W2GAN: RECOVERING AN OPTIMAL TRANSPORT MAP WITH A GAN

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding and improving Generative Adversarial Networks (GAN) using notions from Optimal Transportation (OT) theory has been a successful area of study, originally established by the introduction of the Wasserstein GAN (WGAN). An increasing number of GANs incorporate OT for improving their discriminators, but that is so far the sole way for the two domains to cross-fertilize. We consolidate the bridge between GANs and OT with one model: W2GAN, where the discriminator approximates the second Wasserstein distance. This model exhibits a twofold connection: the discriminator implicitly computes an optimal map and the generator follows an optimal transport map during training. Perhaps surprisingly, we also provide empirical evidence that other GANs also approximately following the Optimal Transport.

1 INTRODUCTION

In Generative Adversarial Networks (GAN) (Goodfellow et al., 2014), the role of the discriminator $\mathbf{D}$ is to find a way to distinguish two random variables: a real one $\mathbf{X}$ and a fake one $\mathbf{G}(\mathcal{Z})$, where $\mathbf{G}$ is a generator and $\mathcal{Z}$ a latent space whose distribution is known. It does so by approximating a quantity that depends on the two distributions, $\mathcal{L}(\mathbf{G}(\mathcal{Z}), \mathbf{X})$, for instance f-divergences (Mao et al., 2016; Nowozin et al., 2016). The performance of the discriminator hence critically relies on the properties of $\mathcal{L}$. For instance, the first version of GAN can be summarized by the minimax game:

$$\min_{\mathbf{G}} \max_{\mathbf{D}} \mathbb{E}_{x\sim\mathbf{P}_{\mathcal{X}}}[\log(\mathbf{D}(x))] + \mathbb{E}_{z\sim\mathbf{P}_{\mathcal{Z}}}[\log(1-\mathbf{D}(\mathbf{G}(z))] \tag{1}$$

When the discriminator is trained to optimality, the generator's objective is equivalent to minimizing the Jensen-Shannon divergence between $\mathbf{G}(\mathcal{Z})$ and $\mathcal{X}$. However, in practice, training the discriminator to the optimum before each generator update often leads to problems due to the Jensen-Shannon divergence's lack of continuity, particularly when the support of the two distributions are disjoint.

To tackle this problem, Arjovsky et al. (2017) proposed using the first Wasserstein distance as an alternative to f-divergence. The smoothness of the resulting discriminator comes directly from the definition of this metric, and in practice induces stability and robustness in the GAN training. In fact, the objective of the discriminator in WGAN and its extensions (Arjovsky et al., 2017; Miyato et al., 2018; Gulrajani et al., 2017b) corresponds to a particular instance of the Kantorovitch problem, which is the main problem at stake in Optimal Transport (OT) theory. The Kantorovitch problem consists of finding an optimal plan for "transporting" one distribution onto another one according to a predefined point-wise cost function. Borrowing these notions from OT theory have proved to be useful for improving both understanding and performance of GANs.

On the other hand, computing an *optimal map*, a map between two distributions with a minimal cost of "transport", constitutes another formidable problem in the OT community (Peyré & Cuturi, 2018). An optimal map has many important implications such as computing barycenters of distributions (Rabin et al., 2012; Bonneel et al., 2014) or realizing domain transfer (Alexandre, 2018). It also allows deep understanding for relating probability measures and their evolution trough gradient flows (Villani, 2008; Ambrosio et al., 2008). Computing them is thus an important challenge as it allows simulations that may help theoretical investigations, or in a more applied perspective to compute statistics over distributions. Recently, regularized versions of the Kantorovitch problem were proposed to provide

improved algorithms for computing the optimal map between (semi-)discrete distributions (Cuturi, 2009; Genevay et al., 2016). This approach has also been recently adopted in the GAN community to improve generative models. Sanjabi et al. (2018) used the optimal transport map as a new way to train a GAN discriminator, and Salimans et al. (2018) proposed replacing the discriminator with the Sinkhorn algorithm, which can compute any Wasserstein metric.

While recent GAN models incorporate notions from OT for defining the discriminator objective, it remains unclear whether it is possible to reverse the contribution, that is having a GAN that would recover an Optimal Transport map. This is motivated by the fact that both generators in GAN and optimal transport map (Seguy et al., 2017) play the role of generative models. In this paper, we construct a GAN generator whose training dynamic takes place on an optimal transport map.

Our motivations for seeking a connection between Generative Adversarial Networks and the Optimal Transport map are twofold. First, as an application of GANs to the estimation of the optimal transport in high-dimenisonal spaces. The well-known success of GAN in generative modelling suggests that such a generator designed might provide a good approximation of the optimal transport map, especially for high dimensional continuous distributions where the problem becomes especially challenging.

Second, we wish to leverage the properties of Optimal Transport theory to improve and extend the GAN-based generative modeling. In particular, there has been consideral recent interest in using GANs to perform unsupervised domain transfer (Zhu et al., 2017; Almahairi et al., 2018). However, to date, these methods generally rely on architectural features and heuristics to ensure meaningful domain transfer. Optimal Transport theory provides a means of formalizing the mapping between the domain distributions and placing this line of inquiry on a more rational foundations. In this work, we seek to take the first steps on this path by establishing a new connection between GANs and the optimal transport map.

To this end, we consider a discriminator based on the second Wasserstein distance. As with the majority of GANs relying on OT metrics, we approximate this distance using its dual formulation (called the dual of the Kantorovitch problem in OT theory). What motivates the use of this metric in particular is that one can recover a unique optimal transport between two distributions by solving the dual of its Kantorovitch problem (Ambrosio & Gigli, 2013). Hence our discriminator might be used either for training a generator in a GAN context, or to compute an optimal transport map directly. As far as we know, it is the first demonstation of a GAN achieving reasonable generative modelling results and an approximation of the optimal transport map between two continuous distributions.

Interestingly, the generator of our GAN is not learning an arbitrary mapping between the latent space and the target distribution. We show that, at convergence, it is reproducing an optimal transport map. We also provide insights on the way the generator evolves during training: it follows the Wasserstein-2 geodesic between its initialization and the target distribution. Finally, although the theoretical arguments do not scale immediately to the case of the other GANs, the experiments suggest that the generator trained using WGAN and its extensions also approximate an optimal transport map.

The paper is organized as follow:

- In Section 2, we provide the needed theoretical background of Optimal Transport theory.
- In section 3, we detail our new model. We first detail how training only the discriminator of W2GAN gives a natural algorithm to compute an optimal transport map. Then we present the GAN setup. Finally, we provide theoretical insight on how the generator in GAN is essentially recovering an optimal transport map.
- In section 4, we provide an empirical evaluation on both synthetic and real data.

2 BACKGROUND

Optimal Transport (OT) theory (Villani, 2008; Santambrogio, 2010; Ambrosio & Gigli, 2013) introduces a natural quantity to distinguish two probability measures. Given two probability measures μ and ν on the euclidean space $\mathbb{R}^d$, the original Monge problem is to find a map T that "transports" the μ distribution on ν, and minimizes the cost of the transport, which is point-wise defined by a

fixed cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. $c(x, y)$ can be seen as the cost for transporting a unit from x to y. The problem is summarized by:

$$\inf_{T \in \mathcal{A}_T} \int_{\mathbb{R}^d} c(x, T(x)) d\mu(x) \tag{2}$$

where $\mathcal{A}_T$ is the set of all maps from $\mathbb{R}^d$ to $\mathbb{R}^d$ that respect the marginal ν, more formally written as $T_{\#}\mu = \nu$. $T_{\#}\mu$ is called the push-forward measure of μ and is defined as $T_{\#}\mu(A) = \mu(T^{-1}(A))$ for any Borel set $A \in \mathbb{R}^d$.

Unfortunately, (2) often does not admit solution as $\mathcal{A}_T$ might be empty. To circumvent this issue, one considers the so-called Kantorovitch relaxation of this problem:

$$\inf_{\pi \in \mathcal{A}(\mu,\nu)} \int_{x,y \in \mathbb{R}^d} c(x, y) d\pi(x, y) \tag{3}$$

where $\mathcal{A}(\mu, \nu)$ is the set of joint distribution whose first and second marginals are equal to μ and ν respectively. π is called the transport plan between μ and ν. If π is deterministic — specifically, for any x, there is a unique y such that $\pi(x, y) > 0$ — then it is a transport map as defined in Monge problem. It suffices to define such a map $T(x)$ by the only y respecting $\pi(x, y) > 0$. However, we could instead consider π to be a "one-to-many" transport: for each x, they might be several y such that $\pi(x, y) > 0$. While Monge's development had the problem of non-existence of the transport map, the Kantorovitch relaxation $\mathcal{A}(\mu, \nu)$ is never empty. In particular, it always contains the independent joint distribution $\mu \times \nu$. In many situations, under mild assumption on c (semi-lower continuity and bounded from below), there always exists a minimizer of (3), so that the infimum might be replaced by a minimum (Ambrosio & Gigli, 2013). We will denote $V_c(\mu, \nu)$ as the value of this minimum. We distinguish between a transport *map* T achieving the minimum in (2) and a transport *plan* π minimizing (3). Note that for most applications, we are more interested in obtaining an approximation of the optimal transport map than a transport plan.

When the cost c is a distance, V_c exactly matches the famous $W_1(\mu, \nu)$, which is called the first Wasserstein distance between μ and ν. When c is a distance to a power of some positive integer p, $V_c^{1/p}$ is denoted W_p and is commonly called the pth Wasserstein distance. An important result is that those are actually respecting the axioms of a distance over distributions (Ambrosio & Gigli, 2013). For the sake of simplicity, we will abusively extend the term "Wasserstein distance" to any V_c, for any cost function c, even though not all choices of c lead to V_c being a metric over probability distributions.

One could also work with the dual of problem (3), which is proven to lead to the same value $V_c(\mu, \nu)$. We write this dual as:

$$\begin{aligned} \sup_{\phi,\psi \in \mathcal{A}^*(\mu,\nu)} \quad & \int_{x \in \mathcal{X}} \phi(x) d\mu(x) + \int_{y \in \mathcal{Y}} \psi(y) d\nu(y) \\ \mathcal{A}^*(\mu, \nu) := \quad & \{(\phi, \psi) : \mathbb{R}^d \to \mathbb{R} / \ \forall x, y \in \mathcal{X} \times \mathcal{Y}, \phi(x) + \psi(y) \leqslant c(x, y)\} \end{aligned} \tag{4}$$

We will also denote as $V^*(\phi, \psi)$ the value $\int_{x \in \mathbb{R}^d} \phi(x) d\mu(x) + \int_{y \in \mathbb{R}^d} \psi(y) d\nu(y)$. A pair (ϕ, ψ) maximizing V^* are called Kantorovitch potentials. Figure 2 in the Appendix provides an illustration of the shape of an optimal transport map T solving (2).

The discriminator of the Wasserstein GAN (WGAN) (Arjovsky et al., 2017) computes the dual formulation (4), with ν being the "true" distribution $\mathbf{G}(\mathcal{Z})$ (defined by training examples) and μ being the "fake" distribution (defined by samples drawn from the generator). In practice, the constraint may be enforced by the addition of a constraint-violation penalty term to the objective Gulrajani et al. (2017a). Although introduced for other reasons, Cuturi (2009) proposed an efficient penalized version of (3) whose resulting optimization problem is called the entropic-regularized Optimal Transport. We shall only provide its dual form:

$$\sup_{\phi,\psi} V^*(\phi, \psi) + \lambda_{ineq} \mathcal{L}_{ineq}(\phi, \psi) \tag{5}$$

where $\mathcal{L}_{ineq}$ penalizes (ϕ, ψ) when they violate the constraint in $A(\mu, \nu)$:

$$\mathcal{L}_{ineq}(\phi, \psi) := -\mathbb{E}_{x,y \in (\mathbb{R}^d)^2} \left[e^{\frac{\phi(x) + \psi(y) - c(x,y)}{\lambda_{ineq}}} \right]$$

Other penalties were theoretically explored by Blondel et al. (2018), including the widely used L_2-penalty (Seguy et al., 2018)

$$\mathcal{L}_{ineq}(\phi, \psi) := -\mathbb{E}_{x,y\in(\mathbb{R}^d)^2}\left[(\phi(x) + \psi(y) - c(x, y))_+^2\right]$$

Objective (5) was shown to asymptotically recover the value of (4) when $\lambda_{ineq} \to 0$.

Due to its suitability for training parametric models such as neural network, we exploit this optimization objective to compute Wasserstein distances W_p and V_c. As it is provable (Berman, 2013) that the 'differential' of $W_p(., \nu)$ with respect to its first variable is $-\phi$ (where ϕ is a Kantorovitch potential), we may conclude that a GAN framework where the discriminator accurately computes (4) through (5) provides a means of training a generative model. The details of such a framework are developed in the next section.

An other perspective is to recover an approximation of the solution of (2) (resp. (3)), namely an optimal transport map (resp. plan), given optimal solutions of (4) (i.e. Kantorovitch potentials). Getting an approximate transport plan might be done by considering a specific relationship that holds between the solutions of the dual and the primal in regularized versions of the Optimal Transport (Cuturi, 2009; Xie et al., 2018). As for an optimal transport map, Seguy et al. (2018) provides a two-step procedure, that is, they first compute a regularized transport plan, and then calculate an approximate transport map by doing a barycentric projection of the plan. Nonetheless, as we will see in the following proposition, there is no distinction between an optimal plan and its corresponding map in the case of the Wasserstein distances W_p with $p > 1$. In addition, the primal-dual relationship is straightforward. Getting an optimal transport map solving (2) is a by-product of computing Kantorovitch potentials by solving (4). We summarize rigorously this statement in the following:

Proposition 1. *Fix $p \geqslant 2$ and the cost $c(x, y) := \frac{\|x-y\|^p}{p}$. Then there is one unique optimal transport map π solving* (3)*. It is deterministic, that is it corresponds to an optimal transport plan T solving* (2)*: $\pi = (Id \times T)$. Also, the Kantorovitch potentials ϕ, ψ are unique up to a translation and the following relation holds:*

$$\forall x \in \textit{supp}(\mu), T(x) = x - \|\nabla\phi(x)\|^{\frac{1}{p-1}-1}\nabla\phi(x) \tag{6}$$

3 THE SECOND-WASSERSTEIN GAN

Although the framework presented below is valid for any W_p, for simplicity of presentation, we shall concentrate on the case where the discriminator asymptotically computes W_2 and where the optimal transport is taken with respect to the W_2 problem. Essentially, whether we want to solve (2) or eventually train a GAN generator, the procedure solves an approximation of (4) using the optimisation objective (5). Then if one needs to solve either the Monge problem (2) or the Kantorovitch relaxation (3), it suffices to use the general correspondence of the primal and dual problems in (6). In the following we detail how to solve the Monge optimal transport problem (2). We then develop the W2GAN framework: a GAN based on the use of the second Wasserstein distance. Notice that in both cases the discriminator objective is eventually refined thanks to additional terms. We removed their justification in the Appendix 6.2 for keeping the following models as simple as possible.

3.1 APPROXIMATING THE MONGE MAP

We are given two datasets (X_i) and (Y_i) which we assume to be samples of independent random variables $X \sim \mu$ and $Y \sim \nu$. We recall that, in the case of the optimal transport problem associated to W_2 (i.e. when the cost c is the square of the euclidean distance), there exists an optimal map T for (2), and it is equivalent to the optimal plan of (3). Proposition 1 says that it is enough to solve (4) to obtain T. Our procedure thus simply consists in iterating gradient ascent on (5), which is close to (4). Hence we learn the Kantorovitch potentials ϕ, ψ which we implement as neural networks. We see that their objective is a proper optimization loss. Indeed, we rephrase (5) as

$$\sup_{\phi,\psi} \mathbb{E}(\phi(X) + \psi(Y)) - \lambda_{ineq}\mathbb{E}\left[\left(\phi(X) + \psi(Y) - \frac{\|X - Y\|_2^2}{2}\right)_+^2\right] =: \sup_{\phi,\psi} \mathcal{L}_D(\phi, \psi) \tag{7}$$

We could similarly have chosen the regularization of entropy. After iterating gradient ascent on (7) until convergence, we simply compute the map T using the identity given by (6)

$$\forall x \in \text{supp}(\mu), T(x) = x - \nabla\phi(x) \tag{8}$$

The whole procedure is given in Algorithm 1.

3.2 A GAN BASED ON A 2-WASSERSTEIN METRIC

The context is similar to the previous one. We are given a dataset (X_i) which is assumed to be sampled from a random variable X in the euclidean space $\mathbb{R}^d$. We often call X the true distribution. The generator G is a map from a certain (low dimensional) space endowed with a random variable $\mathcal{Z}$ whose distribution $P_{\mathcal{Z}}$ is fixed beforehand, onto $\mathbb{R}^d$. We often call $G(\mathcal{Z})$ the fake or generated distribution. We want to learn a generator G by minimizing $W_2(X, G(\mathcal{Z}))$ (which is an abuse of notation for $W_2(p_X, p_{G(\mathcal{Z})})$). Now because W_2 has neither tractable computation in the case of continuous distributions, nor reachable gradient, we rely on an approximation of it. That is, a discriminator we denote (ϕ, ψ) is trained to compute $W_2(p_X, p_{G(\mathcal{Z})})$. Hence the objective of the discriminator, for a given G is:

$$\sup_{\phi,\psi} \mathbb{E}(\phi(G(Z))+\psi(X))-\lambda_{ineq}\mathbb{E}\left[\left(\phi(G(Z)) + \psi(X) - \frac{\|G(Z) - X\|_2^2}{2}\right)_+^2\right] =: \sup_{\phi,\psi} \mathcal{L}_D(\phi, \psi, G(Z)) \tag{9}$$

Plugging in the generator objective, we can summarize the whole procedure as

$$\sup_{\phi,\psi} \inf_{G} \left[\mathbb{E}(\phi(G(Z)) + \psi(X))\right] - \lambda_{ineq}\mathbb{E}\left[\left(\phi(G(Z)) + \psi(X) - \frac{\|G(Z) - X\|_2^2}{2}\right)_+^2\right] \tag{10}$$

Notice that it is not exactly a minimax game, as the generator does not update thanks to the penalty term on the discriminator, which we only think as a regularization for encoding the c-inequality of the dual potentials. This is similar to the approach taken for the gradient penalty or lipschitzness of the discriminator in Gulrajani et al. (2017a) and Petzka et al. (2017). On the other hand, some methods prefer to consider the penalty as part of the asymptotic metric computed by the discriminator, that is a regularized Wasserstein distance. Then the generator is updated with respect to the penalty term also (Sanjabi et al., 2018). The whole work flow is detailed in Algorithm 2.

3.3 THE DYNAMICS OF WASSERSTEIN GANS ARE RELATED TO THE OPTIMAL TRANSPORT MAP

Using the two previous problems (i.e the generative modelling and the computation of the optimal transport map between two distributions) we are able to provide a new insight on GANs that relies on a Wasserstein distance. Our analysis is valid for WGAN and its extensions, but benefits from a better theory in the case of W2GAN. What we observe is that during training, a generator $G(Z)$ is basically following the "line" defined the optimal transport map between its initialization $G_0(Z)$ and the target distribution X whose probability measures we denote μ_0 and μ_X respectively.

We divide our analysis into two cases.

Case 1: We consider the ideal case of a perfect minmax game between generator and discriminator. The assumptions are that they are free of parameters in that they perfectly cover the L_2 space of functions. We assume infinitesimal discretization of the update steps, that is, we consider a time-dependent differential equation governing the generated distribution. In this case, we can show that the time dependent generated distribution μ_t evolves exponentially fast toward the target distribution μ_X and evolves according to an optimal transport map between μ_0 and μ_X.

Case 2: We switch to the case of finite parameters with small updates for the generator, where we show a similar dynamic of the generated distribution although additional terms appearing in the update equation of the generator hinders to conclude rigorously in such case.

Finally, we provide low-dimensional experiments to supports our claims.

CASE 1: NON-PARAMETRIC GENERATOR

Consider the context of W2GAN, where the discriminator (ϕ, ψ) asymptotically computes exactly the squared second Wasserstein distance $W_2^2(G(Z), X)$, i.e we suppose it is trained infinitely many times

at each update of the generator, and we forget about the bias induced by the c-inequality constraint being encoded in the objective as a penalty (equivalently, we assume $\lambda_{ineq} = 0$). We also assume G to be free of parameters and to evolve continuously, thus writing $G(t, Z)$ as a time-dependent random variable where G is in the space of L_2 functions and t is a fictive variable. We also assume that absolutely continuous measures of $\mathcal{P}(\mathbb{R}^d)$ admits representation. That is, for all such μ, there exists a function G such that $\mu = G\#P_Z$. In this case we may equivalently work on the corresponding generated probability distribution of interest, which we denote as $\mu_t := p_{G(t,Z)}$. Naturally, we consider the gradient flow:

$$\dot{\mu_t} = -\nabla F(\mu) := -\nabla W_2^2(\mu_t, \mu_X). \tag{11}$$

One would need to introduce the definition of gradient flow in the space of probability measures and in particular the notion of gradients with respect to probability measures and velocities of time-dependent measures in order to fully express the meaning of the above. We refer to Ambrosio et al. (2008) for a comprehensive overview.

One could also have defined such a flow in the W_1 case, but in the case of the W_2 metric, we can refine (11) thanks to Ambrosio et al. (2003) and obtain:

$$\dot{\mu_t} = T_t - Id \tag{12}$$

where T_t is the unique optimal transport map, Solving (2) between μ_t and μ_X. According to the gradient flow (12), locally, the generated distribution evolves towards μ_X by following the Wasserstein 2 geodesic. In fact, this is a global behaviour, from Ambrosio et al. (2008):

Theorem 1. *Denote T the optimal transport map between μ_0 and μ_X. Then we have that the gradient flow solving* (11) *is uniquely determined:*

$$\mu_t = [e^{-t}Id + (1 - e^{-t})T]\#\mu_0 \tag{13}$$

Hence a consequence (also from Ambrosio et al. (2008)) is that the generated distribution evolves exponentially fast towards μ_X:

Corollary 1.

$$\forall t \geqslant 0, W_2^2(\mu_t, \mu_X) = e^{-2t}W_2^2(\mu_0, \mu_X) \tag{14}$$

The evolution in (13) suggests that the generated distribution follows the Wasserstein-2 geodesic between μ_0 and μ_X. That means the training dynamic of the generator "draws" the optimal transport map between μ_0 and μ_X. At the end of training, the generator $G(\infty, .)$ provides a certain optimal transport map. For each z, the 'arrows' joining $G(0, z)$ and $G(\infty, z)$ together constitute the optimal map. Figure 3 in the Appendix helps visualizing this analysis.

CASE 2: PARAMETRIZED GENERATOR.

We now turn to the more realistic case of a parametrized generator $G(\theta, Z)$. Lui et al. (2017) already made the conjecture that a generator trained with the second Wasserstein distance would have its parameters updated towards the direction of an optimal transport. Let us prove their statement in the case of our model. In the following, $J_f(u)$ denotes the Jacobian of a function f at point u. Let $G(\theta, z)$ be our parametrized generator, which we assume to admit derivatives w.r.t both the parameters θ and Z. Also, we introduce $\mu_\theta := \mathbb{P}_{G(\theta,Z)} = G(\theta, .)\#\mathbb{P}_Z$ where $\mathbb{P}_Z$ is the measure known beforehand on the latent variable Z. We consider the context of alternating gradient descent with learning rate $\alpha > 0$ for the generator. We naturally associate the fictive time variable t, so that we consider the discrete update equation:

$$\theta_{t+1} = \theta_t - \alpha \frac{\partial W_2^2(\mu_{\theta_t}, \mu_X)}{\partial \theta} \tag{15}$$

Proposition 2. At each generator update, we assume the discriminator ϕ, ψ to achieve $\sup_{\phi,\psi} V^*(\phi, \psi, \mu_\theta, \mu_X) = W_2^2(\mu_\theta, \mu_X)$ where $V^*(\phi, \psi, \mu_\theta, \mu_X) := \int_{\mathbb{R}^d} \phi(x) d\mu_\theta(x) + \int_{\mathbb{R}^d} \psi(x) d\mu_X$ is the value function of the dual Kantorovitch problem (4). Then (15) admits a tractable form as:

$$\frac{\partial W_2^2}{\partial \theta} = \mathbb{E}(J^T_{\phi(G(\theta,Z))}(\theta)) = \mathbb{E}(\nabla_\theta \phi(G(\theta, Z)))$$

Moreover, the gradient ascent dynamics of the generator are linked to the optimal transport map T solving (2) between μ_θ and μ_X as:

$$\forall z, G(\theta_{t+1}, z) = G(\theta_t, z) + \alpha J_{G(.,z)}(\theta_t) \times \mathbb{E}\left(J^T_{G(.,Z)}(\theta_t) \times [T(G(\theta_t, Z)) - G(\theta_t, Z)]\right) + o(\alpha) \tag{16}$$

Equation (16) is close to the one conjectured in Lui et al. (2017). Following their approach, for the sake of clarity, we consider the case where the latent variable Z is a constant z. Then we read (16) as

$$G(\theta_{t+1}, z) = G(\theta_t, z) + \alpha J_{G(.,z)}(\theta_t) J^T_{G(.,z)}(\theta_t) \times [T(G(\theta_t, z)) - G(\theta_t, z)]$$

As similarly observed in Lui et al. (2017), if we ignore the possibly significant effect of $J_{G(.,z)}(\theta_t) J^T_{G(.,z)}(\theta_t)$, the term $T(G(\theta_t, z)) - G(\theta_t, z)$ enforces G to be updated on the optimal transport plan joining μ_θ and μ_X.

Based on this intuition, we formulate the assumption that for a general latent space Z, updating θ from time t to time $t+1$ leads to a generated distribution $\mu_{\theta_{t+1}}$ moving along the optimal transport between μ_{θ_t} and μ_X. In other words, denoting T_t as such an optimal transport, $\mu_{\theta_{t+1}}$ is in the set:

$$S_t := \{[sId + (1-s)T_t]\#\mu_{\theta_t} | s \in [0;1]\}$$

It is known (Villani, 2008) that S_t describes the unique Wasserstein 2 geodesic between μ_{θ_t} and μ_X. Considering two consecutives sets S_k and S_{k+1}, because from our assumption μ_{k+1} is in S_k, that implies $S_{k+1} \subset S_k$. Hence we may conclude that the generated distribution evolves incrementally toward μ_X in the set

$$S_0 = \{[sId + (1-s)T]\#\mu_0 | s \in [0;1]\}$$

where μ_0 is the initial generated distribution and T the optimal transport between μ_0 and μ_X.
In particular, we get two insights from this analysis:

- The dynamic of the generator during training is taking place on the optimal transport path between two distributions.
- At the end of training, the generator recovers an optimal map between the initial generator and the target distribution.

Nonetheless, the dynamics of the generator is easily perturbed by biased updates. Imagine, for instance, that there exists $k > 0$, such that μ_{k+1} is not on the optimal path between μ_k and μ_X, i.e it is not in S_k. Then, from now on, the generator will evolve ideally on the Optimal Transport map joining μ_{k+1} and μ_X, and thus will no longer evolve in S_k and S_0. In some sense, the learning dynamics of the generator is a Markov process in that it "forgets" the previous updates. A wrong update leads to a generator that no longer evolving exactly on the optimal map between the initial μ_0 and the target μ_X. It would certainly be of interest to have a mathematical framework that accounts for this variance in the evolution of the generator, but this is beyond the scope of this paper.

One could design a similar analysis in the case of the first Wasserstein distance W_1 and apply it to WGAN (Arjovsky et al., 2017), WGAN-GP (Gulrajani et al., 2017b) and WGAN-LP (Petzka et al., 2017) for instance. However, here is no strict equivalent of the previous theorems when the cost is the distance itself (recall that W_p^p stands for the optimal value of (2) or (3) when we take $c(x,y) := \frac{\|x-y\|^p}{p}$). In particular, the analysis hardly depends on the relationship (6) between Kantorovitch potential ϕ, ψ and Monge map T solving (2). In the Wasserstein 1 case, T may be proven to exist and is related to a certain Kantorovitch potential f by $\frac{T(x)}{\|T(x)\|} = -\nabla f(x)$ (Ambrosio, 2002). Hence one gets the direction of an OT map thanks to the knowledge of the dual optimal variable, but not its norm. Second, one does not have uniqueness of the optimal map, making it difficult to predict on which Wasserstein-1 geodesic the generator is moving during training. Still, it is interesting to understand that it is moving on one of the geodesic, hence possibly following an optimal transport map.

Finally, in Section 4, we show that the low-dimensional experiments support this analysis. The generator seems to reproduce an optimal transport map. Interestingly, WGAN-GP and WGAN-LP also seem to learn an optimal map rather than a random mapping between the latent distribution and the target distribution.

4 EXPERIMENTS

In this section, we verify our results empirically on several synthetic and real datasets. Using a discrete computation of the Optimal Transport, we use it as a benchmark to evaluate:

- The optimal map computed by our discriminator when trained using (8).
- The optimal map generated by other models specifically designed for approximating OT.
- The map drawn by our generator at the end of training.
- The map drawn by generators of other GANs at the end of training.[1]

4.1 2-DIMENSIONAL SYNTHETIC DATA

We consider learning optimal maps between synthetic datasets in 2-dimensional space, which allows us to visualize the learned optimal map. Each dataset is composed of two parts: X and Y, and we learn optimal maps from X to Y. We consider three datasets (data samples are shown in Table 1). (i) **4-Gaussians**: X and Y are both mixtures of 4 Gaussians with equal mixture weights. The mixture centers of Y is closer together than that of X. (ii) **Checkerboard**: X and Y are both mixtures of uniform distributions over 2D squares, of size 4 and 5 respectively, with equal mixture weights. The mixture centers of the two distributions form an alternating "checkerboard" pattern. (iii) **2-Spirals**: X and Y are uniform distributions over spirals that are rotations of each other.

In Table 1 we show optimal mappings (black arrows) based on our approach for computing OT with the discriminator (*W2-OT*) and two other baselines: (i) *Discrete OT*: a discrete assignment algorithm, which we apply on the fixed number of data samples in Table 1. Note that this algorithm works only for discrete data, and we use it here as benchmark optimal map for the shown samples. (ii) *Barycentric-OT*: a two-step algorithm for computing OT between continuous or discrete distributions introduced by Seguy et al. (2018). The algorithm is based on computing a regularized optimal plan then estimating a parameteric OT based on a barycentric projection. We notice that our closed-form W2-OT can largely recover the optimal map between X and Y, especially in the checkerboard and gaussian case. It seems to get the right direction in the 2-D spiral problem, though overlapping. The Barycentric-OT approach seems to work well but note that this algorithm learns a parameteric mapping through a two-step process. It seems more unstable on the checkerboard experiment and mode-collapses in the 4-gaussians case.

In Table 2 we show the recovered optimal map by the generator using our proposed W2GAN approach, and compare it to three baselines: Jenson-Shannon GAN (JS-GAN) (Goodfellow et al., 2014), WGAN-GP (Gulrajani et al., 2017a) and WGAN-LP (Petzka et al., 2017). We can see that both W2GAN and WGAN-LP can successfully recover OT between in all 3 datasets, which matches our expectation as both faithfully approximate the defined metric between distributions. In comparison, WGAN-GP performs poorly, and we observed that training becomes especially unstable as $G(Z)$ approaches Y, sometimes moving away from a good solution. This happens even with more training of the discriminator and smaller generator learning rates. This supports Petzka et al. (2017)'s argument that the gradient penalty in WGAN-GP prevents the discriminator from converging. The original GAN framework based on the Jensen-Shannon divergence also surprisingly arrives at the same OT solution in two datasets: 4-Gaussians and 2-Sprials. This may be because the OT solution is in a sense the "simplest" solution for a generator to find. On the other hand, it does not reproduce OT in the checkerboard case.

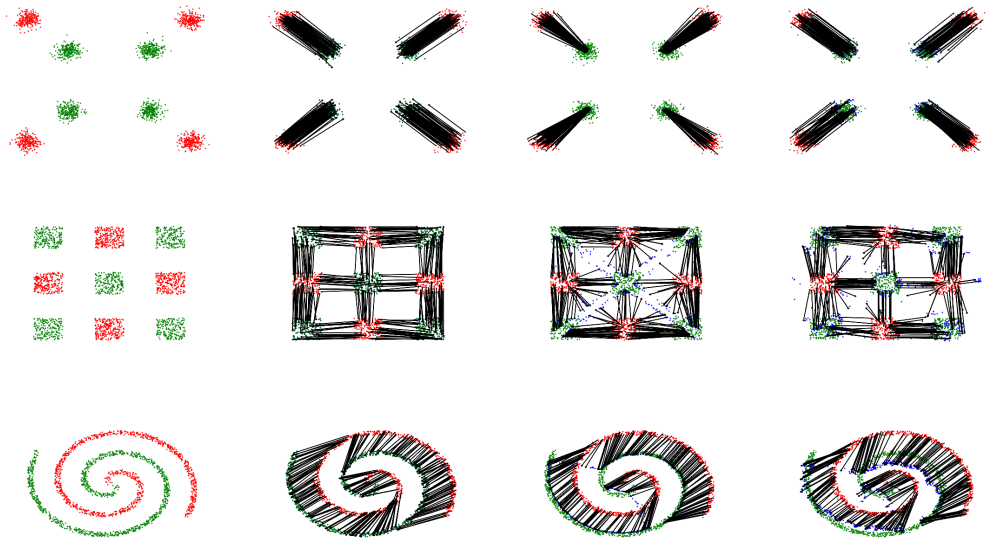
4.2 HIGH-DIMENSIONAL DATA

We experiment two settings with high-dimensional data. In the first we consider the task of mapping multivariate Guassian[2] to MNIST as proposed in Seguy et al. (2018). We show in Figure 2-(a) the results of mapping multivariate Gaussian samples to MNIST samples. As expected, our W2GAN framework performs well compared to the Barycentric-OT framework of Seguy et al. (2018).

In the second we consider the generation task of CIFAR-10 samples. We use a standard convolutional architecture similar and similar learning settings to (Gulrajani et al., 2017a). We can see in Figure

[1] Training details can be found in Appendix 6.4

[2] With mean and covariance matrix estimated with maximum-likelihood on MNIST.



(a) Data Samples (b) Discrete OT (c) Barycentric-OT (d) W2-OT (ours)

Table 1: True data samples and learned optimal maps (black arrows). (a) Samples of X (red) and Y (green) in three synthetic datasets. (b) Optimal map computed using an optimal assignment algorithm *only between fixed data samples* (c) Optimal map using parametric barycentric-OT (Seguy et al., 2018). (d) Optimal map using our proposed W2-OT approach. Note that both (c) and (d) are between distributions of X and Y.

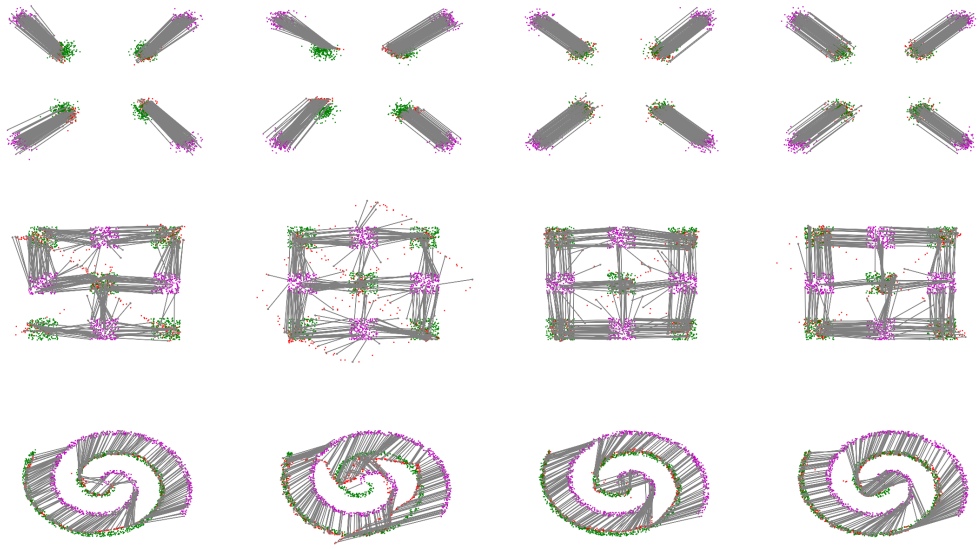
2-(b) that our W2GAN can perform reasonably well as a generative model, while have a strong theoretical advantages.

(a) (left) MV-Gaussian samples (right) corresponding MNIST samples (b) CIFAR-10 samples

Figure 1: Generated samples from W2GAN in two benchmark datasets: (a) MNIST (b) CIFAR-10.

5 Conclusion And Future Works

We believe this work offers a new perspective on GANs: a way to characterize the dynamics of the generator during training. In particular we realized that Wasserstein GAN and its extensions do not learn just any generator that fits the target distribution. It learns the generator that follows an optimal map between its initialized distribution and the target distribution. To establish this, we built the W2GAN model, which is able both to approximate an optimal transport map and to serve as a new GAN model. We then connect how the corresponding generator is trained with the behaviour of the optimal map. In the Appendix 6.5, we raise interesting questions about our model left for future



(a) JS-GAN (b) WGAN-GP (c) WGAN-LP (d) W2GAN (ours)

Table 2: Mappings learned by the generator in three synthetic datasets for JS-GAN, WGAN-GP, WGAN-LP and W2GAN. The generator maps (gray arrows) samples from Z (magenta) to samples $G(Z)$ (red) such that it matches distribution of Y (green). WGAN-LP and W2GAN can successfully recover optimal map between X and Y in all three datasets. The generator is the identity at the beginning of training, thus the analysis 3.3 implies the generator is learning a map between X and Y.

works. In particular, how to generalize this analysis of the generator dynamic during training to other GANs? How to confirm that asymptotically, when the term of regularization goes to 0, the gradient of the regularized optimal potential ϕ for (5) does indeed recover the relationship (8) with the optimal map T solving (2)?

REFERENCES

Raphaël Alexandre. Sommaire. 2018.

Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *arXiv preprint arXiv:1802.10151*, 2018.

L Ambrosio. arXiv : math / 0304389v1 [math . AP] 24 Apr 2003 Optimal Transport Maps in Monge-Kantorovich Problem. III, 2002.

Luigi Ambrosio and Nicola Gigli. *A User's Guide to Optimal Transport*, pp. 1–155. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savar. Gradient flows with metric and differentiable structures, and applications to the wasserstein space, 2003.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savar. *Gradient flows: in metric spaces and in the space of probability measures*. Birkhuser, 2008.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Robert J. Berman. Lecture notes on optimal transportation - with an eye towards complex geometry. 2013.

Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and Sparse Optimal Transport. 84, 2018.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein Barycenters of Measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2014. ISSN 15737683. doi: 10.1007/s10851-014-0506-3.

Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of Entropic Schemes for Optimal Transport and Gradient Flows. pp. 1–28, 2015. ISSN 10957154. doi: 10.1137/15M1050264. URL http://arxiv.org/abs/1512.02783.

R Cominetti and J San Martfn. Asymptotic analysis of the exponential penalty trajectory in linear programming. 67:169–187, 1994.

Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transportation distances. pp. 1–13, 2009.

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. 32(2):685–693, 22–24 Jun 2014. URL http://proceedings.mlr.press/v32/cuturi14.html.

Aude Genevay, Gabriel Peyré, Marco Cuturi, and Francis Bach. Stochastic Optimization for Large-scale Optimal Transport. (Nips), 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved Training of Wasserstein GANs. 2017a. ISSN 00308870. doi: 10.1016/j.aqpro.2013.07.003. URL http://arxiv.org/abs/1704.00028.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5767–5777. Curran Associates, Inc., 2017b. URL http://papers.nips.cc/paper/7159-improved-training-of-wasserstein-gans.pdf.

Kry Yik Chau Lui, Yanshuai Cao, Maxime Gazeau, and Kelvin Shuangjian Zhang. Implicit Manifold Learning on Generative Adversarial Networks. 2017. URL http://arxiv.org/abs/1710.11260.

Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, and Zhen Wang. Multi-class generative adversarial networks with the L2 loss function. *CoRR*, abs/1611.04076, 2016. URL http://arxiv.org/abs/1611.04076.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *ICLR*, 2018.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 271–279. Curran Associates, Inc., 2016.

Henning Petzka, Asja Fischer, and Denis Lukovnicov. On the regularization of wasserstein gans. *CoRR*, abs/1709.08894, 2017. URL http://arxiv.org/abs/1709.08894.

Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. 2018. URL http://arxiv.org/abs/1803.00567.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6667 LNCS:435–446, 2012. ISSN 03029743. doi: 10.1007/978-3-642-24785-9_37.

Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018.

Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D. Lee. On the convergence and robustness of training GANs with regularized optimal transport. *arXiv preprint arXiv:1802.08249*, 2018.

Filippo Santambrogio. Euclidean , Metric , and Wasserstein Gradient Flows : an overview. pp. 1–65.

Filippo Santambrogio. Introduction to Optimal Transport Theory. pp. 1–16, 2010. doi: 10.1017/CBO9781107297296.002. URL http://arxiv.org/abs/1009.3856.

Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-Scale Optimal Transport and Mapping Estimation. (1781):1–15, 2017. URL http://arxiv.org/abs/1711.02283.

Vivien Seguy, Bharath Bhushan Damodaran, Rmi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2018.

Cédric Villani. Optimal transport, old and new. *A Series of comprehensive Studies in Mathematics*, 2008.

Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. A fast proximal point method for wasserstein distance. *CoRR*, abs/1802.04307, 2018. URL http://arxiv.org/abs/1802.04307.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.

6 Appendix

6.1 Proofs

Proof of proposition 1. This result is a particular case of a well-known correspondence between Kantorovitch potentials and optimal transport map. In fact, when the cost c is such that $c(x, y) = h(x - y)$ and h is strictly convex, one has the existence and uniqueness of an optimal Monge map T solving (2) and a specific relationship with the Kantorovith potential ϕ, ψ solving (4) Villani (2008); Santambrogio:

$$T(x) = x - (\nabla h)^{-1}(\nabla \phi(x))$$

Fix $p \geqslant 2$ and consider the case of the p-Wasserstein distance. Then $c(x, y) = h(x - y)$ where $h(x) := \frac{1}{p}\|x\|^p$. A norm is always convex by triangle inequality, and any $x \to x^p$ is strictly convex and increasing on $\mathbf{R}_+$, so the previous result provides the uniqueness of the optimal transport map T. It only remains to invert the gradient of h. A quick calculation gives:

$$\forall y \in Im\nabla h, \nabla h^{-1}(y) = \|y\|^{\frac{1}{p-1}-1} y$$

as we supposed the norm to be the euclidean one. Plugging this into the first expression, we obtain the desired result. □

Proof of proposition 2. Proving the first part of proposition 2 is exactly similar as the way it is done for the Wasserstein-1 case in Arjovsky et al. (2017). We recall that the main ingredients of the proof is first to convey an envelop theorem to obtain that

$$\nabla_\theta W_2^2(\mu_{\theta_t}, \mu_X) = \nabla_\theta \mathbb{E}[\phi(G(\theta, Z))]$$

and second to use a dominated convergence argument to invert the expectation and the gradient operator in the right-hand side of the above.
For the second part of the proof, we rewrite the gradient ascent equation (15) for the generator parameter θ:

$$\theta_{t+1} = \theta_t - \alpha \nabla_\theta W_2^2(\mu_{\theta_t}, \mu_X) = \theta_t - \alpha \mathbb{E}[\nabla_\theta \phi(G(\theta_t, Z))] = \theta_t - \alpha \mathbb{E}[J^T_{\phi(G(\theta_t, Z))}(\theta_t)]$$

$$= \theta_t - \alpha \mathbb{E}[((\nabla\phi(G(\theta_t, Z)))^T J_{G(.,Z)}(\theta_t))^T] = \theta_t - \alpha \mathbb{E}[(J_{G(.,Z)}(\theta_t)^T \nabla\phi(G(\theta_t, Z))]$$

Therefore, for a fix latent variable z,

$$G(\theta_{t+1}, z) = G(\theta_t - \alpha \mathbb{E}[J_{G(.,Z)}(\theta_t)^T \nabla\phi(G(\theta_t, Z))], z)$$

Then by first order Taylor expansion:

$$G(\theta_{t+1}, z) = G(\theta_t, z) - \alpha J_{G(.,z)}(\theta_t)\mathbb{E}[J_{G(.,Z)}(\theta_t)^T \nabla\phi(G(\theta_t, Z)] + o(\alpha)$$

We conclude using the hypothesis that ϕ is maximizing the dual of Kantorovitch and is related to an optimal map T trough (8). □

6.2 Exploiting a primal-dual relationship and interpolating the constraint to improve the model

In the following, we explain two important modifications of the objective functions for the discriminator. Those bring our model closer to the theory of Optimal Transport. They both apply for both GAN and OT goals. In order to unify notations, we consider the two random variables X and Y to be discriminated by ϕ, ψ. In the GAN context, one may think of X as the generated distribution $G(Z)$ and Y as the true distributions. In both cases the discriminator objective is divided into two parts. First, a main objective $\mathcal{L}_{OT}(\phi, \psi, X, Y) := \mathbb{E}(\phi(X) + \psi(Y))$ which corresponds to the dual of the Kantorovitch problem. Second, the inequality constraint $\mathcal{L}_{ineq}(\phi, \psi, X, Y) := -\lambda_{ineq}\mathbb{E}[(\phi(X) + \psi(Y) - \frac{\|X-Y\|_2^2}{2})_+^2]$. Recall from the introduction that one could prefer an other choice of regularization such as the entropy penalty. Hence the overall objective for the discriminator:

$$\sup_{\phi,\psi} \mathcal{L}_{OT}(\phi, \psi, X, Y) + \mathcal{L}_{ineq}(\phi, \psi, X, Y)$$

The first idea is to take advantage of relation (8), which we know to be true between an optimal potential ϕ and an optimal map T. In fact, Optimal transport theory asserts that the inequality constraint should be exactly saturated where there is some transport (Villani, 2008):

Theorem 2. *Consider any lower-semicontinuous cost function c and a optimal transport plan for* (3)*, and Kantorovitch potentials ϕ, ψ for* (4)*. Then,*

$$\forall x, y, (x, y) \in supp(\pi) \Longrightarrow \phi(x) + \psi(y) = c(x, y) \tag{17}$$

In our case, that is when the cost is a pth power of the euclidean distance, we know from Proposition 1 than an optimal transport plan is actually an optimal transport map T, and we dispose of a relationship with the corresponding Kantorovitch potential ϕ. Hence an immediate consequence of the above for the case of the square of the euclidean distance is:

Corollary 2. *For the Kantorovitch problem* (2) *related to $c(x, y) := \frac{\|x-y\|_2^2}{2}$, that is the computation of the second Wasserstein distance, given Kantorovitch potentials ϕ, ψ:*

$$\forall x \in supp(\mu), \phi(x) + \psi(x - \nabla\phi(x)) = \frac{\|\nabla\phi(x)\|_2^2}{2} \tag{18}$$

Thus we suggest enforcing our discriminator to abide by relation (18). Notice that the previous corollary admits an exact symmetric relationship involving the gradient of ψ. This is done by adding the following penalty during training:

$$\mathcal{L}_{eq}(\phi, \psi, X, Y) := -\lambda_{eq}\left[\left(\phi(X) + \psi(X - \nabla\phi(X)) - \frac{\|\nabla\phi(X)\|_2^2}{2}\right)^2 \right. \tag{19}$$

$$\left. + \left(\phi(Y - \nabla\psi(Y)) + \psi(Y) - \frac{\|\nabla\psi(Y)\|_2^2}{2}\right)^2\right] \tag{20}$$

We call it the c-equality penalty term as it tries to enforce the dual functions to saturate to c-inequality constraint at the right locations. Hence the overall objective of the discriminator:

$$\sup_{\phi,\psi} \mathcal{L}_{OT}(\phi, \psi, X, Y) + \mathcal{L}_{ineq}(\phi, \psi, X, Y) + \mathcal{L}_{eq}(\phi, \psi, X, Y) \tag{21}$$

Remark 1. It is interesting to bridge this objective with the one used in GAN relying on the first Wasserstein distance. In WGAN-GP, the discriminator is asked to have gradient exactly equal to 1. Translating that into the Optimal Transport theory, it is tantamount to have potentials ϕ, ψ saturating the c-inequality. That is, this gradient penalty term is similar to our $\mathcal{L}_{eq}$. Importantly though, Petzka et al. (2017) raised the fact that it is not a valid practice according to theory to ask the gradient norm of the discriminator to be 1 everywhere. Instead, WGAN-LP is a model where the gradient's norm is enforced to be less than 1, which is then translated into Optimal Transport perspective as potentials ϕ, ψ respecting the c-inequality. Their penalty term is thus similar to our $\mathcal{L}_{ineq}$. Our model takes advantage of the two forms. The difference is that our $\mathcal{L}_{eq}$ is completely justified -at least because it enforces a relationship which is true at optimum- thanks to the theory existing in the W_2 case.

A second modification concerns the way we encode the penalty enforcing the inequality constraint to be respected by ϕ, ψ. In fact, we modify $\mathcal{L}_{ineq}$ in order to bring it closer to the theory. Looking at the definition of $\mathcal{A}^*(\mu, \nu)$, the set of constraint of the dual of Kantorovitch problem (4), we see that the c-inequality constraint should be respected point wise everywhere in the ambient space where the distributions are defined. Instead, the entropy regularization or the L^2 one only enforce it for pairs x, y that live in the supports of the two distributions. We want to reduce this bias by somehow enforcing the inequality on a broader set. It is sufficient that the potentials respect the inequality constraint point wise on a convex compact set Ω containing the support of the two distributions. Hence we suggest enforcing it on the convex envelop of the two supports: $\Omega :=$Conv(supp(μ) $\bigcup$ supp(ν)). To do so, we define two i.i.d random variables $\tilde{X}$ and $\tilde{Y}$ which follow the same law as $\epsilon X + (1-\epsilon)Y$ where $\epsilon \sim \mathcal{U}([0,1])$. Hence the overall objective for ϕ, ψ:

$$\sup_{\phi,\psi} \mathcal{L}_{OT}(\phi, \psi, X, Y) + \mathcal{L}_{ineq}(\phi, \psi, \tilde{X}, \tilde{Y}) + \mathcal{L}_{eq}(\phi, \psi, X, Y) \tag{22}$$

Remark 2. This interpolation idea has already been used in GANs relying on the first Wasserstein distance, such as WGAN-GP and WGAN-LP in their gradient penalty. This practice was more motivated by better results. Here we provided a theoretical argument in favor of such practice. On the other hand, models trying to broaden GANs to higher order Wasserstein distances and/or to compute Optimal transport map (Sanjabi et al., 2018; Seguy et al., 2018; Salimans et al., 2018) in a similar manner only enforced the constraint on the support of the distributions.

Remark 3. We motivated the definition of $\tilde{X} := \epsilon X + (1-\epsilon)Y$ so that it basically takes value in the convex envelop of X and Y. But in all generality, it is wrong that only convex interpolation between two points allows to recover all Conv(supp(μ) $\bigcup$ supp(ν)). By Carathodory theorem, one would need to interpolate at most between d+1 points to recover all of it. Although not impossible at all, we wanted to keep the model simple and not depend on the dimension of the latent space $\mathbb{R}^d$.

6.3 ALGORITHMS

Algorithm 1 Computing an optimal transport map with $D = (\phi, \psi)$ and an L^2 regularization

Require: Two random variables X and Y from which we can sample.
Require: $\lambda_{\text{eq}}, \lambda_{\text{ineq}}, bD$.
Require: Initial critic parameters w_0.
  while w has not converged **do**
    Initialize losses $\mathcal{L}_{OT}$,$\mathcal{L}_{eq}$ and $\mathcal{L}_{ineq}$ to 0.
    for $i = 1, ..., bD$ **do**
      Sample real data $x, x' \sim \mathbf{P}_X$, $y, y' \sim \mathbf{P}_Y$ and $\epsilon_1, \epsilon_2 \sim \mathcal{U}([0,1])$.
      $\mathcal{L}_{OT} \leftarrow \mathcal{L}_{OT} - \frac{1}{bD}(\phi(x) + \psi(y))$
      $\mathcal{L}_{eq} \leftarrow \mathcal{L}_{eq} + \frac{\lambda_{eq}}{bD}([\phi(x) + \psi(x - \nabla\phi(x)) - \frac{\|\nabla\phi(x)\|_2^2}{2}]^2 + [\phi(y - \nabla\psi(y)) + \psi(y) - \frac{\|\nabla\psi(y)\|_2^2}{2}]^2)$
      $\tilde{x} \leftarrow \epsilon_1 x + (1-\epsilon_1)y$, $\tilde{y} \leftarrow \epsilon_2 x' + (1-\epsilon_2)y'$
      $\mathcal{L}_{ineq} \leftarrow \mathcal{L}_{ineq} + \frac{\lambda_{ineq}}{bD}(\phi(\tilde{x}) + \psi(\tilde{y}) - \frac{\|\tilde{x}-\tilde{y}\|_2^2}{2})_+^2$
    end for
    $\mathcal{L}_D \leftarrow \mathcal{L}_{OT} + \mathcal{L}_{ineq} + \mathcal{L}_{eq}$
    update (ϕ, ψ) with respect to $\mathcal{L}_D$.
  end while

Algorithm 2 W2GAN with $D = (\phi, \psi)$ and L^2-regularization

Require: latent space Z and true distribution X from which we have an available sample procedure.
Require: $\lambda_{\text{eq}}, \lambda_{\text{ineq}}, n_{\text{critic}}, bD, p$.
Require: Initial critic parameters w_0,
initial generator parameters θ_0.
while θ has not converged **do**
Initialize generator and discriminator losses $\mathcal{L}_D$,$\mathcal{L}_G$ to 0.
for $t = 1, ..., n_{\text{critic}}$ **do**
Initialize losses $\mathcal{L}_{OT}$,$\mathcal{L}_{eq}$ and $\mathcal{L}_{ineq}$ to 0.
for $i = 1, ..., bD$ **do**
Sample real data $x, x' \sim \mathbf{P}_X$, latent variable $y, y' \sim \mathbf{P}_{G(Z)}$ and $\epsilon_1, \epsilon_2 \sim \mathcal{U}([0, 1])$.
$\mathcal{L}_{OT} \leftarrow \mathcal{L}_{OT} - \frac{1}{bD}(\phi(x) + \psi(y))$
$\mathcal{L}_{eq} \leftarrow \mathcal{L}_{eq} + \frac{\lambda_{eq}}{bD}([\phi(x) + \psi(x - \nabla\phi(x)) - \frac{\|\nabla\phi(x)\|_2^2}{2}]^2 + [\phi(y - \nabla\psi(y)) + \psi(y) - \frac{\|\nabla\psi(y)\|_2^2}{2}]^2)$
$\tilde{x} \leftarrow \epsilon_1 x + (1 - \epsilon_1)y$, $\tilde{y} \leftarrow \epsilon_2 x' + (1 - \epsilon_2)y'$

$\mathcal{L}_{ineq} \leftarrow \mathcal{L}_{ineq} + \frac{\lambda_{ineq}}{bD}(\phi(\tilde{x}) + \psi(\tilde{y}) - \frac{\|\tilde{x} - \tilde{y}\|_2^2}{2})_+^2$
end for
$\mathcal{L}_D \leftarrow \mathcal{L}_{OT} + \mathcal{L}_{ineq} + \mathcal{L}_{eq}$
update (ϕ, ψ) with respect to $\mathcal{L}_D$.
end for
$\mathcal{L}_G \leftarrow \mathcal{L}_{OT}$.
update G with respect to $\mathcal{L}_G$.
end while

6.4 ARCHITECTURE

Below we describe the architecture details of our experiments.

For the 2D synthetic data experiments, the learning rate used for Barycentric-OT is 0.005 (although we did not notice that the learning rate influenced the solution quality significantly). For the GAN experiments, learning rates were chosen from the set $\{0.00001, 0.0005, 0.00005, 0.00001\}$. For the Wasserstein-based GANs, the number of discriminator updates per generator update is chosen from the set $\{5, 10, 20\}$. This is set to 1 in the Jensen-Shannon based GAN by default. λ_{gp} for both WGAN-LP and WGAN-LP is set to 10, following their conventions. For W2-OT and W2GAN, we set $\lambda_{eq} = \lambda_{ineq} = 200$. Additionally, in order to balance out ϕ and ψ in W2GAN, we add an additional regularizer: $\|\nabla\phi(X) - \nabla\psi(T(X)\|^2 + \|\phi(\nabla\phi(X) - \nabla\psi(T(X)\|^2$ with a small weight $\lambda_{balance} = 0.01$. To enforce that $G_0(Z) = Z$, we parameterize G by $G(Z) = H(Z) + Z$, where $H(Z)$ is initialized to be close to 0. H is parameterized by 4 fully connected hidden layers of size 128, with ReLU activations and batch norm in between the layers, and 1 fully connected final layer. ϕ and ψ are each parameterized by 2 fully connected layers with ReLU activations in between, and 1 fully connected final layer.

For multivariate to Gaussian experiments, we use the same architecture as (Seguy et al., 2018), composed of 2 fully-connected layers for both generator and discriminator. We also use similar initialization as the 2D experiments.

For CIFAR-10 experiments, we use a conventional architecture of two convolution layers followed by a fully connected layer similar to the convolutional architecture of (Gulrajani et al., 2017a). Training hyper-parameters are also based on (Gulrajani et al., 2017a).

6.5 REMAINING QUESTIONS AND FUTURE WORKS

In the following we detail questions raised by our model.

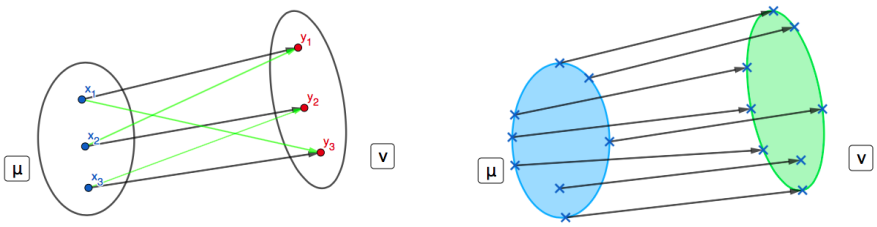
- A crucial relationship used in our model is (8), relating an optimal map T solving the Monge problem (2) with the Kantorovitch potential ϕ solving the dual (4). We recall that this link is $T(x) = x - \nabla\phi(x)$ for any x in the support of the initial distribution. In practice though, we

get this potential by solving a regularized version of (4), for instance (5) or (21). Assuming for simplicity we deal with (5), we face a dual problem which basically removes the hard constraint defining (4) by including a penalty term $\mathcal{L}_{ineq}$ in the main objective, the emphasis of which we control with the hyperparameter λ_{ineq}. Denoting $\phi_{\lambda_{ineq}}$ the corresponding solution -which may be proven to be unique up to translation- it seems to constitute a hard challenge to prove that $\phi_{\lambda_{ineq}} \to \phi$ as $\lambda_{ineq} \to 0$. The case of the entropy penalty, that is when we define the regularization $\mathcal{L}_{ineq}(\phi, \psi) := -\mathbb{E}_{x,y \in (\mathbb{R}^d)^2}[(\phi(x) + \psi(y) - c(x,y))^2_+$ has been widely studied. In a discrete setting, the convergence might be obtained as a result of Cominetti & Martfn (1994), but their analysis do not straightforwardly scale to the case of absolutely continuous probability measures. What we would find more useful is to prove the convergence $\nabla\phi_{\lambda_{ineq}} \to \nabla\phi$, as it would mean that our approximation of the optimal transport map T using relationship (8) is valid. Notice that the entropy regularized dual also admits a primal formulation, which solution $\pi_{\lambda_{ineq}}$ may be proven to converge in some sense to the unique transport map π solving (3) (Carlier et al., 2015). No similar results have been found about the dual variables for the moment. Also, the case of other penalties as the L_2 one remain open problems.

- An other crucial theoretical need for strengthening our analysis would be to strengthen the discussion in section 3.3 about the parametrized generator and prove the assumption that the generated distribution locally evolves toward an optimal transport map.

- In practice, what are the new possibilities enabled by the analysis of the generator's dynamic and the W2GAN model? First, It would be of interest to characterize the dynamic of GANs relying of f-divergences in a similar manner. In the Wasserstein case, it is possible that knowing the generator is learning an optimal correspondence could be useful in a Domain Transfer situation. On the other hand, as the W2GAN model approximates the second Wasserstein distance, it enables tackling the formidable challenge of computing Wasserstein barycenters of distributions (Cuturi & Doucet, 2014). In fact, it seems likely that an adversarial and parametric method would perform well in the task of generating all Wasserstein interpolations between distributions in high-dimension, a considerable problem addressed by Xie et al. (2018) for instance.

6.6 Additional figures



(a) The discrete case. (b) The continuous case.

Figure 2: The Monge problem. **(a)** A discrete example of the Monge problem (2) for distributions in $\mathbb{R}^2$. The μ distribution consists in three equally weighted diracs in x_1, x_2 and x_3, while the ν one is represented by y_1, y_2 and y_3 in the same way. Black arrows denote the actual optimal transport map T. The green arrows together also define a map from μ onto ν, but it is not optimal. **(b)** A continuous example of the Monge problem (2). μ and ν are uniform distributions on the blue and green ellipsoids respectively. The optimal transport map T is defined for any point in the support of μ, and we see how it transports some points onto ν's support with the arrows.

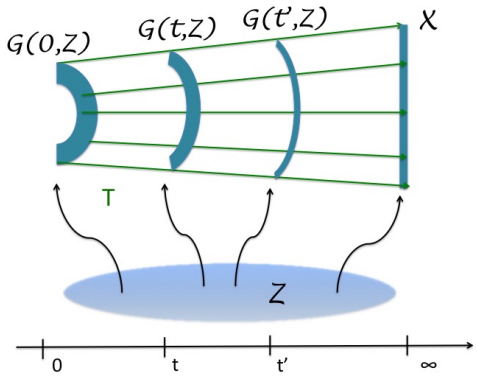


Figure 3: The time evolving generated distribution minimizing its Wasserstein distance with the true distribution X. The latent space is fixed and we denote it as Z. Green arrows give the shape of the optimal transport map T between the initial distribution $G(0, Z)$ and the true distribution X. During training, $G(t, Z)$ does not follow an arbitrary path for converging toward X. It follows the Wasserstein 2 geodesic between $G(0, Z)$ and X described by T.