Report: Neural Optimal Transport

Mathis Hardion
MVA, Télécom Paris
mathis.hardion@telecom-paris.fr

Antoine Olivier
MVA, École des ponts ParisTech
antoine.olivier@eleves.enpc.fr

Abstract

Optimal transport (OT) has become a cornerstone of machine learning, and is widely used in generative modeling. Neural networks are an increasingly popular way of scaling OT to large and high-dimensional data. The paper studied in this report introduces a novel way to do so in the context of weak OT, by reformulating the dual problem appropriately. This report discusses the interest of (weak) OT in generative modeling, the advantages and disadvantages of the proposed method, and provides avenues for further reasearch. Experiments are performed on synthetic 2D datasets and on MNIST/KMNIST. A link between weak OT and entropic OT is highlighted and a possible adaption of the paper's method to that case is proposed. A theoretical result showing that appropriate weak costs tend to Monge's OT is shown, an interpretation of weak OT as a "stochastic Monge" problem is provided and a further generalization as "stochastic Kantorovitch" and its dual are derived.

I. Introduction

This report discusses the main contributions of [18], their strengths, shortcomings and possible extensions. Optimal Transport (OT) has recently gained a lot of traction in computational mathematics [28, 25], as it provides a distance between probability measures enjoying many desirable properties and notably faithfulness to the ground metric. In machine learning and more specifically generative modeling, the majority of methods use estimates of the Wasserstein distance as loss functions to train models, with [3, 29] introducing the use of neural networks (NNs) to do so on large-scale data, and many works extending on the idea ([13, 21, 9, 11, 24, 10], etc.). More recently, multiple articles [8, 27, 5, 22] investigated the use of the optimal transport plan itself as a generator, with promising results. However, those have key limitations: [27, 5] only apply to the quadratic cost, and assume the existence of a Monge/Brenier map, [8] requires high dimensional sampling, and [22] struggles to scale to large/complex datasets. Moreover, methods using NNs to compute transport maps can be complex to train, see [19] for a review. Therefore, the studied paper [18] aims at generalizing the approach to weak OT (WOT) costs [12] and obtain a scalable procedure.

The resulting approach, while not a WGAN, still ties with the methodology seen in session 2 of the course: it uses the dual formulation of (weak) OT, C-transform and learns both the dual variable (which can be seen as a 'discriminator' but does not need to be 1-Lipschitz like in WGAN) and the OT plan, which is used as conditional 'generator' but solves OT which is not the case of the WGAN. It also uses Stochastic Gradient Ascent-Descent (SGAD), but WGAN solves an $\inf \sup$ problem where [18] solves a $\sup \inf$ problem instead, which is different in general.

This report is organized as follows: section II introduces OT, WOT and discusses their relevance within the generative modeling context, section III adresses the paper's main contributions under a critical light, section IV challenges the method against both synthetic and real data to assert its strengths and shortcomings, and section V provides extra theoretical results and potential directions of further research.

II. OT and WOT in generative modeling

For a Polish space $\mathcal{X}$, $\mathcal{P}(\mathcal{X})$ denotes the set of probability measures on $\mathcal{X}$. We take $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ Polish spaces and for $\alpha \in \mathcal{P}(\mathcal{X})$, $\beta \in \mathcal{P}(\mathcal{Y})$ we denote the set of couplings (i.e. measures with marginals α, β) by $\Pi(\alpha, \beta) \coloneqq \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \mid \pi(\cdot \times \mathcal{Y}) = \alpha, \pi(\mathcal{X} \times \cdot) = \beta\}$. We may also denote the two marginals of $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ by π_1, π_2 respectively. For a map T, the pushforward operator is denoted by $T_\sharp$. For a cost function $c : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$, we recall Monge's OT formulation [23]:

$$\mathbb{M}(\alpha, \beta) \coloneqq \inf_{T_\sharp \alpha = \beta} \int_{\mathcal{X}} c(x, T(x)) d\alpha(x), \tag{1}$$

which finds an OT map $T : \mathcal{X} \longrightarrow \mathcal{Y}$ which transfers mass with minimal cost. Within our context, we are not interested in the cost by itself, but rather the optimal map. Having access to it can be useful in problems such as unpaired translation/style transfer or inpainting, since it moves its input to a 'close' point in the target distribution's support, which makes it faithful to the original image (see e.g. [27]). However, such a map may not exist, and thus (1) was later relaxed by Kantorovitch [16] to allow splitting of mass and guarantee existence of minimizers (see e.g. [25, Remark 2.13]):

$$\mathbb{K}(\alpha, \beta) \coloneqq \inf_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y). \tag{2}$$

If the optimal plan is of the form $(\mathrm{Id}, T)_\sharp \alpha$, then we recover (1). Even when a Monge map does not exist, one can use the OT plan as substitute: instead of a deterministic map, one can use the stochastic map

$$\tilde{T} : \left| \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathcal{P}(\mathcal{Y}) \\ x & \longmapsto & \pi(\cdot|x) \end{array} \right.$$

where $\pi(\cdot|x) \in \mathcal{P}(\mathcal{Y})$ is the distribution conditional to x. Then, for an input x, one can sample $\tilde{T}(x) \in \mathcal{P}(\mathcal{Y})$, which also allows for more sample diversity than a deterministic map. If needed, one can also compute/estimate $\bar{T} : x \mapsto \int_{\mathcal{Y}} y d\pi(y|x)$ as a deterministic map.

A further generalization of OT was recently proposed in [12], which considers a cost not between coupled points but directly between a point and the distribution of points coupled to it, i.e. for $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$ (named weak cost) the weak OT (WOT) is defined as

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

Note that (2) is recovered when C is the "strong" cost $C(x, \mu) = \int_{\mathcal{Y}} c(x, y) d\mu(y)$, since for $\pi \in \Pi(\alpha, \beta)$, $d\pi(x, y) = d\pi(y|x) d\alpha(x)$. The added flexibility of WOT in the case of interest where $\pi(\cdot|x)$ will be sampled is that one can add regularization terms to enforce desired behaviors for this generator directly through the definition of C. For instance, [18] suggests the γ-weak cost defined as

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

which tends to make the generator have higher variance and thus more sample diversity.

In order to solve OT problems, many methods rely on their dual formulation which often enjoy simpler optimization procedures (we once again refer to [19] for a review). The duality of WOT was studied by [30] which shows that under mild conditions, the dual form of (3) is

$$\mathbb{T}(\alpha, \beta) = \sup_{f \in \mathscr{C}_{b,s}(\mathcal{Y})} \left(\int_{\mathcal{X}} f^C(x) d\alpha(x) + \int_{\mathcal{Y}} f(y) d\beta(y) \right), \tag{5}$$

where $\mathscr{C}_{b,s}(\mathcal{Y})$ denotes the set of real-valued, continuous, upper bounded, not very rapidly growing functions on $\mathcal{Y}$, and the weak C-transform of f is defined as

$$f^C : \left| \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \inf\limits_{\mu \in \mathcal{P}(\mathcal{Y})} C(x, \mu) - \int_{\mathcal{Y}} f(y) d\mu(y). \end{array} \right. \tag{6}$$

Note that with the strong cost, one recovers the Kantorovitch duality [31, Section 5].

III. Contributions of the paper, strengths and shortcomings

The paper's main contributions consists in reformulating the dual WOT (5), and essentially recover some generalization of noise outsourcing [15, Theorem 5.10] allowing for simpler representation of the plan π, enabling derivation of a maximin formulation which is then solved by approximating the variables with NNs, which are shown to be universal approximators for plans. The paper places itself in the case where $\mathcal{X}, \mathcal{Y}$ are subsets of euclidean spaces, and use an atomless distribution ζ on another subset $\mathcal{Z}$ of an euclidean space, which we refer to as outsourcing distribution. With such a distribution, they show the following reformulation of the weak C-transform (6) as a consequence of [28, Cor 1.29]:

$$\forall x, f^C(x) = \inf_{t:\mathcal{Z}\to\mathcal{Y}} C(x, t_\sharp\zeta) - \int_{\mathcal{Z}} f(t(z))d\zeta(z). \tag{7}$$

This allows the authors to derive the following reformulation of dual WOT:

$$\mathbb{T}(\alpha,\beta) = \sup_{f\in\mathscr{C}_{b,s}(\mathcal{Y})} \left(\int_{\mathcal{Y}} f(y)d\beta(y) + \inf_{T:\mathcal{X}\times\mathcal{Z}\to\mathcal{Y}} \int_{\mathcal{X}} \left(C\left(x, T(x,\cdot)_\sharp\zeta\right) - \int_{\mathcal{Z}} f(T(x,z))d\zeta(z) \right) d\alpha(x) \right). \tag{8}$$

The intuition is that $T(x,\cdot)_\sharp\zeta$ corresponds to the conditional $\pi(\cdot|x)$, and thus we recover noise outsourcing in the sense that the map T implicitly represents π. And indeed, [18] shows that when T realizes an OT plan, it minimizes the inner term in (8). The converse is also shown to be true when the weak cost is strictly convex in its second argument.

This formulation of WOT has major practical advantages: first, the outsourcing distribution can be any atomless distribution over a euclidean space of any dimension, meaning one can choose it to be easy to sample and tune its dimensionality based on the complexity of the task at hand. Second, the representation of π through a map allows one to directly parametrize and optimize it, i.e. the loss directly relates to the plan, rather than indirectly through the dual variables in the case of entropic OT [8]. However, the formulation is a maximin problem, which is proposed to be solved with a Stochastic Gradient Ascent Descent (SGAD), an algorithm which can be very slow in practice (cf. section IV), and for which convergence rates are not so straightforward to obtain (see e.g. [4] for recent results). Additionally, the main criticism of entropic OT methods presented by the paper is that they recover a biased plan, however the use of the γ-weak cost also adds a bias by enforcing higher variance.

Nevertheless, this straightforward objective yields favorable results for one-to-many translation by necessitating only a single hyperparameter γ, compared to other approaches such as Aug-CycleGAN [1] and M-UNIT [14], which typically require multiple hyperparameters.

IV. Experiments

IV.1 Toy 2D multimodal problem

We first study the method on 2D ($\mathcal{X} = \mathcal{Y} = \mathbb{R}^2$) synthetic data, where the input distribution is discrete i.e. $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, and the target is made of uniform circles i.e. $\frac{1}{m}\sum_{j=1}^m (y_j + r(\cos(2\pi\cdot), \sin(2\pi\cdot)))_\sharp \mathcal{U}[0,1]$. The cost used is the square distance to the mean, i.e. $C(x,\mu) \coloneqq \left\|x - \int_{\mathcal{X}} y d\mu(y)\right\|^2$. We thus expect to recover the centers of the different circles. The results are shown figure 1, where we see that the methods does not recover all modes, despite long training. It appears the method may struggle in multimodal cases, as is often the case for sampling algorithms. Code was adapted from the one provided by [18].

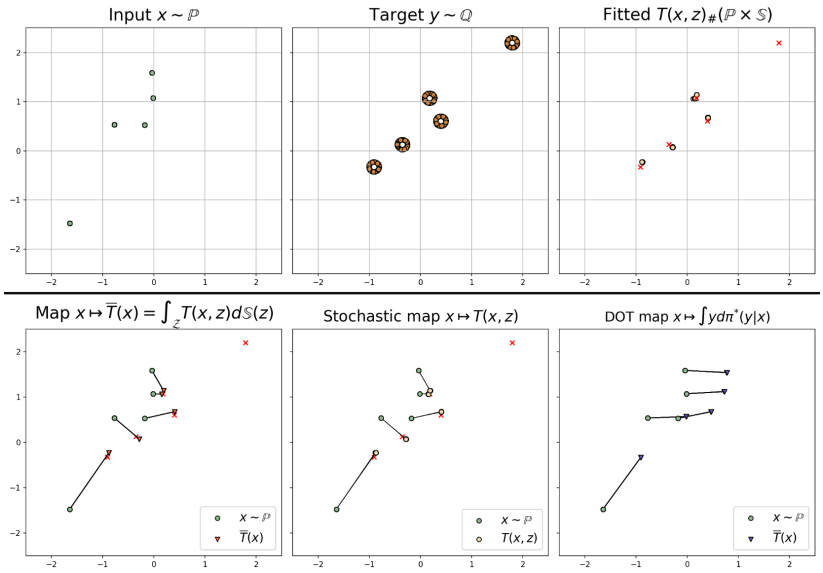
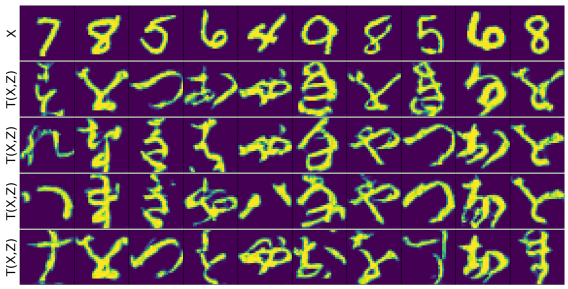


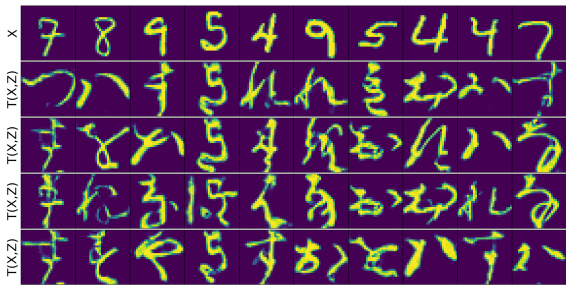
Figure 1: Toy 2D example with discrete input and multimodal target. Note how the algorithm fails to recover the upper right mode.

IV.2 Optimal Transport between MNIST and KMNIST

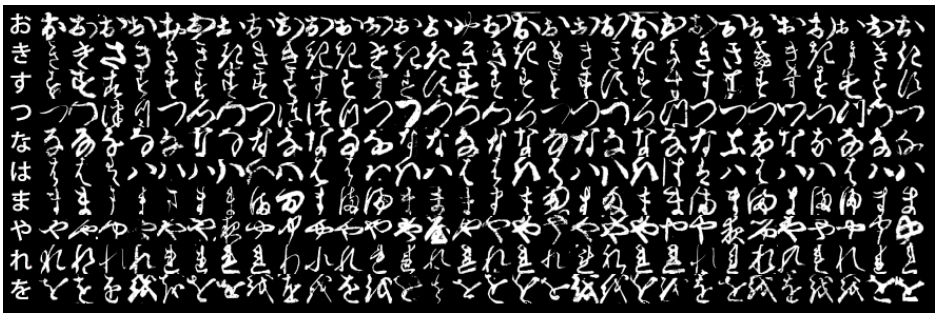
In our next experiment, we decided to compute optimal transport between the datasets MNIST [20] and KMNIST (Kuzushiji-MNIST) [6], using the methodology of the paper. To accomplish this task, The stochastic map T was approximated using a U-Net neural network architecture and the potential f was approximated using a convolutional neural network. The training process was a modified version of the training setting proposed by the paper, aiming to achieve faster results on what we assume to be a simpler task. More details can be found in appendix C.



(a) Random Images



(b) Random Test Images



(c) Sample from the KMNIST Dataset

Figure 2: Result of the experiment: In Figure 2a and 2b, the first row corresponds to the input x given to the model T, and the next four rows correspond to the output for different samples z.

Throughout the training process, we observed distinct phases of development in the model T. Initially, our model was independent from z (non-stochastic), likely influenced by the *dynamic weak cost*

(appendix C) suggested by the paper. The model also appeared to approximate the identity function. As training progressed, the dependence on z began to manifest, and T gradually diverged from the identity. By the end of training, it became capable of generating realistic-looking hiragana characters from a MNIST image. However, it is worth noting that the model also frequently generated images with a hiragana-like structure that did not correspond to any real hiragana, which might be attributed to insufficient training.

V. Further research

V.1 Entropic Optimal Transport as WOT

Entropic Optimal Transport (EOT) has become a popular alternative to OT due to lower computational costs thanks to Sinkhorn's algorithm [7]. It consists in regularizing the OT cost with an entropic term, i.e. it is defined as

$$\mathbb{S}_\varepsilon(\alpha,\beta) = \inf_{\pi\in\Pi(\alpha,\beta)} \left(\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y) + \varepsilon \mathrm{KL}\left(\pi||\alpha\otimes\beta\right)\right), \tag{9}$$

where $\mathrm{KL}\left(\pi||\rho\right) \coloneqq \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\rho}(x,y)\right) d\pi(x,y)$ if π is absolutely continuous w.r.t. ρ (denoted $\pi \ll \rho$) and $+\infty$ otherwise. We now show that WOT can also recover EOT when we fix β, for the cost

$$C_\varepsilon(x,\mu) \coloneqq \int_{\mathcal{Y}} c(x,y)d\mu(y) + \varepsilon \mathrm{KL}\left(\mu||\beta\right). \tag{10}$$

Indeed, since any $\pi \in \Pi(\alpha,\beta)$ such that the cost in (9) 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

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

From a practical perspective, the consideration of EOT as WOT with the cost C_ε will not yield any computational advantages since the first term in (10) is estimated using Monte Carlo just as in [18], and the second term may be more complex to estimate, requiring methods such as k-nearest neigbors density estimation [32]. However, the regularization term may help enforcing that generated samples from $\pi(\cdot|x)$ appear 'likely' to be in the target distribution, since it tends to reduce $\mathrm{KL}\left(\pi(\cdot|x)||\beta\right)$. An extra challenge is that one would need to enforce the support of $\pi(\cdot|x)$ to be close to that of β at the start, so that the estimated KL does not diverge and provide unstable behaviors. This could be achieved by initially considering the strong cost by itself and adding the regularization after enough training steps.

V.2 Approximate Monge with WOT

The γ-weak cost (4) proposed by [18] allows to increase the generator's variance and thus sample diversity as discussed previously. Now consider the opposite: by switching the sign in front of γ, one instead enforces lower variance of the generator, and intuitively for large γ, on average over x, $\pi(\cdot|x)$ will become almost deterministic i.e. close to some dirac mass $\delta_{T(x)}$. Thus, we could expect to recover some approximation of a Monge map. More formally, we at least have the following result.

Proposition 1. *For some nonnegative cost c on $\mathcal{X}\times\mathcal{Y}$, define*

$$\tilde{C}_\gamma : x,\mu \mapsto \int_{\mathcal{Y}} c(x,y)d\mu(y) + \gamma \mathrm{Var}\left(\mu\right). \tag{11}$$

Then, it holds that

$$\inf_{\pi\in\Pi(\alpha,\beta)} \lim_{\gamma\to+\infty} \int_{\mathcal{X}} \tilde{C}_\gamma(x,\pi(\cdot|x))d\alpha(x) = \inf_{T_\sharp\alpha=\beta} \int_{\mathcal{X}} c(x,T(x))d\alpha(x). \tag{12}$$

Note that the limit in (12) is well defined since the considered quantity is nondecreasing in γ, and also that the infimum on the right hand side may be $+\infty$ i.e. no assumption is made about the existence of a Monge map. The proof is provided in appendix A. The main limitations of this result are that first, we would prefer the inf and lim to be swapped, and second, in our case of interest we would additionally like convergence of the OT plan to some $(\mathrm{Id}, T^*)_\sharp \alpha$ in the case where an OT map T^* exists. That way, one would have guarantees about the optimization procedure proposed in [18]. Further research could therefore try to extend this result and obtain these properties. The first one could most likely be achieved by ensuring the total cost is lower semi-continuous with respect to π in an appropriate sense, while the second one may prove more challenging or potentially impossible since the cost $\tilde{C}_\gamma$ is now concave due to the concavity of the variance operator, meaning a minimizer would not be unique. We provide a toy illustration in appendix B, where convergence is nevertheless still observed, and the resulting map is close to deterministic as expected.

V.3 WOT as "stochastic Monge"

By comparing (1) and (3), one can see they are quite similar: the deterministic map $T : \mathcal{X} \to \mathcal{Y}$ is replaced with the stochastic map $\mathscr{T} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ with $\mathscr{T} : x \mapsto \mathscr{T}_x \coloneqq \pi(\cdot|x)$. The constraint $T_\sharp \alpha = \beta$ is replaced by $\pi \in \Pi(\alpha, \beta)$, which can be reformulated as follows:

$$\begin{aligned}
\begin{cases} \pi_1 = \alpha \\ \pi_2 = \beta \end{cases} &\iff \begin{cases} \pi_1 = \alpha \\ \forall A, \int_{\mathcal{X}} \pi(A|x) d\pi_1(x) = \beta(A) \end{cases} \\
&\iff \begin{cases} \pi_1 = \alpha \\ \int_{\mathcal{X}} \mathscr{T}_x(A) d\alpha(x) = \beta(A) \end{cases} \\
&\iff \begin{cases} \pi_1 = \alpha \\ \int_{\mathcal{P}(\mathcal{Y})} \mu(A) d(\mathscr{T}_\sharp \alpha)(\mu) = \beta(A) \end{cases}
\end{aligned}$$

Where $\mathscr{T}_\sharp \alpha \in \mathcal{P}(\mathcal{P}(\mathcal{Y}))$ (and $\mathcal{P}(\mathcal{Y})$ can be made into a Polish space, see [2, Remark 7.1.7]). Thus, defining for $\Phi \in \mathcal{P}(\mathcal{P}(\mathcal{Y}))$ its expectation $\mathbb{E}[\Phi] \in \mathcal{P}(\mathcal{Y})$ as the measure such that

$$\forall A, \mathbb{E}[\Phi](A) \coloneqq \int_{\mathcal{P}(\mathcal{Y})} \mu(A) d\Phi(\mu), \tag{13}$$

the constraint reads

$$\begin{cases} \pi_1 = \alpha \\ \pi_2 = \beta \end{cases} \iff \begin{cases} \pi_1 = \alpha \\ \mathbb{E}[\mathscr{T}_\sharp \alpha] = \beta \end{cases} \tag{14}$$

Therefore, every measurable stochastic map $\mathscr{T} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ verifying $\mathbb{E}[\mathscr{T}_\sharp \alpha] = \beta$ is uniquely linked with a coupling $\pi \in \Pi(\alpha, \beta)$ defined by its first marginal $\pi_1 \coloneqq \alpha$ and its conditional distribution $\pi(\cdot|x) \coloneqq \mathscr{T}_x$. We have thus shown that WOT can be reformulated as what we refer to as "stochastic Monge problem":

Proposition 2. *The following equality holds:*

$$\mathbb{T}(\alpha, \beta) = \inf_{\mathbb{E}[\mathscr{T}_\sharp \alpha] = \beta} \int_{\mathcal{X}} C(x, \mathscr{T}_x) d\alpha(x). \tag{15}$$

where the infimum is taken over measurable $\mathscr{T} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ verifying $\mathbb{E}[\mathscr{T}_\sharp \alpha] = \beta$.

With this result, the next obvious step is to apply the Kantorovitch relaxation to (15) considering it may be nonconvex for general costs (like (10) or (11)). One can first identify $x \in \mathcal{X}$ with $\delta_x \in \mathcal{P}(\mathcal{X})$, which can be done by replacing $C(x, \nu)$ with $\tilde{C}(\delta_x, \nu) \coloneqq \int_{\mathcal{X}} C(\tilde{x}, \nu) d\delta_x(\tilde{x})$, extended naturally as $\tilde{C}(\mu, \nu) = \int_{\mathcal{X}} C(x, \nu) d\mu(x)$, but we next consider a general cost $C : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$. Then, we would obtain the relaxation which we refer to as "stochastic Kantorovitch problem":

Definition 1. *The stochastic Kantorovitch cost between two probability measures α, β is defined as*

$$\mathbb{SK}(\alpha, \beta) \coloneqq \inf_{\substack{\mathbb{E}[\Gamma_1] = \alpha \\ \mathbb{E}[\Gamma_2] = \beta}} \int_{\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})} C(\mu, \nu) d\Gamma(\mu, \nu) \tag{16}$$

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.

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 (15). Additionally, as in (2), the fact that Γ is a probability measure is implied by the marginal constraints: if $\mathbb{E}[\Gamma_1] = \alpha$, one has

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

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 marginal constraints. This allows one to define duality the same as for (2): denoting for a function f and measure μ $\langle f, \mu \rangle \coloneqq \int f d\mu$ for notational convenience, and 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 (13) is equivalent to

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

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

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

Where we denote $\langle f, \cdot \rangle \oplus \langle g, \cdot \rangle : (\mu, \nu) \mapsto \langle f, \mu \rangle + \langle g, \nu \rangle$. Finally, the dual formulation of (16) reads as follows.

Proposition 3. *Assuming duality holds, the dual Stochastic Kantorovitch reads*

$$\mathbb{SK}(\alpha, \beta) = \inf_{\langle f, \cdot \rangle \oplus \langle g, \cdot \rangle \leq C} \langle f, \alpha \rangle + \langle g, \beta \rangle . \tag{17}$$

Further research would study under which circumstances duality holds, and possibly find a reformulation using a generalized C-transform of sorts, although that may prove challenging due to the form of the constraints. If such a formulation is possible, it may be feasible to extend [18]'s method to the stochastic Kantorovitch cost, which could allow computation of more general transport plans.

References

[1] A. 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 J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 195–204. URL: `https://proceedings.mlr.press/v80/almahairi18a.html`.

[2] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Birkhäuser, 2008.

[3] M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 214–223. URL: https://proceedings.mlr.press/v70/arjovsky17a.html.

[4] A. 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 F. Ruiz, J. Dy, and J.-W. van de Meent. Vol. 206. Proceedings of Machine Learning Research. PMLR, Apr. 2023, pp. 172–235. URL: https://proceedings.mlr.press/v206/beznosikov23a.html.

[5] S. Chaudhari, S. Pranav, and J. M. 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. DOI: 10.1109/ICASSP49357.2023.10097266.

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

[7] M. Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *Advances in Neural Information Processing Systems*. Ed. by C. Burges et al. Vol. 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

[8] M. Daniels, T. Maunu, and P. 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. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/6c2e49911b68d315555d5b3eb0dd45bf-Paper.pdf.

[9] I. Deshpande, Z. Zhang, and A. 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] Y. Dukler et al. "Wasserstein of Wasserstein Loss for Learning Generative Models". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, June 2019, pp. 1716–1725. URL: https://proceedings.mlr.press/v97/dukler19a.html.

[11] A. Genevay, G. Peyré, and M. Cuturi. "Learning Generative Models with Sinkhorn Divergences". In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. Ed. by A. Storkey and F. Perez-Cruz. Vol. 84. Proceedings of Machine Learning Research. PMLR, Apr. 2018, pp. 1608–1617. URL: https://proceedings.mlr.press/v84/genevay18a.html.

[12] N. Gozlan et al. "Kantorovich duality for general transport costs and applications". In: *Journal of Functional Analysis* 273.11 (2017), pp. 3327–3405. ISSN: 0022-1236. DOI: https://doi.org/10.1016/j.jfa.2017.08.015. URL: https://www.sciencedirect.com/science/article/pii/S0022123617303294.

[13] I. 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. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf.

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

[15] O. Kallenberg. *Foundations of Modern Probability*. 1st ed. Probability and Its Applications. Springer New York, NY, 1997, pp. XII, 523. DOI: 10.1007/b98838.

[16] L. Kantorovitch. "On the Translocation of Masses". In: *Management Science* 5.1 (1958), pp. 1–4. ISSN: 00251909, 15265501. URL: http://www.jstor.org/stable/2626967 (visited on 03/15/2024).

[17] D. P. Kingma and J. Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[18] A. Korotin, D. Selikhanovych, and E. Burnaev. "Neural Optimal Transport". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=d8CBRlWNkqH.

[19] A. 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. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/7a6a6127ff85640ec69691fb0f7cb1a2-Paper.pdf.

[20] Y. LeCun and C. Cortes. "MNIST handwritten digit database". In: (2010). URL: http://yann.lecun.com/exdb/mnist/.

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

[22] A. Makkuva et al. "Optimal transport mapping via input convex neural networks". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 6672–6681. URL: https://proceedings.mlr.press/v119/makkuva20a.html.

[23] G. Monge. *Mémoire sur la théorie des déblais et des remblais*. Imprimerie royale, 1781. URL: https://books.google.fr/books?id=IG7CGwAACAAJ.

[24] H. Petzka, A. Fischer, and D. 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. URL: https://openreview.net/forum?id=B1hYRMbCW.

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

[26] O. Ronneberger, P. Fischer, and T. Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

[27] L. Rout, A. Korotin, and E. Burnaev. "Generative Modeling with Optimal Transport Maps". In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=5JdLZg346Lw.

[28] F. 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. DOI: 10.1007/978-3-319-20828-2.

[29] V. 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. URL: https://openreview.net/forum?id=B1zlp1bRW.

[30] J. B. Veraguas, M. Beiglboeck, and G. Pammer. "Existence, duality, and cyclical monotonicity for weak transport costs". In: *Calculus of Variations and Partial Differential Equations* 58 (Nov. 2019). DOI: 10.1007/s00526-019-1624-y.

[31] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN: 9783540710509. URL: https://books.google.fr/books?id=hV8o5R7_5tkC.

[32] Q. Wang, S. R. Kulkarni, and S. Verdu. "Divergence Estimation for Multidimensional Densities Via k-Nearest-Neighbor Distances". In: *IEEE Transactions on Information Theory* 55.5 (2009), pp. 2392–2405. DOI: 10.1109/TIT.2009.2016060.

Appendix

A. Proof of proposition 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

$$\int_{\mathcal{X}} \tilde{C}_\gamma(x, \pi(\cdot|x)) \xrightarrow[\gamma \to +\infty]{} \begin{cases} \int_{\mathcal{X}} c(x, T(x)) d\alpha(x) & \text{if } \pi(\cdot|x) = \delta_{T(x)} \ \alpha\text{-a.e. for some } T : \mathcal{X} \longrightarrow \mathcal{Y} \\ +\infty & \text{otherwise.} \end{cases} \tag{18}$$

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

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

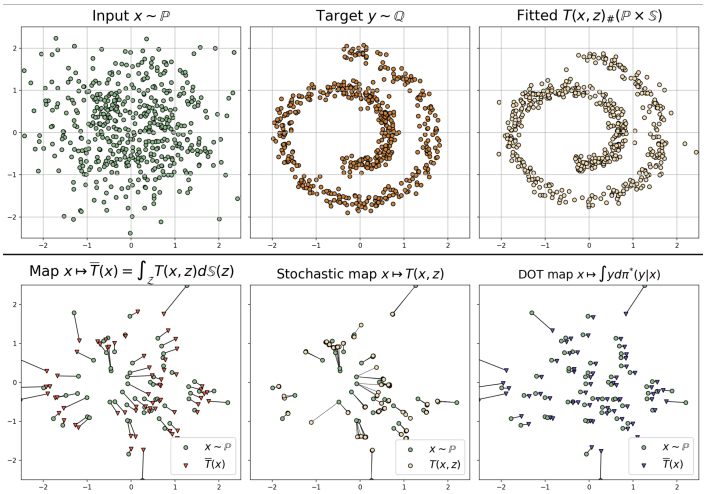
i.e. $T_\sharp \alpha = \beta$. Conversely, if there is some measurable T such that $T_\sharp \alpha = \beta$, the coupling π defined by its first marginal $\pi_1 = \alpha$ and conditional $\pi(\cdot|x) := \delta_{T(x)}$ does indeed verify $\pi \in \Pi(\alpha, \beta)$ using the

same computation: for any continuous bounded f,

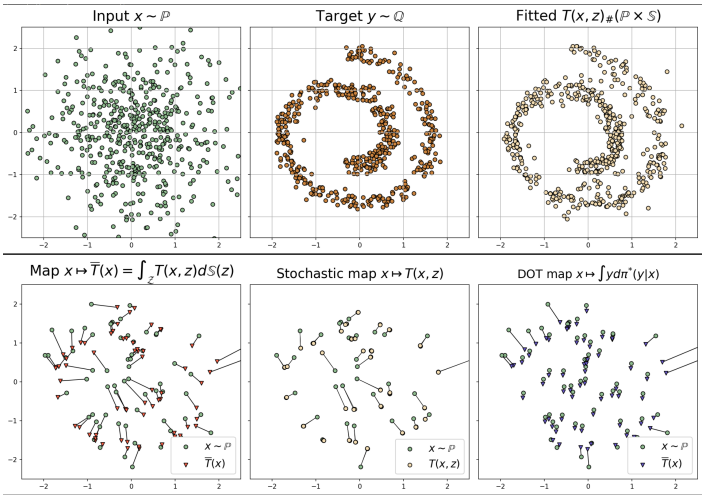
$$\begin{aligned}\int_{\mathcal{Y}} f(y)d\pi_2(y) &= \int_{\mathcal{X}\times\mathcal{Y}} f(y)d\pi(x,y)\\ &= \int_{\mathcal{X}}\int_{\mathcal{Y}} f(y)d\pi(y|x)d\alpha(x)\\ &= \int_{\mathcal{X}} f(T(x))d\alpha(x)\\ &= \int_{\mathcal{Y}} f(y)d\beta(y).\end{aligned}$$

As a result, the (possibly empty) set of πs for which the limiting cost in (18) is finite is exactly described by the set of measurable maps T verifying $T_\sharp\alpha = \beta$, hence proving (12) i.e. propostion 1.

B. Toy illustration of proposition 1



(a) $C_\gamma, \gamma = 1$



(b) $\tilde{C}_\gamma, \gamma = 10$

Figure 3: Toy experiment illustrating the difference between the γ-weak cost enforcing higher variance 3a and our alternative definition penalizing it instead 3b. One can see that as expected, we recover a basically deterministic map in the latter case.

C. Training Details for experiment IV.2

Pre-processing. We rescale the images to have values between $[-1, 1]$.

Neural networks. We employ a straightforward CNN architecture for the potential function f, the details of which are available in the code provided. We use UNet [26] as the stochastic transport map $T(x, z)$. The noise z is simply an additional input channel , i.e., the dimension of the noise equals the image size (28×28). We use high-dimensional Gaussian noise with axis-wise $\sigma = 0.1$.

Optimization. We utilize the Adam optimizer [17] with default beta values for both T_θ and f_ω. The learning rate is set to $lr = 10^{-4}$, and the batch size is $|X| = 64$. We sample $|Z_x| = 4$ noise samples per each image x in batch. We conduct $k_T = 10$ inner iterations (iterations of the T optimization step per optimization step of f). The model is trained for approximately 10,000 epochs.

Dynamic weak cost We train the algorithm with the gradually changing γ of the γ-weak cost. Starting from $\gamma = 0$, we linearly increase it to the desired value $\frac{2}{3}$ during 3K first iterations of f_ω.