Project 10: Learning an Optimal Transport Brenier Map

Félix Fourreau & Mathis Wauquiez

École des Ponts ParisTech (ENPC) & ENS Paris-Saclay

{felix.fourreau, mathis.wauquiez}@eleves.enpc.fr

March 21, 2025

Outline

Optimal Transport: Motivation

- Monge's optimal transport :

$$\inf_{T:\, T_{\#}p_X = p_Y} \mathbb{E}_{x \sim p_X}[c(x, T(x))]$$

- When $c(x, y) = \|x - y\|^2$, Brenier's Theorem applies.

Brenier's Theorem

Theorem: Let p_X and p_Y be probability measures on $\mathbb{R}^n$ with finite second moments, with p_X absolutely continuous. Then, there exists a unique convex function $\varphi : \mathbb{R}^n \to \mathbb{R}$ (up to an additive constant) such that:

$$T(x) = \nabla\varphi(x) \quad \text{for } p_X\text{-almost every } x.$$

- **Key point:** The optimal transport map can be expressed as the gradient of a convex function.

Why Learn the Gradient Directly?

- Traditional methods (ICNN/ICGN) enforce convexity via complex architecture .
- New approach: design networks that *directly* learn a monotone gradient mapping.
- Ensures network Jacobian is positive semidefinite (PSD) $\Rightarrow$ gradient of a convex function.

Cascaded Monotone Gradient Network (C-MGN)

- Multi-layer architecture with shared weight matrices.
- Defined by:

$$
\begin{aligned}
z_0 &= Wx + b_0, \\
z_l &= Wx + \sigma_l(z_{l-1}) + b_l, \quad l = 1, \ldots, L-1, \\
\text{C-MGN}(x) &= W^T \sigma_L(z_{L-1}) + V^T V x + b_L.
\end{aligned}
$$

- σ_l: non-decreasing (e.g., ReLU) ensuring PSD Jacobians :
-

$$
\mathrm{J}_{\text{C-MGN}}(x) = W^\top \left(\sum_{\ell=1}^{L} \prod_{i=\ell}^{L} J_{\sigma_i}(z_{i-1}) \right) W + V^\top V.
$$

Modular Monotone Gradient Network (M-MGN)

- Uses K parallel layers:

$$z_k = W_k x + b_k, \quad k = 1, \dots, K,$$

$$\text{M-MGN}(x) = a + V^T V x + \sum_{k=1}^{K} s_k(z_k)\, W_k\, \sigma_k(z_k).$$

- Can be rewritten as:

$$\text{M-MGN}(x) = a + W^\top \sigma(Wx + b)$$

- **Key:** Activation and scaling functions are chosen non-decreasing.

Loss Functions: Kullback–Leibler Divergence

- Defined as:

$$D_{KL}(p \parallel q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx.$$

- Closed-form for multivariate normals:

$$D_{KL}(\mathcal{N}(\mu_0, \Sigma_0) \parallel \mathcal{N}(\mu_1, \Sigma_1)) = \frac{1}{2}\Big[\mathrm{tr}(\Sigma_1^{-1}\Sigma_0)$$

$$+ (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0 - d + \log \tfrac{\det \Sigma_1}{\det \Sigma_0}\Big]$$

- **Limitation:** Only controls mean and covariance.

Loss Function : Dual Wasserstein Distance

- Formulated as:

$$W(\mu, \nu) = \sup_{f \in \mathrm{Lip}_1} \left\{ \mathbb{E}_{x \sim \mu}[f(x)] - \mathbb{E}_{y \sim \nu}[f(y)] \right\}$$

- Trains an adversarial critic (3-layer MLP, hidden dim = 64, leakyReLU) with gradient penalty.
- Better captures distribution divergence beyond mean/covariance.

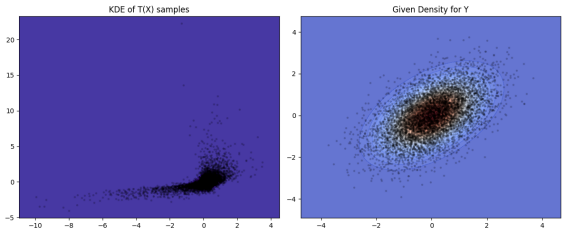
Experiments Overview

- **Gaussian Mapping:** Match two Gaussian distributions.
- **Gaussian Mixture Model (GMM) Mapping:** Transfer between GMMs.
- **Pixel Distribution Mapping:** Transform daytime to sunset pixel distributions.
- **MNIST Generation:** Generative model for handwritten digits.

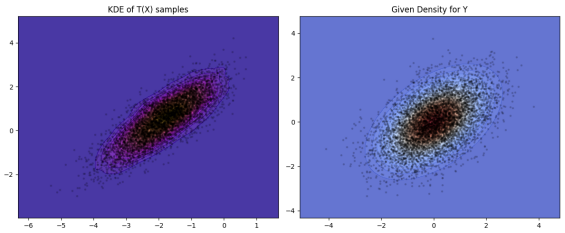
Gaussian Mapping Experiment

- **Setup:** Source and target Gaussians with means μ_1, μ_2 and covariances Σ_1, Σ_2.
- **Observation:**
  - KL loss aligns mean & covariance but may miss finer details.
  - Wasserstein loss yields a closer match.
- **Optimizer:** Adam, learning rate 10^{-4}.

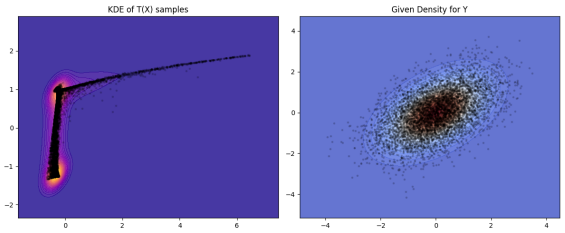
Gaussian Mapping Experiment : Results



(a) MMGN

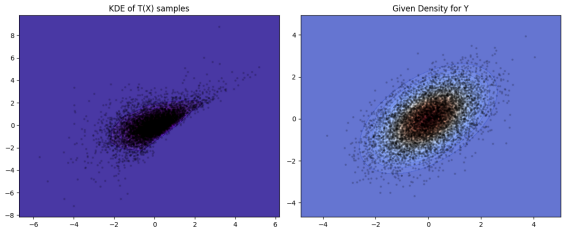


(b) CMGN

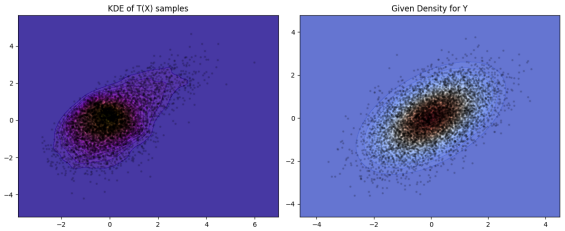


(c) ICNN

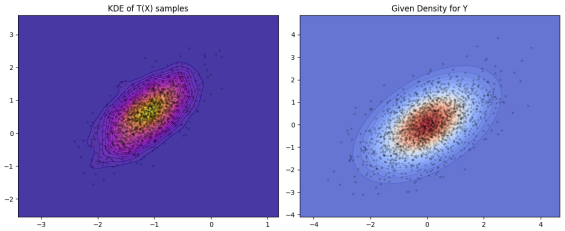
Figure: Results using the Kullback-Leibler divergence



(a) MMGN



(b) CMGN



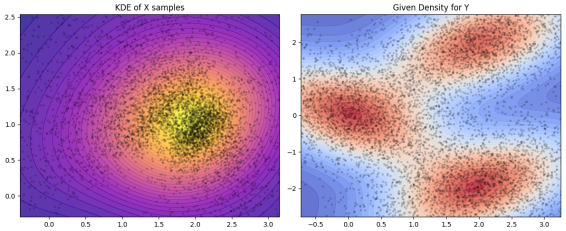
(c) ICNN

Figure: Results using the adversarial Wasserstein distance

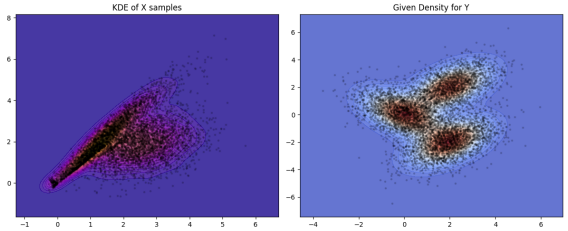
GMM Mapping Experiment

- **Setup:** Source GMM (2 components) vs. target GMM (3 components).
- **Example:** Source centered at $(-2, 0)$ and $(2, 0)$; target at $(-3, 1)$, $(0, -1)$, and $(3, 1)$.
- **Results:**
  - KL divergence perfectly aligns mean/covariance.
  - Wasserstein loss better captures overall distribution.
  - CMGN generally outperforms MMGN.

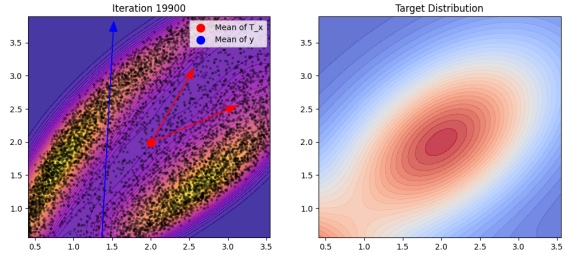
GMM Mapping Experiment: Results



(a) MMGN

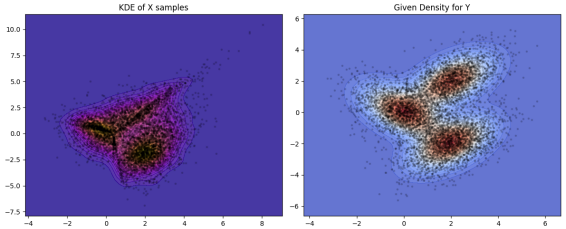


(b) CMGN

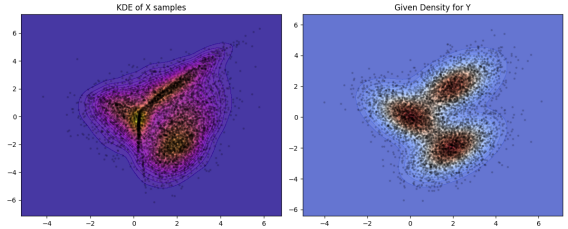


(c) ICNN

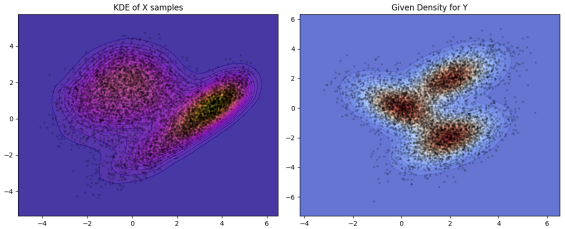
Figure: Results using the Kullback-Leibler divergence for GMM mapping



(a) MMGN



(b) CMGN



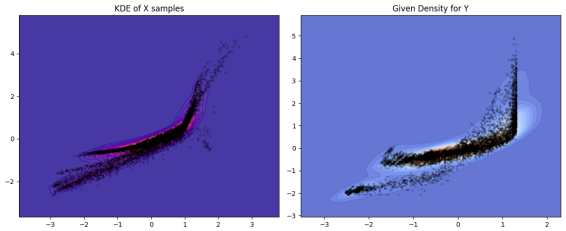
(c) ICNN

Figure: Results using the adversarial Wasserstein distance for GMM mapping

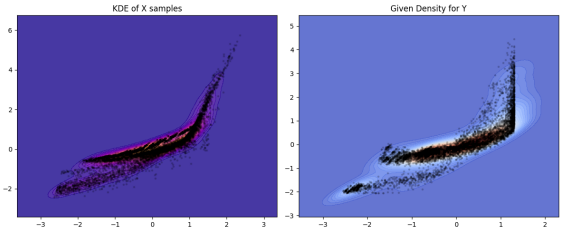
Pixel Distribution Mapping Experiment

- **Task:** Map pixel distribution of a daytime image to a sunset image.
- **Key requirement:** Preserve image structure.
- **Results:**
  - CMGN yields near-perfect mapping.
  - Demonstrates optimal transport for arbitrary distributions.
  - Training time: Approximately 10 seconds on a GTX 1050 Ti.

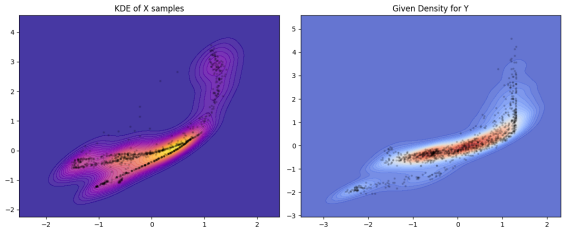
Pixel Distribution Mapping Experiment : Result



(a) MMGN



(b) CMGN



(c) ICNN

Figure: Mapped and target R and G channels

(a) MMGN

(b) CMGN

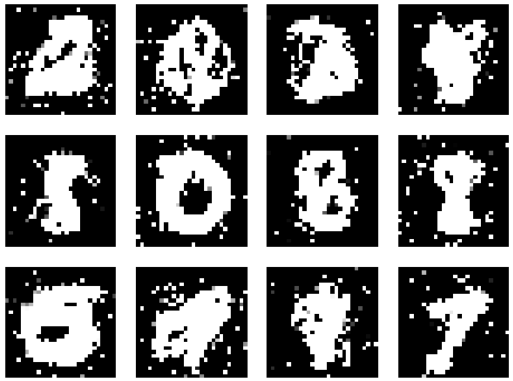
(c) ICNN

Figure: Results using the adversarial Wasserstein distance

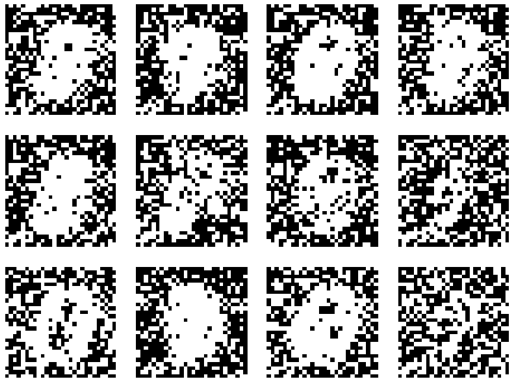
MNIST Generation Experiment

- **Task:** Train a generative network from a Gaussian latent space to reproduce MNIST digits.
- **Approach:** Adversarial Wasserstein loss.
- **Observations:**
  - CMGN produces reasonable samples with noise.
  - MMGN underperforms due to its simpler structure.
  - Compared to ICNN, both models are less sharp and class-consistent.

MNIST Generation Experiment



(a) CMGM



(b) MMGN

(c) ICNN

Figure: Results using the adversarial Wasserstein distance for Gaussian Mapping

Conclusion and Future Work

- Presented two architectures (C-MGN and M-MGN) that directly learn a monotone gradient mapping.
- Guarantee convexity via non-decreasing activation functions.
- Experiments show promising results in low-dimensional settings, with limitations on large-scale image generation.
- **Future:** Develop more complex architectures to handle higher-dimensional data.

Questions

Questions?