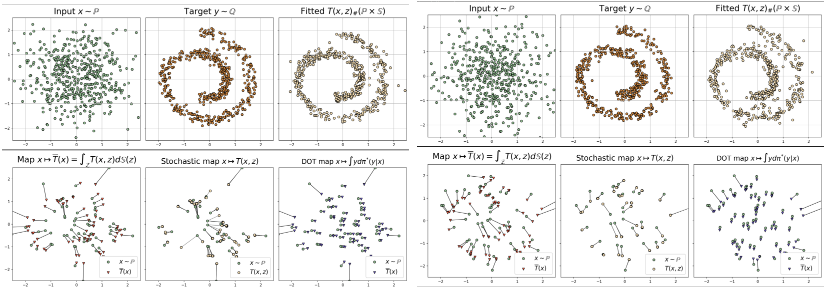
Proof of (1)

i.e. $T_\sharp\alpha = \beta$. Conversely, if there is some measurable T such that $T_\sharp\alpha = \beta$, the coupling π defined by its first marginal $\pi_1 = \alpha$ and conditional $\pi(\cdot|x) \coloneqq \delta_{T(x)}$ does indeed verify $\pi \in \Pi(\alpha, \beta)$ using the same computation: for any continuous bounded f,

$$\begin{aligned}\int_{\mathcal{Y}} f(y)d\pi_2(y) &= \int_{\mathcal{X}\times\mathcal{Y}} f(y)d\pi(x,y)\\ &= \int_{\mathcal{X}}\int_{\mathcal{Y}} f(y)d\pi(y|x)d\alpha(x)\\ &= \int_{\mathcal{X}} f(T(x))d\alpha(x)\\ &= \int_{\mathcal{Y}} f(y)d\beta(y).\end{aligned}$$

As a result, the (possibly empty) set of πs for which the limiting cost in (7) is finite is exactly described by the set of measurable maps T verifying $T_\sharp\alpha = \beta$, hence proving (1).

Toy illustration of Monge as limit of WOT



(A) $C_\gamma, \gamma = 1$ (B) $\tilde{C}_\gamma, \gamma = 10$

Figure 6: Toy experiment illustrating the difference between the γ-weak cost enforcing higher variance 6a and our alternative definition penalizing it instead 6b. One can see that as expected, we recover a basically deterministic map in the latter case.

MATHÉMATIQUES VISION APPRENTISSAGE

Proof of EOT as WOT

Since any $\pi \in \Pi(\alpha, \beta)$ such that the cost in EOT is finite verifies $\pi \ll \alpha \otimes \beta$, it also holds that for any $x \in \mathcal{X}$, $\pi(\cdot|x) \ll \beta$, as one can easily obtain its Radon-Nikodym derivative: $d\pi(y|x) = \frac{d\pi(x,y)}{d(\alpha \otimes \beta)} d\beta(y)$. Thus, one has

$$\begin{aligned}
\inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}} C_\varepsilon(x, \pi(\cdot|x)) d\alpha(x) &= \inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}} \int_{\mathcal{Y}} \left(c(x,y) + \varepsilon \log \left(\frac{d\pi(y|x)}{d\beta} \right) \right) d\pi(y|x) d\alpha \\
&= \inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X} \times \mathcal{Y}} \left(c(x,y) + \varepsilon \log \left(\frac{d\pi(x,y)}{d(\alpha \otimes \beta)} \right) \right) d\pi(x,y) \\
&= \mathbb{S}_\varepsilon(\alpha, \beta).
\end{aligned}$$

Proof of $\mathbb{SK}$ duality

Note that if the optimal coupling is of the form $(\delta, \mathscr{T})_\sharp \alpha$ with $\delta : x \mapsto \delta_x$, we recover the stochastic Monge OT (4). Additionally, as in $\mathbb{K}$, the fact that Γ is a probability measure is implied by the marginal constraints: if $\mathbb{E}\left[\Gamma_1\right] = \alpha$, one has

$$\begin{aligned} \int_{\mathcal{P}(\mathcal{X})} 1 d\Gamma_1(\mu) &= \int_{\mathcal{P}(\mathcal{X})} \mu(\mathcal{X}) d\Gamma_1(\mu) \\ &= \mathbb{E}\left[\Gamma_1\right](\mathcal{X}) \\ &= \alpha(\mathcal{X}) \\ &= 1, \end{aligned}$$

and therefore

$$\begin{aligned} \Gamma(\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})) &= \Gamma_1(\mathcal{P}(\mathcal{X})) \\ &= 1. \end{aligned}$$

Denoting $\mathcal{M}(\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}))$ the set of nonnegative measures over $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$, one can therefore take the infimum over $\Gamma \in \mathcal{M}(\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}))$ verifying the

Proof of $\mathbb{SK}$ duality

marginal constraints. Assuming duality holds i.e. sup and inf can be swapped, one has

$$\begin{aligned}\mathbb{SK}(\alpha,\beta) &= \inf_{\Gamma}\sup_{f,g}\langle C,\Gamma\rangle + (\langle f,\alpha\rangle - \langle f,\mathbb{E}[\Gamma_1]\rangle) + (\langle g,\beta\rangle - \langle g,\mathbb{E}[\Gamma_2]\rangle)\\ &= \sup_{f,g}\langle f,\alpha\rangle + \langle g,\beta\rangle + \inf_{\Gamma}\langle C,\Gamma\rangle - \langle f,\mathbb{E}[\Gamma_1]\rangle - \langle g,\mathbb{E}[\Gamma_2]\rangle\end{aligned}$$

Where the infimum is over $\Gamma \in \mathcal{M}(\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}))$ and f, g continuous bounded functions on $\mathcal{X}$, $\mathcal{Y}$ respectively. For such functions, the definition of $\mathbb{E}[\Gamma_1]$ is equivalent to

$$\begin{aligned}\langle f,\mathbb{E}[\Gamma_1]\rangle &= \int_{\mathcal{P}(\mathcal{X})}\int_{\mathcal{X}} f(x)d\mu(x)d\Gamma_1(\mu)\\ &= \int_{\mathcal{P}(\mathcal{X})}\langle f,\mu\rangle\, d\Gamma_1(\mu),\end{aligned}$$

Proof of $\mathbb{SK}$ duality

and the same can be said for $\langle g, \mathbb{E}[\Gamma_2]\rangle$, whence

$$\begin{aligned}
\inf_{\Gamma} \langle C, \Gamma\rangle - \langle f, \mathbb{E}[\Gamma_1]\rangle - \langle g, \mathbb{E}[\Gamma_2]\rangle &= \inf_{\Gamma} \int_{\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y})} (C(\mu,\nu) - \langle f, \mu\rangle - \langle g, \nu\rangle) d\Gamma(\mu,\nu) \\
&= \begin{cases} 0 & \text{if } \langle f, \cdot\rangle \oplus \langle g, \cdot\rangle \leq C \\ -\infty & \text{otherwise.} \end{cases}
\end{aligned}$$

References

[1] Amjad Almahairi et al. "Augmented CycleGAN: Learning Many-to-Many Mappings from Unpaired Data". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 195–204.

[2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Birkhäuser, 2008.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 214–223.

[4] Aleksandr Beznosikov et al. "Stochastic Gradient Descent-Ascent: Unified Theory and New Efficient Methods". In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Ed. by Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent. Vol. 206. Proceedings of Machine Learning Research. PMLR, Apr. 2023, pp. 172–235.

[5] Shreyas Chaudhari, Srinivasa Pranav, and José M.F. Moura. "Learning Gradients of Convex Functions with Monotone Gradient Networks". In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2023, pp. 1–5.

[6] Tarin Clanuwat et al. *Deep Learning for Classical Japanese Literature*. Dec. 3, 2018. arXiv: `cs.CV/1812.01718`.

[7] Marco Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. Curran Associates, Inc., 2013.

[8] Max Daniels, Tyler Maunu, and Paul Hand. "Score-based Generative Neural Networks for Large-Scale Optimal Transport". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 12955–12965.

[9] Ishan Deshpande, Ziyu Zhang, and Alexander G. Schwing. "Generative Modeling Using the Sliced Wasserstein Distance". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[10] Yonatan Dukler et al. "Wasserstein of Wasserstein Loss for Learning Generative Models". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, June 2019, pp. 1716–1725.

References

[11] Jean Feydy et al. "Interpolating between Optimal Transport and MMD using Sinkhorn Divergences". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. 2019, pp. 2681–2690.

[12] Aude Genevay, Gabriel Peyré, and Marco Cuturi. "Learning Generative Models with Sinkhorn Divergences". In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. Ed. by Amos Storkey and Fernando Perez-Cruz. Vol. 84. Proceedings of Machine Learning Research. PMLR, Apr. 2018, pp. 1608–1617.

[13] Nathael Gozlan et al. "Kantorovich duality for general transport costs and applications". In: *Journal of Functional Analysis* 273.11 (2017), pp. 3327–3405.

[14] Ishaan Gulrajani et al. "Improved Training of Wasserstein GANs". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017.

[15] Xun Huang et al. *Multimodal Unsupervised Image-to-Image Translation*. 2018. arXiv: 1804.04732 [cs.CV].

[16] Olav Kallenberg. *Foundations of Modern Probability*. 1st ed. Probability and Its Applications. Springer New York, NY, 1997, pp. XII, 523.

[17] L. Kantorovitch. "On the Translocation of Masses". In: *Management Science* 5.1 (1958), pp. 1–4.

[18] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[19] Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. "Neural Optimal Transport". In: *The Eleventh International Conference on Learning Representations*. 2023.

[20] Alexander Korotin et al. "Do Neural Optimal Transport Solvers Work? A Continuous Wasserstein-2 Benchmark". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 14593–14605.

[21] Yann LeCun and Corinna Cortes. "MNIST handwritten digit database". In: (2010).

[22] Huidong Liu, Xianfeng Gu, and Dimitris Samaras. "Wasserstein GAN With Quadratic Transport Cost". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019.

[23] Ashok Makkuva et al. "Optimal transport mapping via input convex neural networks". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 6672–6681.

[24] G. Monge. *Mémoire sur la théorie des déblais et des remblais*. Imprimerie royale, 1781.

References

[25] Henning Petzka, Asja Fischer, and Denis Lukovnikov. "On the regularization of Wasserstein GANs". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[26] Gabriel Peyré and Marco Cuturi. "Computational Optimal Transport". In: *Foundations and Trends in Machine Learning* 11.5-6 (2019), pp. 355–607.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

[28] Litu Rout, Alexander Korotin, and Evgeny Burnaev. "Generative Modeling with Optimal Transport Maps". In: *International Conference on Learning Representations*. 2022.

[29] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians. Calculus of Variations, PDEs, and Modeling*. 1st ed. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser Cham, 2015, pp. XXVII, 353.

[30] Vivien Seguy et al. "Large Scale Optimal Transport and Mapping Estimation". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[31] Julio Backhoff Veraguas, Mathias Beiglboeck, and Gudmund Pammer. "Existence, duality, and cyclical monotonicity for weak transport costs". In: *Calculus of Variations and Partial Differential Equations* 58 (Nov. 2019).

[32] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.

[33] Qing Wang, Sanjeev R. Kulkarni, and Sergio Verdu. "Divergence Estimation for Multidimensional Densities Via k-Nearest-Neighbor Distances". In: *IEEE Transactions on Information Theory* 55.5 (2009), pp. 2392–2405.