On the Existence of Optimal Transport Gradient for Learning Generative Models

Antoine Houdard, Arthur Leclaire, Nicolas Papadakis, Julien Rabin

On the Existence of Optimal Transport Gradient for Learning Generative Models

Antoine Houdard[1], Arthur Leclaire[1], Nicolas Papadakis[1], and Julien Rabin[2]

[1] *Univ. Bordeaux, CNRS, IMB, UMR 5251, France*
[2] *Normandie Univ., UniCaen, ENSICAEN, CNRS, GREYC, UMR 6072, France*

Abstract

The use of optimal transport cost for learning generative models has become popular with Wasserstein Generative Adversarial Networks (WGAN). Training of WGAN relies on a theoretical background: the calculation of the gradient of the optimal transport cost with respect to the generative model parameters. We first demonstrate that such gradient may not be defined, which can result in numerical instabilities during gradient-based optimization. We address this issue by stating a valid differentiation theorem in the case of entropic regularized transport and specify conditions under which existence is ensured. By exploiting the discrete nature of empirical data, we formulate the gradient in a semi-discrete setting and propose an algorithm for the optimization of the generative model parameters. Finally, we illustrate numerically the advantage of the proposed framework.

1 Introduction

Generative models are efficient tools to synthesize plausible samples that look similar to a given data distribution. With the emergence of deep neural networks, generative models such as Variational AutoEncoders [17] or Generative Adversarial Networks (GAN) [12] now provide state-of-the-art results for most of machine learning methods dedicated to restoration and edition of signal, image and video data.

Wasserstein GAN. A popular instance of the GAN framework is given by Wasserstein GAN, or WGAN [2]. The generative model of a WGAN is trained to provide a synthetic distribution that is close to a given data distribution with respect to an optimal transport cost, e.g. the 1-Wassertein distance as in [2].

Several improvements and extensions of the original WGAN framework have been proposed in the literature. Most works have taken advantage of the

particular case of 1-Wassersein distance. Following Rubinstein-Kantorovitch duality, the 1-Wassertein distance can be approximated using a 1-Lipschitz discriminator network. The weight clipping strategy, originally suggested in [2] to obtain a Lipschitz network, leads to convergence issues when training a WGAN. Many technical improvements have therefore been proposed to both enforce the Lipschitz constraint of the discriminator network and stabilize the training [13, 23, 21].

Other extensions come from the generalization to p-Wasserstein distances or regularized optimal transport costs. Among these extensions, the method of [19] considers generic convex costs for optimal transport and relies on low dimentional discrete transport problems on batches during the learning. In [18], the case of the 2-Wasserstein distance is tackled thanks to input convex neural networks [1] and a cycle-consistency regularization. In order to have a differentiable distance, the use of entropic regularization of optimal transport has been proposed in different ways. The Sinkhorn algorihm [4] is plugged to compute regularized optimal transport between minibatches in [11], while a regularized WGAN loss function is considered in [24]. The semi-discrete formulation of optimal transport has also been exploited to propose generative models based on 2-Wasserstein distance [14] or strictly convex costs [3].

Differentiation of optimal transport in WGAN training. During the training of all the aforementioned WGAN methods, an optimal transport cost is minimized with respect to the generator parameters. Using Fenchel's duality, the parameter estimation problem is reformulated as a min-max problem. Parameter estimation is then performed with an alternating procedure that requires the gradient expression of the optimal transport cost. The computation of such a gradient involves the differentiation of a maximum and the use of an envelop theorem. *However, we argue that this envelop theorem may not stand in the general case.* More importantly, failure cases can occur even under the theoretical assumptions made in [2].

In this work, we therefore propose an in-depth analysis of the gradient computation for optimal transport through the study of failure cases and their practical consequences. We demonstrate that a stronger envelop theorem holds in the case of entropic regularization of optimal transport. Finally, exploiting the discrete nature of the training data (corresponding to most practical cases), we single out the semi-discrete setting of optimal transport. In this setting, we derive an algorithm for the optimization of the generative model parameters. Contrary to previous works as [24], we pay particular attention to the possible singularities that may prevent us from using the usual envelope theorem, and examine the impact of such irregularities in the learning process.

Problem statement. Considering the empirical measure ν associated with a discrete dataset $\{y_1, \dots, y_n\}$ in a compact $\mathcal{Y}$, we aim at inferring a (continuous) generative model $g_\theta : \mathcal{Z} \to \mathcal{X}$ (defined from a latent space $\mathcal{Z}$ to a compact $\mathcal{X}$ and depending on parameter θ) whose output distribution μ_θ best fits ν. The

function g_θ pushes a distribution ζ on the latent space $\mathcal{Z}$ so that μ_θ is the push-forward measure $\mu_\theta = g_\theta \sharp \zeta$ (defined by $g_\theta \sharp \zeta(B) = \zeta(g_\theta^{-1}(B))$). The goal is thus to compute a parameter θ that minimizes the optimal transport cost

$$\mathrm{OT}_c(\mu_\theta, \nu) = \inf_{\pi \in \Pi(\mu_\theta, \nu)} \int c(x, y) d\pi(x, y), \tag{1}$$

where $c : \mathcal{X} \times \mathcal{Y} \to \mathbf{R}$ is a Lipschitz cost function and $\Pi(\mu_\theta, \nu)$ is the set of probability distributions on $\mathcal{X} \times \mathcal{Y}$ having marginals μ_θ and ν. The direct minimization of $\mathrm{OT}_c(\mu_\theta, \nu)$ with respect to θ is a difficult task. However, whenever $\mathcal{X}$ and $\mathcal{Y}$ are compact, duality holds and the so-called semi-dual formulation of optimal transport yields [25]

$$\mathrm{OT}_c(\mu_\theta, \nu) = \max_{\psi \in L^\infty(\mathcal{Y})} \int_{\mathcal{X}} \psi^c(x) d\mu_\theta(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y), \tag{2}$$

where $\psi \in L^\infty(\mathcal{Y})$ is called a *dual potential* and

$$\psi^c(x) = \min_{y \in \mathcal{Y}} \left[c(x, y) - \psi(y) \right] \tag{3}$$

is the c-transform of ψ. Any dual potential satisfying the max in (2) will be referred to as a *Kantorovitch potential*. Denoting as

$$F(\psi, \theta) = \int_{\mathcal{X}} \psi^c(x) d\mu_\theta(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y),$$

the differentiation with respect to θ of the optimal transport cost $\mathrm{OT}_c(\mu_\theta, \nu)$ can be related to the differentiation under the maximum of the quantity

$$W_c(\theta) = \max_{\psi \in L^\infty(\mathcal{Y})} F(\psi, \theta). \tag{4}$$

We now introduce the following result from [2] that will be used all along the paper.

Theorem 1 (Envelop theorem)**.** *Let θ_0 and ψ_0^* verifying $W_c(\theta_0) = F(\psi_0^*, \theta_0)$. If W_c and $\theta \mapsto F(\psi_0^*, \theta)$ are both differentiable at θ_0, then*

$$\nabla W_c(\theta_0) = \nabla_\theta F(\psi_0^*, \theta_0). \tag{5}$$

The proof of this *weak* version of the envelop theorem is straightforward.

Proof. Let θ_0 and ψ_0 be as in the hypothesis of the theorem. Let us define

$$H : \theta \mapsto F(\psi_0^*, \theta) - W(\theta). \tag{6}$$

For all θ, $H(\theta) \leq 0$ from the definition of W and $H(\theta_0) = 0$ from definition of ψ_0. Since we assume H differentiable at θ_0, we get $\nabla H(\theta_0) = 0$ and the result follows. □

However, there may exist no couple (θ_0, ψ_0^*) for which (5) holds, even in cases that would seem favorable. This scenario can indeed occur for W_c being differentiable everywhere, or for F admitting partial derivative in θ for almost every θ and ψ.

Outline. The main objective of this paper is to identify what may prevent the existence of the right-hand side quantity $\nabla_\theta F(\psi_0, \theta_0)$ in (5). In Section 2, we study a simple counterexample where $F(\psi_0^*, \cdot)$ is never differentiable at θ_0 and we discuss the impact of this property on the actual Wasserstein GAN procedure. We then demonstrate in Section 3 that the entropic regularization of optimal transport allows the application of a stronger version of the envelop theorem and ensures the differentiability of the regularized optimal transport cost. The formulation of the gradient is nevertheless difficult to exploit in practice, as it requires to estimate an expectation over the whole target distribution ν. We therefore propose in Section 4 to take advantage of the discrete nature of the target data $\{y_1, \dots, y_n\}$ to derive a feasible algorithm in the semi-discrete setting of optimal transport. We finally illustrate the benefits of the proposed framework through numerical experiments in Section 4.2.

2 Case study on a synthetic example

In this section, we present an example that brings down the theoretical assumption made in [2]. In Section 2.1 we demonstrate that the envelop Theorem 1 does not hold for this example. More precisely, we show that even if the gradient of the optimal transport cost with respect to the parameter exists for any θ, the gradient of the function F does not exist for any θ. In Section 2.2, we emphasize that for this example, the generative model satisfies the hypothesis made in [2] and show that it can lead to convergence instabilities during the training.

2.1 An example with discrete measures

Let us consider a simple optimal transport problem between a Dirac δ_θ located at $\theta \in \mathbf{R}^2$ and a sum of two Diracs at positions $y_1 \neq y_2 \in \mathbf{R}^2$ (see Figure 2). This setting corresponds to a latent code $\zeta = \delta_0$ and a generator g_θ defined by $g_\theta(z) = z - \theta$ for all $z \in \mathcal{Z}$. We show that in this case the hypotheses of Theorem 1 are satisfied for no θ_0. More precisely we show the following result.

Proposition 1. *Let $\mu_\theta = g_\theta \sharp \zeta = \delta_\theta$ and $\nu = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}$ and consider a cost $c(x, y) = \|x - y\|^p$, $p \geq 1$ then*

- *$\theta \mapsto W_c(\theta) = \mathrm{OT}_c(\mu_\theta, \nu)$ is differentiable at any $\theta \notin \{y_1, y_2\}$ for $p = 1$, and at any θ for $p > 1$,*
- *for any θ_0 and any $\psi_0^* \in \mathrm{argmax}_\psi F(\psi, \theta_0)$, the function $\theta \mapsto F(\psi_0^*, \theta)$ is **not** differentiable at θ_0.*

Hence relation (5) never stands.

Proof. The dual formulation of optimal transport writes

$$OT_c(\mu_\theta, \nu) = \max_{\psi \in \mathbf{R}^2} F(\psi, \theta) \tag{7}$$

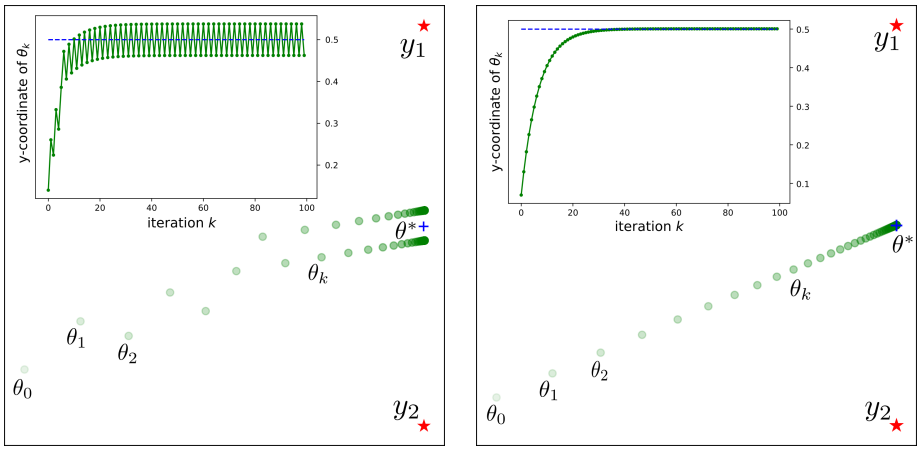

Figure 1: Plot of the trajectory of the parameter θ^k during optimization of the generative model $g_\theta(z) = z - \theta$ for two training points $\{y_1, y_2\}$. Left: as predicted by Proposition 1, the process does not converge for the optimal transport OT_c with quadratic cost $c = \|.\|^2$. Right: as supported by Theorem 5, the training converges to the solution $\theta^* = (0, 0.5)$ when considering regularized optimal transport OT_c^λ. See Section 3.4 for more details.

with $F(\psi, \theta) = \min_{i=1,2} [c(\theta, y_i) - \psi_i] + \frac{\psi_1 + \psi_2}{2}$. We can therefore write

$$F(\psi, \theta) = \begin{cases} c(\theta, y_1) + \frac{\psi_2 - \psi_1}{2} & \text{if } c(\theta, y_1) - \psi_1 \leq c(\theta, y_2) - \psi_2 \\ c(\theta, y_2) + \frac{\psi_1 - \psi_2}{2} & \text{if } c(\theta, y_2) - \psi_2 \leq c(\theta, y_1) - \psi_1. \end{cases}$$

Which yields $2F(\theta, \psi) \leq c(\theta, y_1) + c(\theta, y_2)$, where equality is reached for $\psi_1 = c(\theta, y_1)$ and $\psi_2 = c(\theta, y_2)$. As a consequence, the optimal transport cost writes

$$\mathrm{OT}_c(\mu_\theta, \nu) = \frac{c(\theta, y_1) + c(\theta, y_2)}{2}, \tag{8}$$

and it is differentiable at any $\theta \notin \{y_1, y_2\}$, and even at $\theta \in \{y_1, y_2\}$ for $p > 1$. This shows the first point of the proposition.

Let us now fix θ_0. One can show that

$$\operatorname*{argmax}_\psi F(\psi, \theta_0) = \{(c(\theta_0, y_1), c(\theta_0, y_2)) + C, C \in \mathbf{R}\}.$$

Next we fix $\psi^* \in \operatorname{argmax}_\psi F(\psi, \theta_0)$, so that $\psi_2^* - \psi_1^* = c(\theta_0, y_2) - c(\theta_0, y_1)$. We define the Laguerre cells for ψ^*

$$L_i(\psi^*) = \{x \mid \forall k \neq i,\ c(x, y_i) - \psi_i^* < c(x, y_k) - \psi_k^*\}$$

as well as the boundary set

$$H(\psi^*) = \{x \mid c(x, y_1) - \psi_1^* = c(x, y_2) - \psi_2^*\}.$$

The function F thus writes

$$F(\psi, x) = \begin{cases} c(x, y_1) + \frac{\psi_2^* - \psi_1^*}{2} & \text{if } x \in L_1(\psi^*) \\ c(x, y_2) + \frac{\psi_1^* - \psi_2^*}{2} & \text{if } x \in L_2(\psi^*) \\ c(x, y_1) + \frac{\psi_2^* - \psi_1^*}{2} = c(x, y_2) + \frac{\psi_1^* - \psi_2^*}{2} & \text{if } x \in H(\psi^*) \end{cases}$$

Notice that both Laguerre cells $L_1(\psi^*)$ and $L_2(\psi^*)$ are open sets. The map $x \mapsto F(\psi^*, x)$ is differentiable on these cells and we have

$$\forall x \in L_i(\psi^*),\ \nabla_x F(\psi^*, x) = \nabla_x c(x, y_i). \tag{9}$$

Therefore, in the neighborhood of a point $x \in H(\psi^*)$, the restrictions of $x \mapsto F(\psi^*, x)$ on $L_1(\psi^*)$ and $L_2(\psi^*)$ are smooth but their gradients do not agree at the boundary in-between. As a consequence, $\theta \mapsto F(\psi^*, \theta)$ is not differentiable at any $\theta_0 \in H(\psi^*)$ which shows the second point of the proposition. □

2.2 Consequences on the convergence of WGAN

Wasserstein GAN approaches exploit the gradient $\nabla_\theta F$ in (5) to train a generative model g_θ. Relation (5) is stated in [2] under the following hypothesis relying on the Lipschitz properties of the generator.

Hypothesis 1. *$g : \Theta \times \mathcal{Z} \to \mathcal{X}$ satisfies Hypothesis 1 if there exists $L : \Theta \times \mathcal{Z} \to \mathbf{R}_+$ such that $\forall \theta$, there exists a neighborhood Ω of θ such that $\forall \theta' \in \Omega$ and $\forall z, z' \in \mathcal{Z}$*

$$\|g(\theta, z) - g(\theta', z')\| \leq L(\theta, z)\|(\theta, z) - (\theta', z')\| \tag{10}$$

$$\text{and for all } \theta \qquad \mathbf{E}_{Z \sim \zeta}\left[L(\theta, Z)\right] := L(\theta) < \infty. \tag{11}$$

In our previous example of Section 2.1, we demonstrate that the envelop theorem does not hold for the generator $g_\theta(z) = z - \theta$. This generator nevertheless satisfies Hypothesis 1 as it is locally Lipschitz with constant $L(\theta, z) = 1$. Hence, **the example we designed raises a flaw of Theorem 3 in** [2], although this result is used to justify the existence of gradients for the training of most of WGAN-based methods.

Let us now emphasize an important point. The envelop theorem used in [2] requires both terms of (5) to exist simultaneously. In fact, the differentiability of $F(\psi, \cdot)$ for a given ψ is only guaranteed at almost every θ. Therefore, if we consider the optimal potential ψ_0^* corresponding to a given θ_0 the function $F(\psi_0^*, \cdot)$ may not be differentiable at θ_0. This is exactly what happens in our synthetic example.

Such issue occurs in any case where the generated distribution is not absolutely continuous with respect to the Lebesgue measure. This is typically the case when the latent space $\mathcal{Z}$ of a generative model is lower dimensional than the ambient data space $\mathcal{X}$.

3 Gradient of the regularized optimal transport cost

In this section, we propose to exploit the differential properties of the entropic regularization of optimal transport, in order to have a well-posed training procedure for WGAN. More precisely, we show that under some assumptions on the generator g_θ, we can compute the gradient of the regularized optimal transport cost between the generated measure μ_θ and the target measure ν. In Section 3.1, we first recall results on the entropic regularization of optimal transport. In particular, we formulate the regularized optimal transport cost as the maximum of a function $F_c^\lambda(\psi, \theta)$ over ψ. Then, we show in Section 3.2 the differentiability of F_c^λ with respect to θ, under the Hypothesis 1 on the generator g_θ. In Section 3.3 we demonstrate the differentiability of the optimal transport cost under different hypotheses, relating the gradient of the optimal transport cost to the gradient of F_c^λ. In Section 3.4 we finally apply these results to the synthetic example from Section 2.1.

3.1 Definitions and requirements

We consider from now on the entropic regularization of optimal transport.

Definition 1 (OT cost with entropic regularization [9])**.** *For $\lambda > 0$, the regularized OT cost is defined by*

$$\mathrm{OT}_c^\lambda(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int c(x, y) d\pi(x, y) + \lambda H(\pi | \mu \otimes \nu) \tag{12}$$

where

$$H(\pi | \mu \otimes \nu) = \int \left(\log \left(\frac{d\pi(x, y)}{d(\mu(x)\nu(y))} \right) - 1 \right) d\pi(x, y) + 1 \tag{13}$$

is the relative entropy of the transport plan π w.r.t. the product measure $\mu \otimes \nu$.

Note that $\lambda > 0$ makes the problem (12) strictly convex, whereas $\lambda = 0$ corresponds to the non-regularized case from (1). In the following, we may either refer to OT_c or OT_c^0. We consider the relative entropy as in [9] which differs from the regularized formulation originally introduced in [4]. As shown in [9], the dual formulation of the problem with relative entropy can be expressed as the maximum of an expectation. In order to write the semi-dual formulation as in (2), we introduce the regularized c, λ-transform:

Definition 2 (regularized c, λ-transforms)**.** *Let $\psi : \mathcal{Y} \to \mathbf{R}$ we define the c, λ-transform for $\lambda > 0$ by*

$$\psi^{c,\lambda}(x) = -\lambda \log \left(\int_{\mathcal{Y}} \exp \left(\frac{\psi(y) - c(x, y)}{\lambda} \right) d\nu(y) \right). \tag{14}$$

This formula with $\lambda = 0$ yields the c-transform ψ^c introduced earlier. The semi-dual formulation then writes.

Proposition 2 (Semi-dual formulation of regularized transport)**.** *The primal problem* (12) *is equivalent to the semi-dual problem*

$$\mathrm{OT}_c^\lambda(\mu,\nu) = \max_{\psi\in L^\infty(\mathcal{Y})} \int_{\mathcal{X}} \psi^{c,\lambda}(x)d\mu(x) + \int_{\mathcal{Y}} \psi(y)d\nu(y). \tag{15}$$

We then recall the following theorem that will be used in the next section.

Theorem 2 (Existence and uniqueness of the dual solution [9])**.** *Providing $c \in L^\infty(\mathcal{X}\times\mathcal{Y})$, the dual problem* (15) *admits a solution $\psi^* \in L^\infty(\nu)$ which is unique $\nu-$a.e. up to an additive constant. Any solution ψ^* will be referred to as a Kantorovitch potential.*

As in the un-regularized case, we consider $\mu_\theta = g_\theta\sharp\zeta$ defined through a generative model. Defining

$$F_c^\lambda(\theta,\psi) = \mathbf{E}_{Z\sim\zeta}[\psi^{c,\lambda}(g_\theta(Z))] + \mathbf{E}_{Y\sim\nu}[\psi(Y)], \tag{16}$$

our main goal is to differentiate with respect to the parameter θ the quantity

$$W_c^\lambda(\theta) = \mathrm{OT}_c^\lambda(\mu_\theta,\nu) = \max_{\psi\in L^\infty(\mathcal{Y})} F_c^\lambda(\theta,\psi). \tag{17}$$

Before going further let us state the following Lemma about regularity of the Kantorovitch potentials

Lemma 1. *If c is uniformly continuous on $\mathcal{X}\times\mathcal{Y}$, there exists a function $\omega : \mathbf{R}_+ \to \mathbf{R}_+$ not-decreasing, continuous at 0 with $\omega(0) = 0$ called* modulus of continuity *of c such that $\forall x, x' \in \mathcal{Y}$ and $\forall y, y' \in \mathcal{Y}$,*

$$|c(x,y) - c(x',y')| \leq \omega(\|x-x'\| + \|y-y'\|).$$

In this case, any Kantorovicth potential shares this same modulus of continuity.

This lemma contains two different points to demonstrate. The first one can be shown using a standard analysis result stating that any uniformly continuous function admits a modulus of continuity. The second one is specific to optimal transport theory. Let us demonstrate the two propositions separately.

Proposition 3. *Let $(\mathcal{M}, |\cdot|)$ be a metric space and $f : \mathcal{M} \to \mathbf{R}$ be a uniformly continuous function. Then f admits a modulus of continuity $\omega : \mathbf{R}_+ \to \mathbf{R}_+$ conitnuous at 0 with $\omega(0) = 0$.*

Proof. From uniform continuity, let $\delta > 0$ be such that $\|a-b\| < \delta$ implies $|f(a) - f(b)| < 1$. Let us take ω defined for all $t \leq \delta$ by $\omega(t) = \sup\{|f(a)-f(b)| \mid \|a-b\| \leq t\}$ and $\omega(t) = 2\sup_{x\in\mathcal{M}}\|f(x)\|$ whenever $t > \delta$. From this definition, one obtains ω positive, non-decreasing and the fact that $\forall a, b \in \mathcal{M}$, $|f(a)-f(b)| \leq \omega(\|a-b\|)$ and $\omega(0) = 0$. Let us now show that ω is continuous at 0, that is to say that $\omega(t) \to 0$ when $t \to 0$. Since ω is positive and monotone, the limit exists and we denote $l = \lim_{t\to 0}\omega(t)$. Let $\varepsilon > 0$, the uniform continuity yields that there exists $\eta > 0$ such that $\|a-b\| < \eta$ implies $|f(a)-f(b)| < \varepsilon$. Then for any $t < \min(\delta,\eta)$ we have $\omega(t) \leq \varepsilon$ and therefore $l \leq \varepsilon$. This shows that $l = 0$ and concludes the proof. □

Proposition 4. *Let ω be a modulus of continuity of the cost c, then any Kantorovitch potential of OT_c^λ shares the same modulus of continuity as c.*

Proof. Let ψ be a solution of the semi-dual problem of entropic regularized optimal transport (Eq. (13) in the main paper). There exists $\varphi \in L^\infty(\mathcal{X})$ such that $\psi = \varphi^{c,\lambda}$. Then from the definition of ω in Lemma 1, we have for all $x \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$:

$$\varphi(x) - c(x,y) = \varphi(x) - c(x,y) + c(x,y') - c(x,y') \leq \omega(\|y - y'\|) + \varphi(x) - c(x,y').$$

This leads to

$$\exp\left(\frac{\varphi(x) - c(x,y)}{\lambda}\right) \leq \exp\left(\frac{\omega(\|y-y'\|) + \varphi(x) - c(x,y')}{\lambda}\right).$$

Taking the expectation for $x \sim \mu$ yields

$$\mathbf{E}_{X\sim\mu}\left[\exp\left(\tfrac{\varphi(X)-c(X,y)}{\lambda}\right)\right] \leq \exp\left(\tfrac{\omega(\|y-y'\|)}{\lambda}\right)\mathbf{E}_{X\sim\mu}\left[\exp\left(\tfrac{\varphi(X)-c(X,y')}{\lambda}\right)\right].$$

Finally applying $-\lambda \log$, we get

$$\varphi^{c,\lambda}(y') - \omega(\|y - y'\| \leq \varphi^{c,\lambda}(y).$$

Switching the roles of $\varphi^{c,\lambda}(y)$ and $\varphi^{c,\lambda}(y')$, we finally obtain the desired result:

$$|\varphi^{c,\lambda}(y') - \varphi^{c,\lambda}(y)| \leq \omega(\|y - y'\|).$$

□

Finally, the following theorem permits to control the variations of W_c^λ with the variations of $\theta \to g_\theta$

Theorem 3. *If $c \in \mathscr{C}^1(\mathcal{X} \times \mathcal{Y})$ and if $\mathcal{X}$ and $\mathcal{Y}$ are compact, there exists κ such that c is κ-Lipschitz and for all θ_1, θ_2*

$$|W_c^\lambda(\theta_1) - W_c^\lambda(\theta_2)| \leq \kappa \mathbf{E}_{Z\sim\zeta}\left[\|g_{\theta_1}(Z) - g_{\theta_2}(Z)\|\right]. \tag{18}$$

Proof. As c is a $\mathscr{C}^1$ function on a compact set, it is κ-Lipschitz with $\kappa = \sup_{\mathcal{X}\times\mathcal{Y}} \|Dc(x,y)\|$, where $Dc(x,y)$ is the differential of c at (x,y). Let us first prove that for all θ_1, θ_2

$$|W_c^\lambda(\theta_1) - W_c^\lambda(\theta_2)| \leq \kappa W_1(\mu_{\theta_1}, \mu_{\theta_2}), \tag{19}$$

where $W_1(\mu_{\theta_1}, \mu_{\theta_2})$ is the 1-Wasserstein distance between μ_{θ_1} and μ_{θ_2} (i.e. the Wasserstein distance associated with the cost $(x,y) \mapsto \|x - y\|$). Taking ψ_1 a Kantorovitch potential for $W_c^\lambda(\theta_1)$ and ψ_2 a Kantorovitch potential for $W_c^\lambda(\theta_2)$, we can write

$$\begin{aligned}
&W_c^\lambda(\theta_1) - W_c^\lambda(\theta_2) \\
&\quad = \mathbf{E}_{\mu_{\theta_1}}\left[\psi_1^{c,\lambda}(X)\right] + \mathbf{E}_\nu\left[\psi_1(Y)\right] - \mathbf{E}_{\mu_{\theta_2}}\left[\psi_2^{c,\lambda}(X)\right] - \mathbf{E}_\nu\left[\psi_2(Y)\right] \\
&\quad = \mathbf{E}_{\mu_{\theta_1}}\left[\psi_1^{c,\lambda}(X)\right] - \mathbf{E}_{\mu_{\theta_2}}\left[\psi_1^{c,\lambda}(X)\right] \\
&\quad + \left(\mathbf{E}_{\mu_{\theta_2}}\left[\psi_1^{c,\lambda}(X)\right] + \mathbf{E}_\nu\left[\psi_1(Y)\right] - \mathbf{E}_{\mu_{\theta_2}}\left[\psi_2^{c,\lambda}(X)\right] - \mathbf{E}_\nu\left[\psi_2(Y)\right]\right).
\end{aligned}$$

By optimality of ψ_2 for $W_c^\lambda(\theta_2)$, the sum of terms in the last parenthesis is non-positive. By switching θ_1 and θ_2 in the previous formula, we get

$$\mathbf{E}_{\mu_{\theta_1}}\left[\psi_2^{c,\lambda}(X)\right] - \mathbf{E}_{\mu_{\theta_2}}\left[\psi_2^{c,\lambda}(X)\right] \leq W_c^\lambda(\theta_1) - W_c^\lambda(\theta_2) \leq \mathbf{E}_{\mu_{\theta_1}}\left[\psi_1^{c,\lambda}(X)\right] - \mathbf{E}_{\mu_{\theta_2}}\left[\psi_1^{c,\lambda}(X)\right]$$

and thus

$$|W_c^\lambda(\theta_1) - W_c^\lambda(\theta_2)| \leq \sup_{\psi \in L^\infty(\mathcal{Y})} \left|\mathbf{E}_{\mu_{\theta_1}}\left[\psi^{c,\lambda}(X)\right] - \mathbf{E}_{\mu_{\theta_2}}\left[\psi^{c,\lambda}(X)\right]\right| \tag{20}$$

From Proposition 2 all the $\psi^{c,\lambda}$ shares the modulus of continuity of the cost c, which is κ-Lipschitz and noted $\psi^{c,\lambda} \in Lip_\kappa$. Hence we can restrict the supremum to κ-Lipschitz functions and we get that

$$|W_c^\lambda(\theta_1) - W_c^\lambda(\theta_2)| \leq \sup_{\varphi \in Lip_\kappa} \left|\mathbf{E}_{\mu_{\theta_1}}\left[\varphi(X)\right] - \mathbf{E}_{\mu_{\theta_2}}\left[\varphi(X)\right]\right| \tag{21}$$

$$\leq \kappa \sup_{\varphi \in Lip_1} \left|\mathbf{E}_{\mu_{\theta_1}}\left[\varphi(X)\right] - \mathbf{E}_{\mu_{\theta_2}}\left[\varphi(X)\right]\right| = \kappa W_1(\mu_{\theta_1}, \mu_{\theta_2}). \tag{22}$$

Then using the transport plan $\gamma = (g_{\theta_1}\sharp\zeta, g_{\theta_2}\sharp\zeta)$, which is admissible for $W_1(\mu_{\theta_1}, \mu_{\theta_2})$, we get

$$W_1(\mu_{\theta_1}, \mu_{\theta_2}) \leq \mathbf{E}_{Z\sim\zeta}\left[\|g_{\theta_1}(Z) - g_{\theta_2}(Z)\|\right],$$

which gives the result. □

Hence a Lipschitz regularity on g implies the same regularity on W_c^λ. In view of establishing the differentiability of W_c^λ, we first study the differentiability of $F_c^\lambda(\theta, \psi)$ with respect to θ in the next section.

3.2 Regularity of F_c^λ

Before coming to our object of interest, the gradient of $W_c^\lambda(\theta)$ defined in (17) with respect to the generator parameters θ, we first study the regularity of F_c^λ defined by (14). Compared to the un-regularized case, the entropic regularization changes the minimum in the c-transform (3) into a soft-minimum in the c, λ-transform (14). With this additional regularity property, we show the following differentiability theorem on the functional F_c^λ.

Theorem 4. *Let $c \in \mathscr{C}^1(\mathcal{X} \times \mathcal{Y})$ for $\mathcal{X}$ and $\mathcal{Y}$ compact and $g : \Theta \times \mathcal{Z}$ satisfying Hypothesis 1. Then for almost every θ_0, for any $\psi \in L^\infty$ and $\lambda > 0$ the function $\theta \mapsto F_c^\lambda(\theta, \psi)$ in (16) is differentiable at θ_0 and*

$$\nabla_\theta F_c^\lambda(\theta_0, \psi) = \mathbf{E}_{Z\sim\zeta}\left[\left(\partial_\theta g(\theta_0, Z)\right)^T \nabla\psi^{c,\lambda}(g(\theta_0, Z))\right]. \tag{23}$$

Moreover, if g is also $\mathscr{C}^1$, then $F_c^\lambda(\cdot, \psi)$ is $\mathscr{C}^1$ on Θ.

Proof. Let $\psi \in L^\infty$. The cost c being $\mathscr{C}^1$ and $\mathcal{Y}$ a compact, the dominated convergence theorem gives that the c, λ-transform $\psi^{c,\lambda}$ (14) is $\mathscr{C}^1$ and

$$\nabla \psi^{c,\lambda}(x) = \frac{\int_{\mathcal{Y}} \exp\left(\frac{\psi(y)-c(x,y)}{\lambda}\right) \nabla_x c(x,y) d\nu(y)}{\int_{\mathcal{Y}} \exp\left(\frac{\psi(y)-c(x,y)}{\lambda}\right) d\nu(y)}. \tag{24}$$

From Hypothesis 1, g is differentiable at a.e. (θ, z). This implies that for almost every θ, g is differentiable at (θ, z) for a.e. z (and therefore admits a partial differential w.r.t. θ). We set such a θ_0. Let $p(\theta, z) := \psi^{c,\lambda}(g(\theta, z))$ so that

$$F_c^\lambda(\theta, \psi) = \mathbf{E}_{Z\sim\zeta}[p(\theta, Z)] + \int \psi d\nu. \tag{25}$$

For any ψ, since $\psi^{c,\lambda}$ is differentiable everywhere, there exists a neighborhood Ω_1 of θ_0 such that, for any $\theta \in \Omega_1$, $p(\theta, z)$ admits a partial differential at θ for almost every z. This partial differential writes

$$q(\theta, z) = \partial_\theta(\psi^{c,\lambda}(g(\theta, z))) = (\partial_\theta g(\theta, z))^T \nabla \psi^{c,\lambda}(g(\theta, z)). \tag{26}$$

Since g satisfies Hypothesis 1, there also exists a neighborhood Ω_2 of θ_0 such that, for all $\theta \in \Omega_2$ and any $z \in \mathcal{Z}$, $\|\partial_\theta g(\theta, z)\| \leq L(\theta_0, z)$. Since $\mathcal{X}$ is compact, we have $C := \sup_{\mathcal{X}} \|\nabla \psi^{c,\lambda}\| < \infty$. Thus we get $\|q(\theta, z)\| \leq CL(\theta_0, z)$ with $\mathbf{E}[L(\theta_0, Z)] < \infty$. Besides, for $\theta \in \Omega_1 \cap \Omega_2$ and any $z \in \mathcal{Z}$,

$$\frac{\|p(\theta, z) - p(\theta_0, z) - \langle q(\theta_0, z), \theta - \theta_0\rangle\|}{\|\theta - \theta_0\|} \tag{27}$$

$$\leq \sup_{\theta \in \Omega_1 \cap \Omega_2} \|\partial_\theta g(\theta, z)\| C + \|q(\theta_0, z)\| \tag{28}$$

$$\leq 2L(\theta_0, z)C < \infty \tag{29}$$

We can thus apply the dominated convergence theorem that yields that $\theta \mapsto F_c^\lambda(\theta, \psi)$ is differentiable at θ_0 with $\nabla_\theta F_c^\lambda(\theta_0, \psi) = \mathbf{E}[q(\theta_0, Z)]$, which is the desired formula.

When g is $\mathscr{C}^1$, this is true for any $\theta_0 \in \Theta$ and since $\psi^{c,\lambda}$ is also $\mathscr{C}^1$, we get that $F_c^\lambda(\cdot, \psi)$ is also $\mathscr{C}^1$. □

Discussion. Thanks to the entropic regularization, Theorem 4 is valid for almost every θ and for **any** ψ. This is an essential point for latter purpose, as it ensures that the result stays true for a couple (θ, ψ^*) where ψ^* is a Kantorovitch potential for θ. In the un-regularized case $\lambda = 0$, the proof of Theorem 4 does not hold and additional assumptions on g are required. Contrary to the c, λ-transform $\psi^{c,\lambda}$, the c-transform (3) is indeed only differentiable almost everywhere. In this case, we only have that for any ψ, $\theta \mapsto F_c^0(\theta, \psi)$ is differentiable for almost every θ, with the 'almost every θ' depending on ψ. Notice that this is the actual result proved in [2].

Another point to discuss is the $\mathscr{C}^1$ assumption on the cost c in Theorem 4. This assumption does not cover the cost $c(x, y) = \|x - y\|$ from the original

WGAN framework. This issue can be treated by taking care of the points where $c(x, y)$ is not differentiable. To do so, we can assume that for almost every θ, the generator $g_\theta \sharp \zeta$ does not put mass on the atoms of the measure ν. This assumption is for instance true if $g_\theta \sharp \zeta$ is absolutely continuous w.r.t. the Lebesgue measure. Hence, in order to keep the our statements as simple as possible, we still rely on the cost assumption $c \in \mathscr{C}^1(\mathcal{X} \times \mathcal{Y})$ in the following.

3.3 Differentiation of W_c^λ

We can now state our main result ensuring the differentiability of $W_c^\lambda(\theta)$ with respect to the generator parameters θ. We first demonstrate this claim for $\mathscr{C}^1$ generators g.

Theorem 5. *Let c be a $\mathscr{C}^1(\mathcal{X} \times \mathcal{Y})$ cost, for $\mathcal{X}$ and $\mathcal{Y}$ compact. If the generator g satisfies Hypothesis 1 with $g \in \mathscr{C}^1(\Theta \times \mathcal{Z}, \mathcal{X})$ then the mapping*

$$W_c^\lambda : \theta \mapsto \mathrm{OT}_c^\lambda(g_\theta \sharp \zeta, \nu) \tag{30}$$

is $\mathscr{C}^1$ and we have, for any $\theta \in \Theta$,

$$\begin{aligned} \nabla_\theta W_c^\lambda(\theta) &= \nabla_\theta F_c^\lambda(\theta, \psi_*) \\ &= \mathbf{E}_{Z \sim \zeta} \left[\left(\partial_\theta g(\theta, z) \right)^T \nabla \psi_*^{c,\lambda}(g(\theta, z)) \right], \end{aligned} \tag{31}$$

where $\psi_* \in \operatorname{argmax}_\psi F_c^\lambda(\psi, \theta)$, and

$$\nabla \psi^{c,\lambda}(x) = \frac{\mathbf{E}_{Y \sim \nu} \left[\exp\left(\frac{\psi(Y) - c(x,Y)}{\lambda} \right) \nabla_x c(x, Y) \right]}{\mathbf{E}_{Y \sim \nu} \left[\exp\left(\frac{\psi(Y) - c(x,Y)}{\lambda} \right) \right]}. \tag{32}$$

The demonstration of this theorem is based on the application of the following result.

Proposition 5 (A.1 from [22])**.** *Let $w(\theta) = \max_\psi h(\psi, \theta)$ and assume that*

1. *there exists $\psi^* : \theta \mapsto \psi^*(\theta)$ defined on a neighborhood of θ_0 s.t. $h(\psi^*(\theta), \theta) = w(\theta)$ and ψ^* continuous at θ_0*
2. *h is differentiable in θ at $(\psi^*(\theta_0), \theta_0)$ and $\nabla_\theta h$ continuous in (ψ, θ) at $(\psi^*(\theta_0), \theta_0)$*

Then w is differentiable at θ_0 and

$$\nabla w(\theta_0) = \nabla_\theta h(\psi^*(\theta_0), \theta_0). \tag{33}$$

We also need the following technical lemma.

Lemma 2. *There exists a continuous selection of Kantorovitch potentials, that is, a function $\psi_* : \Theta \to \mathscr{C}(\mathcal{Y})$ which is continuous (for the uniform norm on $\mathscr{C}(\mathcal{Y})$) and such that, for any θ, $\psi_*(\theta) \in \operatorname{argmax}_\psi F_c^\lambda(\theta, \psi)$.*

Proof. Let $\theta_0 \in \Theta$. First notice that the Kantorovitch potential ψ solving (17) is unique (up to a constant) since $\lambda > 0$ and c is bounded [9]. We set an arbitrary $y_0 \in \mathcal{Y}$. For all $\theta \in \Theta$, let us consider ψ_θ to be the Kantorovitch potential such that $\psi_\theta(y_0) = 0$. Since the cost c is continuous on $\mathcal{X} \times \mathcal{Y}$ compact, it is absolutely continuous. Lemma 1 also provides that the cost has a bounded modulus of continuity ω that is shared with any Kantorovich potential ψ_θ. This implies that the set $\{\psi_\theta\}_{\theta\in\Omega}$ is equicontinuous and that for all θ, ψ_θ is bounded (independently of θ) by $\sup_{y\in\mathcal{Y}} \omega(|y - y_0|)$. The Arzela-Ascoli theorem therefore implies that $\{\psi_\theta\}_{\theta\in\Omega}$ is relatively compact.

Now, by contradiction, assume that $\theta \mapsto \psi_\theta$ is not continuous at a point θ_0. Then there exists $\varepsilon > 0$ and a sequence $(\theta_n) \in \Omega^{\mathbf{N}}$, such that $\theta_n \to \theta_0$ and

$$\|\psi_{\theta_n} - \psi_{\theta_0}\|_\infty > \varepsilon. \tag{34}$$

Since $\{\psi_\theta\}_{\theta\in\Omega}$ is relatively compact, we can extract a subsequence $\theta_{r(n)}$ such that $\psi_{\theta_{r(n)}}$ converges uniformly towards a function f in $\mathscr{C}(\mathcal{Y})$. It follows that $\psi^{c,\lambda}_{\theta_{r(n)}}$ also converges towards $f^{c,\lambda}$. Let us denote $\mu_n = g_{\theta_n}\sharp\zeta$. This measure μ_n weakly converges towards μ_0 and we can therefore write

$$\mathrm{OT}^\lambda_c(\mu_n, \nu) = \int_{\mathcal{X}} \psi^{c,\lambda}_{\theta_{r(n)}}(x) d\mu_n(x) + \int_{\mathcal{Y}} \psi_{\theta_{r(n)}}(y) d\nu(y) \tag{35}$$

$$\underset{n\to\infty}{\longrightarrow} \int_{\mathcal{X}} f^{c,\lambda}(x) d\mu_0(x) + \int_{\mathcal{Y}} f(y) d\nu(y). \tag{36}$$

Since $\mathrm{OT}^\lambda_c(\mu_n, \nu) \to \mathrm{OT}^\lambda_c(\mu_0, \nu)$, we get that f is a Kantorovitch potential for $\mathrm{OT}^\lambda_c(\mu_0, \nu)$. With $f(y_0) = \lim \psi_{\theta_{r(n)}}(y_0) = 0$ and the uniqueness up to a constant of Kantorovitch potentials we get $f = \psi_{\theta_0}$ which gives the contradiction and concludes the proof. □

Proof of Theorem 5. The demonstration follows from the application of Proposition 5. Lemma 2 gives the first hypothesis for applying Proposition 5. Theorem 4 for $g \in \mathscr{C}^1(\Theta \times \mathcal{Z})$ gives the second one and leads to the expression of $\nabla_\theta F^\lambda_c(\theta, \psi)$. □

Case of g not necessarily $\mathscr{C}^1$. When g only satisfies Hypothesis 1, the gradient of F^λ_c is not necessarily $\mathscr{C}^1$ as seen in Section 3.2. Therefore F^λ_c may not satisfy the second hypothesis of Proposition 5, which is required to show the existence of the gradient in Theorem 5. However we can still give a weaker result in this case, following the same sketch of proof as in [2]. We have already stated in Theorem 4 that for any ψ, the gradient of $F^\lambda_c(\psi, \cdot)$ exists for almost every θ. Therefore, we only need to show the existence of the gradient of W^λ_c almost everywhere to ensure that Theorem 1 holds for almost every θ.

Let then demonstrate that W^λ_c is differentiable for almost every θ. Recall that g satisfies Hypothesis 1. From relation (18) in Theorem 3, we get that for

all θ_1, there exists a neighborhood Ω of θ_1 such that for all $\theta_2 \in \Omega$

$$\begin{aligned} |W_c^\lambda(\theta_1) - W_c^\lambda(\theta_2)| &\leq \kappa \mathbf{E}_{Z\sim\zeta}\left[L(\theta_1, Z)\|\theta_1 - \theta_2\|\right] \\ &\leq \kappa L(\theta_1)\|\theta_1 - \theta_2\|. \end{aligned} \tag{37}$$

The function W_c^λ is thus locally Lipschitz and differentiable for almost every θ by Rademacher theorem.

3.4 Back to the synthetic example

We now study the synthetic example from Section 2 within the regularized framework. The regularized optimal transport between $\mu_\theta = \delta_\theta$ and $\nu = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}$ writes in this case $W_c^\lambda(\theta) = \max_{\psi\in\mathbf{R}^2} F_c^\lambda(\psi, \theta)$ with $F_c^\lambda(\psi, \theta) = \frac{\psi_1+\psi_2}{2} - \lambda \log\left(\frac{1}{2}\left(\exp\left(\frac{\psi_1 - c(\theta,y_1)}{\lambda}\right) + \exp\left(\frac{\psi_2 - c(\theta,y_2)}{\lambda}\right)\right)\right)$. The optimal ψ^* maximizing $F_c^\lambda(\psi, \theta)$ satisfies $\psi_1^* - \psi_2^* = c(\theta, y_1) - c(\theta, y_2)$ and as in the unregularized case (8) we obtain

$$\mathrm{OT}_c^\lambda(\mu_\theta, \nu) = \frac{c(\theta, y_1) + c(\theta, y_1)}{2}.$$

Hence both regularized and un-regularized problem share the same solution. In the un-regularized case, Proposition 1 states that the gradient of W_c cannot be related to the gradient of F_c. On the other hand, Theorem 5 stands for the regularized setting W_c^λ.

Let us now highlight the numerical influence of the regularization. Starting from θ^0 and for a given time step $\tau > 0$, we consider the iterative algorithm

$$\begin{cases} \psi^k \in \mathrm{argmax}_\psi F_c^\lambda(\psi, \theta^k) \\ \theta^{k+1} = \theta^k - \tau \nabla_\theta F_c^\lambda(\psi^k, \theta^k). \end{cases} \tag{38}$$

Recall that the gradient of $\nabla_\theta F_c^\lambda(\psi^k, \theta^k)$ does not exist for $\lambda = 0$. In this case, we propose to approximate ψ^k with the gradient ascent procedure proposed in [10] and then obtain the gradient via back-propagation. This corresponds to the alternate procedure in the WGAN framework of [2].

We illustrate the behavior of Algorithm (38), with $\tau = 0.1$, for both un-regularized and regularized settings in Figure 2, where we take $y_1 = (0, 0)$ and $y_2 = (0, 1)$ and the $\mathscr{C}^1$ cost $c(x, y) = ||x-y||^2$ corresponding to the 2-Wasserstein distance. In such a setting, the optimal generator $g_\theta(z) = z - \theta$ is obtained for $\theta = \frac{y_1+y_2}{2} = (0, 0.5)$. In the un-regularized case $\lambda = 0$, the gradient $\nabla_\theta F_c^\lambda(\psi^k, \theta^k)$ alternates between directions $(\theta - y_1)$ and $(\theta - y_2)$. The parameter θ^k thus oscillates around $(0, 0.5)$. On the other hand, for the regularization parameter $\lambda = 0.1$, the gradient is well defined and θ^k converges monotonously towards $(0, 0.5)$.

4 Learning a generative model with regularized optimal transport

We finally come to practical considerations. The gradient formula (31) given in Theorem 5 takes the form of an expectation. In order to minimize the regularized optimal transport cost with respect to the generator parameter θ, we perform a stochastic gradient optimization with $(\partial_\theta g(\theta, z))^T \nabla \psi_*^{c,\lambda}(g(\theta, z))$, the term inside the expectation. Such framework involves two limitations: (i) an optimal Kantorovitch potential $\psi_*^{c,\lambda}$ has to be approximated at each iteration; (ii) the stochastic gradient formula requires the computation of an expectation on the *whole* data distribution ν. The authors of [10] showed that in the case of a discrete target measure ν, the first issue can be addressed with a stochastic gradient ascent on ψ. Moreover, if ν is discrete, the second point amounts to compute a mean over the dataset. This step is thus feasible, as the associated computational cost is linear with the number of data.

When facing concrete applications, the target dataset is actually discrete. We thus propose to formulate the optimal transport in a semi-discrete way in Section 4.1. We then present a numerical procedure in the spirit of the WGAN approach and some numerical examples in Section 4.2.

4.1 Semi-discrete formulation

We consider a finite dataset $\{y_1, \dots, y_n\}$, associated to the discrete target measure $\nu = \frac{1}{n}\sum_i \delta_{y_i}$. In this setting, the *semi-discrete* formulation of the regularized cost W_c^λ is

$$W_c^\lambda(\theta) = \max_{\psi \in \mathbf{R}^n} \mathbf{E}_z\left[\psi^{c,\lambda}(g_\theta(z))\right] + \frac{1}{n}\sum_i \psi_i, \tag{39}$$

with $\psi^{c,\lambda}(x) = -\lambda \log\left(\frac{1}{n}\sum_i \exp\left(\frac{\psi_i - c(x,y_i)}{\lambda}\right)\right)$. From Theorem 5, the gradient of W_c^λ writes

$$\nabla_\theta W_c^\lambda(\theta) = \mathbf{E}_z\left[q(\theta, \psi^*, z)\right], \tag{40}$$

with

$$q(\theta, \psi^*, z) = (\partial_\theta g(\theta, z))^T \sum_i \eta_i(g_\theta(z)) \nabla_\theta c(g_\theta(z), y_i) \tag{41}$$

and

$$\eta_i(x) = \frac{\exp\left(\frac{\psi_i^* - c(x,y_i)}{\lambda}\right)}{\sum_j \exp\left(\frac{\psi_j^* - c(x,y_j)}{\lambda}\right)} \tag{42}$$

This formulation has four benefits: (i) the formulation (39) is known to be concave on ψ and an optimum can be approximated with a stochastic gradient ascent procedure [10]; (ii) the dual potential ψ does not need to be encoded with a neural network; (iii) the formulation holds for any cost c that is $\mathscr{C}^1$; and (iv) the formula (40) provides a stochastic gradient that can be computed numerically.

4.2 Numerical illustrations

We propose in Algorithm 1 a numerical scheme based on a stochastic gradient descent of the regularized optimal transport cost. To perform the stochastic gradient descent on θ, we use the Adam optimizer [16] of the Pytorch library with default parameters, and a learning rate $lr = 10^{-4}$. We train the network for $N = 4000$ iterations. The estimation of ψ is done with $N_\psi = 200$ iterations of the stochastic gradient ascent algorithm of [10] dedicated to entropy regularized optimal transport and we use batches of $K = 100$ samples.

Algorithm 1 Learning Generative Model with stochastic gradient of semi-discrete entropic optimal transport

Inputs: regularization parameter λ, cost function c, sample size K, number of iterations N and N_ψ, training set $\{y_1, \dots y_n\}$
Output: estimated generative model parameter θ^N
Initialisation: $\psi^0 = 0$, random initialization of θ^0
for $k = 1$ **to** N **do**
- Estimate ψ^k with N_ψ iterations of the stochastic gradient ascent method in [10] on a batch of size K
- Draw a batch of K samples z from ζ
- Update θ^{k+1} with a stochastic gradient descent in the direction $q(\theta^k, \psi^k, z)$ given in (41)

end for

We now consider the application of Algorithm 1 to the learning of a generative model on the MNIST dataset with the cost $c(x, y) = \|x - y\|^2$. The generative model we considered in our experiments consists of four fully connected layers starting from a latent variable $z \in \mathcal{Z} \subset \mathbf{R}^{128}$ to the image space $\mathcal{X} \subset \mathbf{R}^{784}$, with intermediate dimensions 256, 512 and 1024. In this setting, the learning of the generator with a GPU Nvidia K40m takes approximately 2 hours. We show some generated digits in Figure 2 for different regularization parameters λ. This experiment shows that the proposed framework is able to learn a complex generative model, provided that the regularization parameter is sufficiently small. With high values of the regularization parameters such as $\lambda = 10^{-1}$, the obtained generator realizes a compromise between all data points and concentrates to a mean image. This can be explained with the gradient expression (41) which, for $\lambda \to \infty$, pushes the generator towards a uniform average of the data points. This also corroborates our observation that for high λ, the generator stabilizes in early iterations.

Last, observe that, as in WGAN approaches, we only use $N_\psi \ll n$ iterations for updating ψ^k. In practice, this number should depend on the dataset size n in order to properly approximate the exact solution of the semi-discrete optimal transport problem. Another point is that ψ depends on n parameters. When n is large, an interesting approximation could be to implicitly represent ψ with a shallow neural network as in [26].

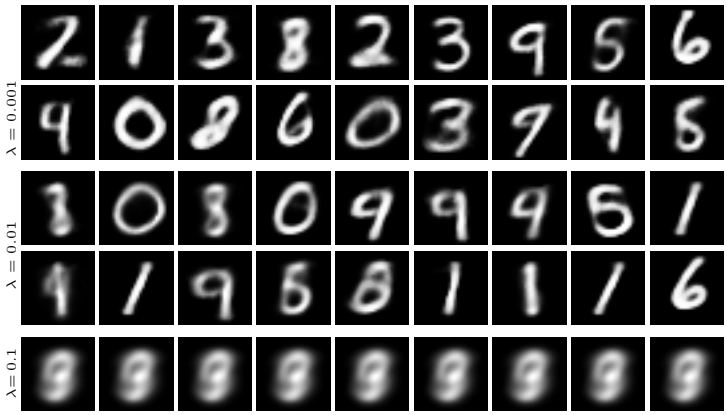

Figure 2: Random samples from generative models learned on the MNIST dataset with Alg. 1, for 3 regularization parameters λ.

5 Conclusion and discussion

We have demonstrated that using optimal transport cost to train generative models, as popularized by the Wasserstein GAN framework [2], raises theoretical issues when computing the gradient, even when assuming strong regularity properties on the generative model itself. This flaw, illustrated on a toy example, can be circumvented by regularizing the optimal transport cost with entropy. The entropic regularization of optimal transport cost [4] indeed enjoys interesting properties, such as an explicit dual formulation, fast computation and robustness to outliers [5]. As illustrated in experiments, and consistently with former works in the literature, the entropic smoothing may however yields an oversmoothed solution [6]. This can be cancelled by compensating the regularized cost bias, as demonstrated in [8] and illustrated in [15]. The resulting Sinkhorn divergence is nevertheless challenging to compute when training generative models, as the generated distribution is continuous. Another interesting point is the extension of our analysis to other regularizations of optimal transport [7] or gradient penalty [13].

We took advantage in this work of the discrete nature of the target distribution, defined as a collection of training samples, to propose a simple optimization algorithm based on the stochastic gradient of the semi-discrete formulation of the regularized optimal transport. As a corollary, the proposed framework is not restricted to the $1-$Wasserstein cost function anymore, as mostly done in the literature. Recently, some works have been considering the training of generative models with other representation of the training set, for instance using differentiable data augmentation [27]. The question of learning generative model with regularized optimal transport between continuous distributions, as recently studied in [20], is an interesting perspective we leave for future work.

References

[1] Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155. PMLR, 2017.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[3] Yucheng Chen, Matus Telgarsky, Chao Zhang, Bolton Bailey, Daniel Hsu, and Jian Peng. A gradual, semi-discrete approach to generative network training via explicit wasserstein minimization. In *International Conference on Machine Learning*, pages 1071–1080. PMLR, 2019.

[4] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[5] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014.

[6] Marco Cuturi and Gabriel Peyré. A smoothed dual approach for variational wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

[7] Arnaud Dessein, Nicolas Papadakis, and Jean-Luc Rouas. Regularized optimal transport and the ROT mover's distance. *Journal of Machine Learning Research*, 19, 2018.

[8] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019.

[9] Aude Genevay. *Entropy-regularized optimal transport for machine learning*. PhD thesis, Paris Sciences et Lettres, 2019.

[10] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pages 3440–3448, 2016.

[11] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.

[12] 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*, pages 2672–2680, 2014.

[13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein GANs. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[14] Antoine Houdard, Arthur Leclaire, Nicolas Papadakis, and Julien Rabin. Wasserstein generative models for patch-based texture synthesis. *arXiv preprint arXiv:2007.03408*, 2020.

[15] Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Debiased sinkhorn barycenters. In *International Conference on Machine Learning*, pages 4692–4701. PMLR, 2020.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[17] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, 2014.

[18] Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. *arXiv preprint arXiv:1909.13082*, 2019.

[19] Huidong Liu, GU Xianfeng, and Dimitris Samaras. A two-step computation of the exact GAN wasserstein distance. In *International Conference on Machine Learning*, pages 3159–3168. PMLR, 2018.

[20] Arthur Mensch and Gabriel Peyré. Online sinkhorn: Optimal transport distances from sample streams. *Advances in Neural Information Processing Systems*, 33, 2020.

[21] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[22] Daisuke Oyama and Tomoyuki Takenawa. On the (non-) differentiability of the optimal value function when the optimal solution is unique. *Journal of Mathematical Economics*, 76:21–32, 2018.

[23] Henning Petzka, Asja Fischer, and Denis Lukovnicov. On the regularization of wasserstein GANs. In *International Conference on Learning Representations*, 2018.

[24] 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.

[25] Filippo Santambrogio. Optimal transport for applied mathematicians. *Progress in Nonlinear Differential Equations and their applications*, 87, 2015.

[26] Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. In *International Conference in Learning Representations*, 2018.

[27] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient GAN training. *arXiv preprint arXiv:2006.10738*, 2020.