Generative Models for Images Project - Neural Optimal Transport

Clément Weinreich David Heurtel-Depeiges
Master MVA

March 25, 2024

1 Introduction

Optimal Transport (OT) theory is a mathematical framework that seeks the most efficient way of transporting mass from one distribution to another by minimizing a given cost. Due to its ability to quantify the geometric discrepancy between probability distributions, OT has become a common tool in machine learning where one often has to deal with collections of samples that can be interpreted as probability distributions. Since the publication of GAN by Goodfellow et al., 2014, generative modeling has been subject to considerable improvements. In particular, WGAN (Arjovsky et al., 2017) has opened the door to the use of OT in this field, making it a powerful tool for large-scale generative modeling tasks. Most methods expand on WGANs and use OT for the loss of generative models (Gulrajani et al., 2017, Liu et al., 2019). However, limited attention has been paid to the use of OT map as the generative model itself. In this line of research, Rout et al., 2022 proposed to compute deterministic (one-to-one) OT plans for the Wasserstein-2 distance. Building upon this work, Korotin et al., 2023b proposed a scalable algorithm (NOT: Neural Optimal Transport) to compute deterministic and stochastic (one-to-many) OT plans for strong and weak costs using deep neural networks.

In this project, we focus on this method from a theoretical and empirical point of view to provide a comprehensive understanding of the subject. We begin by presenting the necessary mathematical background to approach the topic (section 2). The introduced notations and concepts are then employed to describe NOT and explain how weak OT can be used for generative modeling and reconstruction tasks (section 3). The same section is equipped with experiments on toy and real data to validate the approach. Then we highlight the limits of using weak OT for learning generative models by explaining how NOT can fail and find "fake" OT plans (section 4). Finally, the study concludes with a discussion of the method and the various aspects raised in the previous sections (section 5). Our code can be publicly accessed on GitHub `https://github.com/david-heurtel-depeiges/NOT_project_MVA`.

2 Background on Optimal Transport and Weak Optimal Transport

Let $\mathcal{X}$ and $\mathcal{Y}$ two Polish spaces and $\mathcal{P}(\mathcal{X}), \mathcal{P}(\mathcal{Y})$ the respective sets of probability distributions on them. Let $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ two probability distributions, and $\Pi(\mu, \nu)$ the set of probability distributions on $\mathcal{X} \times \mathcal{Y}$ having marginals μ and ν For a strong cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the primal formulation of the strong OT cost given by Monge is

$$\text{Cost}(\mu, \nu) = \inf_{T_{\#}\mu = \nu} \int_{\mathcal{X}} c(x, T(x)) d\mu(x) \tag{1}$$

where the minimum is taken over measurable functions (or transport maps) $T : \mathcal{X} \to \mathcal{Y}$ that maps μ to ν, and $T_{\#}$ represents the associated push-forward operator. The optimal OT map is T^*. This formulation is not symmetric and does not allow mass splitting (there may be no T that satisfies $T_{\#}\mu = \nu$). A relaxation of this formulation given by Kantorovich is

$$\text{Cost}(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) \tag{2}$$

where the minimum is computed across all transport plans π. The optimal transport plan is denoted by $\pi^* \in \Pi(\mu, \nu)$. If π^* can be represented as $[\mathrm{id}_{\mathcal{X}}, T^*]_{\#}\mu \in \Pi(\mu, \nu)$ for T^* an optimal map, then T^* minimizes Equation 1, and the plan is called deterministic (otherwise it is stochastic) (Villani et al. 2009). Weak OT extends this formulation with a weak cost $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$, i.e., a function that takes a point $x \in \mathcal{X}$ and a distribution of $y \in \mathcal{Y}$ as input, one can define the weak OT cost (Gozlan et al. 2017) between μ, ν as

$$\mathrm{Cost}(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}} C(x, \pi(\cdot|x)) d\pi(x) \tag{3}$$

Using the weak cost $C : (x, \nu) \to \int_Y c(x, y) d\nu(y)$, the problem defined in Equation 3 becomes equivalent to the strong OT problem defined in Equation 2. A proof of this statement is given in subsection A.1. The weak OT problem is an extension of strong OT that allows the cost to depend on a distribution rather than on a single point. It models scenarios where the transport cost is not determined by single point-to-point mappings but rather by distributions of mass around those points.

A typical weak OT cost for $\mathcal{X} = \mathcal{Y} = \mathbb{R}^D$ is the γ-weak Wasserstein-2 cost for $\gamma \geq 0$ denoted $\mathcal{W}_{2,\gamma}$, corresponding to formulation (3) with the γ-weak quadratic cost

$$C_{2,\gamma}(x, \nu) = \int_{\mathcal{Y}} \frac{1}{2} ||x - y||^2 d\nu(y) - \frac{\gamma}{2} \mathrm{Var}(\nu) \tag{4}$$

where $\mathrm{Var}(\nu)$ is the variance of ν:

$$\mathrm{Var}(\nu) = \int_{\mathcal{Y}} ||y - \int_{\mathcal{Y}} y' d\nu(y')||^2 d\nu(y) = \frac{1}{2} \int_{\mathcal{Y} \times \mathcal{Y}} ||y - y'||^2 d\nu(y) d\nu(y'). \tag{5}$$

Note here that using $\gamma = 0$ makes the transport cost strong. When used within Equation 4, the variance term acts as a regularization, it measures the spread of $\nu = \pi(\cdot|y)$ around its mean. Using $\gamma > 0$ encourages transportation plans that do not spread mass too thinly. This cost provides a balance between the cost of moving mass (integral term) and the dispersion of mass (variance term).

When the cost $C(x, \nu)$ is lower-bounded, convex in ν, and lower semi-continuous, it is called appropriate. For appropriate costs, Veraguas et al., 2019 prove that the minimizer π^* of Equation 3 always exists. Since the functional $\mathrm{Var}(\mu)$ is concave in μ and non-negative, the weak quadratic cost defined in Equation 4 is appropriate when $\gamma \in [0, 1]$ (so it is lower-bounded). For appropriate costs, Veraguas et al., 2019 also show that the weak OT cost (3) admits the dual formulation

$$\mathrm{Cost}(\mu, \nu) = \sup_f \int_{\mathcal{X}} f^C(x) d\mu(x) + \int_{\mathcal{Y}} f(y) d\nu(y), \tag{6}$$

where f are upper-bounded, continuous, and not rapidly decreasing functions, and

$$f^C(x) = \inf_{\nu \in \mathcal{P}(\mathcal{Y})} \left\{ C(x, \nu) - \int_{\mathcal{Y}} f(y) d\nu(y) \right\} \tag{7}$$

is the weak C-transform[1].

3 Application to generative modeling and reconstruction

3.1 Neural Optimal Transport method

The authors Korotin et al., 2023b present an algorithm to implicitly learn an OT plan π^* using neural networks. The method relies on the use of a latent atomless distribution $\mathbb{S} \in \mathcal{P}(\mathcal{Z})$ where $\mathcal{Z} \subset \mathbb{R}^S$ (e.g. $\mathbb{S} = \mathrm{Unif}([0, 1]^Z)$ or $\mathbb{S} = \mathcal{N}(0, I_Z)$). To derive the algorithm, the authors reformulate the dual form while considering this latent distribution, which leads to a saddle point optimization problem that allows to recover a stochastic OT map. In this section, we review the main steps that allow us to

[1] This formulation extends the c-transform seen in class where the infimum was taken directly on $y \in \mathcal{Y}$.

build the foundations of the algorithm. For the next parts, it is assumed that μ, ν are supported on subsets $\mathcal{X} \subset \mathbb{R}^M$ and $\mathcal{Y} \subset \mathbb{R}^N$ respectively. The approach starts from reformulating the C-transform to get an analogous form of its integral in the dual form (6) allowing to get the desired saddle point optimization problem.

For a measurable map $t : \mathcal{Z} \to \mathcal{Y}$, the C-transform (7) can be reformulated as

$$f^C(x) = \inf_t \left\{ C(x, t_{\#}\mathbb{S}) - \int_{\mathcal{Z}} f(t(z)) d\mathbb{S}(z) \right\}. \tag{8}$$

This formulation works thanks to Lemma 1 of Korotin et al., 2023b which shows that there exists a measurable map $t : \mathcal{Z} \to \mathcal{Y}$ satisfying $\nu = t_{\#}\mathbb{S}$ as $\mathcal{Z} \subset \mathbb{R}^S$, $\mathcal{Y} \subset \mathbb{R}^N$ and $\mathbb{S}$ is atomless. In this formulation, $C(x, t_{\#}\mathbb{S})$ represents the cost of moving mass from a point x in $\mathcal{X}$ to the distribution that results from pushing from $\mathbb{S}$ the map t. The map t defines how each point z from the support of $\mathbb{S}$ is transported to a point in the target space $\mathcal{Y}$. Essentially, it "reshapes" the atomless distribution $\mathbb{S}$ to fit the target space in a way that respects the cost C. This reformulation of the C-transform allows us to consider the cost of transporting mass from a single point x not to a single point y in the target space but to a distribution around y which is specified by the pushforward measure $t_{\#}\mathbb{S}$. As mentioned in section 2, it is particularly relevant when the target measure is not a single one-to-one mapping from the source space, but a more complex distribution that requires a stochastic representation. Integrating this formulation of the C-transform gives

$$\int_{\mathcal{X}} f^C(x) d\mu(x) = \inf_T \int_{\mathcal{X}} \left(C\left(x, T(x, \cdot)_{\#}\mathbb{S}\right) - \int_{\mathcal{Z}} f(T(x, z)) d\mathbb{S}(z) \right) d\mu(x) \tag{9}$$

where the minimization is performed over all measurable map $T : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$. The map T represents the stochastic transport plan that specifies how the mass from point x is distributed over the target space $\mathcal{Y}$ according to the distribution $\mathbb{S}$ transformed by T. The integration of the C-transform provides a way to compute the overall cost of transporting mass from μ to ν across all points in $\mathcal{X}$, taking into account the stochastic nature of the transport plan T and cost C. It allows for optimizing over transport plans that may spread the mass from each point in $\mathcal{X}$ over a set of points in $\mathcal{Y}$. Substituting the integrated C-transform into the dual formulation (6) automatically leads to a max-min optimization problem (or a saddle point optimization problem)

$$\text{Cost}(\mu, \nu) = \sup_f \inf_T \int_{\mathcal{Y}} f(y) d\nu(y) + \int_{\mathcal{X}} \left(C(x, T(x, \cdot)_{\#}\mathbb{S}) - \int_{\mathcal{Z}} f(T(x, z)) d\mathbb{S}(z) \right) d\mu(x) \tag{10}$$

where the the functional under $\sup_f \inf_T$ is denoted by $\mathcal{L}(f, T)$.

For two probability distributions, μ, and ν, T^* is considered as a stochastic OT map if it corresponds to an optimal transport plan π^*. The key argument of the method is Lemma 4 of Korotin et al., 2023b which shows that such maps T^* can be obtained by solving the previous inner $\inf_T$ problem for optimal f^*. For every optimal potential $f^* \in \arg\sup_f \inf_T \mathcal{L}(f, T)$, it holds that

$$T^* \in \arg\inf_T \mathcal{L}(f^*, T) \tag{11}$$

When using the γ-weak quadratic cost (4), Gozlan and Juillet, 2020 show that a maximizer f^* of the dual form (6) exists. Hence, it is possible to extract optimal maps T^* from optimal saddle points (f^*, T^*) obtained by optimizing Equation 10.

3.2 Learning with NOT in practice

In practice, the saddle point problem can be approached with neural networks $f_\omega : \mathbb{R}^N \to \mathbb{R}$ (potential network) and $T_\theta : \mathbb{R}^M \times \mathbb{R}^S \to \mathbb{R}^N$ (mapping network) to parameterize f and T. In practice, it is also necessary to have an empirical estimator $\hat{C}$ of $C(x, T(x, \cdot)_{\#}\mathbb{S})$. For the γ-weak quadratic cost, one can derive an unbiased Monte-Carlo estimator from a random batch $Z \sim \mathbb{S}$ of size $|Z| \geq 2$:

$$\begin{aligned} C_{2,\gamma}(x, T(x, \cdot)_{\#}\mathbb{S}) &= \frac{1}{2} \int_{\mathcal{Z}} ||x - T(x, z)||^2 d\mathbb{S}(z) - \frac{\gamma}{2} \text{Var}(\mathbb{S}) \\ &\approx \frac{1}{2|Z|} \sum_{z \in Z} ||x - T(x, z)||^2 - \frac{\gamma}{2} \hat{\sigma}^2 = \hat{C}_{2,\gamma}(x, T(x, Z)) \end{aligned} \tag{12}$$

where $\hat{\sigma}^2$ is the corrected batch variance. Hence, learning a stochastic OT map T_θ from μ to ν can be done by iterating on batches $Y \sim \nu$, $X \sim \mu$, and for each $x \in X$ a batch $Z_x \sim \mathbb{S}$. At each iteration, we update f_ω using gradient ascent with $\frac{\partial \mathcal{L}(f_\omega, T_\theta)}{\partial \omega}$. And for K_T inner iterations we sample new batches $X \sim \mu$, $Z_x \sim \mathbb{S}$ and update T_θ using gradient descent with $\frac{\partial \mathcal{L}(f_\omega, T_\theta)}{\partial \theta}$.

3.3 2D toy dataset

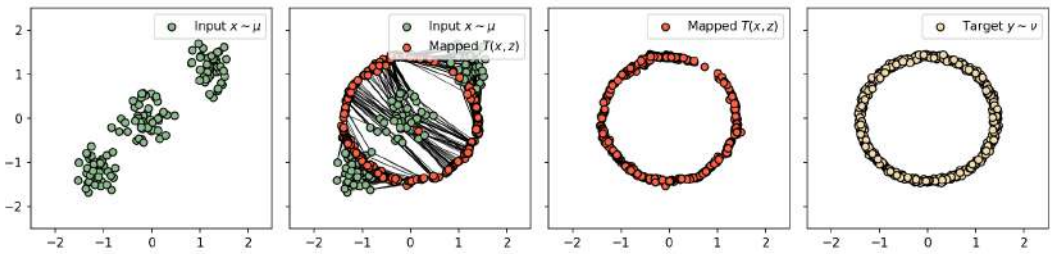

Figure 1: Learned map $T_\theta(x, z)$ between a mixture of 3 isotropic Gaussians and a noised circle using the γ-weak quadratic cost with $\gamma = 0.75$.

To validate the performance of NOT we learned transport maps between 2D toy distributions. We employed MLPs for f_ω and T_θ, further details on optimization are given in Appendix C. Figure 1 shows the final mapping from a mixture of 3 isotropic Gaussians to a noised circle. Overall the pushforward distribution $T_\#(\mu \times \mathbb{S})$ seems to match the target distribution ν. A more challenging setting with tighter μ illustrates a weakness of the NOT algorithm (instabilities, especially for higher values of γ, see Figure 5 and further discussion in section 4).

3.4 One-to-many image translation

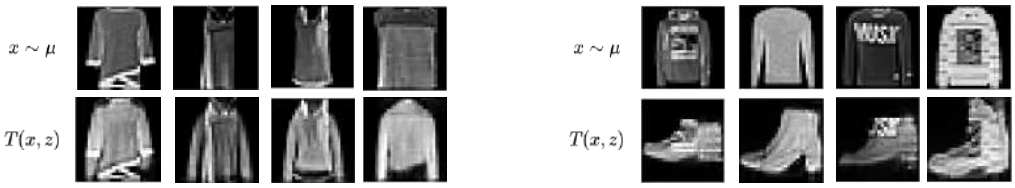

Figure 2: Unpaired image translations with $T_\theta(x, z)$ between T-shirt and Coat (left), and between Pullover and Ankle boot (right) using γ-weak quadratic cost with $\gamma = 1$.

We applied the NOT algorithm to image datasets in the context of unpaired image translations. All these experiments were performed using CNNs. The architectures and details about optimization are given in Appendix C. We first consider two subsets of the Fashion-MNIST (Xiao et al. 2017) dataset for μ and ν. Figure 2 shows the translation from T-shirt to Coat, and from Pullover to Ankle boot. Furthermore, Figure 6 given in the appendix shows the effect of γ on the structure of the learned stochastic map. Going from $\gamma = 0$ (strong cost) with a conditional collapse we observe increasing variety in style as γ grows, up to $\gamma = 1.25$ where significant differences appear. Even though this effect remains limited due to the gray-scale nature of this dataset, one can observe a variance-similarity trade-off controlled by the parameter γ. We also performed the same task between two classes of images in the CIFAR10 (Krizhevsky 2012) dataset. However, due to training instabilities, the results seem convincing. Some samples are shown in Figure 3, and more quantitative results are given in Figure 7 (in appendix).

3.5 Image Reconstruction

In addition to reproducing toy datasets and one-to-many image translation applications, we test the NOT framework on an image reconstruction task. In image reconstruction, the goal is to recover an

image from a corrupted version of it, where the corrupted image is obtained by applying a degradation operator (blurring, masking, down-sampling) and, optionally, adding noise. For most operators, the problem is ill-posed and requires some form of regularization. In deep generative modeling, this is achieved either by using a generative prior (diffusion model with guidance, joint posterior maximization in the latent space of a VAE or a GAN) or a denoiser (Plug-and-Play methods like ADMM). As a side application of the NOT framework, we propose to use the optimal transport plan learning algorithm to directly learn a stochastic-OT map that transports the corrupted image to the clean one.

More formally, given a target image distribution ν, a degradation operator $\mathcal{A}$ and a noise level σ, the corrupted image distribution μ is given by:

$$\mu = \mathcal{A}_{\#}\nu + \mathcal{N}(0, \sigma^2) \tag{13}$$

We then learn a stochastic OT map T that transports the corrupted image distribution μ to the clean one ν. The cost functions we chose to use are the γ-weak quadratic cost and the γ-weak kernel cost (see Seq. 4.2). We try this idea on the MNIST dataset for a proof of concept on a super-resolution and to explore hyper-parameter choices. We then apply the method to the flowers-256 dataset (Nilsback and Zisserman, 2008) to see how it performs on higher-dimensional data (Appendix B).

One should note that, contrary to usual image reconstruction methods, the NOT framework is not designed per se to sample either the posterior mean, posterior distribution or the maximum a posteriori estimate. If training is successful, reconstructed images belong to the support of the target distribution ν but there is no guarantee that the reconstructed images are representative of the posterior distribution. For this reason, not every cost function is adapted to every degradation operator. For super-resolution, costs based on the L_2 norm seem appropriate. For other image reconstruction tasks, like deblurring, one might want to use, as suggested by Korotin et al., 2023b, other basis for cost functions (with a data-consistency term for instance).

Toy dataset example Before applying the method for image reconstruction, we first try a low-dimensional reconstruction task on 2D toy data. We consider a simple Gaussian Mixture Model dataset with a non-linear degradation operator (projection on the unit circle). The standard L_2 distance seems appropriate to this "un-projection" task. As shown in Figure 11, NOT effectively succeed in reconstructing the initial distribution.

Results on MNIST We down-sample 32×32 images to 16×16 resolution before adding noise and up-sampling them (so that μ and ν have support of equal dimension). We then train different models for weak costs and weak kernel costs. Models are evaluated using the Peak Signal to Noise Ratio.

	$\gamma = 0$	$\gamma = 1/2$	$\gamma = 1$
NOT	22.6	22.2	22.8
KNOT	22.7	21.9	22.3

Table 1: Peak Signal to Noise Ratio (PSNR) ↑ on MNIST super-resolution for the different methods

Quantitatively, we observe that stochastic maps learned under the weak kernel cost underperform their weak quadratic cost counterpart (the PSNR is lower). However, qualitatively, samples generated using the weak kernel cost appear to be of higher quality, both with respect to the base image and as MNIST digits in general.

This difference may be due to the choice of the PSNR as a comparison metric. Maybe focusing on the FID or SSIM metrics would have yielded quantitative results more in line with our qualitative analysis.

In addition, we found that training with the weak quadratic cost was significantly less stable, even for distributions that are quite close to each other.

4 Limits of weak OT for learning generative models

4.1 Existence of "fake" optimal solutions

Thanks to Equation 11, the NOT algorithm allows to extract an optimal stochastic OT map T^* from the solution (f^*, T^*). However, the arg inf set for an optimal f^* may not only contain the optimal stochastic transport maps but also other stochastic functions (such as fake solutions), which means that not all T^* are optimal stochastic OT maps. Korotin et al., 2023a even show that it is especially true for the γ-weak quadratic cost, (4) as the arg inf set might contain fake solutions T^*. Fake solutions correspond to learned maps that do not satisfy the desired property of transporting one distribution to another optimally according to the cost function. These solutions are not distribution-preserving, for $f^* \in \arg\sup_f \inf_T \mathcal{L}(f, T)$ and $T^\dagger \in \arg\inf_T \mathcal{L}(f^*, T)$ with $T^\dagger$ denoting a fake solution, then we have $T^\dagger_\#(\mu \times \mathbb{S}) \neq \nu$. Hence, even though the NOT algorithm finds a saddle point, it does not guarantee that the corresponding transport plan T^* is valid; it could be a fake solution that does not properly transport the distribution μ to ν. As noticed by Korotin et al., 2023b, the emergence of fake solutions can be noticed during optimization if the method exhibits fluctuations between various non-optimal mappings, rather than converging to a stable solution. We independently observe this phenomenon on other tasks and datasets as illustrated in Figure 8 (in appendix).

Moreover, Korotin et al., 2023a (Theorem 1 and Corollary 1) proves the existence of fake saddle points (f^*, T^*), in which T^* is not a stochastic OT map, because it does not ensure the full transport of μ to ν. The main ideas of the proofs are given in subsection A.1.

4.2 Strictly convex costs and kernel costs

Knowing that the $\arg\inf_T \mathcal{L}(f^*, T)$ set might contain fake solution, the authors found a condition to overcome this limit: for strictly convex costs $C(x, \nu)$ in ν, all the solutions of the saddle point problem provide stochastic OT maps, resolving the problem of fake solutions. This Lemma is proved by Korotin et al., 2023b in Appendix F. The weak quadratic costs are simply convex in ν but not strictly convex, hence the solutions are indeed not guaranteed to be stochastic OT maps. A plausible explanation for the occurrence of fake solutions in the context of weak costs $C(x, \nu)$ could be the presence of regions in $\mathcal{P}(\mathcal{Y})$ where $C(x, \nu)$ is not strictly convex in ν. To solve this issue, one could think about adding a strictly convex in ν regularization term $R(\nu)$ to the weak cost $C(x, \mu)$, or more directly considering strictly convex costs. Korotin et al., 2023a propose kernel costs that are strictly convex and can easily be estimated from samples.

Let $\mathcal{H}$ be a Hilbert space (feature space) and $u : \mathcal{Y} \to \mathcal{H}$ a function (feature map). The PDS (Positive Definite Symmetric) kernel $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ with the feature map u is denoted by $k(y, y') = \langle u(y), u(y') \rangle_{\mathcal{H}}$. For $\nu \in \mathcal{P}(\mathcal{Y})$, the PDS kernel $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is said to be characteristic if its kernel mean embedding $u(\nu) = \int_{\mathcal{Y}} u(y) d\nu(y) \in \mathcal{H}$ is injective (one-to-one mapping). Injectivity in this context means that each distribution in $\mathcal{P}(\mathcal{Y})$ can be uniquely represented in the RKHS, allowing distinct distributions to be distinguished based on their embeddings. In this setting, they define the γ-weak quadratic cost between features

$$C_{u,\gamma}(x, \nu) = \frac{1}{2} \int_{\mathcal{Y}} ||u(x) - u(y)||^2_{\mathcal{H}} d\nu(y) - \frac{\gamma}{2} \left[\frac{1}{2} \int_{\mathcal{Y} \times \mathcal{Y}} ||u(y) - u(y')||^2_{\mathcal{H}} d\nu(y) d\nu(y') \right]. \tag{14}$$

This cost can be reformulated using only the kernel function k which removes the need of knowing the explicit feature map u. Using the PDS kernel k, the expression $||u(y) - u(y')||^2_{\mathcal{H}}$ can be reformulated as the difference in kernel evaluations $k(y, y) - 2k(y, y') + k(y', y')$. Hence, we can rewrite the previous cost as

$$C_{k,\gamma}(x, \nu) = \frac{1}{2} k(x, x) - \int_{\mathcal{Y}} k(x, y) d\nu(y) + \frac{1 - \gamma}{2} \int_{\mathcal{Y}} k(y, y) d\nu(y) + \frac{\gamma}{2} \int_{\mathcal{Y} \times \mathcal{Y}} k(y, y') d\nu(y) d\nu(y'), \tag{15}$$

which is a generic γ-weak kernel cost. Note here that this cost generalizes the γ-weak quadratic cost when $\mathcal{H} = \mathbb{R}^D$, $u(x) = x$, with the respective bilinear kernel $k(x, y) = \langle u(x), u(y) \rangle = \langle x, y \rangle$. The γ-weak quadratic cost directly uses the squared Euclidean distance implying an explicit feature map,

and the variance of ν is also explicitly computed. The γ-weak kernel cost uses a kernel to implicitly define the feature space, and the variance and interactions within ν are captured implicitly through the kernel function evaluations.

Korotin et al., 2023a prove that γ-weak kernel costs are appropriate, hence for k a continuous PDS kernel and $\gamma \in [0, 1]$, an OT plan π^* for the cost $C_{k,\gamma}$ exists and the previously discussed dual formulation of Equation 6 holds true. Moreover, if k is a characteristic kernel and $\gamma \in (0, 1]$, then the OT plan π^* for cost $C_{k,\gamma}(x, \nu)$ is unique. This is because if k is characteristic, then $C_{k,\gamma}(x, \nu) = C_{u,\gamma}(x, \nu)$ is strictly convex in ν. In this context, they provide their key Theorem 2 which states that under the previous hypothesis and using the cost $C_{k,\gamma}(x, \nu)$, for any optimal potential $f^* \in \arg\sup_f \inf_T \mathcal{L}(f, T)$ we have $T^* \in \arg\inf_T \mathcal{L}(f^*, T)$ if and only if $T^*(x, \cdot)_{\#}\mathbb{S} = \pi^*(y|x)$ holds true μ-almost surely for all $x \in \mathcal{X}$. The proof relies on the strict convexity of $C_{k,\gamma}(x, \nu)$. This implies that every optimal saddle point (f^*, T^*) provides a stochastic OT map T^*, which effectively is distribution-preserving. However, this important theorem implicitly assumes the existence of a maximizer f^* of the dual form (6) for the γ-weak kernel cost. In this context, the proof of Gozlan and Juillet, 2020 discussed for (11) does not hold anymore. It is guaranteed to find a stochastic OT map T^* if we have f^*, but there are no theoretical guarantees for its existence. A remaining unanswered question thus subsists in the precise conditions for the existence of such maximizers.

As the bilinear kernel is not characteristic, this theorem does not apply which makes the γ-weak quadratic cost sensible to fake solutions. In practice, one can consider the family of distance-induced kernels $k(x, y) = \frac{1}{2}(||x||^\alpha + ||y||^\alpha - ||x - y||^\alpha)$ which provides $||u(x) - u(x')||^2_{\mathcal{H}} = k(x, x) - 2k(x, x') + k(x', x') = ||x - x'||^\alpha$. With this kernel, the γ-weak kernel cost (14,15) can be simplified into

$$C_{k,\gamma}(x, \nu) = C_{u,\gamma}(x, \nu) = \frac{1}{2}\int_{\mathcal{Y}} ||x - y||^\alpha d\nu(y) - \frac{\gamma}{2}\left[\frac{1}{2}\int_{\mathcal{Y}\times\mathcal{Y}} ||y - y'||^\alpha d\nu(y)d\nu(y')\right]. \tag{16}$$

Using $\alpha = 2$ leads to the bilinear kernel and we recover the γ-weak quadratic cost of Equation 4 (writing the variance term as in Equation 5 makes it clearer). However, Sejdinovic et al., 2013 proved that using $\alpha = 1$ yields a PDS and characteristic kernel. Thus it guarantees that the solution of the saddle point optimization problem is a stochastic OT map. Finally, an unbiased Monte-Carlo estimator $\hat{C}_{k,\gamma}$ for $x \in \mathcal{X}$ and a batch $Z \sim \mathbb{S}$ can be derived in the same manner which makes it possible to use the cost in practice.

4.3 Experiments with the weak kernel cost

4.3.1 2D Toy dataset

We consider the same challenging 2D Toy dataset as in subsection 3.3. In Figure 9, one can observe that for the same setup and value of γ, but using the γ-weak kernel cost $C_{k,\gamma}$, the optimization is not fluctuating between non-optimal mappings and the pushforward $T_{\#}(\mu \times \mathbb{S})$ is closely matching ν. Using this cost effectively prevents oscillating between saddle points that provide fake solutions. Even in this challenging setting, the γ-weak kernel cost remains stable for different values of γ as shown in the appendix in Figure 5 (right). According to these qualitative results, from $\gamma = 0.5$ to $\gamma = 1.25$ the stochastic OT map $T_\theta(x, z)$ correctly transport distribution μ to ν, which was not possible using $C_{2,\gamma}$.

4.3.2 One-to-many image translation

For the unpaired image translation task, Figure 10 presents a comparison of 3 examples of pullover to ankle boot translations, showcasing the differences between the strong cost, weak quadratic cost, and weak kernel cost. One can observe that using the γ-weak kernel cost $C_{k,\gamma}$ instead of $C_{2,\gamma}$ seems to improve the variety of samples for the same values of γ. Due to the low number of samples per class for these datasets, it is challenging to compute the FID and quantitatively assess these results. However, qualitatively it seems that the γ-weak kernel cost improves the stability and helps NOT to converge to a more stable solution. Figure 3 illustrates the comparison between bird-to-car image translation from the CIFAR10 dataset using two different costs. The results, after training for 40K iterations with the only variation being the cost, show more stability with $C_{k,\gamma}$. The same behavior can be observed in the additional results in Figure 7 (in appendix).

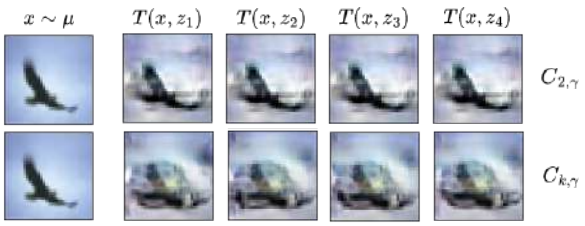

Figure 3: Unpaired image translation with $T_\theta(x, z)$ between birds and cars in the CIFAR10 dataset, using the γ-weak quadratic cost (first row) and the γ-weak kernel cost (second row) with $\gamma = 0.5$.

5 Discussion

Impact Korotin et al. 2023b introduces a novel Optimal Transport algorithm, that allows to learn stochastic or deterministic OT maps using deep neural networks. While WGANs only use OT as a metric between generated and target distributions, the NOT algorithm is more general and can be used for image-to-image tasks such as translation.

Limitation In the NOT framework, not all solutions of the saddle point optimization problem are valid stochastic OT maps, with correct marginals. The γ-weak quadratic cost can lead to fake solutions and a multitude of non-optimal mappings, leading to training instabilities. Korotin et al. 2023a then introduces a new cost, the γ-weak kernel cost, which is strictly convex and ensures that the solution of the saddle point optimization problem is a stochastic OT map. We empirically verify that this cost is more stable and produces better qualitative results. However, as noticed by Korotin et al. 2023a, although the γ-weak kernel cost is strictly convex in ν, leading to a correct optimal transport plan if f^* exists, there are no theoretical guarantees as to that existence.

Reconstruction We test the use of the NOT algorithm for image reconstruction tasks (in an unpaired setting), namely super-resolution on both a subset of MNIST and the Flowers-256 dataset (for a higher dimensional problem). Although NOT maps achieve better PSNR, KNOT samples seem of higher quality. Overall, as expected, variety in generated samples increase with γ, while staying coherent with features of the noisy observation. On the Flowers dataset, we observe that both costs lead to sensitive hyper-parameter choices with non-satisfying results.

Perspective With both NOT and KNOT methods, the authors open the door to a wide variety of cost functions, suggesting the use of perceptual losses for example. In the case of KNOT an open problem remains the existence of f^* for kernel costs. Overall, exploring theoretical guarantees depending on cost functions could lead to new and interesting applications of the NOT algorithm for other imaging tasks.

References

J.-J. Alibert, G. Bouchitté, and T. Champion. A new class of costs for optimal transport planning. *European Journal of Applied Mathematics*, 30(6):1229–1263, 2019.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan, 2017.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014.

N. Gozlan and N. Juillet. On a mixture of brenier and strassen theorems. *Proceedings of the London Mathematical Society*, 120(3):434–463, Mar. 2020. ISSN 1460-244X. doi: 10.1112/plms.12302. URL http://dx.doi.org/10.1112/plms.12302.

N. Gozlan, C. Roberto, P.-M. Samson, and P. Tetali. Kantorovich duality for general transport costs and applications. *Journal of Functional Analysis*, 273(11):3327–3405, 2017. ISSN 0022-1236.

doi: https://doi.org/10.1016/j.jfa.2017.08.015. URL https://www.sciencedirect.com/science/article/pii/S0022123617303294.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans, 2017.

A. Korotin, D. Selikhanovych, and E. Burnaev. Kernel neural optimal transport, 2023a.

A. Korotin, D. Selikhanovych, and E. Burnaev. Neural optimal transport, 2023b.

A. Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

H. Liu, X. Gu, and D. Samaras. Wasserstein gan with quadratic transport cost. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4832–4841, 2019.

M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

L. Rout, A. Korotin, and E. Burnaev. Generative modeling with optimal transport maps, 2022.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5), Oct. 2013. ISSN 0090-5364. doi: 10.1214/13-aos1140. URL http://dx.doi.org/10.1214/13-AOS1140.

J. B. Veraguas, M. Beiglboeck, and G. Pammer. Existence, duality, and cyclical monotonicity for weak transport costs, 2019.

C. Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

A Additional mathematical details

A.1 Proofs

Weak OT cost generalizes strong OT cost: Using the weak cost $C : (x, \nu) \to \int_Y c(x, y)d\nu(y)$, the problem defined in Equation 3 becomes equivalent to the strong OT problem defined in Equation 2

Proof. Incorporating this cost into Equation 3 gives

$$\begin{aligned} Cost(\mu, \nu) &= \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} c(x, y)d\pi(\cdot|x)(y) \right) d\pi(x) \\ &= \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}} \int_{\mathcal{Y}} c(x, y)d\pi(y|x)d\pi(x) \\ &= \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)d\pi(x, y) \quad \text{(Strong OT formulation)} \end{aligned}$$

The last line is obtained as $d\pi(y|x)d\pi(x)$ is the disintegration of π with respect to x. □

Main ideas of Korotin et al., 2023a's Theorem 1 / Corollary 1: Korotin et al., 2023a proves the existence of fake saddle points (f^*, T^*), in which T^* is not a stochastic OT map. First, they rely on the fact that for the γ-weak quadratic cost (4) on $\mathcal{X} = \mathcal{Y} = \mathbb{R}^D$, Alibert et al., 2019 proved that there exists a continuously differentiable convex function $\phi^* : \mathbb{R}^D \to \mathbb{R}$ such that $\pi^* \in \Pi(\mu, \nu)$ is optimal if and only if $\int_{\mathcal{Y}} y d\pi^*(y|x) = \nabla\phi^*(x)$ holds true μ-almost surely. So for the transport to be optimal, the average location to which x is moved (considering all possible destinations y) must be exactly $\nabla\phi^*(x)$. In that case, ϕ^* is called an optimal restricted potential. Under certain conditions, their Theorem 1 characterizes the $\arg\inf_T \mathcal{L}(f^*, T)$ set by showing that any $T^\dagger$ within this set must align its output's

expectation with the gradient of the optimal restricted potential, $\mathbb{E}_{z\sim\mathbb{S}}[T^\dagger(x,z)] = \nabla\phi^*(x)$, μ-almost everywhere. However, this condition alone does not ensure that $T^\dagger$ respects the entire structure of an optimal transport plan, only that it matches at the first-order moments driven by $\nabla\phi^*$. The Corollary 1 then leverages this characterization to assert the existence of optimal saddle points where the associated T^* does not fulfill the role of a stochastic OT map, essentially because it does not ensure the full transport of the distribution μ to ν as required. This arises because the weak quadratic cost framework can validate mappings that align with $\nabla\phi^*$ in expectation but may diverge in terms of actual distribution transportation, thus leading to fake solutions.

B Additional experiments

Super-Resolution on Flowers-256 Images from the Flowers dataset (Nilsback and Zisserman, 2008) are cropped to resolution 256 to form distribution μ. ν is then obtained using $4\times$ dow-sampling and adding noise. Both the NOT and KNOT methods proved unstable in learning optimal transport maps. In most cases, artifacts appeared during training (grid artifacts for UNet architecture for example). We did not find robust enough sets of hyper-parameters to achieve correct reconstruction and compare methods. Samples from the learned OT reconstruction map can be found in Figure 4. The stochastic OT map is able to remove the added noise in the down-sampled images and is able to reconstruct some lost features but we are far from state-of-the-art methods.

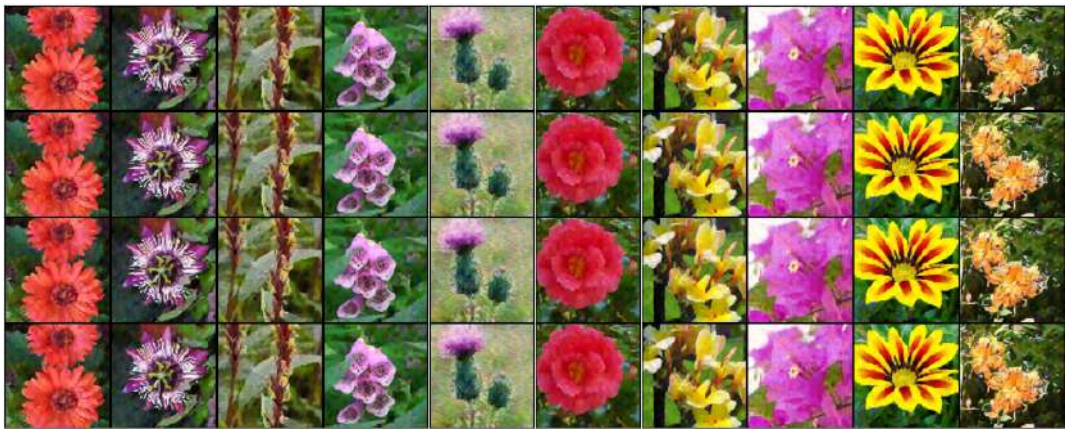

Figure 4: Samples from a noisy-super resolution reconstruction map

C Experimental details

C.1 2D Toy datasets

For these experiments, we considered MLPs with two hidden layers of dimension 128 and ReLU activations for f_ω and T_θ. The stochastic map $T_\theta(x,z)$ takes as input $x \in \mathbb{R}^2$ and $z \in \mathbb{R}^2$ as a single vector, with $z \sim \mathbb{S}$ where $\mathbb{S}$ is a standard normal distribution. We do 10 inner iterations for T_θ, use Adam optimizer for both networks with a learning rate of 10^{-4}, and a batch size of 64. For the weak cost, 4 noise samples z are sampled per image x. We train for 10K updates of f_ω, which takes approximately 7 minutes on an NVIDIA V100 GPU.

C.2 Unpaired image translations

We employed CNNs for potential and mapping networks. T_θ consists in 7 convolutional layers with 128 kernels of size 5×5 that are interleaved with ReLU non-linearities. The stochastic map $T_\theta(x,z)$ takes as input an image x with c channels and a standard normal noise z of the same shape with 1 channel. The potential network f_ω consists of 5 convolutional layers interleaved with ReLU activations and Average pooling layers. Other optimization details are similar to subsection C.1. We train for 3K

updates of f_ω for Fashion-MNNIST which takes approximately 35 minutes on an NVIDIA V100 GPU. For the CIFAR dataset, we trained for 40K updates of f_ω, taking 30 hours on the same GPU.

C.3 Super-resolution

For MNIST, digit 2 we employed CNNS for potential and mapping networks, as for the unpaired image translation. We tested variations on the number of channels and depth of the networks. We trained for 3K updates to 10k updates, depending on the phenomenon we wanted to observe, with 1K updates taking 6 minutes on an RTX 3090.

For the Flowers-256, we tried both UNet and standard CNN architectures, both failing at successfully solving the reconstruction problem. Training was significantly slower, due to image size, with 1K update taking an hour on the same GPU.

D Large figures

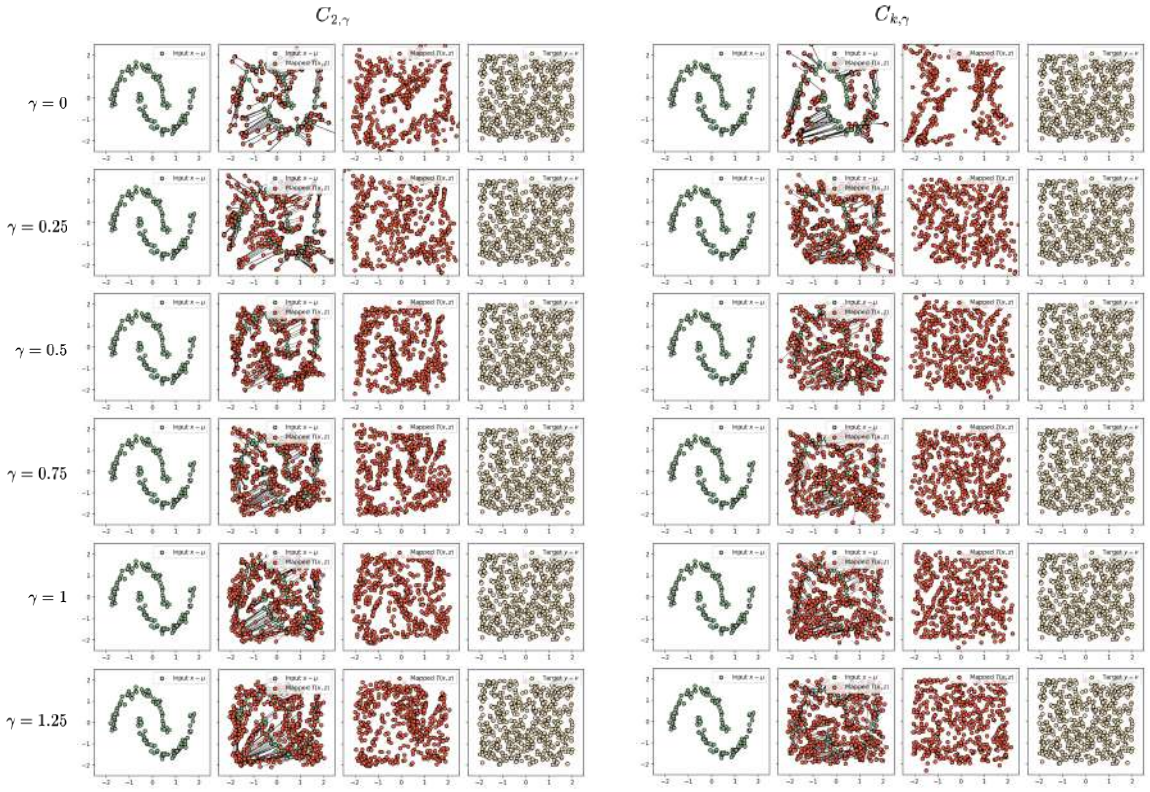

Figure 5: Comparing learned maps between μ (two interleaved half circles) and $\nu \sim \text{Unif}([-2, 2]^2)$, for different values of γ between 0 and 1.25, using the γ-weak quadratic cost (left) and the γ-weak kernel cost (right). Mapping from these two distributions is challenging as μ is particularly less dispersed than ν, which makes the optimization fluctuate between saddle points (as observed in the left subfigure).

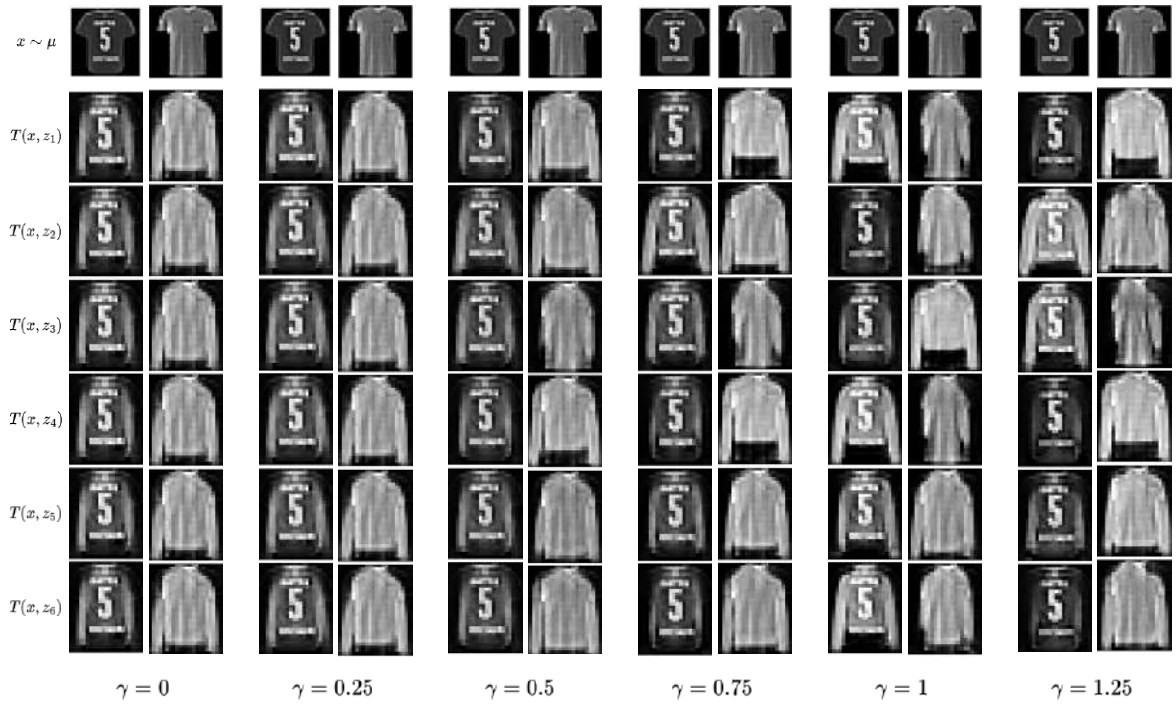

Figure 6: Unpaired image translation in the Fashion-MNIST dataset between T-shirt and Coat classes using the γ-weak quadratic cost with different values of γ, showcasing the variance-similarity trade-off in the samples controlled by γ. With $\gamma = 0$ (strong cost), we effectively observe a conditional collapse where the learned map $T_\theta(x, z)$ is independent of z, and no variety can be observed in the samples. Progressively increasing γ introduces variety into the samples while maintaining a similarity with $x \sim \mu$. However, for $\gamma = 1, 1.25$, we observe mapped inputs that differ more significantly from the source. For instance, the jacket with the 5 on the back notably changes in contrast compared to the original dark t-shirt.

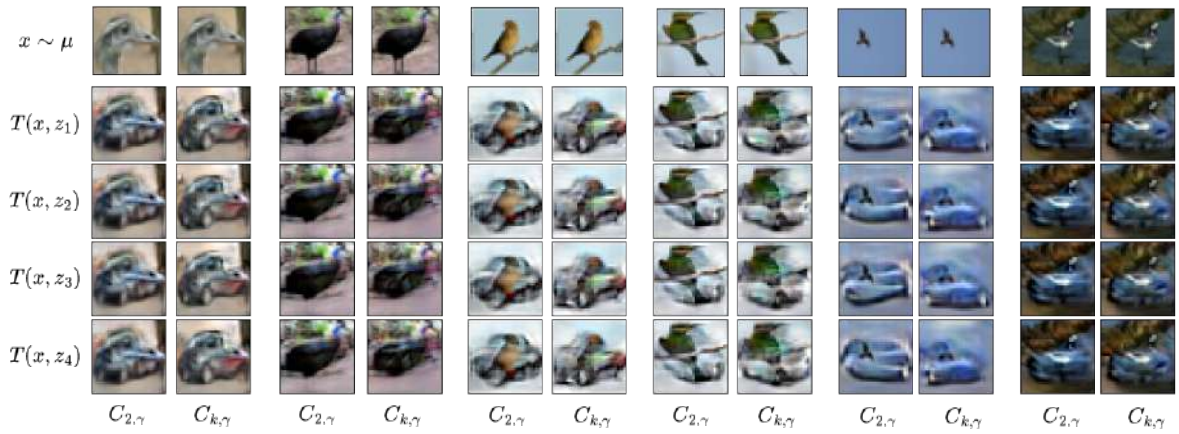

Figure 7: Unpaired image translation in the CIFAR10 dataset between bird and cars classes, with the first column of each example corresponding to the γ-weak quadratic cost $C_{2,\gamma}$, and the second column corresponding to the γ-weak kernel cost $C_{k,\gamma}$ (both using $\gamma = 0.5$). It is important to note that most results are failed translations, thus these examples are cherry-picked. A failure case is often observed when the background behind the bird is not uniform as shown in the right-most example. This is due to poor optimization conditions with notable instabilities.

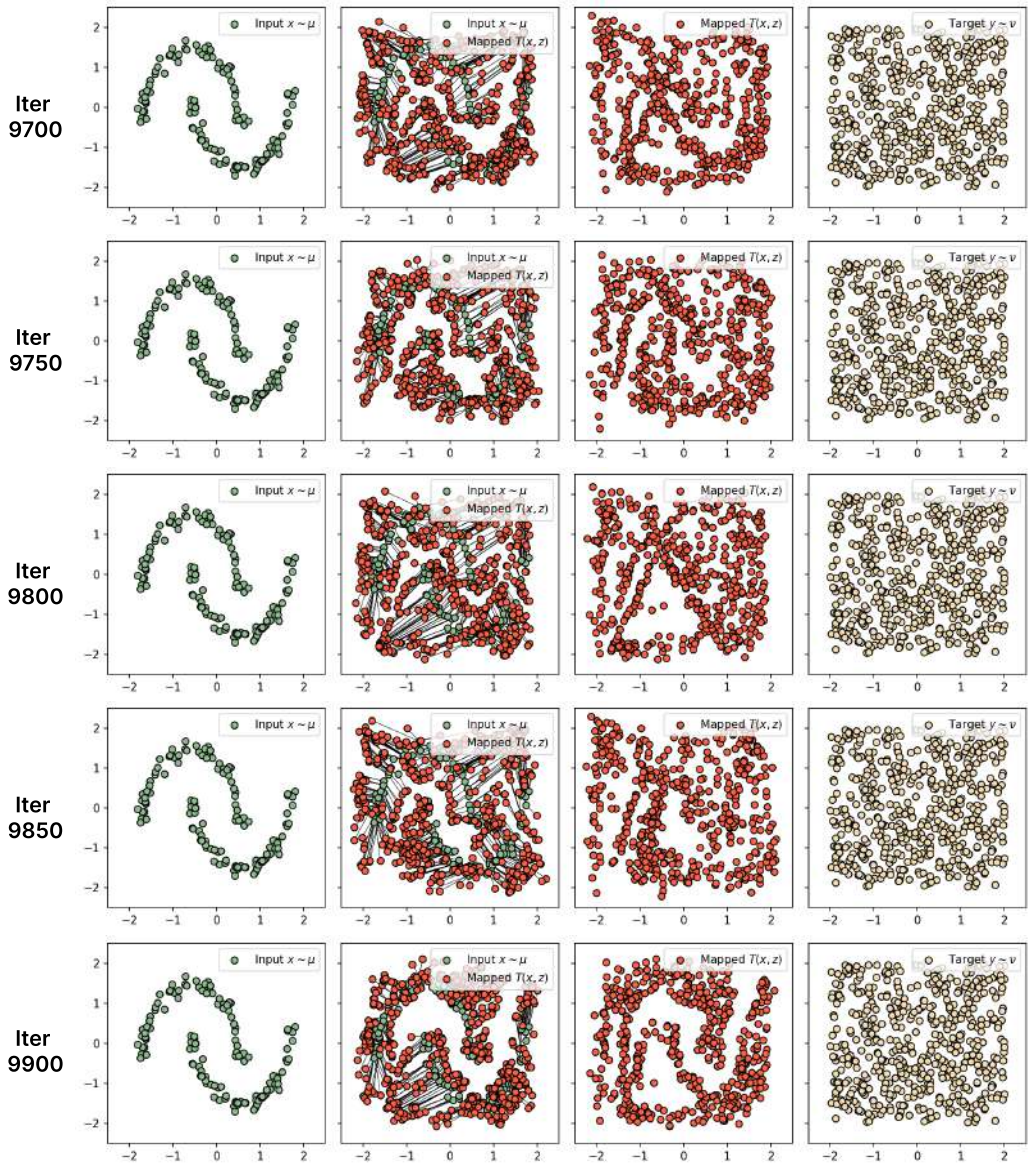

Figure 8: Evolution of the optimization at different steps on 2D toy datasets using the γ-weak quadratic cost $C_{2,\gamma}$ with $\gamma = 1$. One can observe fluctuations between non-optimal mappings, i.e., fake solutions.

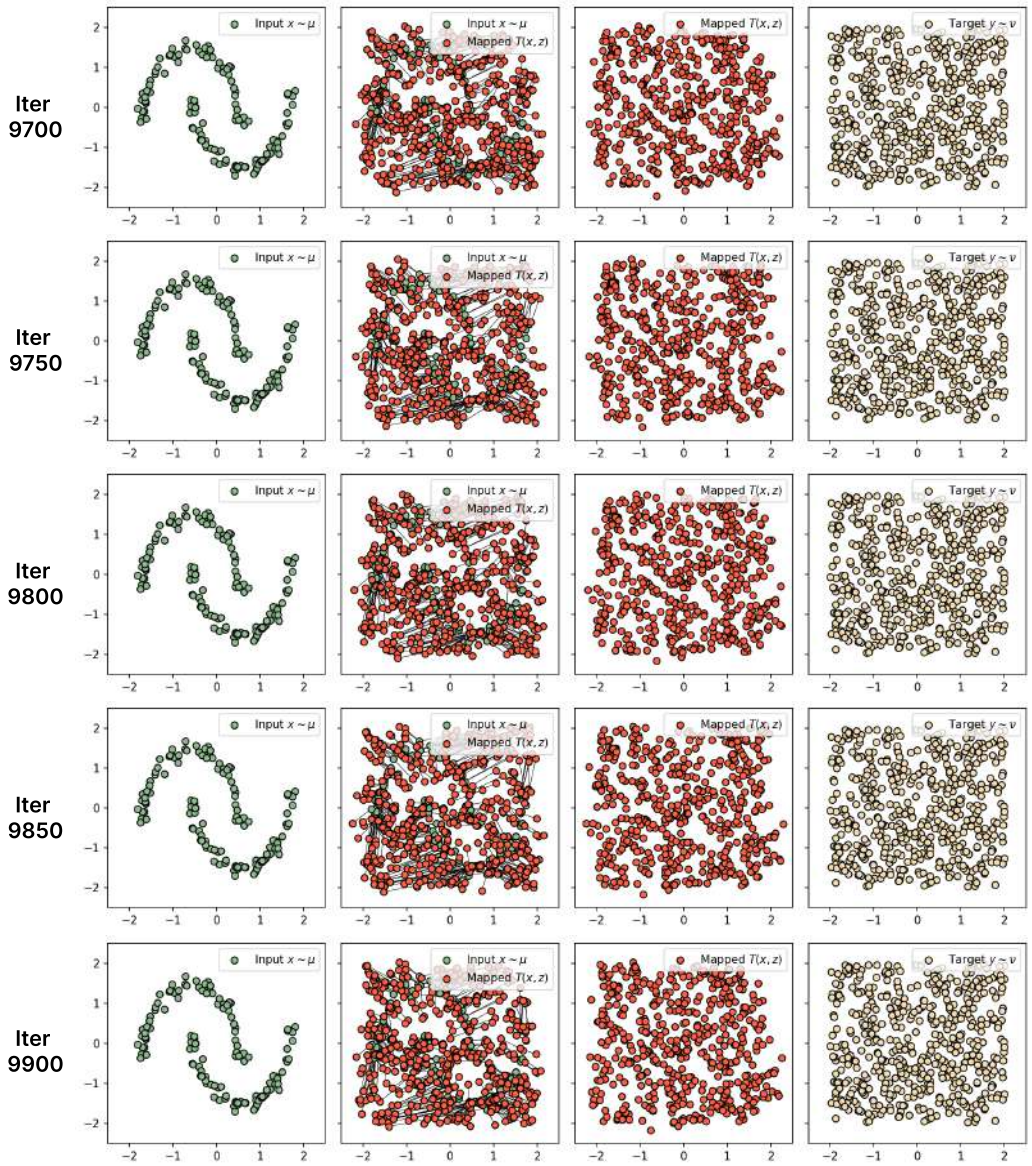

Figure 9: Evolution of the optimization at different steps on 2D toy datasets using the γ-weak kernel cost $C_{k,\gamma}$ with $\gamma = 1$, in the exact same configuration as in Figure 8. The optimization is no more fluctuating between non-optimal mappings but converges to a stable solution.

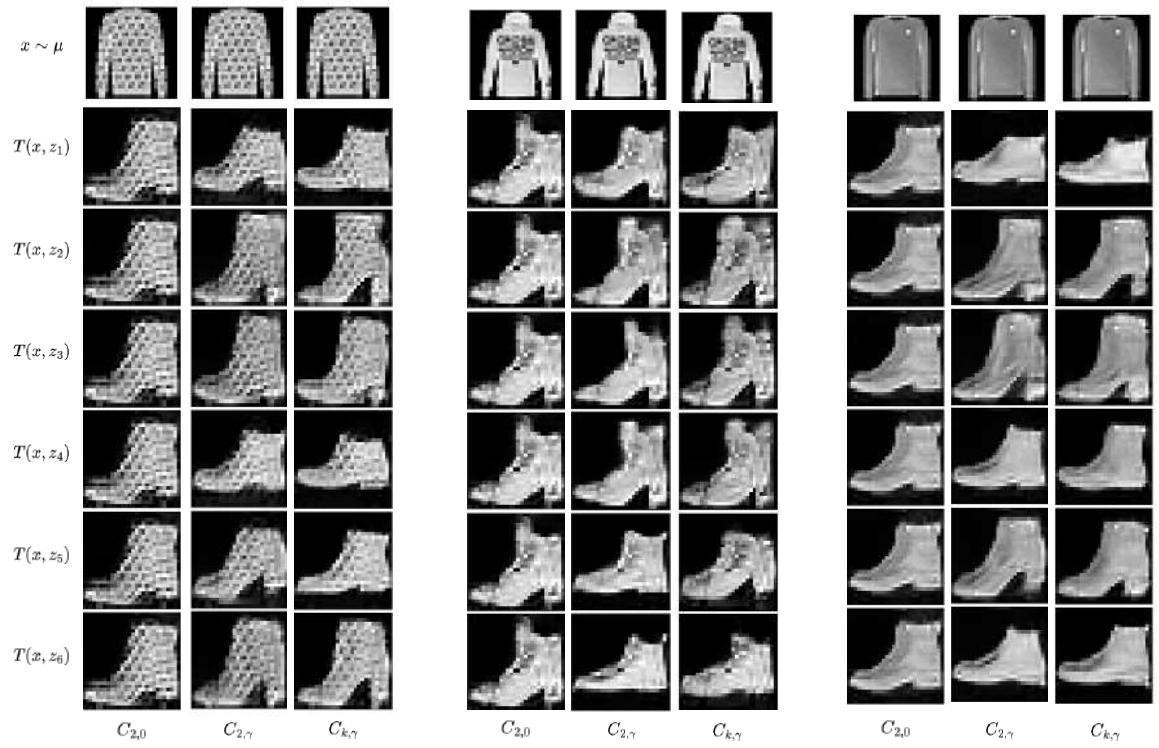

Figure 10: Unpaired image translation in the Fashion-MNIST dataset between Pullover and Ankle boot classes with three different costs: strong quadratic cost $C_{2,0}$, γ-weak quadratic cost $C_{2,\gamma}$ and γ-weak kernel cost $C_{k,\gamma}$.

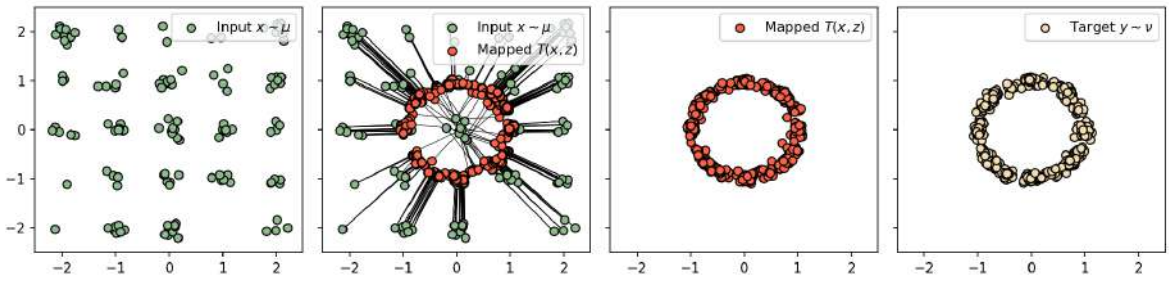

Figure 11: Learned map $T_\theta(x, z)$ between the Gaussian mixture model to its projection on the unit circle.

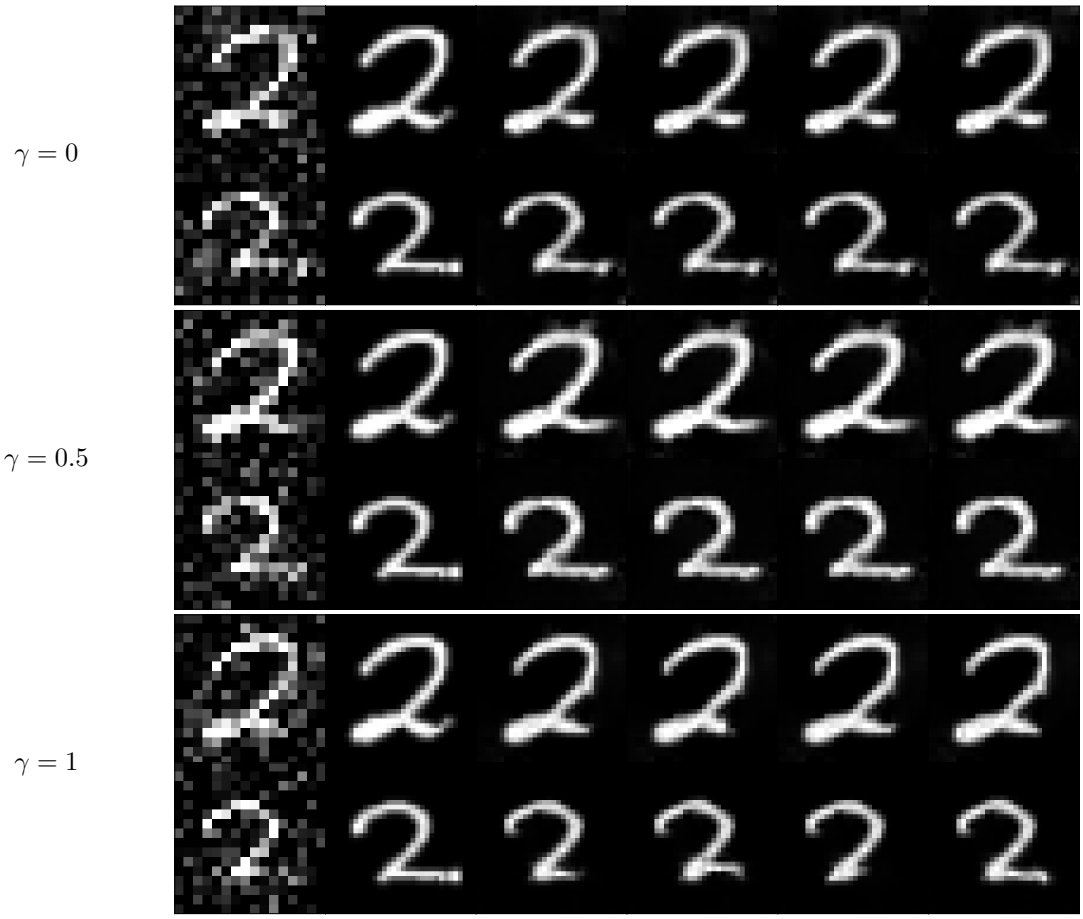

Figure 12: Reconstruction of images from the MNIST-2 dataset using the γ-**weak quadratic cost** for $\gamma = 0, 0.5, 1$ (top-to-bottom)

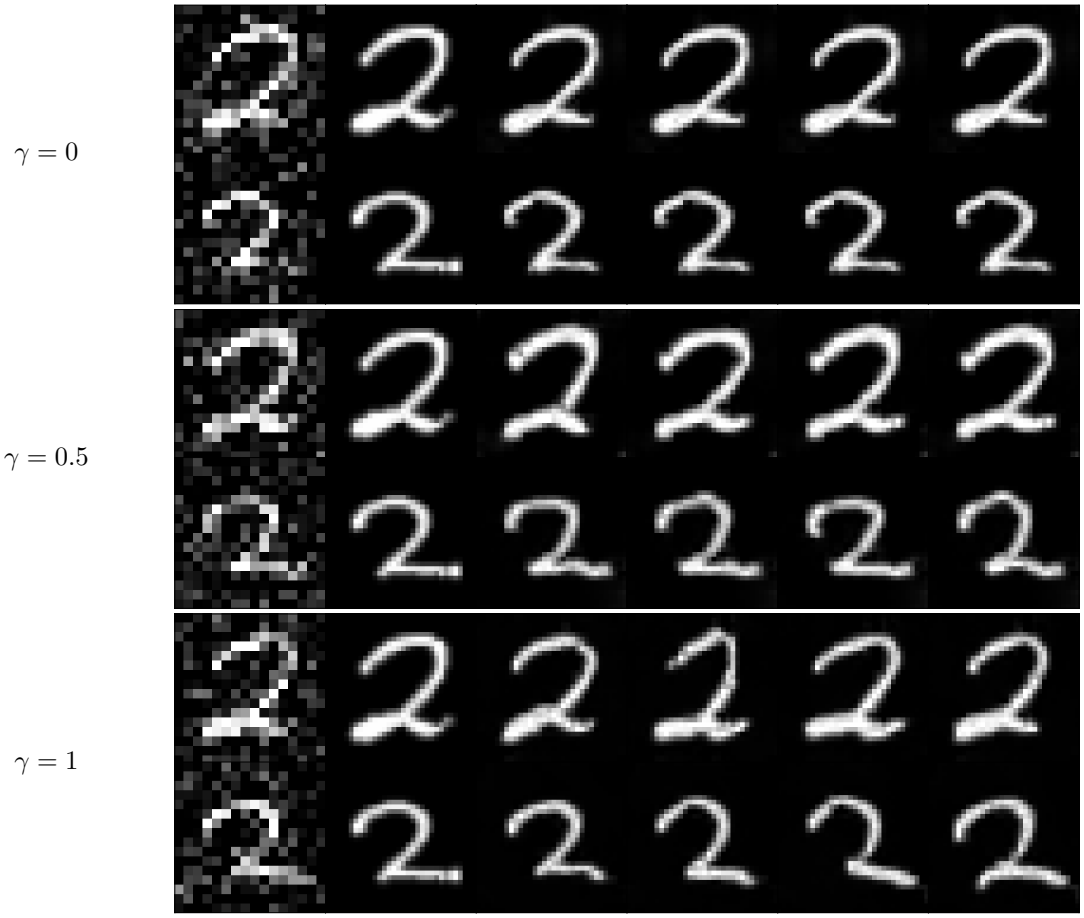

Figure 13: Reconstruction of images from the MNIST-2 dataset using the γ**-weak kernel cost** for $\gamma = 0, 0.5, 1$ (top-to-bottom). Notive the increase in variety while keeping features that are coherent with the downsampled image.