On the Gradient Formula for learning Generative Models with Regularized Optimal Transport Costs

Anonymous authors
Paper under double-blind review

Abstract

The use of optimal transport costs for learning generative models has become popular with Wasserstein Generative Adversarial Networks (WGANs). Training a WGAN requires the differentiation of the optimal transport cost with respect to the parameters of the generative model. In this work, we provide sufficient conditions for the existence of a gradient formula in two different frameworks: the case of semi-discrete optimal transport (i.e. with a discrete target distribution) and the case of regularized optimal transport (i.e. with an entropic penalty). Both cases are based on a semi-dual formulation of the optimal transport cost, and the gradient formula involves a solution of this semi-dual problem. Our study makes a connection between the gradient of the WGAN loss function and the Laguerre diagrams associated to semi-discrete transport maps, which helps to forge an intuition on the generator updates by gradient descent. The learning problem is addressed with an alternate algorithm, whose behavior is examined on several synthetic low-dimensional examples and on the high-dimensional case of MNIST digits generation. We exhibit that this alternate algorithm is in general not convergent, but that, in most cases, depending on the choice of hyperparameters of the optimization strategy, it can stabilize close to a configuration that provides a good enough solution for the generative learning problem. We also analyze the impact of entropic regularization, which can improve the convergence speed in a low-dimensional setting, while introducing a bias that often damages the quality of the resulting generative model.

1 Introduction

Generative modeling is at the heart of various problems in data science, either to approximate the data distribution in order to draw new samples, or to interpolate the data points. Beyond the purpose of image synthesis or editing, adopting such a generative model can also be used to reconstruct or restore corrupted data (Bora et al., 2017; Hand & Joshi, 2019; Hyder & Asif, 2020; Heckel & Soltanolkotabi, 2020; Menon et al., 2020; Shamshad & Ahmed, 2020; Damara et al., 2021; Leong, 2021) or to propose a geometric structure for the data that may reveal some interpretable dimensions (Radford et al., 2015; Shen et al., 2020). Supervised or not, learning generative models on large datasets thus opens new perspectives on the resolution of inverse problems.

Given the empirical distribution ν of the data supported on a compact set $\mathcal{Y} \subset \mathbb{R}^d$, estimating a generative network consists in solving

$$\operatorname*{arg\,min}_{\theta \in \Theta} \mathcal{L}(\mu_\theta, \nu) \tag{1}$$

where μ_θ is the distribution of generated samples (parameterized by a θ in an open subset $\Theta \subset \mathbb{R}^q$) with support included in a compact set $\mathcal{X} \subset \mathbb{R}^d$, and where $\mathcal{L}$ is a loss function between probability distributions. The distribution μ_θ is often considered to be the law of a random variable $g_\theta(Z)$ where g_θ is a neural network and Z a random variable, and samples of the model can then be obtained by passing new realizations of Z through the network g_θ. These distributions are often built upon features computed from samples and data points (such as the latent space of a variational auto-encoder (Kingma & Welling, 2014)) which may be integrated in the loss function (1). In this context, we face the long-standing problem of quantifying the discrepancy between probability distributions in a relevant and efficient manner.

1.1 Wasserstein Generative models

In the seminal work of Goodfellow et al. (2014) on adversarial training of generative networks, the considered loss is related (in a dual sense) to the Jensen-Shannon divergence between feature distributions. The major innovation of such a framework is that these features are simultaneously learnt from the dataset by training a binary classification network which competes against the generative network to discriminate between data points and generated samples. Arjovsky et al. (2017) later remarked that the Jensen-Shannon divergence has major flaws that directly impact the learning of GANs such as the convergence and robustness, and then proposed to use optimal transport (OT) costs instead, leading to new generative models called Wasserstein GANs (WGANs).

Wasserstein distance The OT cost between probability distributions μ_θ and ν is defined by

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

where $\Pi(\mu_\theta, \nu)$ is the set of probability distributions on $\mathcal{X} \times \mathcal{Y}$ having marginals μ_θ and ν, while the ground cost $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a continuous function ($c(x, y)$ represents the elementary cost between locations $x \in \mathcal{X}$ and $y \in \mathcal{Y}$). A simple and popular choice is the Euclidean distance between points to the power $p \geq 1$, *i.e.* $c(x, y) = \|x - y\|^p = \left(\sum_{i=1}^d x_i^2\right)^{p/2}$, for which $W(.,.)^{\frac{1}{p}}$ defines the well-known p-Wasserstein distance. Another possible, but more complex choice, is to define the cost function as a metric in feature space.

As we will recall later, the OT cost (2) admits a dual formulation (Santambrogio, 2015)

$$W(\mu_\theta, \nu) = \sup_{(\varphi, \psi) \in \mathcal{K}_c} \int \varphi d\mu_\theta + \int \psi d\nu, \tag{3}$$

where (φ, ψ) is a couple of dual variables that belongs to the set

$$\mathcal{K}_c = \{(\varphi, \psi) \in \mathscr{C}(\mathcal{X}) \times \mathscr{C}(\mathcal{Y}), \text{ subject to } \varphi(x) + \psi(y) \leq c(x, y) \quad \mu \otimes \nu \text{ a.e. } \} \tag{4}$$

where $\mathscr{C}(\mathcal{X})$ indicates the set of real continuous functions on $\mathcal{X}$ and $\otimes$ the product of two measures. Optimizing one of the dual variables in (3) amounts to taking the c-transform

$$\psi^c(x) = \min_{y \in \mathcal{Y}} c(x, y) - \psi(y), \tag{5}$$

thus leading to another expression of the OT cost, often called "semi-dual formulation":

$$W(\mu_\theta, \nu) = \sup_{\psi \in \mathscr{C}(\mathcal{Y})} \int \psi^c d\mu_\theta + \int \psi d\nu. \tag{6}$$

Solving the constrained dual problem (3) is a difficult (and possibly infinite-dimensional) optimization problem. One possibility to get an unconstrained optimization problem on (φ, ψ) is to rely on the entropic regularization of OT (Chizat, 2017; Peyré & Cuturi, 2019). The entropy-regularized OT admits a similar semi-dual formulation (6) except that a smoothed version of the minimum (called softmin) is used in the computation of the c-transform (5). In the discrete setting, the regularized OT cost can be computed efficiently with the Sinkhorn algorithm (Peyré & Cuturi, 2019), that exhibits geometric convergence. When one of the distribution is continuous (as in equation 1), one has to rely on stochastic algorithms (Genevay et al., 2016), which are provably convergent, but considerably slower. First attempts of using entropy-regularized OT for generative modeling have been made in (Genevay et al., 2018; Seguy et al., 2018) and we pursue here this investigation.

Learning with Wasserstein losses Once chosen the loss function $\mathcal{L}$, the minimization problem (1) can be solved with a gradient-based algorithm, for example a stochastic gradient descent or the ADAM algorithm (Kingma & Ba, 2015). Hence the main topic of this paper is the computation of the gradients of (1) with respect to θ in the case where $\mathcal{L}$ is the OT cost with or without entropic regularization. As we will prove later, the gradient of (1) is directly linked to the dual variable introduced in the dual formulation (3). Indeed, we will give conditions ensuring that the gradient at a point θ_0 can be expressed as

$$\nabla_\theta \big(W(\mu_\theta, \nu)\big)_{|\theta=\theta_0} = \nabla_\theta \left(\int \varphi_* d\mu_\theta \right)_{|\theta=\theta_0} \tag{7}$$

where (φ_*, ψ_*) is an optimal dual variable for $W(\mu_{\theta_0}, \nu)$, *i.e.* a solution of the dual problem (3).

Such formula was proved in (Arjovsky et al., 2017) for the 1-Wasserstein cost, with the hypothesis that both sides of the equality exist. This proof was adapted by Sanjabi et al. (2018) to the case of regularized OT costs. Both these proofs are based on some version of the so-called "envelope theorem" (also called Danskin's theorem in the context of convex optimization), which allows to differentiate under the maximum. This theorem requires some regularity assumptions that should be carefully checked. As we will see in Section 2.4, in the discrete setting, there exist some irregular cases where these assumptions are not sufficient to make formula (7) licit.

To sum up, the first goal of this paper is to provide a new set of hypotheses that validates (7) and to show how these results apply to generative models parameterized by neural networks.

1.2 Related works

The computation of the gradient in (7) involves the optimal dual variable φ_*. In order to compute the update direction of the generator parameters, one thus needs a dual solver which is designed to approximate the optimal dual variable in a given subclass of functions (which may cause relatively large errors, a point that we will discuss later in the paper.) Many attempts have been proposed in the literature, with several ways to parameterize the problem (3).

The method of Arjovsky et al. (2017), being based on the 1-Wasserstein distance, only requires one dual variable that is constrained to be 1-Lipschitz. In practice, this dual variable is parameterized by a neural network, and the Lipschitz constraint is enforced by weight clipping (WGAN-WC). On a similar formulation, Gulrajani et al. (2017) suggest to impose Lipschitzness by including a gradient penalty in the dual loss (WGAN-GP).

In contrast, Seguy et al. (2018) consider regularized OT costs with a generic cost function. This leads to an unconstrained dual problem, but with two dual variables. In practice, the authors choose to parameterize both dual variables with neural networks. Closely related, the work by Sanjabi et al. (2018) is based on the same formulation of WGAN training with regularized OT and also relies on the neural network parameterization of the dual variables. The authors study the convergence and stability of the training procedure, the convergence being proved for the case of discrete distributions. In particular, they give the expression of the gradient of (7) (under a primal form) in the case of discrete regularized OT. This gradient expression is exploited to perform WGAN training by stochastic gradient descent. Notice also that Liu et al. (2019) applied a similar regularized OT framework on empirical distributions (i.e. discrete distributions obtained from samples) to learn a generator, with the particularity that the cost function depends on a set of simultaneously-learned features.

Closer to our framework, Chen et al. (2019) consider the semi-dual formulation of OT (6). In their work, the dual variable ψ is optimized with the stochastic algorithm proposed in (Genevay et al., 2016). Contrary to Seguy et al. (2018), they do not parameterize the dual variable ψ with a neural network, which hinders the applicability of their method to a very large dataset. Indeed, as shown in (Leclaire & Rabin, 2021), the convergence speed of the ASGD algorithm used for optimizing ψ decreases when either the dimension d of samples or the dimension of vector ψ (equal to the number of training points) increases. In (Chen et al., 2019), the corresponding primal solution π (which, in that case, is supported on a graph) is then used to perform a gradient step on the generative model using the estimated transportation cost. As a result, the whole algorithm of Chen et al. (2019) is not expressed as a direct minimization of (1). Our work, by contrast, shows that the proposed gradient calculation for (7) makes it possible to train the generator g_θ and the dual variable ψ with an alternate min-max procedure on the dual cost (3).

Closely related to (Chen et al., 2019), Mallasto et al. (2019a) propose to parameterize the dual variable ψ by a neural network and to obtain the second one φ with an approximated c-transform computed on mini-batches. With such an approximated c-transform, the pair of dual variables may not satisfy the constraint (4), which led the authors to integrate in the loss a penalty on $c(x, y) - \varphi(x) - \psi(y)$. A comparative study on the different ground costs is also realized. One benefit of this approach is that it scales up to a very large database, while keeping a relatively precise way to estimate the Wasserstein cost. On batch strategy, let us also mention the work by Fatras et al. (2020) who consider an alternative Wasserstein cost that is inherently defined as an expectation over mini-batches; this stochastic approximation introduces an estimation bias which is shown in practice to regularize the transportation problem.

There have been several works (Mallasto et al., 2019b; Stanczuk et al., 2021; Korotin et al., 2021; 2022) that question the performance of OT dual solvers and its impact on WGAN learning. Mallasto et al. (2019b) compare several algorithms to estimate the dual variables, namely WGAN-WC, WGAN-GP and two methods denoted as "c-transform" and "(c, ε)-transform" (those two methods being similar to the algorithm proposed in the present paper, but with a batchwise computation of the c-transform). Mallasto et al. (2019b) show that these two last methods, while better estimating the OT cost, do not improve the visual quality of the generative model when used in WGAN learning (producing very blurry images for the CelebA database). These findings are confirmed by Stanczuk et al. (2021) who give more explanations related to biased estimation of Wasserstein gradients, false Wasserstein minima supported on barycenters, and the limitations of the Euclidean distance computed on natural images. Our experimental study shows that, with the simpler MNIST database, a WGAN learning based on c-transform computations on all datapoints (and not batchwise) can produce sharp images while relying on a provably-convergent estimation algorithm for the Wasserstein cost. Also, our theoretical analysis of gradients allows to better understand the impact of the errors made by the OT dual solver: the performance of the dual solver does not guarantee the sharpness of generated images, but plays a role in making the generative distribution cover the whole database.

The authors of (Korotin et al., 2021; 2022) pursue these investigations by proposing benchmarks for OT solvers (for the 2-Wasserstein and 1-Wasserstein costs respectively) using absolutely continuous measures μ, ν that are specifically designed to know explicitly the groundtruth OT map from μ to ν. Contrary to our method, the OT solvers they compare are all based on various parameterizations of the dual problem with neural networks, whose optimization requires to draw some batches of μ and ν. Such a mini-batch optimization is known to regularize the OT cost and to induce an estimation bias. In order to stay as close as possible to the original OT cost, we do not apply a batch procedure on ν, thus restricting our experimental section to small datasets.

The theoretical results of the present paper rely heavily on concepts related to semi-discrete OT, that is, OT between an absolutely continuous distribution μ and a discrete distribution ν supported on a finite set. After that Aurenhammer et al. (1998) made the connection between semi-discrete OT and a concave optimization problem, many numerical solutions have been proposed to solve semi-discrete OT, see Kitagawa et al. (2017) and references therein. We will regularly use the concepts of Laguerre diagram (illustrated in Fig. 2) and c-transforms of functions defined on a finite set. In addition to the proof of our main results, these tools will prove very useful to examine the practical behavior of the proposed learning algorithm. The concept of c-transform is also central in the construction of the benchmark of (Korotin et al., 2022) (where "MinFunnel" functions can be seen, for $c(x, y) = \|x-y\|$, as c-transforms attached to a set of discrete points). For the numerical simulations of the present paper, we will use the stochastic algorithm for semi-discrete OT proposed by Genevay et al. (2016). Finally, let us mention that Lei et al. (2019) studied some connections between WGAN learning and several formulations of the semi-discrete OT problem, recalling its equivalence with other well-known problems from convex geometry. In this perspective, the present paper provides new insights on the connection between the WGAN and OT problems, which pursue the observations formulated in (Genevay et al., 2017).

1.3 Contributions and outline

The main purpose of this paper is to propose a complete set of hypotheses that ensure the validity of the gradient formula (referred to as equation (7) above and (Grad-OT) in a more general form below). Our approach is not restricted to the cost $c(x, y) = \|x - y\|$ (inducing the 1-Wasserstein distance) and involves weak regularity hypotheses on the cost and the generator. Based on this gradient formula, we consider a stochastic algorithm for learning a generative model that can be understood as an alternate optimization algorithm on the semi-dual cost.

The main contributions of this paper are the following ones.

- For unregularized OT, we prove the formula (Grad-OT) in the semi-discrete setting, that is, when ν is a distribution supported on a finite set. We first prove the formula for x-regular costs (Theorem 3) and extend it to less regular costs (Theorem 4). Both these results are based on a technical assumption that for any θ, μ_θ does not charge a particular subset called the Laguerre interface.
- For regularized OT, we prove (Grad-OT) for x-regular costs (Theorem 5) with no assumption on ν. We also adapt this result for the differentiation of the Sinkhorn divergence (Theorem 6).
- We provide a counterexample (Proposition 2) that illustrates how (Grad-OT) can fail for unregularized OT when the technical hypothesis is not met.
- On the practical side, we provide experiments that illustrate the behavior of the alternate optimization algorithm for generative model learning (Section 6). First, with the toy example of Proposition 2, we examine the trajectory of the optimization algorithm, and explain why it does not get trapped in a singular point with no gradient. Second, with various synthetic cases in dimension 2, we exhibit different possible behaviors for the alternate algorithm, depending on the choice of gradient step size strategies: this algorithm often stabilizes but it may also oscillate or converge to sub-optimal configurations. Third, we apply the algorithm to the learning of a WGAN for MNIST digits generation. In this high-dimensional case, we examine the stability of the loss function, and underline the impact of the entropic regularization both on numerical and visual results.

We emphasize that the theorems ensuring (Grad-OT) are based on weak regularity assumptions for the cost and the generator that are very often met in practice. These regularity hypotheses are true with the quadratic cost $c(x, y) = \|x - y\|^2$ and with generators given as neural networks with $\mathscr{C}^1$ activation functions. Also, in the case of unregularized OT, the technical hypothesis is true except for degenerate positioning of the support of μ_θ. Therefore, our results apply to most practical cases of WGAN training, and explain why the algorithms for WGAN learning generally do not get caught in singular points with inexisting gradients.

In Section 2 we recall the complete framework for WGAN learning, and in particular well-known results on the dual formulations of OT costs. Section 3 is dedicated to the proof of the gradient formula (Grad-OT) in the case of unregularized ($\lambda = 0$) semi-discrete OT (i.e. when ν is discrete). In Section 4, we extend the gradient formula to the case of entropy-regularized OT (with no assumption on ν). Section 5 draws a relation of the obtained formula with derivatives in the sense of distributions. Finally, Section 6 contains numerical experiments obtained with an alternate optimization algorithm for WGAN learning.

2 The Wasserstein GAN Problem

In this section, we first introduce the primal, dual and semi-dual formulations of the OT cost. The OT problem is presented in the general case including entropic regularization with parameter $\lambda \geq 0$, thus encompassing the unregularized case $\lambda = 0$. Next, we introduce the generative learning problem as well as the regularity hypothesis on the generator, that will be used in the next sections to show the existence of gradients of the OT cost with respect to θ. We close this section with a counter-example of measures μ_θ and ν for which the desired gradient formula (7) does not hold. We also formulate several remarks that will help to forge an intuition on the behavior of the alternate algorithm used for practical WGAN learning.

First, let us give some notations that are used in the whole paper. We refer to Appendix A for a list of the most important notations used in the paper. Let $\mathcal{X}, \mathcal{Y}$ be compact subsets of $\mathbb{R}^d$, and let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous function. Let Θ be an open subset of $\mathbb{R}^q$ (that will be used to parameterize the generator g_θ).

We say that $\varphi : \mathcal{X} \to \mathbb{R}$ is $\mathscr{C}^1$ on $\mathcal{X}$ if it is $\mathscr{C}^1$ (i.e. differentiable with continuous derivatives) on an open neighborhood of $\mathcal{X}$.

Definition 1 (x-regularity)**.** *We say that c is* ***x-regular*** *if there is $L > 0$ and an open set U with $\mathcal{X} \subset U \subset \mathbb{R}^d$ such that for any $y \in \mathcal{Y}$, $c(\cdot, y)$ can be extended to a L-Lipschitz $\mathscr{C}^1$ function on U.*

Let us illustrate this definition with the cost $c(x, y) = \|x - y\|^p$ on $\mathbb{R}^d$, with $p \geq 1$, which is of constant use in the applications. For $p > 1$, it is clear that c is a smooth function on $\mathbb{R}^d \times \mathbb{R}^d$, and therefore x-regular (because $\nabla_x c$ is a continuous function on $\mathbb{R}^d \times \mathbb{R}^d$ and is thus bounded on $U \times \mathcal{Y}$ for any bounded open set U containing $\mathcal{X}$). However, for $p = 1$, the cost $c(x, y) = \|x - y\|$ is not x-regular as soon as $\mathcal{X}$ and $\mathcal{Y}$ are not disjoint: indeed, for a fixed $y \in \mathcal{Y}$, $x \mapsto \|x - y\|$ is not differentiable at y.

Convention for notation of derivatives. For a real-valued univariate function φ, the gradient of φ will simply be denoted by $\nabla\varphi$. However, for multivariate functions, we will always use an index to indicate the variable of differentiation: for example $\nabla_x c$ refers to the gradient w.r.t. the first argument of c.

2.1 Optimal Transport, Primal and Dual Problems

Let μ, ν be two probability measures on $\mathcal{X}, \mathcal{Y}$ respectively.

Definition 2 (Primal formulation)**.** *For $\lambda \geqslant 0$, the regularized optimal transport cost is defined by*

$$W_\lambda(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c \, d\pi + \lambda \, \mathrm{KL}(\pi|\mu \otimes \nu) \tag{8}$$

where $\Pi(\mu, \nu)$ is the set of probability distributions on $\mathcal{X}\times\mathcal{Y}$ with marginals μ and ν, and where $\mathrm{KL}(\pi|\mu\otimes\nu) = \int \log(\frac{d\pi}{d\mu\otimes\nu})d\pi$ if π admits a density $\frac{d\pi}{d\mu\otimes\nu}$ w.r.t. $\mu \otimes \nu$ and $+\infty$ otherwise.

Theorem 1 (Dual formulation (Santambrogio, 2015; Genevay, 2019; Feydy et al., 2019))**.** *Strong duality holds in the sense that*

$$W_\lambda(\mu, \nu) = \max_{\varphi\in\mathscr{C}(\mathcal{X}),\psi\in\mathscr{C}(\mathcal{Y})} \int \varphi(x)d\mu(x) + \int \psi(y)d\nu(y) - \int m_\lambda\big(\varphi(x) + \psi(y) - c(x, y)\big)d\mu(x)d\nu(y) \tag{9}$$

where, for $\lambda = 0$, $m_0(t) = 0$ if $t \geq 0$, and $+\infty$ otherwise, and for $\lambda > 0$, $m_\lambda(t) = \lambda(e^{\frac{t}{\lambda}} - 1)$ A solution (φ, ψ) of this dual problem is called a pair of Kantorovich potentials. When $\lambda > 0$, the solutions of the dual problem are uniquely defined almost everywhere up to an additive constant (i.e. if (φ, ψ) is a solution, then any solution can be written $(\varphi - k, \psi + k)$ with $k \in \mathbb{R}$). Also, when $\lambda > 0$, the primal problem admits a unique solution

$$d\pi(x, y) = \exp\left(\frac{\varphi(x) + \psi(y) - c(x, y)}{\lambda}\right) d\mu(x)d\nu(y). \tag{10}$$

If $\lambda = 0$ the primal solution π of (8) may not be absolutely continuous w.r.t. $\mu \otimes \nu$ anymore. In that case, under weak assumptions (for example, when μ, ν admits second-order moments and μ is absolutely continuous), one can show (Santambrogio, 2015) that the support of π is actually supported on the graph of an optimal transport map T^* i.e., the optimal π is the probability distribution of $(X, T^*(X))$ where X has distribution μ (and in particular $T^*(X)$ has distribution ν).

For $\psi \in \mathscr{C}(\mathcal{Y})$, let us define the regularized c-transform as in (Feydy et al., 2019)

$$\forall x \in \mathcal{X}, \quad \psi^{c,\lambda}(x) = \operatorname*{softmin}_{y\in\mathcal{Y}} c(x, y) - \psi(y) \tag{11}$$

where the softmin operation is defined as

$$\operatorname*{softmin}_{y\in\mathcal{Y}} u(y) = \begin{cases} \min\limits_{y\in\mathcal{Y}} u(y) & \text{if } \lambda = 0, \\ -\lambda \log \displaystyle\int e^{-\frac{u(y)}{\lambda}} d\nu(y) & \text{if } \lambda > 0. \end{cases} \tag{12}$$

We also define the analoguous operators for the x-variable (and for simplicity, we use the same notation for c-transforms of x-functions or y-functions). It must be noted that the regularized c-transform $\psi^{c,\lambda}$ also depends on ν even if ν is omitted in the notation.

Given a pair of dual variables $(\varphi, \psi) \in \mathscr{C}(\mathcal{X}) \times \mathscr{C}(\mathcal{Y})$, one can see that taking c-transforms

$$\begin{cases} \tilde{\psi} = \varphi^{c,\lambda} \\ \tilde{\varphi} = \psi^{c,\lambda} \end{cases}, \tag{13}$$

leads to a new pair $(\tilde{\varphi}, \tilde{\psi})$ of dual variables that have a better dual cost (9) than (φ, ψ). Therefore, the dual problem can always be restricted to c-concave functions, that is, functions that can be written as c-transforms. This is an important point since c-concave functions inherits some regularity from the cost function (see Appendix B) and can be naturally extrapolated to any $x \in \mathcal{X}$.

Theorem 2 (Semi-dual formulation (Genevay, 2019))**.** *The dual problem (9) is equivalent to the following semi-dual problem*

$$W_\lambda(\mu, \nu) = \sup_{\psi \in \mathscr{C}(\mathcal{Y})} \int \psi^{c,\lambda}(x) d\mu(x) + \int \psi(y) d\nu(y). \tag{14}$$

A solution ψ of the semi-dual problem is called a Kantorovich potential. In other words, ψ is a Kantorovich potential if and only if $(\psi^{c,\lambda}, \psi)$ is a pair of Kantorovich potentials. By symmetry, we can also formulate a semi-dual problem on the dual variable φ.

Let us consider the exponential scalings of the dual variables $a = e^{\frac{\varphi}{\lambda}}, b = e^{\frac{\psi}{\lambda}}$. With this notation, the coupled fixed point equations (13) can be reformulated as a single fixed point equation on a (or b). The corresponding operator for a (or b) can be shown to be a contractive operator on the unit sphere of L^∞_+ equipped with the Hilbert metric. In the discrete case where $\mathcal{X}, \mathcal{Y}$ are both finite, iterating this contractive operator corresponds exactly to the Sinkhorn algorithm (Cuturi, 2013).

Remark 1 (On the domain of definition of c-transforms)**.** *In the primal formulation (8) of OT, this is obvious that the domains $\mathcal{X}, \mathcal{Y}$ can always be restricted to the respective supports of μ, ν. This is less obvious in the semi-dual formulation (14) for $\lambda = 0$, because the c-transform $\psi^{c,0}$ is impacted by the values of ψ on the whole set $\mathcal{Y}$ (and not just ν almost everywhere, as emphasized in (Villani, 2009, Remark 5.5)). But of course, this is still true, because one can take the dual of the primal formulation restricted to the true supports of μ, ν.*

2.2 Learning a Generative Network

With the main OT concepts now defined, we can turn to the problem of learning a Wasserstein generative adversarial network. Estimating a WGAN from an empirical data distribution ν consists in minimizing

$$h_\lambda(\theta) = W_\lambda(\mu_\theta, \nu), \tag{15}$$

where the generated distribution μ_θ is assumed to be the distribution of $g_\theta(Z)$, with Z a random variable. Denoting by ζ the probability distribution of Z on the measurable space $\mathcal{Z}$, we therefore have that μ_θ is the image measure of ζ by the generator g_θ, also known as the pushforward $\mu_\theta = g_\theta \sharp \zeta$. More precisely, the notation g_θ refers to $g(\theta, \cdot)$ where $g : \Theta \times \mathcal{Z} \to \mathbb{R}^d$ is a function defined on the product of the open set $\Theta \subset \mathbb{R}^q$ with $\mathcal{Z}$. In the following, we give different sets of conditions on g that allow to compute the derivatives of h_λ.

All the results of this paper are related to the behavior of the function

$$I(\varphi, \theta) = \int_{\mathcal{X}} \varphi d\mu_\theta, \quad (\varphi \in \mathscr{C}(\mathcal{X}),\ \theta \in \Theta). \tag{16}$$

Indeed, the semi-dual expression of optimal transport gives

$$h_\lambda(\theta) = \max_{\psi \in \mathscr{C}(\mathcal{Y})} I(\psi^{c,\lambda}, \theta) + \int_{\mathcal{Y}} \psi d\nu. \tag{17}$$

In order to study the gradient of h_λ, we thus define $F_\lambda : \mathscr{C}(\mathcal{Y}) \times \Theta \to \mathbb{R}$ with

$$\forall \psi \in \mathscr{C}(\mathcal{Y}), \forall \theta \in \Theta, \quad F_\lambda(\psi, \theta) = I(\psi^{c,\lambda}, \theta) = \int_{\mathcal{X}} \psi^{c,\lambda} d\mu_\theta = \mathbb{E}[\psi^{c,\lambda}(g_\theta(Z))], \tag{18}$$

where the expectation is taken with respect to the probability distribution ζ of Z.

Combining all previous definitions, the problem we tackle writes

$$W_\lambda(\mu_\theta, \nu) = h_\lambda(\theta) = \max_{\psi \in \mathscr{C}(\mathcal{Y})} H_\lambda(\psi, \theta) \tag{19}$$

with

$$H_\lambda(\psi, \theta) = F_\lambda(\psi, \theta) + \int_{\mathcal{Y}} \psi d\nu, \tag{20}$$

and our objective is to study the relationship between $\nabla h_\lambda(\theta)$ and $\nabla_\theta F_\lambda(\psi, \theta)$. In the unregularized case, we use simpler notations, h, W, H, and F for h_0, W_0, H_0 and F_0 respectively, dropping the index λ. In accordance with our convention used for differentiation, ∇h_λ refers to the gradient w.r.t. the only argument of h_λ (which is θ), and $\nabla_\theta F_\lambda$ refers to the gradient w.r.t. the second argument of F_λ (which is also θ).

As we will see later, computing the derivatives of h_λ boils down to differentiating under the max, which is allowed by the so-called envelope theorem. This result appears under different forms in the literature (Oyama & Takenawa, 2018). In Appendix C, we recall the version of the envelope theorem that is used in the proofs of the following sections.

However, if we temporarily admit the differentiability of all terms, the computation of the gradient is straightforward:

Proposition 1. *Let θ_0 and ψ_0 satisfying $h_\lambda(\theta_0) = H_\lambda(\psi_0, \theta_0)$.*
If h_λ and $\theta \mapsto F_\lambda(\psi_0, \theta)$ are both differentiable at θ_0, then

$$\nabla h_\lambda(\theta_0) = \nabla_\theta F_\lambda(\psi_0, \theta_0). \tag{Grad-OT}$$

Proof. First, notice that $H_\lambda(\psi_0, \cdot)$ and $F_\lambda(\psi_0, \cdot)$ differ by a constant, and thus have same gradients. Using (17), for any θ, $h_\lambda(\theta) \geqslant H_\lambda(\psi_0, \theta)$ with equality if $\theta = \theta_0$. Therefore, the function $\theta \mapsto h_\lambda(\theta) - H_\lambda(\psi_0, \theta)$ has a minimum at $\theta = \theta_0$, and its gradient at θ_0 vanishes. This gives $\nabla h_\lambda(\theta_0) = \nabla_\theta H_\lambda(\psi_0, \theta_0)$ and thus the desired result. □

Now, showing the existence of $\nabla_\theta F_\lambda(\psi_0, \theta_0)$ consists in differentiating under the expectation in (18). For that, we need the following technical hypothesis.

Definition 3 (Hypothesis (G_Θ))**.** *For $\theta_0 \in \Theta$, we say that $g : \Theta \times \mathcal{Z} \to \mathcal{X}$ satisfies Hypothesis (G_{θ_0}) if there exists a neighborhood V of θ_0 and $K \in L^1(\mathcal{Z})$ such that almost surely $\theta \mapsto g(\theta, Z)$ is $\mathscr{C}^1$ on V with differential $\theta \mapsto D_\theta g(\theta, Z)$ and*

$$\forall \theta \in V, \quad \zeta\text{-a.s.} \quad \|g(\theta, Z) - g(\theta_0, Z)\| \leqslant K(Z)\|\theta - \theta_0\|. \tag{21}$$

We say that g satisfies Hypothesis (G_Θ) if g satisfies Hypothesis (G_{θ_0}) for any $\theta_0 \in \Theta$.

Remark 2. *A sufficient condition for Hypothesis (G_Θ) is that almost surely, $\theta \mapsto g(\theta, Z)$ is $\mathscr{C}^1$ on Θ and that there exists $K \in L^1(\mathcal{Z})$ such that*

$$\forall \theta, \theta' \in \Theta, \quad \zeta\text{-a.s.} \quad \|g(\theta', Z) - g(\theta, Z)\| \leqslant K(Z)\|\theta' - \theta\|. \tag{22}$$

Notice that Θ is an arbitrary open set, and can thus be reduced to localize the problem in a neighborhood of a fixed point $\theta \in \Theta$. The interest of Definition 3 is that it does not impose uniformity in θ_0 (for both the neighborhood V and the upper bound $K(Z)$).

Remark 3. *Note that Hypothesis (G_Θ) is true as soon as g is $\mathscr{C}^1$ on $\Theta \times \mathcal{Z}$ with $\mathcal{Z} \subset \mathbb{R}^s$ compact. Indeed, in this case, for any $\theta_0 \in \Theta$, there is $r > 0$ such that $V = \overline{B}(\theta_0, r)$ is included in Θ and then (21) holds with the constant upper bound $K = \sup_{V \times \mathcal{Z}} \|D_\theta g\|$. In particular, this is true when g is parameterized by a neural network with $\mathscr{C}^1$ activation functions, and with an input noise Z supported on a bounded subset of $\mathbb{R}^s$ (for example Z uniform on $[-1, 1]^s$).*

2.3 Previous results on the differentiability of OT costs

Outside the recent context of WGAN learning, the regularity and gradient of the entropy-regularized OT cost $p \mapsto W_\lambda(p,q)$ (with $\lambda > 0$) were expressed in (Cuturi & Peyré, 2016, Prop. 2.3) in the case of discrete disributions (p,q). In the same context, the Gâteaux-differentiability of $(p,q) \mapsto W_\lambda(p,q)$ was proved in (Feydy et al., 2019), which was later extended to continuous differentiability in (Bigot et al., 2019, Prop. 2.3). If (p_θ) is a parametric family of distribution supported on the *same* finite set, applying the chain rule gives the differentiability of $\theta \mapsto W_\lambda(p_\theta, q)$.

The gradient formula (Grad-OT) is given in the seminal paper on WGAN (Arjovsky et al., 2017) with the assumption that both sides of the equality exist. This formula (extended to more general costs) has been exploited in several papers related to WGAN learning, for example (Liu et al., 2019). In the context of discrete regularized OT, one can find in (Sanjabi et al., 2018, Appendix C) a gradient formula expressed through the primal formulation, with a short proof limited to discrete regularized OT.

2.4 A telling counter-example

We now show that the differentiation with respect to θ under the max in (17) can fail, even for regular generators g, thus discarding the formula (Grad-OT). To illustrate this point, let us consider a simple unregularized OT problem (i.e. $\lambda = 0$) between a Dirac δ_θ located at $\theta \in \mathbb{R}^d$ and a sum of two Diracs at positions $y_1 \neq y_2 \in \mathbb{R}^d$. This setting can be obtained by setting $g_\theta(z) = \theta$ (for any distribution of Z).

Proposition 2. *Let* $\mu_\theta = \delta_\theta$ *and* $\nu = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}$.

For $p \geq 1$, *consider the cost* $c(x,y) = \|x-y\|^p$ *where* $\|\cdot\|$ *is the Euclidean norm on* $\mathbb{R}^d$. *Then*

- $h(\theta) = W(\mu_\theta, \nu)$ *is differentiable at any* $\theta \notin \{y_1, y_2\}$ *for* $p = 1$, *and at any* θ *for* $p > 1$,
- *for any* $\psi_{*0} \in \arg\max_\psi H(\psi, \theta_0)$, *the function* $\theta \mapsto F(\psi_{*0}, \theta)$ *is* ***not*** *differentiable at* θ_0.

Hence relation (Grad-OT) never stands.

Proof. The only distribution on $\mathcal{X} \times \mathcal{Y}$ having marginals μ_θ, ν is $\pi = \frac{1}{2}\delta_{(\theta,y_1)} + \frac{1}{2}\delta_{(\theta,y_2)}$. Therefore, recalling that $h(\theta) = W(\mu_\theta, \nu)$ and from the definition of the primal problem (8), we have

$$h(\theta) = \frac{1}{2}c(\theta, y_1) + \frac{1}{2}c(\theta, y_2) \tag{23}$$

which gives the first point.

Next, one can explicitly solve the dual problem, which, from the equivalent semi-dual formulation (14) and the definition (18) of F, reduces to the following optimization problem with respect to $(\psi(y_1), \psi(y_2)) \in \mathbb{R}^2$

$$\max_{\psi \in \mathbb{R}^2} H(\psi, \theta) \quad \text{with} \quad H(\psi, \theta) = \psi^c(\theta) + \frac{\psi(y_1) + \psi(y_2)}{2}, \tag{24}$$

where, for any ψ, the c-transform (12) writes

$$\psi^c(\theta) = \begin{cases} c(\theta, y_1) - \psi(y_1) & \text{if } c(\theta, y_1) - \psi(y_1) \leqslant c(\theta, y_2) - \psi(y_2), \\ c(\theta, y_2) - \psi(y_2) & \text{otherwise.} \end{cases} \tag{25}$$

Therefore,

$$H(\psi, \theta) = \begin{cases} c(\theta, y_1) + \frac{1}{2}(\psi(y_2) - \psi(y_1)) & \text{if } \psi(y_2) - \psi(y_1) \leqslant c(\theta, y_2) - c(\theta, y_1), \\ c(\theta, y_2) - \frac{1}{2}(\psi(y_2) - \psi(y_1)) & \text{otherwise.} \end{cases} \tag{26}$$

For a fixed θ_0, $H(\cdot, \theta_0)$ is maximal at any ψ_{*0} such that

$$\psi_{*0}(y_2) - \psi_{*0}(y_1) = c(\theta_0, y_2) - c(\theta_0, y_1). \tag{27}$$

Besides, from (26), one sees that $H(\psi, \cdot)$ is made of two pieces whose gradients can be computed explicitly for the cost $c(x, y) = \|x - y\|^p$:

$$\forall \theta \neq y_j, \quad \nabla_x c(\theta, y_j) = p\|\theta - y_j\|^{p-2}(\theta - y_j), \tag{28}$$

and, when $p > 1$ we also have $\nabla_x c(\theta, y_j) = 0$ for $\theta = y_j$. Therefore, as illustrated on Fig. 1, the gradients of the two pieces agree at no point θ: for any θ, $\nabla_x c(\theta, y_1) \neq \nabla_x c(\theta, y_2)$. In particular, the function $H(\psi, \cdot)$ is not differentiable at the interface $A_\psi = \{ \theta \in \mathbb{R}^d \mid c(\theta, y_2) - c(\theta, y_1) = \psi(y_2) - \psi(y_1) \}$. As $F(\psi, \cdot)$ only differs from $H(\psi, \cdot)$ by the constant $\frac{\psi(y_1)+\psi(y_2)}{2}$, it is also not differentiable on A_ψ But then, for ψ_{*0} satisfying (27), θ_0 lies on the interface $A_{\psi_{*0}}$, which gives the second point. ☐

Remark 4. *Let us now examine the regularized case. With the same measures μ_θ, ν defined in Proposition 2, for $\lambda > 0$, one has*

$$F_\lambda(\psi, \theta) = \psi^{c,\lambda}(\theta) = -\lambda \log \left(\frac{1}{2} e^{-\frac{c(\theta, y_1) - \psi(y_1)}{\lambda}} + \frac{1}{2} e^{-\frac{c(\theta, y_2) - \psi(y_2)}{\lambda}} \right), \tag{29}$$

and from the primal formulation, one can also get directly that

$$h_\lambda(\theta) = \frac{1}{2} c(\theta, y_1) + \frac{1}{2} c(\theta, y_2). \tag{30}$$

It is clear that, for any ψ, the function $\theta \mapsto F_\lambda(\psi, \theta)$ is differentiable on $\mathbb{R}^d$ for $p > 1$ and on $\mathbb{R}^d \setminus \{y_1, y_2\}$ for $p = 1$. Therefore, in the regularized case $\lambda > 0$, $F_\lambda(\psi, \cdot)$ has no singularity on the interface A_ψ. For a fixed θ_0, the maximization of $H(\cdot, \theta_0)$ leads to the same solutions satisfying (27), i.e. any ψ_{*0} such that

$$e^{-\frac{c(\theta, y_1) - \psi(y_1)}{\lambda}} = e^{-\frac{c(\theta, y_2) - \psi(y_2)}{\lambda}}. \tag{31}$$

It follows that Formula (Grad-OT) holds in that case:

$$\nabla h_\lambda(\theta_0) = \frac{1}{2} \nabla_x c(\theta, y_1) + \frac{1}{2} \nabla_x c(\theta, y_2) = \nabla_\theta F_\lambda(\psi_{*0}, \theta_0). \tag{32}$$

This will be confirmed by Theorem 5 below. Let us also notice that using regularized OT does not allow to cope with all differentiability issues. Indeed, for $p = 1$, h_λ is not differentiable at y_1, y_2 even for $\lambda > 0$. This illustrates the need of a regularity hypothesis on the cost, even in the case of regularized OT.

Again, we would like to highlight the main pitfall in the previous proof illustrated in Fig. 1: the measure δ_θ puts some mass on a thin subset where $\theta \mapsto F(\psi_{*0}, \theta)$ is not smooth. This counterexample can then be extended to other situations where the measure μ_θ is not suported on a single point. For example, for the same ν, the gradient problem still happens if μ_θ is the distribution of $\theta + Zu$ where Z is a uniform random variable on $[-1, 1]$ and where $u \in \mathbb{R}^d$ is orthogonal to $y_2 - y_1$ (μ_θ is the uniform distribution on a segment orthogonal to $[-1, 1]$ centered at θ). In the following section, we will adopt an hypothesis on the generator that avoids this objection.

3 Gradient formula in the unregularized semi-discrete setting

In this section, we prove the formula (Grad-OT) in the semi-discrete case, *i.e.* when the target measure ν is supported on a finite set of points. Therefore, in this section, $\mathcal{X} \subset \mathbb{R}^d$ is compact and $\mathcal{Y}$ is finite with J points, so that $\mathscr{C}(\mathcal{Y})$ identifies to $\mathbb{R}^J$. We only consider the case $\lambda = 0$ since more general results are given for regularized OT in the next section. The analysis developed in this section builds on concepts related to semi-discrete OT (Aurenhammer et al., 1998; Kitagawa et al., 2017), and in particular the Laguerre diagram whose definition will be recalled below (and which is sometimes also called power diagram as in (Mérigot, 2011)). The Laguerre diagram will permit to formulate a condition on the generator g_θ that serves for computing the gradient of h by differentiating under the max in (19).

We recall the notations $W = W_0$, $h = h_0$, and $F = F_0$ from the definitions (8), (15) and (18) with $\lambda = 0$. Also, for $\lambda = 0$, we simply write $\psi^c = \psi^{c,0}$ the c-transform of a function ψ.

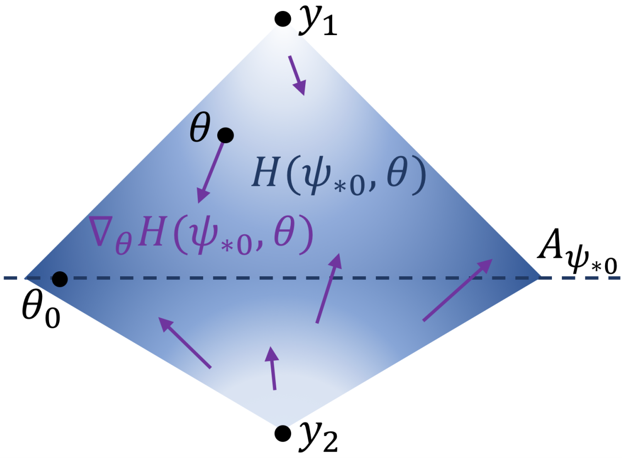


Figure 1: **Counter-example for the gradient formula.** This diagram illustrates the counter-example of Proposition 2 for $p = 2$. The notations used in this diagram are detailed in the proof. The function $\theta \mapsto H(\psi_{*0}, \theta)$ is drawn in blue background. For the optimal ψ_{*0} corresponding to θ_0, the function $\theta \mapsto H(\psi_{*0}, \theta)$ is not differentiable on the interface $A_{\psi_{*0}}$, where θ_0 lies.

Since $\mathscr{C}(\mathcal{Y})$ identifies to $\mathbb{R}^J$, we study the function $F : \mathbb{R}^J \times \Theta \to \mathbb{R}$ defined by

$$F(\psi, \theta) = \int_{\mathcal{X}} \psi^c d\mu_\theta = \mathbb{E}[\psi^c(g_\theta(Z))]. \tag{33}$$

3.1 The Laguerre diagram and the semi-discrete transport maps

In this section, we recall the definitions of concepts (transport maps, Laguerre diagrams) which are classically used in semi-discrete OT Aurenhammer et al. (1998); Kitagawa et al. (2017). We will also give two lemmas in order to compute the gradient of c-transform functions.

To any $\psi \in \mathbb{R}^J$, we associate a Laguerre diagram, which is a collection $(\mathtt{L}_\psi(y))_{y\in\mathcal{Y}}$ of Laguerre cells defined by

$$\mathtt{L}_\psi(y) = \{\ x \in \mathcal{X} \mid \forall y' \neq y,\ c(x, y) - \psi(y) < c(x, y') - \psi(y')\ \}. \tag{34}$$

We also define the Laguerre interface by

$$A_\psi = \mathcal{X} \setminus \bigcup_{y\in\mathcal{Y}} \mathtt{L}_\psi(y). \tag{35}$$

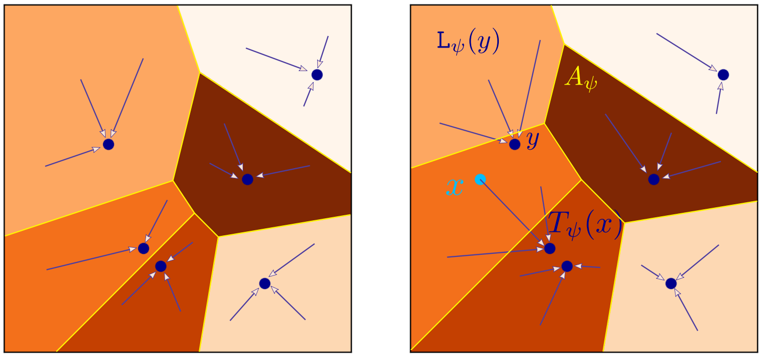
It is the set of points for which there is a "tie" in the minimum defining $\psi^c(x)$ (and it can be thought of as the union of boundaries of all Laguerre cells). For example, if $c(x, y) = \|x - y\|^2$ in $\mathbb{R}^d$, A_ψ is contained in a union of hyperplanes whose directions are orthogonal to the segments $[y_1, y_2]$, $y_1, y_2 \in \mathcal{Y}$. In particular A_ψ has zero Lebesgue measure, and thus the Laguerre diagram form a partition $\mathbb{R}^d$ up to a negligible set.

By construction, for $x \in \mathcal{X} \setminus A_\psi$, we can uniquely define the transport map

$$T_\psi(x) = \operatorname*{arg\,min}_{y\in\mathcal{Y}} c(x, y) - \psi(y). \tag{36}$$

If μ is a probability measure on $\mathcal{X}$, one can verify that for any $\psi \in \mathbb{R}^J$, T_ψ is an optimal transport map between μ and $T_\psi\sharp\mu$. Conversely, in a setting where μ is absolutely continuous on $\mathcal{X}$, solving the OT problem from μ to ν is in fact equivalent to find $\psi \in \mathbb{R}^J$ such that $T_\psi\sharp\mu = \nu$. We refer to Kitagawa et al. (2017) for more detailed formulations of these statements.

The two following lemmas will be useful to study the regularity of F and h.



Voronoi diagram Laguerre diagram

Figure 2: **Laguerre diagram and semi-discrete transport maps.** We display here two colored partitions corresponding to Laguerre diagrams associated to a set $\mathcal{Y}$ with 6 points (drawn in dark blue), and with the quadratic cost $c(x,y) = \|x-y\|^2$ on $[0,1]^2$. On the left we display the Laguerre diagram for $\psi = 0$, which corresponds to the classical Voronoi diagram. On the right we display the Laguerre diagram for a specific value of ψ. The map T_ψ is drawn as blue arrows. We also display in yellow the Laguerre interface A_ψ. Notice that, for a Laguerre diagram, a point y does not necessarily belong to its Laguerre cell $\mathtt{L}_\psi(y)$.

Lemma 1. *The map $(\psi, x) \mapsto T_\psi(x)$ is locally constant on $\mathbb{R}^J \times A_\psi$.*

Proof. Let $(\psi, x) \in \mathbb{R}^J \times A_\psi$ and let $y = T_\psi(x)$. By definition of Laguerre cells,

$$c(x,y) - \psi(y) - \min_{z \in \mathcal{Y} \setminus \{y\}} (c(x,z) - \psi(z)) < 0. \tag{37}$$

The left-hand side is a function that is jointly continuous in (ψ, x) and is < 0 at (ψ, x). Thus, there is a neighborhood W of (ψ, x) where it stays negative, that is,

$$\forall (\psi', x') \in \mathsf{W}, \quad c(x', y) - \psi'(y) - \min_{z \in \mathcal{Y} \setminus \{y\}} (c(x', z) - \psi'(z)) < 0. \tag{38}$$

Therefore $T_{\psi'}(x') = y$ on W. $\square$

Lemma 2. *Assume that $\mathcal{Y}$ is finite with J points and that c is x-regular (see Definition 1). Let $\psi \in \mathbb{R}^J$. Then ψ^c is $\mathscr{C}^1$ on $\mathcal{X} \setminus A_\psi$ and*

$$\forall x \in \mathcal{X} \setminus A_\psi, \quad \nabla \psi^c(x) = \nabla_x c(x, T_\psi(x)). \tag{39}$$

Proof. First, one can notice that $\mathcal{X} \setminus A_\psi = \bigcup_{y \in \mathcal{Y}} \mathtt{L}_\psi(y)$ is an open subset of $\mathcal{X}$ because the Laguerre cells $\mathtt{L}_\psi(y)$ are open (thanks to the continuity of c). Besides, if $x \in \mathcal{X} \setminus A_\psi$, $y = T_\psi(x)$ is well-defined and $x \in \mathtt{L}_\psi(y)$. Thus $\mathtt{L}_\psi(y)$ is an open neighborhood of x on which we have

$$\forall u \in \mathtt{L}_\psi(y), \quad \psi^c(u) = c(u, y) - \psi(y) \ .$$

Therefore, on $\mathtt{L}_\psi(y)$, ψ^c is as regular as $c(\cdot, y) = c(\cdot, T_\psi(x))$ and has the same gradient. $\square$

Remark 5. *In a more general OT setting expressed in (Santambrogio, 2015, Theorem 1.47) with a $\mathscr{C}^1$ cost, if ψ^c is differentiable μ-almost everywhere and if T is a map satisfying $\nabla_x c(x, T(x)) = \nabla \psi^c(x)$, then T is an optimal transport map between μ and $T\sharp\mu$. The last lemma shows that the situation is simpler in the semi-discrete case where the differentiability of ψ^c can be shown easily.*

3.2 Regularity of F

Lemma 3. *Assume that $\mathcal{Y}$ is finite with J points and that c is x-regular. Assume that g satisfies Hypothesis (G_{θ_0}) (Definition 3) at some $\theta_0 \in \Theta$ and V the associated neighborhood of θ_0. Assume also that for any $\psi \in \mathbb{R}^J$, $\mu_\theta(A_\psi) = 0$, i.e. we have $g(\theta, Z) \in \mathcal{X} \setminus A_\psi$ almost surely.*

Then, for any $\psi \in \mathbb{R}^J$, $\theta \mapsto F(\psi, \theta)$ is differentiable at θ_0 and

$$\nabla_\theta F(\psi, \theta_0) = \mathbb{E}[D_\theta g(\theta_0, Z)^T \nabla \psi^c(g(\theta_0, Z))] \tag{40}$$

where $D_\theta g$ is the partial differential of g with respect to θ.

Proof. For $\theta \in \Theta$, let us denote

$$f(\psi, \theta, Z) = \psi^c(g(\theta, Z)) \tag{41}$$

so that we get $F(\psi, \theta) = \mathbb{E}[f(\psi, \theta, Z)]$ from (33). The hypotheses on g and Lemma 2 ensure that $f(\psi, \cdot, Z)$ is almost surely differentiable at θ_0 and thanks to the chain-rule, we have

$$\nabla_\theta f(\psi, \theta_0, Z) := D_\theta g(\theta_0, Z)^T \nabla \psi^c(g(\theta_0, Z)). \tag{42}$$

Besides, since c is x-regular, the c-transforms are L-Lipschitz thanks to Definition 1 and Lemma 6. Therefore, for any $\theta \in V$,

$$\|f(\psi, \theta, Z) - f(\psi, \theta_0, Z)\| \leqslant L\|g(\theta, Z) - g(\theta_0, Z)\| \leqslant LK(Z)\|\theta - \theta_0\| \quad \text{a.s.} \tag{43}$$

Besides, replacing θ by $\theta_0 + t(\theta - \theta_0)$ for $t \in \mathbb{R}$ and letting $t \to 0$, we also get

$$\|\nabla_\theta f(\psi, \theta_0, Z).(\theta - \theta_0)\| \leqslant LK(Z)\|\theta - \theta_0\| \quad \text{a.s.} \tag{44}$$

In particular, $\mathbb{E}[\nabla_\theta f(\psi, \theta_0, Z)]$ exists. Therefore, we have for any $\theta \in V$,

$$\frac{\|f(\psi, \theta, Z) - f(\psi, \theta_0, Z) - \nabla_\theta f(\psi, \theta_0, Z).(\theta - \theta_0)\|}{\|\theta - \theta_0\|} \leqslant 2LK(Z) \quad \text{a.s.} \tag{45}$$

When $\theta \to \theta_0$, the left-hand-side tends almost surely to zero, and thus, the dominated convergence theorem ensures that

$$\mathbb{E}\left[\frac{\|f(\psi, \theta, Z) - f(\psi, \theta_0, Z) - \nabla_\theta f(\psi, \theta_0, Z).(\theta - \theta_0)\|}{\|\theta - \theta_0\|}\right] \longrightarrow 0 \tag{46}$$

and in particular,

$$F(\psi, \theta) - F(\psi, \theta_0) - \mathbb{E}[\nabla_\theta f(\psi, \theta_0, Z)].(\theta - \theta_0) = o(\|\theta - \theta_0\|). \tag{47}$$

This proves that $F(\psi, \cdot)$ is differentiable at θ_0 with

$$\nabla_\theta F(\psi, \theta_0) = \mathbb{E}[\nabla_\theta f(\psi, \theta_0, Z)] = \mathbb{E}[D_\theta g(\theta_0, Z)^T \nabla \psi^c(g_{\theta_0}(Z))]. \tag{48}$$

□

3.3 Gradient of the loss function

Theorem 3. *Assume that $\mathcal{Y}$ is finite with J points and that c is x-regular.*
Assume that

1. *for any $\theta \in \Theta$, the Kantorovich potential ψ_* for $W(\mu_\theta, \nu)$ (defined in Theorem 2) is unique up to additive constants.*
2. *for any $\theta \in \Theta$ and any $\psi \in \mathbb{R}^J$, $\mu_\theta(A_\psi) = 0$, i.e. we have $g(\theta, Z) \in \mathcal{X} \setminus A_\psi$ a.s.*
3. *g satisfies Hypothesis (G_Θ) in Definition 3.*

Then $h(\theta) = W(\mu_\theta, \nu)$ is differentiable at any $\theta \in \Theta$ and

$$\nabla h(\theta) = \nabla_\theta F(\psi_*, \theta) = \mathbb{E}\left[D_\theta g(\theta, Z)^T \nabla \psi_*^c(g_\theta(Z))\right]. \tag{49}$$

where all terms are well-defined.

Proof. The proof consists in applying the envelope Theorem 7 (recalled in Appendix C) in order to show that h is differentiable at a fixed $\theta_0 \in \Theta$. Let us denote by ψ_{*0} the corresponding Kantorovich potential for $W(\mu_{\theta_0}, \nu)$.

First, we need to build a selection of Kantorovich potentials that is continuous at θ_0. Since we assumed that for any $\theta \in V$, the Kantorovich potential for $W(\mu_\theta, \nu)$ is unique up to additive constants, there is a unique $\psi_{*\theta} \in \arg\max F(\cdot, \theta)$ such that $\psi_{*\theta}(y_1) = 0$ (where y_1 is an arbitrary point in $\mathcal{Y}$). The continuity of $\theta \mapsto \psi_{*\theta}$ then follows from Lemma 8 and the fact that $\theta \mapsto \mu_\theta$ is weak-$\star$ continuous at θ_0. Indeed, Hypothesis 3 implies that when $\theta \to \theta_0$, $\mathbb{E}[\|g_\theta(Z) - g_{\theta_0}(Z)\|] \to 0$, which means that $g_\theta(Z) \to g_{\theta_0}(Z)$ in $L^1(\mathcal{Z}, \mathbb{R}^d)$ and thus in distribution.

The next step is to show that $\nabla_\theta F(\psi, \theta)$ exists in the neighborhood of (ψ_{*0}, θ_0), which is guaranteed by Lemma 3.

Finally, we have to demonstrate that $(\psi, \theta) \mapsto \nabla_\theta F(\psi, \theta)$ is continuous at (ψ_{*0}, θ_0). For that, we show that the function $f(\psi, \theta, Z)$ introduced in Equation (41) has a simple expression in the neighborhood of (ψ_{*0}, θ_0). Indeed, for ζ-almost all z, we can define $y = T_{\psi_{*0}}(g_{\theta_0}(z))$. Lemma 1 gives a neighborhood $\mathsf{W}_1 \times \mathsf{W}_2$ of $(\psi_{*0}, g_{\theta_0}(z))$ where $T_\psi(x) = y$. By continuity of $g(\cdot, z)$, $U = \mathsf{W}_1 \times (g(\cdot, z)^{-1}(\mathsf{W}_2))$ is a neighborhood of (ψ_{*0}, θ_0) such that $\forall (\psi, \theta) \in U$, $T_\psi(g_\theta(z)) = y$. Thus,

$$\forall (\psi, \theta) \in U, \quad \nabla_\theta f(\psi, \theta, z) = D_\theta g(\theta, z)^T \nabla \psi^c(g_\theta(z)) = D_\theta g(\theta, z)^T \nabla_x c(g_\theta(z), y). \tag{50}$$

In the neighborhood U, the right-hand side does not depend on ψ anymore. It is also continuous in θ thanks to Hypothesis (G_Θ) and the fact that c is x-regular. This proves that almost surely $\nabla_\theta f(\cdot, \cdot, Z)$ is continuous at (ψ_{*0}, θ_0). Finally, Equation (44) proves that all components of $\nabla_\theta f(\cdot, \cdot, Z)$ are almost surely bounded by $LK(Z) \in L^1(\mathcal{Z})$. Therefore, the dominated convergence theorem ensures that $\nabla_\theta F(\psi, \theta) = \mathbb{E}[\nabla_\theta f(\psi, \theta, Z)]$ is continuous at (ψ_{*0}, θ_0).

We can thus apply the envelope Theorem 7 which gives the desired result. ☐

The hypothesis that c is x-regular (see Definition 1) may not be satisfied in practice for some specific costs, like $c(x, y) = \|x - y\|$ in $\mathbb{R}^d$ related to the 1-Wasserstein distance. We now give a similar result with a relaxed hypothesis that encompasses such non-smooth cost functions.

Theorem 4. *Assume that $\mathcal{Y}$ is finite with J points. Assume that $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is continuous, and that there is a constant $L > 0$, an open set U with $\mathcal{X} \subset U \subset \mathbb{R}^d$ and a set $B \subset \mathcal{X}$ closed in $\mathbb{R}^d$ such that for any $y \in \mathcal{Y}$, $c(\cdot, y)$ can be extended to a L-Lipschitz $\mathscr{C}^1$ function on $U \setminus B$. Assume that g satisfies the three assumptions of Theorem 3, and assume also that $\mu_\theta(B) = 0$ for any $\theta \in \Theta$.*

Then $h(\theta) = W(\mu_\theta, \nu)$ is differentiable at any $\theta \in \Theta$ with the gradient expression (49).

Notice that in the case of the Euclidean distance $c(x, y) = \|x - y\|$ in $\mathbb{R}^d$, for any $y \in \mathcal{Y}$, $c(\cdot, y)$ is $\mathscr{C}^1$ only on $\mathbb{R}^d \setminus \{y\}$. Thus this cost satisfies the condition of the last theorem with $B = \mathcal{Y}$, and the resulting hypothesis on the generator reads as $\mu_\theta(\mathcal{Y}) = 0$ (in addition to the fact that the Laguerre interface A_ψ is μ_θ-negligible).

Proof. One may check that the main parts of the proof given for Theorem 3 are still true with the relaxed hypothesis on the cost. We thus only highlight the minor modifications. First, in Lemma 2, one obtains that the c-transforms ψ^c are only $\mathscr{C}^1$ on $\mathcal{X} \setminus (A_\psi \cup B)$. The proof of Lemma 3 is unchanged because $g_\theta(Z)$ almost surely belongs to $\mathcal{X} \setminus (A_\psi \cup B)$ where the c-transforms are differentiable. Finally, in the last step of the proof of Theorem 3, one should appropriately adapt the neighborhood $W_1 \times W_2$, which is simply done by replacing W_2 by $W_2 \cap B^c$. ☐

Remark 6. *There exist sufficient conditions that ensure the uniqueness of Kantorovich potentials up to additive constants. Indeed, according to (Santambrogio, 2015, Prop. 7.18), the Kantorovich potential for* $W(\mu_\theta, \nu)$ *is unique up to additive constants as soon as the support of* μ_θ *is the closure of a bounded connected open set. Assuming that* $\mu_\theta = g_\theta \sharp \zeta$ *with* ζ *being the uniform distribution on the hypercube* $Q = [-1, 1]^s$ *and* $g_\theta : Q \to \mathbb{R}^d$ *any continuous map, the support of* μ_θ *is exactly* $g_\theta(Q)$ *(see Proposition 3). If moreover the image* $g_\theta(\mathring{Q})$ *of the interior of* Q *is assumed to be open, then* $g_\theta(Q)$ *is the closure of* $g_\theta(\mathring{Q})$ *which is connected, and thus the uniqueness of Kantorovich potentials follows. Let us mention however that in the case where* g_θ *is given by a neural network,* $g_\theta(\mathring{Q})$ *is likely not to be open (for example in the expected case where the image of* g_θ *is included in an hyperplane or a manifold of dimension* $< d$*). It may also be that the uniqueness of Kantorovich potentials is not ensured for every* $\theta \in \Theta$*. Then, if one is only interested in the local behavior of* $\theta \mapsto W(\mu_\theta, \nu)$ *around a given* $\theta_0 \in \Theta$*, one may restrict* Θ *to be an open neighborhood of* θ_0*. Another way to ensure uniqueness of Kantorovich potentials is to work with entropic optimal transport, as we do in the next section.*

Remark 7. *The concepts detailed in this section are helpful to understand the impact on the generator update (e.g. gradient step on* θ*) of using a sub-optimal dual variable* ψ *in the gradient formula (Grad-OT). In other words, we can examine the impact of using the wrong* ψ *in the* ∇_θ*-descent step on*

$$F(\psi, \theta) = \mathbb{E}[\psi^c(g_\theta(Z))].$$

From Lemma 2, we have $\nabla \psi^c(g_\theta(Z)) = \nabla_x c(g_\theta(Z), T_\psi(g_\theta(Z)))$ *(e.g. with* $c(x, y) = \frac{1}{2}\|x - y\|^2$ *on* $\mathbb{R}^d$*,* $\nabla_x c(x, T_\psi(x)) = x - T_\psi(x)$*). If* ψ *is not optimal, the Laguerre diagram is not positioned correctly, and thus, one can understand that descending in the direction* $\nabla \psi^c(g_\theta(Z))$ *pushes* $g_\theta(Z)$ *towards a wrong point* $T_\psi(g_\theta(Z))$*. However, this point* $T_\psi(g_\theta(Z))$ *is still a data point, and thus descending likewise will still help* g_θ *to learn some characteristics of the data points. From this calculation, one can understand that the performance of the dual solver is important not for generating data-like points but more to ensure that the generative network is able to cover the whole database, in a way that respects the target distribution* ν*.*

Remark 8. *The previous gradient calculations also help to understand the relation between OT and WGAN learning. For simplicity, let us consider the case* $c(x, y) = \frac{1}{2}\|x - y\|^2$ *on* $\mathbb{R}^d$*. In this case, moving a point* x *with a gradient step in the direction* $\nabla_x c(x, T_\psi(x))$ *(with step size 1) brings* x *to*

$$x - \nabla_x c(x, T_\psi(x)) = x - (x - T_\psi(x)) = T_\psi(x). \tag{51}$$

In other words, taking this ∇_x*-gradient step at each point is exactly equivalent to applying the current transport map* $T_\psi(x)$*. Therefore, in the semi-discrete case, if* ψ *is adjusted so that* $T_\psi \sharp \mu = \nu$*, then* T_ψ *is a solution of the OT problem, and one can then use it to sample from* ν *(by sampling* $x \sim \mu$ *and moving it with* $y = T_\psi(x)$*). This remark underpins the "Imaginary adversary" claim of Lei et al. (2019).*

However, the WGAN problem consists in finding a measure μ_θ *(or the underlying generator* g_θ*) that solves (1)* ***among a parametric class****. In general,* T_ψ *is not a solution of the WGAN problem because there may be no* θ *such that* $T_\psi = g_\theta$*. And indeed, in the semi-discrete case, it is clearly not relevant to use* $T_\psi \sharp \mu$ *for generative modeling because it would just copy the data points (because* T_ψ *takes values in* $\mathcal{Y}$*). Actually, the gradient formula equation Grad-OT shows how the gradient steps induced by the (sub-optimal) transport map should be repercuted on* θ *and gathered (by* μ_θ*-integration) to perform the update of the generator.*

4 Gradient formula for regularized optimal transport

In this section, we provide differentiability results for h_λ in the case of regularized optimal transport. In this setting, as soon as the cost is regular, the regularized c-transforms are also regular everywhere, while requiring no assumption of the target measure ν, thus not restricted to the semi-discrete case. This makes the situation simpler than the unregularized case. Again, we recall the notations W_λ, h_λ, and F_λ from (8), (15) and (18). In order to obtain the gradient of h_λ, we follow a similar strategy than in Section 3 and study the relation between $\nabla h_\lambda(\theta)$ and $\nabla_\theta F_\lambda(\psi, \theta)$.

4.1 Gradient of the regularized c-transforms

The following lemma was already given in Genevay et al. (2019), but we recall it here for the sake of completeness.

Lemma 4. *Let $\lambda > 0$. Assume that c is x-regular (see Definition 1). Let $\psi \in \mathscr{C}(\mathcal{Y})$.*

Then $\psi^{c,\lambda}$ is $\mathscr{C}^1$ on $\mathcal{X}$ and

$$\forall x \in \mathcal{X}, \quad \nabla \psi^{c,\lambda}(x) = \int_{\mathcal{Y}} \exp\left(\frac{\psi^{c,\lambda}(x) + \psi(y) - c(x,y)}{\lambda}\right) \nabla_x c(x,y) d\nu(y) \tag{52}$$

$$= \frac{\int_{\mathcal{Y}} \exp\left(\frac{\psi(y)-c(x,y)}{\lambda}\right) \nabla_x c(x,y) d\nu(y)}{\int_{\mathcal{Y}} \exp\left(\frac{\psi(y)-c(x,y)}{\lambda}\right) d\nu(y)}. \tag{53}$$

Proof. By definition of the regularized c-transform (see relations (11) and (12)), we have for any $x \in \mathcal{X}$,

$$\forall x \in \mathcal{X}, \quad e^{-\frac{\psi^{c,\lambda}(x)}{\lambda}} = \int_{\mathcal{Y}} e^{\frac{\psi(y)-c(x,y)}{\lambda}} d\nu(y). \tag{54}$$

Since c is x-regular, then for any $y \in \mathcal{Y}$, $x \mapsto e^{\frac{\psi(y)-c(x,y)}{\lambda}}$ is $\mathscr{C}^1$ on a neighborhood of $\mathcal{X}$, and it is Lipschitz with Lipschitz constant $\frac{L}{\lambda}\exp(\frac{\|\psi\|_\infty + \|c\|_\infty}{\lambda})$. Therefore, we can differentiate under the integral to get

$$\forall x \in \mathcal{X}, \quad \nabla_x \left(e^{-\frac{\psi^{c,\lambda}(x)}{\lambda}}\right) = -\frac{1}{\lambda}\int_{\mathcal{Y}} \nabla_x c(x,y) e^{\frac{\psi(y)-c(x,y)}{\lambda}} d\nu(y). \tag{55}$$

Expanding the left-hand side, this is equivalent to

$$\forall x \in \mathcal{X}, \quad \nabla \psi^{c,\lambda}(x) e^{-\frac{\psi^{c,\lambda}(x)}{\lambda}} = \int_{\mathcal{Y}} \nabla_x c(x,y) e^{\frac{\psi(y)-c(x,y)}{\lambda}} d\nu(y), \tag{56}$$

which gives the desired formula (52) for $\nabla \psi^{c,\lambda}$. The second expression (53) follows from using the definition of $\psi^{c,\lambda}$ again. Finally, using (53), one can see that all integrated functions are continuous with respect to x, and bounded. The dominated convergence theorem thus ensures that $\nabla \psi^{c,\lambda}$ is continuous. ☐

4.2 Regularity of F_λ

We recall that Hypothesis (G_Θ) is given in Definition 3.

Lemma 5. *Let $\lambda > 0$. Assume that c is x-regular and that g satisfies Hypothesis (G_Θ).*

Let $\psi \in \mathscr{C}(\mathcal{Y})$. Then the function $\theta \mapsto F_\lambda(\psi, \theta)$ is differentiable on Θ and

$$\forall \theta_0 \in \Theta, \quad \nabla_\theta F_\lambda(\psi, \theta_0) = \mathbb{E}\left[D_\theta g(\theta_0, Z)^T \nabla \psi^{c,\lambda}(g(\theta_0, Z))\right]. \tag{57}$$

Proof. As in the proof of Lemma 3, let us introduce

$$f_\lambda(\psi, \theta, Z) = \psi^{c,\lambda}(g_\theta(Z)) \tag{58}$$

so that $F_\lambda(\psi, \theta) = \mathbb{E}[f_\lambda(\psi, \theta, Z)]$. Using Hypothesis (G_Θ) and Lemma 4, $f_\lambda(\psi, \cdot, Z)$ is differentiable on Θ almost surely, and, thanks to the chain-rule,

$$\nabla_\theta f_\lambda(\psi, \theta, Z) = D_\theta g(\theta, Z)^T \nabla \psi^{c,\lambda}(g_\theta(Z)). \tag{59}$$

Besides, the regularized c-transforms of a x-regular cost being still L-Lipschitz (by Lemma 6), we have an integrable bound for the finite differences of f_λ. Indeed, for a fixed $\theta_0 \in \Theta$, and for any θ in the neighborhood V of θ_0 given by Hypothesis (G_{θ_0}),

$$\|f_\lambda(\psi, \theta, Z) - f_\lambda(\psi, \theta_0, Z)\| \leqslant L\|g(\theta, Z) - g(\theta_0, Z)\| \leqslant LK(Z)\|\theta - \theta_0\| \quad \text{a.s.} \tag{60}$$

The proof can then be ended exactly as the one of Lemma 3. For further use, notice that the previous bound implies

$$|||\nabla_\theta f_\lambda(\psi, \theta, Z)||| = \left|\left|\left|D_\theta g(\theta_0, Z)^T \nabla \psi^{c,\lambda}(g(\theta_0, Z))\right|\right|\right| \leqslant LK(Z) \quad \text{a.s.} \tag{61}$$

where $|||\cdot|||$ is the dual norm associated with $\|\cdot\|$. ☐

4.3 Gradient of the loss with entropic regularization

Theorem 5. *Assume that c is x-regular and that g satisfies* Hypothesis (G_Θ).

Then h_λ is $\mathscr{C}^1$ on Θ and

$$\forall \theta \in \Theta, \quad \nabla_\theta h_\lambda(\theta) = \nabla_\theta F_\lambda(\psi_*, \theta) = \mathbb{E}\left[D_\theta g(\theta, Z)^T \nabla \psi_*^{c,\lambda}(g(\theta, Z))\right], \tag{62}$$

where $\psi_* \in \arg\max_\psi H_\lambda(\psi, \theta)$, and where $\nabla \psi_*^{c,\lambda}$ is given by (52).

Proof. As in Theorem 3, the proof is based on the envelope Theorem 7 applied on $h_\lambda(\theta) = W_\lambda(\mu_\theta, \nu) = \max_\psi H_\lambda(\psi, \theta)$ (see (17) and (19)). With the entropic regularization, the uniqueness (up to additive constants) of the solutions of the dual problem directly provides a continuous selection of Kantorovich potentials (thanks to Lemma 8). Besides, Lemma 5 ensures that $\nabla_\theta F_\lambda(\psi, \theta)$ exists for any $\psi \in \mathscr{C}(\mathcal{Y})$ and any $\theta \in \Theta$. It remains to show that $\nabla_\theta F_\lambda$ is continuous in (ψ, θ). For that, recall that $\nabla_\theta F_\lambda(\psi, \theta) = \mathbb{E}[\nabla_\theta f_\lambda(\psi, \theta, Z)]$ where f_λ is defined in (58), and let us fix $\theta_0 \in \Theta$ and ψ_{*0} an optimal Kantorovich potential for $W_\lambda(\mu_{\theta_0}, \nu)$, and let also $\psi \in \mathscr{C}(\mathcal{Y})$ be arbitrary. Then, for any θ in the neighborhood V of θ_0 given by Hypothesis (G_{θ_0}),

$$\|f_\lambda(\psi, \theta, Z) - f_\lambda(\psi, \theta_0, Z)\| \leqslant L\|g(\theta, Z) - g(\theta_0, Z)\| \leqslant LK(Z)\|\theta - \theta_0\| \quad \text{a.s.} \tag{63}$$

which ensures that all components of $\nabla_\theta f_\lambda(\psi, \theta, Z)$ are almost surely bounded by $LK(Z)$, an integrable bound which does not depend on (ψ, θ). Since $\theta \mapsto g(\theta, Z)$ is almost surely $\mathscr{C}^1$ on V, using (59), the last thing to show is that $(\psi, \theta) \mapsto \nabla \psi^{c,\lambda}(g_\theta(Z))$ is almost surely continuous.

Let us fix $z \in \mathcal{Z}$ for which $g(\cdot, z)$ is $\mathscr{C}^1$ on V and for which (21) holds. Thanks to (56), we can write

$$\nabla \psi^{c,\lambda}(g_\theta(z)) = e^{\frac{\psi^{c,\lambda}(g_\theta(z))}{\lambda}} \int_{\mathcal{Y}} \nabla_x c(g_\theta(z), y) e^{\frac{\psi(y) - c(g_\theta(z), y)}{\lambda}} d\nu(y). \tag{64}$$

For the first term, if $\psi, \chi \in \mathscr{C}(\mathcal{Y})$ and $\theta, \tau \in \Theta$,

$$\begin{aligned}
|\psi^{c,\lambda}(g_\theta(z)) - \chi^{c,\lambda}(g_\tau(z))| &\leqslant |\psi^{c,\lambda}(g_\theta(z)) - \psi^{c,\lambda}(g_\tau(z))| + |\psi^{c,\lambda}(g_\tau(z)) - \chi^{c,\lambda}(g_\tau(z))| && (65)\\
&\leqslant L|g_\theta(z) - g_\tau(z)| + \|\psi^{c,\lambda} - \chi^{c,\lambda}\|_\infty && (66)\\
&\leqslant LK(z)\|\theta - \tau\| + \|\psi - \chi\|_\infty, && (67)
\end{aligned}$$

by virtue of Lemma 7 and thus $(\psi, \theta) \mapsto e^{\frac{\psi^{c,\lambda}(g_\theta(z))}{\lambda}}$ is continuous. For the integral term, $\theta \mapsto \nabla_x c(g_\theta(z), y)$ is continuous on V because c is x-regular and because $\theta \mapsto g(\theta, z)$ is continuous on V. Additionally, $(\psi, \theta) \mapsto e^{\frac{\psi(y) - c(g_\theta(z), y)}{\lambda}}$ is a separable product of two terms which are continuous in ψ and θ respectively. Finally, if ψ is restricted in a neighborhood P_0 of ψ_{*0} bounded by $R > 0$, then all terms under the integral are bounded by $Le^{\frac{1}{\lambda}(R + \|c\|_\infty)}$. Again, the dominated convergence theorem implies that $(\psi, \theta) \mapsto \nabla \psi^{c,\lambda}(g_\theta(z))$ is continuous on $P_0 \times V$. Finally, using the bound (61) and the dominated convergence theorem give that $(\psi, \theta) \mapsto \nabla_\theta F_\lambda(\psi, \theta)$ is continuous on $P_0 \times V$. Therefore, all required conditions are satisfied to apply the envelope theorem on F_λ, which gives the desired formula. □

4.4 Gradient of the Sinkhorn divergence

In this whole paragraph, we assume that $\mathcal{X} = \mathcal{Y}$, that $(x, y) \mapsto c(x, y)$ is $\mathscr{C}^1$ on $\mathcal{X} \times \mathcal{X}$ and symmetric (i.e. $c(x, y) = c(y, x)$ for all $x, y \in \mathcal{X}$) and that it is L-Lipschitz with respect to (x, y):

$$\forall (x, y), (x', y') \in \mathcal{X} \times \mathcal{X}, \quad |c(x, y) - c(x', y')| \leqslant L(\|x - x'\| + \|y - y'\|). \tag{68}$$

Genevay et al. (2018) have shown that learning a generative model based on the Wasserstein cost W_λ induces a bias. For this reason, they propose to use instead the so-called Sinkhorn divergence defined as

$$S_\lambda(\mu, \nu) = W_\lambda(\mu, \nu) - \frac{1}{2}\Big(W_\lambda(\mu, \mu) + W_\lambda(\nu, \nu)\Big). \tag{69}$$

Since we focus here on the regularized WGAN learning problem, we study the function

$$s_\lambda(\theta) = S_\lambda(\mu_\theta, \nu) = W_\lambda(\mu_\theta, \nu) - \frac{1}{2}\Big(W_\lambda(\mu_\theta, \mu_\theta) + W_\lambda(\nu, \nu)\Big) \tag{70}$$

and we extend the previous regularity results to this new criterion. The first term of (70) has already been studied, and the last term does not depend on θ. It thus remains to study the middle term, and for that we rely on the dual formulation of $W_\lambda(\mu_\theta, \mu_\theta)$ given by

$$W_\lambda(\mu_\theta, \mu_\theta) = \max_{\chi, \eta \in \mathscr{C}(\mathcal{X})} \mathcal{F}_\lambda(\chi, \eta, \theta) \tag{71}$$

$$\text{where } \mathcal{F}_\lambda(\chi, \eta, \theta) = \int_{\mathcal{X}} \chi d\mu_\theta + \int_{\mathcal{X}} \eta d\mu_\theta - \lambda \int_{\mathcal{X}\times\mathcal{X}} e^{\frac{\chi(x)+\eta(y)-c(x,y)}{\lambda}} d\mu_\theta(x) d\mu_\theta(y) + \lambda. \tag{72}$$

Again, the problem (71) can be restricted to functions which are regularized c-transforms with respect to μ_θ. Similar to Lemma 5, one can show that such regularized c-transforms are still $\mathscr{C}^1$ on $\mathcal{X}$ and that the gradient can be computed by differentiating under the integral. Besides, because of the symmetry of $W_\lambda(\mu_\theta, \mu_\theta)$, one can show that the couple of optimal potentials (χ_*, η_*) that solve (71) are such that χ_* and η_* are equal up to an additive constant

Based on these observations, we can now state the regularity result for the Sinkhorn divergence. Recall that notations I, F_λ are defined in (16) and (18).

Theorem 6. *Assume that c is x-regular and that g satisfies Hypothesis (G_Θ).*

Then s_λ is $\mathscr{C}^1$ on Θ and

$$\begin{aligned} \forall \theta \in \Theta, \quad \nabla_\theta s_\lambda(\theta) &= \nabla_\theta F_\lambda(\psi_*, \theta) - \nabla_\theta I(\chi_*, \theta) && (73)\\ &= \mathbb{E}\left[D_\theta g(\theta, Z)^T \left(\nabla \psi_*^{c,\lambda}(g(\theta, Z)) - \nabla \chi_*(g(\theta, Z))\right)\right], && (74) \end{aligned}$$

where $\psi_* \in \arg\max_\psi F_\lambda(\psi, \theta)$ and $(\chi_*, \eta_*) \in \arg\max_{(\chi,\eta)} \mathcal{F}_\lambda(\chi, \eta, \theta)$.

Proof. Again, using the result obtained for $\theta \mapsto W_\lambda(\mu_\theta, \nu)$ in the last paragraphs, we only have to study the regularity of $W_\lambda(\mu_\theta, \mu_\theta)$. Once again, this follows from the envelope theorem recalled in Appendix C. For that we let C the set of $\mathscr{C}^1$ and L-Lipschitz functions on $\mathcal{X}$ equipped with the norm $\|\chi\|_\infty + \|\nabla\chi\|_\infty$ and we use the dual expression (71) that can be written

$$\mathcal{F}_\lambda(\chi, \eta, \theta) = I(\chi, \theta) + I(\eta, \theta) + \lambda - \lambda E_\lambda(\chi, \eta, \theta) \tag{75}$$

$$\text{where} \quad E_\lambda(\chi, \eta, \theta) = \mathbb{E}\left[\exp\left(\frac{\Gamma(\chi, \eta, \theta, Z, W)}{\lambda}\right)\right] \tag{76}$$

$$\text{and} \quad \Gamma(\chi, \eta, \theta, z, w) = \chi(g_\theta(z)) + \eta(g_\theta(w)) - c(g_\theta(z), g_\theta(w)), \tag{77}$$

and where W, Z are two independent random variables of distribution ζ. For any $\chi \in C$, differentiating under the integral as in Lemma 5 gives again that $I(\chi, \cdot)$ is differentiable on Θ with gradient

$$\nabla_\theta I(\chi, \theta) = \mathbb{E}\left[D_\theta g(\theta, Z)^T \nabla\chi(g(\theta, Z))\right]. \tag{78}$$

By the same reasoning, one obtains that $E_\lambda(\chi, \eta, \cdot)$ is differentiable on Θ with gradient

$$\nabla_\theta E_\lambda(\chi, \eta, \theta) = \mathbb{E}\left[\frac{\nabla_\theta \Gamma(\chi, \eta, \theta, Z, W)}{\lambda} \exp\left(\frac{\Gamma_\lambda(\chi, \eta, \theta, Z, W)}{\lambda}\right)\right] \tag{79}$$

with

$$\begin{aligned} \nabla_\theta \Gamma(\chi, \eta, \theta, z, w) =& D_\theta g(\theta, z)^T \big(\nabla\chi(g(\theta, z)) + \nabla_x c(g(\theta, z), g(\theta, w))\big) && (80)\\ &+ D_\theta g(\theta, w)^T \big(\nabla\eta(g(\theta, w)) + \nabla_y c(g(\theta, z), g(\theta, w))\big). && (81) \end{aligned}$$

If P_0 is any bounded set of C, using a bound similar to (61) and the dominated convergence theorem shows that $(\chi, \theta) \mapsto I(\chi, \theta)$ is continuous on $P_0 \times V$. Similarly, one can see that Γ is Lispchitz in θ with a (Z, W)-integrable bound

$$|\Gamma(\chi, \eta, \theta, z, w) - \Gamma(\chi, \eta, \tau, z, w)| \leqslant (\|\nabla\chi\|_\infty K(z) + \|\nabla\eta\|_\infty K(w) + L)\|\theta - \tau\|. \tag{82}$$

Since Γ is also bounded by $\|\chi\|_\infty + \|\psi\|_\infty + \|c\|_\infty$ and since the continuity of Γ with respect to (χ, η, θ) can be deduced from (77), this is enough to show that $(\chi, \eta, \theta) \mapsto \nabla E_\lambda(\chi, \eta, \theta)$ is continuous on $P_0 \times V$.

Therefore, we have again all conditions required to apply the envelope theorem: we have a continuous selection of optimal dual variables (χ, η) and the gradient $\nabla_\theta \mathcal{F}_\lambda(\chi, \eta, \theta)$ exists and is continuous in (χ, η, θ), and we can differentiate under the max in (71)

$$\nabla_\theta W_\lambda(\mu_\theta, \mu_\theta) = \nabla_\theta \mathcal{F}_\lambda(\chi_*, \eta_*, \theta), \tag{83}$$

where (χ_*, η_*) is a pair of Kantorovich potentials for $W_\lambda(\mu_\theta, \mu_\theta)$. Finally, notice that by symmetry, there exists $k \in \mathbb{R}$ such that $\chi_* = \eta_* + k$ and by definition $E_\lambda(\chi_*, \eta_*, \theta) = 1$ so that

$$\mathcal{F}_\lambda(\chi_*, \eta_*, \theta) = 2I(\chi_*, \theta) - k \tag{84}$$

and therefore

$$\nabla_\theta \mathcal{F}_\lambda(\chi_*, \eta_*, \theta) = 2\nabla I(\chi, \theta) = 2\mathbb{E}\left[D_\theta g(\theta, Z)^T \nabla\chi_*(g(\theta, Z))\right]. \tag{85}$$

Putting all terms together, we get the desired formula for $\nabla_\theta s_\lambda(\theta)$. □

5 Interpretation with Derivatives in the Sense of Distributions

This section contains a brief discussion on the way to interpret the previously obtained gradient formulae, by taking derivatives in the sense of distributions of $\theta \mapsto \mu_\theta$.

In the proofs above, one can notice that a crucial argument is to examine the regularity of the function

$$I(\varphi, \theta) = \int_\mathcal{X} \varphi d\mu_\theta. \tag{86}$$

In order to better understand the behavior of this function, it is useful to consider the map $\theta \mapsto \mu_\theta$ from Θ to the space $\mathcal{D}'(U)$ of distributions on U, which is the dual of the space $\mathcal{D}(U)$ of compactly-supported $\mathscr{C}^\infty$ functions on U (detailed definitions of these functional spaces can be found in (Hörmander, 2015)). Then, using the duality product on $\mathcal{D}'(U)$, one can rewrite

$$I(\varphi, \theta) = \langle \mu_\theta, \varphi \rangle. \tag{87}$$

Therefore, the regularity of (86) can be understood as the regularity of $\theta \mapsto \mu_\theta$, after evaluation against φ. In order to benefit from the framework of distribution derivatives, it is useful to work on the product $\Theta \times U$.

It is possible to define $T \in \mathcal{D}'(\Theta \times U)$ by setting

$$\forall \Phi \in \mathcal{D}(\Theta \times U), \quad \langle T, \Phi \rangle = \int_\Theta \int_U \Phi(\theta, x)\mu_\theta(dx)d\theta. \tag{88}$$

For any $\Phi \in \mathcal{D}(\Theta \times U)$ whose support is included in a compact $K \times L$,

$$|\langle T, \Phi \rangle| \leqslant \|\Phi\|_\infty \int_\Theta \int_U d\mu_\theta(x)d\theta \leqslant \|\Phi\|_\infty \int_K d\theta \tag{89}$$

since μ_θ is a probability distribution on U. This proves that T defines a 0-th order distribution on $\Theta \times U$. In this context, it is possible to give a meaning to the pointwise evaluation of T at θ, which corresponds to the probability distribution $T(\theta, \cdot) = \mu_\theta$.

Now, the distributional derivatives of T w.r.t. the variable θ_i can be written with the limit in the $\mathcal{D}'$ sense:

$$\frac{\partial T}{\partial \theta_i} = \lim_{h \to 0} \frac{\tau_{-ht_i} T - T}{h} \tag{90}$$

where $(t_1, \dots, t_p)$ is the canonical basis of $\mathbb{R}^p$, and where $\tau_v T$ is the translation of T with vector v, defined by $\langle \tau_v T, \Phi \rangle = \langle T, \Phi(\cdot + v) \rangle$. In other words,

$$\forall \Phi \in \mathcal{D}(\Theta \times U), \quad \left\langle \frac{\partial T}{\partial \theta_i}, \Phi \right\rangle = \lim_{h \to 0} \int_\Theta \int_U \Phi(\theta, x) \; d \left(\frac{\mu_{\theta + ht_i} - \mu_\theta}{h} \right)(x) d\theta. \tag{91}$$

Since T is a distribution of order 0, the partial derivative $\frac{\partial T}{\partial \theta_i}$ is a distribution of order $\leqslant 1$. This explains why the expression of $\frac{\partial T}{\partial \theta_i}$ may involve $\nabla \Phi$.

In the case where μ_θ is the output distribution of a generative network g that satisfies Hypothesis (G_Θ), then for a separable test function $\Phi(\theta, x) = \alpha(\theta)\varphi(x)$ with $\alpha \in \mathcal{D}(\Theta)$ and $\varphi \in \mathcal{D}(U)$, we have

$$\int_\Theta \int_U \alpha(\theta)\varphi(x) \; d \left(\frac{\mu_{\theta + ht_i} - \mu_\theta}{h} \right)(x) d\theta = \int_\Theta \alpha(\theta) \mathbb{E} \left[\frac{\varphi(g(\theta + ht_i, Z)) - \varphi(g(\theta, Z))}{h} \right] d\theta. \tag{92}$$

With arguments similar to the last sections, one can show that this limit is well-defined, which gives

$$\left\langle \frac{\partial T}{\partial \theta_i}(\theta, x), \alpha(\theta)\varphi(x) \right\rangle = \int_\Theta \alpha(\theta) \mathbb{E}[D_{\theta_i} g(\theta, Z)^T \nabla \varphi(g(\theta, Z))] d\theta. \tag{93}$$

In other words, the pointwise evaluation of $\frac{\partial T}{\partial \theta_i}$ at θ is identified to the distribution of order $\leqslant 1$ given by

$$\forall \varphi \in \mathcal{D}(U), \quad \left\langle \frac{\partial T}{\partial \theta_i}(\theta, \cdot), \varphi \right\rangle = \mathbb{E}[D_{\theta_i} g(\theta, Z)^T \nabla \varphi(g(\theta, Z))]. \tag{94}$$

In conclusion, it is possible to interpret the results of the last sections in terms of distributional derivatives of $\theta \mapsto \mu_\theta$ and this explains the apparition of $\nabla \varphi$ in the gradient formulae found in the last sections.

6 Experiments

In this section, we provide an alternate algorithm that tackles the practical minimization of (15) for generator learning with various datasets. We first consider the case of synthetic datasets with a few number of points in dimension 2, which helps to understand and explain the possible trajectories of the alternate algorithm. Then we consider the MNIST dataset (LeCun et al., 1998) composed of 60000 digits in dimension 784. We do not seek to reach state-of-the-art results for generator learning with complex databases, but we focus instead on a practical analysis of the behavior of the proposed algorithm and examine the impact of the regularization parameter and the choice of parameterization for the dual variable.

In this setting, $\mathcal{Y}$ is a finite set of J data points, so that elements $\psi \in \mathscr{C}(\mathcal{Y})$ identifies to vectors in $\mathbb{R}^J$ as in Section 3. On the other hand, μ_θ is the output distribution of the generative network, which can be sampled on demand. In this section, $\lambda \geq 0$.

6.1 Alternate Algorithm

We aim at solving the optimization problem

$$\min_{\theta \in \Theta} h_\lambda(\theta) = \min_{\theta \in \Theta} \max_{\psi \in \mathscr{C}(\mathcal{Y})} H_\lambda(\psi, \theta) \tag{95}$$

$$\text{where } H_\lambda(\psi, \theta) = \int_\mathcal{X} \psi^{c,\lambda} d\mu_\theta + \int_\mathcal{Y} \psi d\nu = F_\lambda(\psi, \theta) + \sum_{y \in \mathcal{Y}} \psi(y)\nu(\{y\}) \; . \tag{96}$$

We adopt an algorithm that alternates between updating θ with one gradient step, and updating ψ with several iterations of a dedicated algorithm.

As we have seen above in (49) and (62), computing the gradient of h_λ requires to compute an optimal dual potential $\varphi_* = \psi_*^{c,\lambda}$, which, in general, cannot be done exactly. Instead, we rely on the stochastic algorithm for semi-discrete OT proposed in (Genevay et al., 2016) to approximate the optimal potential. In the proofs of the previous sections, we have written

$$F_\lambda(\psi, \theta) = \mathbb{E}[f_\lambda(\psi, \theta, Z)] \tag{97}$$

where $f_\lambda(\psi, \theta, z) = \psi^{c,\lambda}(g_\theta(z))$. It has been shown in (Genevay et al., 2016; Houdard et al., 2022) that $H_\lambda(\cdot, \theta)$ defined in (96) is a concave function whose supergradient $\mathcal{D}(\psi, \theta) = \partial_\psi H_\lambda(\psi, \theta)$ can be written as

$$\mathcal{D}(\psi, \theta) = \mathbb{E}[\mathsf{D}(\psi, \theta, Z)] \quad \text{where} \quad \mathsf{D}(\psi, \theta, z) = \partial_\psi \left(f_\lambda(\psi, \theta, z) - \int \psi d\nu \right). \tag{98}$$

This element $\mathsf{D}(\psi, \theta, z) \in \mathbb{R}^J$ can be computed with an explicit formula given in (Genevay et al., 2016; Houdard et al., 2022) (and in practice is implemented by automatic differentiation).

Therefore, for a current θ, we can optimize ψ with a stochastic supergradient ascent

$$\begin{cases} \widetilde{\psi}_k & = \widetilde{\psi}_{k-1} + \frac{\gamma}{k^\alpha} \left(\frac{1}{|B_k|} \sum_{z \in B_k} \mathsf{D}(\widetilde{\psi}_{k-1}, \theta, z) \right) \\ \psi_k & = \frac{1}{k}(\widetilde{\psi}_1 + \cdots + \widetilde{\psi}_k), \end{cases} \tag{99}$$

where $\gamma > 0$ is the learning rate, $\alpha \in (0, 1)$ a parameter, B_k is a batch containing b independent samples of the distribution of Z, and the different batches B_k's are also independent. In ASGD, the variable $\widetilde{\psi}_0$ can be initialized at 0 (cold start), or alternately at the value of ψ obtained by ASGD for the previous θ (warm start).

As recalled in (Genevay et al., 2016) and (Galerne et al., 2018), for $\alpha = 0.5$, the ASGD algorithm has a convergence guarantee in $\mathcal{O}(\frac{\log k}{\sqrt{k}})$ on the function values. After K iterations of the inner loop, we obtain an approximation $\underline{\psi}$ of the optimal dual potential ψ_* for $W_\lambda(\mu_\theta, \nu)$, and we use it to perform the gradient descent step on θ

$$\nabla_\theta h_\lambda(\theta) \approx \nabla_\theta H_\lambda(\underline{\psi}, \theta) = \nabla_\theta F_\lambda(\underline{\psi}, \theta) = \mathbb{E}\left[D_\theta g(\theta_0, Z)^T \nabla \underline{\psi}^{c,\lambda}(g(\theta_0, Z)) \right] \tag{100}$$

Actually, using $\nabla_\theta H_\lambda(\underline{\psi}, \theta)$ as a proxy for $\nabla_\theta h_\lambda(\theta)$ in the gradient descent step can be simply reinterpreted as saying that we perform an alternate optimization on $H_\lambda(\psi, \theta)$, with an inner loop on ψ at each iteration. Again, the expectation in (100) cannot be computed in closed form so that we realize an approximation by taking another batch B' on z:

$$\nabla_\theta H_\lambda(\underline{\psi}, \theta) \approx \frac{1}{|B'|} \sum_{z \in B'} D_\theta g(\theta_0, z)^T \nabla \underline{\psi}^{c,\lambda}(g(\theta_0, z)). \tag{101}$$

The overall algorithm is summarized in Algorithm 1. Notice that in the case of unregularized OT $\lambda = 0$, the algorithm is close to the one proposed by Chen et al. (Chen et al., 2019). But all the computations made in the present paper allow to interpret it as a stochastic alternate optimization algorithm on a fixed cost $H_\lambda(\psi, \theta)$, thus including naturally the regularized case $\lambda > 0$. One benefit of this approach is that the stochastic gradient steps taken on ψ and θ can be implemented by automatic differentiation on a fixed cost H_λ, and the corresponding updates can be implemented with predefined optimizers. In particular, this opens the possibility to adopt other parameterizations of the variable ψ. Indeed, while the usual stochastic algorithm for semi-discrete OT (Genevay et al., 2016) works on the vector $(\psi(y_j)) \in \mathbb{R}^J$, it is also possible to adopt a neural network parameterization ψ_t of ψ as in (Seguy et al., 2018). The update of ψ then translates on an update of the neural network parameters t, which can be done by backpropagating the gradient of

$$t \mapsto f_\lambda(\psi_t, \theta, z) - \int \psi_t d\nu = \psi_t^{c,\lambda}(g_\theta(z)) - \sum_{y \in \mathcal{Y}} \psi_t(y) \nu(\{y\}) \,. \tag{102}$$

Finally, let us mention that a limitation of this approach is that computing the gradients of H_λ requires the differentiation of $f_\lambda(\psi, \theta, z) = \psi^{c,\lambda}(g_\theta(z))$, for a batch of z values. The exact computation of $\psi^{c,\lambda}$ requires to visit all the dataset $\mathcal{Y}$, which is prohibitive for a very large database. An alternative strategy would be to work with a batchwise version of the OT cost, but this introduces an estimation bias as demonstrated by Fatras et al. (2020).

Algorithm 1 Alternate algorithm to learn generative model μ_θ

```
Initialization: $\psi_0 = 0$ and $\theta$ (randomly)
$n = 1$ to $N$
    • Approximate $\psi_n \approx \arg\max H_\lambda(\cdot, \theta)$: inner loop with $K$ iterations of ASGD using supergradient (98) on batches $B_{n,1}, \dots, B_{n,K}$ of size $b$ on $z$ , initialized at $\psi_{n-1}$ (warm start) or at 0 (cold start)
    • Update $\theta$ with one step of ADAM algorithm on $H_\lambda(\psi_n, \cdot)$ using gradient (101) on a batch $B'_n$ of size $b$ on $z$
end for
Output: estimated generative model parameter $\theta$
```

6.2 Generator learning for Synthetic datasets

In this paragraph, we examine the behavior of the alternate Algorithm 1 used to fit a generative model to a small 2D synthetic dataset ($J = 2$ points in Section 6.2.1 and $J = 6$ points in Section 6.2.2). For each experiment, we will give diagrams, to illustrate the position of the dataset $\mathcal{Y}$ and the evolution of the support of μ_θ along the iterations n of algorithm. The square support of these diagrams corresponds to $[0, 1]^2$. The Laguerre diagram associated with a given $\psi \in \mathbb{R}^J$ will be drawn as a colored partition of $[0, 1]^2$. On the right column of the figures, we will display the evolution of the loss $H_\lambda(\psi, \theta)$. Since $H_\lambda(\psi, \theta)$ involves a μ_θ-integral, for plotting we always use a Monte-Carlo estimate obtained with a batch drawn from μ_θ.

We use the settings detailed in Appendix D.1. We use the quadratic cost $c(x, y) = \|x - y\|^2$. The generator parameters θ are here optimized with a conventional stochastic gradient descent (SGD) with fixed learning rate $\eta = 0.2$ (unless otherwise specified). The dual variable ψ is optimized with the ASGD algorithm written in (99) with learning rate γ with the cold start initialization (unless otherwise specified). It is important to notice that, in the following experiments, the number K of ASGD iterations is always fixed ($K = 100$).

6.2.1 2-point dataset

Let us consider a target distribution ν that is uniform on $\mathcal{Y} = \{y_1, y_2\}$. We will fit ν with several generative models detailed in Appendix D.1.

Dirac generator First, let us go back to the example of Proposition 2 and consider the Dirac generator $\mu_\theta = \delta_\theta$. In this case, the gradient descent on θ is actually deterministic. Besides, from the calculations of Section 2.4, we have that the global minimum of $h_\lambda(\theta) = W_\lambda(\mu_\theta, \nu)$ is reached at $\theta_* = \frac{y_1+y_2}{2}$.

The results for the Dirac generator are displayed in Fig. 3. On this diagram one can see the position of the currently estimated Laguerre diagram. Let us recall that, as illustrated in Fig. 1 for a fixed θ, the solution of the dual transportation problem makes the Laguerre interface go through θ. This is not the case when using Algorithm 1, even in this deterministic setting, so that the current point θ does not get caught by the Laguerre interface (in some diagrams of Fig. 3, it can be close but it is not exactly on it). Indeed, with the chosen initialization of θ and the chosen gradient strategies, the combination of gradient steps on ψ and θ has no reason to put the interface exactly on θ. In view of Proposition 2, it means that for a current θ, the ASGD on ψ does not get to the true optimum ψ_* in finite time, and thus the corresponding interface does not pass through θ. Therefore, in this setting the algorithm does not get trapped in a point where the gradient of $\theta \mapsto H_0(\psi, \theta)$ does not exist. This example illustrates that, in practical cases, it is very unlikely to encounter the gradient problem highlighted in Proposition 2. In contrast, the gradient problem would happen if one used optimal gradient step on ψ instead of a fixed step size. Indeed, with such strategy, one iteration of ψ would suffice to put the interface exactly on θ, which prevents from computing the next gradient.

Furthermore, in light of the gradient calculations of the previous sections, one can interpret the behavior of this alternate algorithm. The rows of Fig. 3 exhibit four different behaviors, the three first ones without regularization ($\lambda = 0$) and the last one with $\lambda = 0.1$. In the first row, the alternate algorithm stabilizes, but for a wrong reason: the learning rate γ of ASGD is too small, and thus the interface will not get past θ. Therefore, θ being always in the same Laguerre cell, the gradient descent on θ makes the generator collapse to the corresponding data point y_j (which is not even a local minimum of h_λ).

An opposite phenomenon appears in the second row, where γ is larger: the algorithm now oscillates. In this case, because of the fixed number of iterations of the inner loop, ASGD always makes a non negligible error. This reflects in the position of the interface, which hardly get close to θ, and actually moves periodically on opposite sides of θ. Thus, the point θ will be alternately pushed towards y_1 and y_2. This oscillatory behavior also reflects in the evolution of the loss function. This example shows that, **with a fixed step size strategy, the alternate algorithm may not converge in certain cases**. Note that using a decreasing step, which is often used to mitigate such an oscillating behaviour during stochastic optimization, does not necessarily solve this issue. However, for carefully tuned gradient step strategy, we observe that the algorithm stabilizes around an optimal position (slightly oscillating around θ_* in this row). Using smaller learning rates for θ allows the algorithm to stabilize closer to θ_* as illustrated in the third row.

Finally, for the regularized case shown in the fourth row, one can see that θ quickly gets close to the global optimum $\theta_* = \frac{y_1+y_2}{2}$. Indeed, on the gradient formula (62), one can understand that, at each iteration, the ∇_θ-descent direction points to a weighted average of $y_1 - \theta$ and $y_2 - \theta$, which helps to push θ towards θ_*.

This example also illustrates Remark 8: using directly the transported measure $T_\psi\sharp\mu_\theta$ never solves the WGAN problem because it would just copy y_1 or y_2 whereas the true solution δ_{θ_*} realizes a compromise between all data points. The proposed alternate algorithm is a relevant way to find this compromise, provided that its hyperparameters are carefully tuned. Besides, this example illustrates that entropic regularization may help to reach this compromise faster, while introducing a non negligible bias in the problem as explained in Genevay et al. (2018).

Disk generator Let us now examine a similar example with the same data set $\mathcal{Y} = \{y_1, y_2\}$, and a generator g_θ that induces the uniform distribution $\mu_\theta = \mathcal{U}(D(\theta, r))$ on the disk of center θ and of fixed radius r. In this case, μ_θ is absolutely continuous w.r.t. the Lebesgue measure; thus the SGD on θ is now truly stochastic. Besides, one can check that all hypotheses of Theorem 3 are satisfied in this setting, which validates the gradient formula (Grad-OT) for all values of ψ and θ.

However, even if the gradient formula holds true in that case, different behaviors can still be observed for the alternate algorithm, as illustrated in Fig. 4. All the experiments of Fig. 4 were conducted with $\lambda = 0$. One can observe that, if r is sufficiently small (rows 1 and 2), the algorithm behaves as in the case of the Dirac generator: for large value of γ the algorithm can oscillate around the optimal position $\frac{y_1+y_2}{2}$ (row 1), and for low values of γ, it can make θ collapse to one of the data point (row 2). In this latter case, it is actually the "cold start" initialization $\psi = 0$ which explains the bad positioning of the interface. Indeed, if a "warm start" initialization of ψ is used instead (row 3), then the estimated interface is now able to follow the position of θ more closely. But surprisingly, this makes θ oscillate even more, because the interface will go past θ more often.

Let us also remark that this oscillatory behavior is much reduced when the radius r of the disk is increased (rows 4, 5). Indeed, with the same gradient strategy, if the disk is larger, the interface stays close to $D(\theta, r)$ as soon as it touches it. This illustrates that having a more largely spread distribution μ_θ allows the alternate algorithm to stabilize quicker.

Other generators We finally consider in Fig. 5 other examples of generative model fitting for $\mathcal{Y} = \{y_1, y_2\}$, with more sophisticated generators: one ellipse generator (1-ellipse, row 1), one "mixture of two ellipses" generator (2-ellipse, rows 2 & 3), and a multilayer perceptron (MLP, rows 4 & 5). Those generators are detailed in Appendix D.1. In all rows of Fig. 5, one can observe that the generated distribution tends to concentrate either on $\mathcal{Y}$ or on a one-dimensional structure close to the line (y_1, y_2). The difference between rows 1 and 2 is worth noticing: if possible, the generator better has to collapse on the data points in order

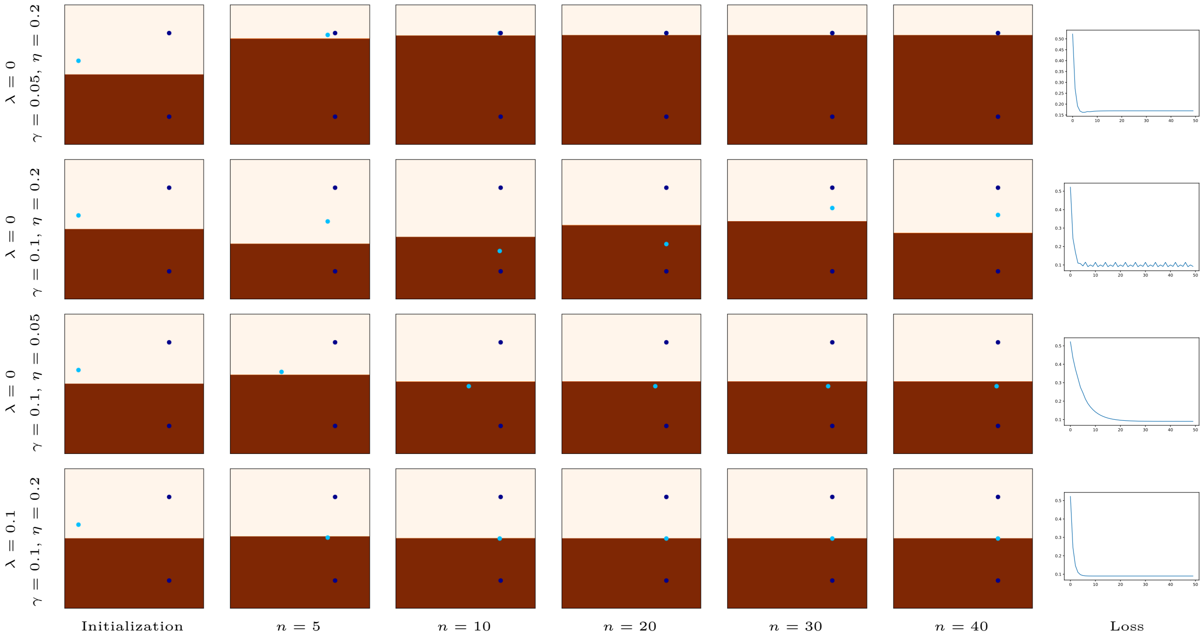


Figure 3: **Generator learning for 2-point dataset $\mathcal{Y}$ and Dirac generator δ_θ.** In the first columns, along the number of iterations n, we display $\mathcal{Y}$ (dark blue) and the current position of θ (light blue). The Laguerre diagram associated to the current value of $\underline{\psi}_\theta$ is displayed in the background with the partition indicated in colors. The last column shows the evolution of the loss $H_\lambda(\underline{\psi}_\theta, \theta)$ along the iterates. For each line, the regularization parameter λ and the learning rates γ, η for ψ and θ (respectively) are indicated on the left. Depending on the choice of parameters and gradient strategy, the alternate algorithm can oscillate or stabilize. See the text for additional comments.

to decrease the W^2 cost. This is possible in the 2-ellipse case (rows 2, 3), but not in the 1-ellipse case. This illustrates that, in addition to the chosen global loss, the choice of generative model impacts very much the practical result of generative model fitting. In particular, the possibility (or not) of "mode collapse" in the resulting generative model cannot be only attributed to the chosen loss function (e.g. the W^2 cost or its regularized version).

6.2.2 6-point dataset, neural generator

In this paragraph, we consider a dataset $\mathcal{Y}$ with 6 points, and we fit a generative model parameterized by a multilayer perceptron (MLP). The parameters of the MLP can be found in Appendix D.1 The results of this experiment are displayed in Fig. 6 below, and in Fig. 11, Fig. 12 of Appendix E. On Fig. 6, one can see that the resulting generative model concentrates on a shape, which is close to the data points for $\lambda = 0$, or degenerates to a barycenter for larger λ. This phenomenon is confirmed by the results of Fig. 11, and agrees with the fact that entropic regularization introduces a non negligible bias in the WGAN fitting problem, as already hightlighted by Genevay et al. (2018). Correcting this bias would require to use the Sinkhorn divergence S_λ instead of W_λ. However, since the source measure μ_θ is absolutely continuous, estimating S_λ (without relying on a batch approximation) does not appear straightforward. For this reason, we do not address WGAN estimation with the Sinkhorn divergence in the present paper.

One can also observe that for MLP generator learning (with a large number of parameters), the loss is often oscillating. The oscillatory behavior can be explained by the same reasons than in Section 6.2.1, and in particular the fixed number of iterations of the inner loop. The instability of the alternate algorithm can also be observed in other examples available in Appendix E. In the first row of Fig. 6, even if the loss oscillates, for $n = 200$, the generated distribution μ_θ realizes a good approximation of ν. This is not the case in the second row in which we drastically lowered the number of iterations of the inner loop to $K = 2$; we then observe a phenomenon similar to the first row of Fig 3. Again, when K is too low, the ASGD does

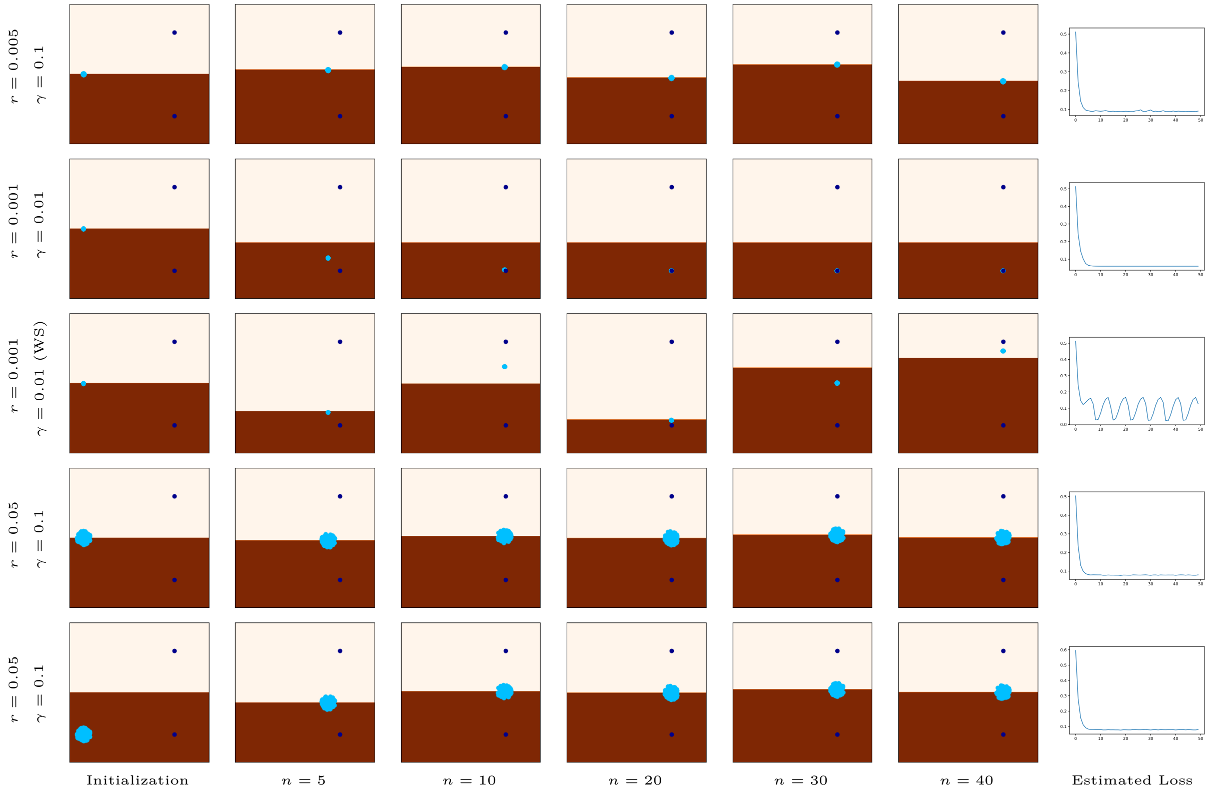


Figure 4: **Generator learning for 2-point dataset $\mathcal{Y}$ and Disk generator centered at θ.** Results are displayed using the same presentation as in Fig. 3, with samples from μ_θ in light blue. For each line, the disk radius r and the learning rate γ are indicated on the left. The regularization parameter is here fixed to $\lambda = 0$. The last column shows the evolution of the estimated loss $H_\lambda(\underline{\psi}_\theta, \theta)$ along the iterates. In the third line, we used a “warm start” (WS) initialization for ψ at each iteration, as opposed to "cold start" in other cases (see Alg. 1). Depending on the choice of parameters and gradient strategy, the alternate algorithm may oscillate or stabilize (possibly to a sub-optimal configuration). See the text for additional comments.

not converge fast enough to irrigate all points of the target distribution ν. Then, if the Laguerre cell of one data point y_j stays empty, then, from equation Grad-OT with $\lambda = 0$, one can see that y_j never appears in the update of θ, and thus nothing pushes μ_θ to visit y_j.

The examples of Fig. 6 may help to get an intuition of the behavior of WGAN learning algorithms when applied to high-dimensional datasets, e.g. datasets of images. First, depending on the choice of hyperparameters, the learning process may stabilize or not, as already noticed by (Stanczuk et al., 2021). But this does not prevent the learned distribution from realizing a good approximation of the data, in the sense that it can produce plausible samples. In this sense, it is useful to interpret the first row of Fig. 6 by imagining that the six datapoints correspond to real images. The learned distribution is able to fit a complex structure underlying the data, made of noisy curves that interpolates between datapoints. The samples obtained on the branch linking two datapoints y_j, y_k can be seen as an interpolation between those two points. With images, such an interpolation of y_j, y_k could be a blurry average (L^2 interpolation) or a better interpolation that can could be linked to a geometric deformation that transforms y_j into y_k. This is what we will observe in the next section when fitting a generative model to the MNIST database. However, we would like to emphasize that, considering the instability of the learning process and the number of parameters involved, it appears quite difficult to ensure *a priori* that a generative model with a given architecture will be appropriate for precise sampling of a given dataset, in the sense that it is able to produce convincing interpolation of the data points while not doing mode collapse nor leaving parts of the dataset aside.

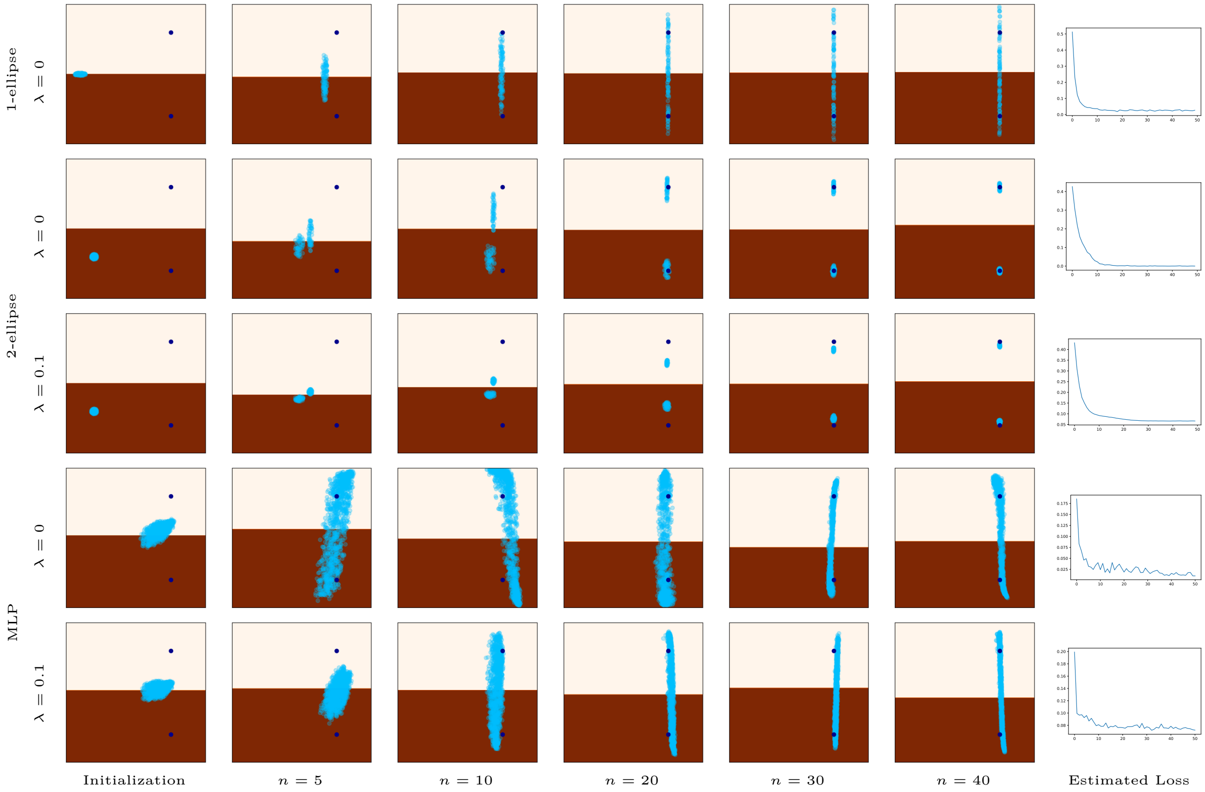


Figure 5: **Generator learning for 2-point dataset $\mathcal{Y}$ and various generators**: 1-ellipse (row 1), 2-ellipse (rows 2 & 3), and MLP (rows 4 & 5). Results are displayed using the same presentation as in Figure 4. The learning rate for ψ is here fixed to $\gamma = 0.1$. See the text comments.

6.3 Generator learning for MNIST dataset

In this paragraph, we address the problem of digits generation by leaning a generative model on the MNIST database. We will discuss the behavior of the alternate Algorithm 1 and examine the impact of the regularization parameter λ both on the visual results and the convergence of the loss function. We also discuss the effect of parameterizing the dual variable ψ by a neural network.

The detailed settings used for these experiments are given in Appendix D.2. The generator parameters θ are here optimized with ADAM algorithm (Kingma & Ba, 2015). The dual variable ψ is optimized with Pytorch implementation of ASGD (because it allows to optimize directly parameters of neural networks), with warm start initialization.

On Fig. 7, we display sampled digits obtained with the generative networks learned with Algorithm 1 run with different settings. One can see that the generators learned with unregularized OT ($\lambda = 0$) produce mostly convincing samples which are slightly more blurry than the images of the database. Some of samples do not exactly resemble a digit but some kind of mixing between different digits, which reflects the fact that the generative network naturally interpolates between the images of the database. This is in agreement with the last paragraph of Section 6.2.2. The two tested architectures for the generator produce comparable results, with a slight advantage for DCGAN. DCGAN indeed provides cleaner samples, thanks to its more complex architecture well adapted to two-dimensional data.

The visual results deteriorate when the regularization parameter λ grows. For very small λ, the results are still comparable to the unregularized case. For larger λ, the outputs of the generative network seem to concentrate on a blurry average of the database,which agrees with the estimation bias of W_λ highlighted by (Genevay et al., 2018). This can be understood by looking at the gradient formula (62) which involves

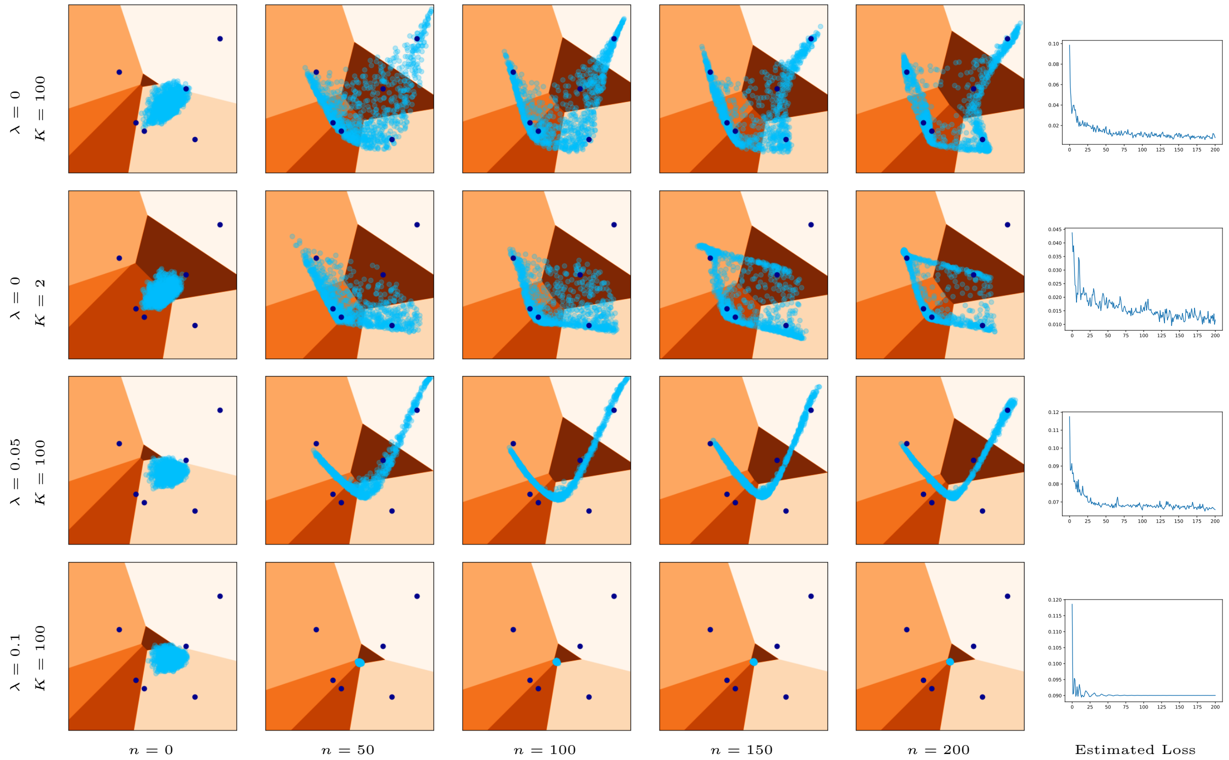


Figure 6: **Generator learning for 6-point dataset $\mathcal{Y}$ with MLP generator.** Results are displayed using the same presentation as in Figure 4. On the left we indicate the regularization parameter λ and the number of iterations K of the ASGD inner loop. See the text for comments.

the gradient of the regularized c-transform given by (52). When $\lambda \to +\infty$, $\nabla\psi^{c,\lambda}(x)$ degenerates to a simple average $\int_{\mathcal{Y}} \nabla_x c(x,y) d\nu(y)$. In other words, with the blur created by entropic regularization on the transport plan, the sampled points are pushed towards all the target points in a mixed manner. In contrast, with unregularized transport, each sampled point x is pushed towards the data point $T_\psi(x)$ assigned by the current OT map, as illustrated in Fig. 2.

To better understand these visual results, we now examine the behavior of the loss function depending on the adopted setting. Fig. 8 shows the evolution of the estimated loss $H_\lambda(\underline{\psi}_\theta, \theta)$ (for the current dual variable $\underline{\psi}_\theta$, with H_λ defined in (96)) along the iterates of Algorithm 1, when the dual variable ψ is either parameterized as a vector in $\mathbb{R}^J$ (SDOT) or a neural network (SDOTNN).

One can see that the loss stabilizes (with small oscillations) in ≈ 500 iterations, and that the limit values obtained with both parameterizations (SDOT and SDOTNN) are very similar. It is interesting to notice that the limit value is even lower with the SDOTNN parameterization: since the adopted multilayer perceptron has here $> 5 \cdot 10^5$ parameters (and is thus much larger than $J = 6 \cdot 10^4$), it is likely that any value $(\psi(y_j))_{1 \leqslant j \leqslant J} \in \mathbb{R}^J$ can be attained with such a parameterization for ψ. Notice also that the loss decreases in a more stable way with the SDOTNN parameterization: this parameterization is indeed likely to be more robust to the individual changes on $\psi(y_j)$ when updating the parameters θ of the generator.

One can notice that, quite surprisingly, the convergence speed does not improve drastically when using a larger regularization parameter λ, in contrast to what we oberved with the Dirac generator of Section 6.2.1. This is confirmed in Fig. 9 where we display results obtained with various regularization parameters λ and the four tested combinations of architectures for the generative network and the dual variable. As expected, increasing the regularization parameter leads to a smoother optimized functional, which reflects in a more stable evolution of the loss. For very small λ ($\leqslant 0.025$), we observe that the regularization does not improve the convergence speed with respect to θ. In this slightly regularized regime, we suggest that the behavior of

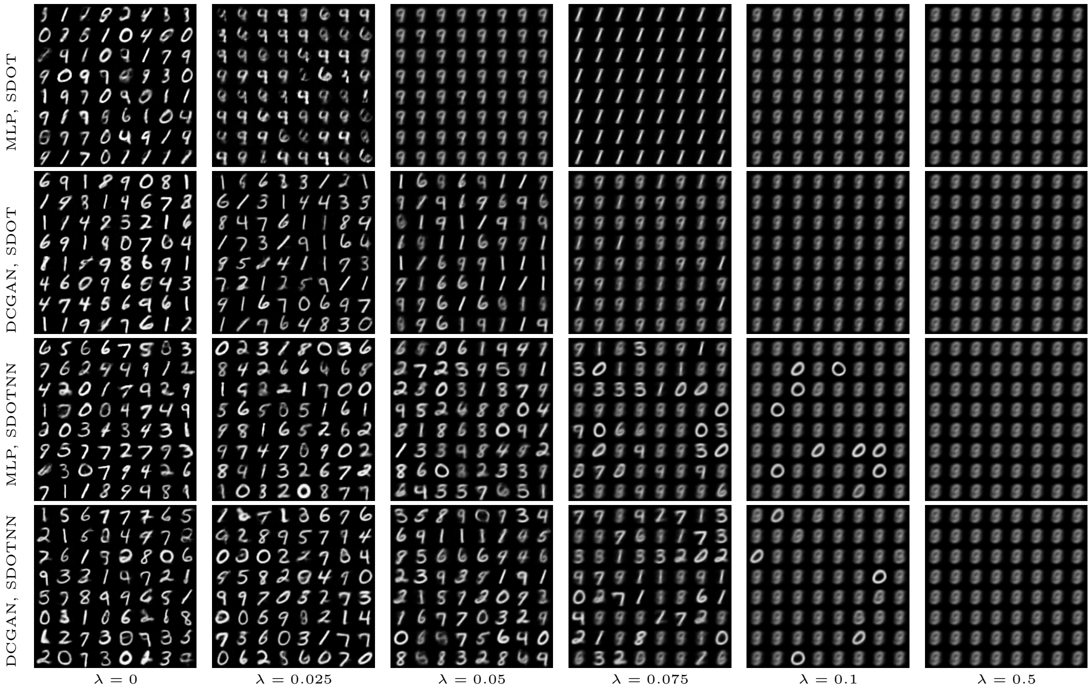
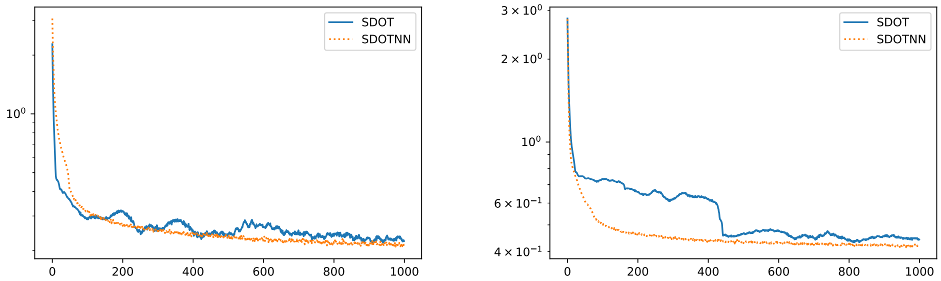


Figure 7: **Generator learning for MNIST dataset.** We compare here several generative networks trained with different architectures for the generator g_θ (MLP or DCGAN) and the dual variable ψ with no parameterization (SDOT) in the two first rows, or MLP parameterization (SDOTNN) in the two last rows, and varying the parameter λ of entropic regularization.

the loss evolution mostly depends on the chosen architecture, and the choice of hyperparameters (step size strategies for ASGD or ADAM).

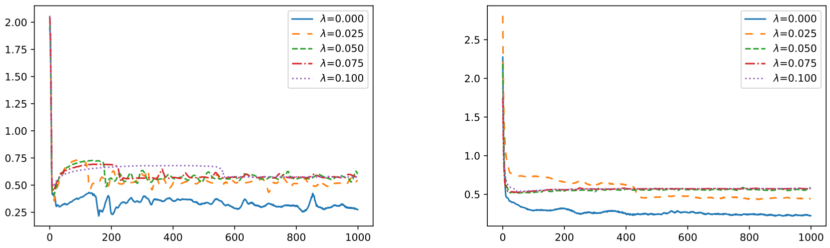


$\lambda = 0$

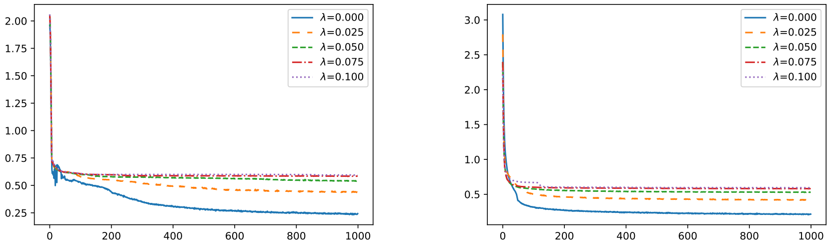
$\lambda = 0.025$

Figure 8: Evolution of the estimated loss $H_\lambda(\underline{\psi}_\theta, \theta)$ along the iterates of Algorithm 1 to learn a DCGAN for generation of MNIST digits. For each iterate θ, the loss is computed by using the current estimate $\underline{\psi}_\theta$ of the dual variable. For two values of the regularization parameter ($\lambda = 0$ on the left and 0.025 on the right), we compare the OT loss values obtained by parameterizing ψ directly as a vector in $\mathbb{R}^J$ (SDOT) or as a neural network (SDOTNN). See the text for comments.

To further analyze the algorithm, for a fixed generator parameter θ, we display in Fig. 10 the evolution of the estimated loss $\psi \mapsto H_\lambda(\psi, \theta)$ (defined in (96)) during the inner ASGD loop used for optimizing the dual variable ψ. In order to complete the comparison, we also include the convergence plot obtained with the ADAM algorithm applied on the same problem. These convergence curves reflect again the slow convergence of the ASGD algorithm. We observed that a careful tuning of the learning rate of ASGD is necessary to



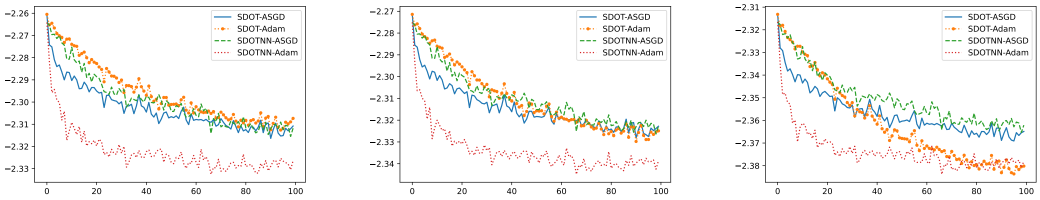
MLP, SDOT DCGAN, SDOT



MLP, SDOTNN DCGAN, SDOTNN

Figure 9: Evolution of the estimated loss $H_\lambda(\underline{\psi}_\theta, \theta)$ along the iterates of Algorithm 1, for the four tested combinations of parameterizations of the generator (MLP or DCGAN) and the dual variable (SDOT or SDOTNN). For each iterate θ, the loss is computed by using the current estimate $\underline{\psi}_\theta$ of the dual variable. Let us recall that the loss function H_λ depends on the regularization parameter λ, which explains why the limit value attained by the algorithm actually increases when $\lambda \to 0$. See the text for additional comments.

obtain a sufficient decrease of the loss. Next, the convergence plots obtained with ASGD are similar with both parameterizations SDOT and SDOTNN. One can observe that for very small regularization, turning to the ADAM algorithm does not improve the convergence speed for the SDOT parameterization. However, we remark that using the ADAM algorithm with the SDOTNN parameterization seems beneficial for all tested regularization parameters: the loss value obtained after 100 iterations is lower than with SDOTNN than with SDOT, and the convergence is much faster.



$\lambda = 0$ $\lambda = 0.001$ $\lambda = 0.005$

Figure 10: Evolution of the estimated loss $\psi \mapsto H_\lambda(\psi, \theta)$ along the iterates of the inner loop of Algorithm 1. Here, the parameter θ of the DCGAN generator is fixed, i.e. we consider a semi-discrete OT problem between a fixed μ_θ and ν. For several values of the regularization parameter, we compare the evolution of the loss when parameterizing ψ directly as a vector in $\mathbb{R}^J$ (SDOT) or as a neural network (SDOTNN). For both parameterizations, the optimization is done using either ASGD with decreasing step size ($\frac{5}{\sqrt{k}}$ for SDOT and $\frac{0.1}{k^{0.8}}$ for SDOTNN) or ADAM (with learning rate 0.001). See the text for comments.

7 Conclusion

In this paper we gave new insights on the theory and practice for learning generative networks with regularized Wasserstein distances. On the theoretical side, we proved a gradient formula for the minimized loss in two different frameworks: in the semi-discrete case (i.e. when the target distribution ν has finite support) without regularization, and in a more general case (with a general ν) with entropic regularization. These results are based on weak regularity hypotheses on the cost and the generator, which are satisfied for many $\mathscr{C}^1$ cost functions and neural network generators with $\mathscr{C}^1$ activation functions. The semi-discrete case is also based on an assumption that the generator does not charge the Laguerre interface, which is the generic case encountered in practice. These hypotheses are helpful to better understand the possible degenerate cases that can be encountered, and we provided such a counterexample.

On the practical side, we showed that an alternate algorithm can approximate the solution of this optimization problem. The inner loop of this algorithm consists in approximating an optimal dual potential for regularized OT with a stochastic optimization algorithm. With experiments on a simple low-dimensional dataset, we demonstrate that, in practice, this alternate optimization algorithm has no reason to fall on the points of inexisting gradient, but that it may exhibit various singular behaviors. In most cases, it stabilizes to a generator that form a good approximation of the target measure. But sometimes, because of the fixed number of iterations for the inner loop, it can also oscillate or get trapped in sub-optimal positions (e.g. mode collapse on a single data point), depending on the choice of hyperparameters. We also illustrated how entropic regularization can help the algorithm to converge faster towards a stable position, while inducing a clear bias in the fitting problem. Together with the theoretical results of the paper, these observations thus help to better understand the practical behavior of such a WGAN learning procedure. In particular, we claim that the sucess or failure of the learning process is not only affected by the chosen loss function (1-Wasserstein, 2-Wasserstein, Jensen-Shannon divergence, etc) but also crucially depends on the choice of generator architecture and the choice of optimization strategy.

Experiments on MNIST digits demonstrate that this algorithm is able to learn, in a high-dimensional setting, a neural network generating relevant images. Convincing visual results are indeed obtained with zero or small regularization parameter λ. For such a small regularization and in high dimension, the smoothing of the targeted loss function is not sufficient to drastically improve the convergence speed of the optimization algorithm for the generator parameters. However, for which concerns the stochastic optimization used to solve the semi-dual OT problem, we observed that it may be beneficial in terms of convergence speed to parameterize the dual variable with a neural network, provided that one uses a well-chosen and carefully tuned algorithm to optimize it. The improvement observed with such a parameterization remains to be explained with a thorough analysis of the ADAM algorithm applied on this semi-discrete OT problem.

Apart from the possible instabilities, the main limitation of the considered algorithm is that the inner loop is based on the computation of a regularized c-transform and thus requires, at each iteration, to visit all data points (in order to find a kind of biased nearest neighbor). In order to scale up to larger database, it has already been proposed (Mallasto et al., 2019a) to approximate the regularized c-transform with a batch strategy. As a perspective, it would be interesting to see if the errors made at each iteration by the ASGD early-stop or by a batch strategy could be controlled in order to get a globally stable optimization process.

References

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

F. Aurenhammer, F. Hoffmann, and B. Aronov. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1):61–76, 1998.

Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electronic Journal of Statistics*, 13(2):5120–5150, 2019.

Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International Conference on Machine Learning*, pp. 537–546. PMLR, 2017.

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*, pp. 1071–1080. PMLR, 2019.

Lénaïc Chizat. *Unbalanced Optimal Transport: Models, Numerical Methods, Applications*. PhD thesis, Université Paris Dauphine, PSL, 2017.

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

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

Muhammad Fadli Damara, Gregor Kornhardt, and Peter Jung. Solving inverse problems with conditional-gan prior via fast network-projected gradient descent. *arXiv preprint arXiv:2109.01105*, 2021.

Kilian Fatras, Younes Zine, Rémi Flamary, Remi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein : asymptotic and gradient properties. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2131–2141. PMLR, 26–28 Aug 2020.

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

Bruno Galerne, Arthur Leclaire, and Julien Rabin. A texture synthesis model based on semi-discrete optimal transport in patch space. *SIAM Journal on Imaging Sciences*, 11(4):2456–2493, 2018.

Aude Genevay. *Entropy-regularized optimal transport for machine learning*. PhD thesis, Université Paris Dauphine, 2019.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in neural information processing systems*, pp. 3440–3448, 2016.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Gan and vae from an optimal transport point of view, 2017. URL `https://arxiv.org/abs/1706.01807`.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In Amos Storkey and Fernando Perez-Cruz (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1608–1617. PMLR, 2018.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1574–1583. PMLR, 2019.

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 C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pp. 5767–5777, 2017.

Paul Hand and Babhru Joshi. Global guarantees for blind demodulation with generative priors. *Advances in Neural Information Processing Systems*, 32, 2019.

Reinhard Heckel and Mahdi Soltanolkotabi. Denoising and regularization via exploiting the structural bias of convolutional generators. In *International Conference on Learning Representations*, 2020.

Lars Hörmander. *The analysis of linear partial differential operators I: Distribution theory and Fourier analysis*. Springer, 2015.

Antoine Houdard, Arthur Leclaire, Nicolas Papadakis, and Julien Rabin. A generative model for texture synthesis based on optimal transport between feature distributions. *Journal of Mathematical Imaging and Vision*, 2022.

Rakib Hyder and M Salman Asif. Generative models for low-dimensional video representation and reconstruction. *IEEE Transactions on Signal Processing*, 68:1688–1701, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *Proceedings of the 3rd International Conference on Learning Representations, ICLR*, 2015. URL `http://arxiv.org/abs/1412.6980`.

Diederik P Kingma and Max Welling. Auto-encoding variational {Bayes}. In *Int. Conf. on Learning Representations*, 2014.

J. Kitagawa, Q. Mérigot, and B. Thibert. A Newton algorithm for semi-discrete optimal transport. *Journal of the European Math Society*, 2017.

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, Alexander Filippov, and Evgeny Burnaev. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. *Advances in Neural Information Processing Systems*, 34:14593–14605, 2021.

Alexander Korotin, Alexander Kolesov, and Evgeny Burnaev. Kantorovich strikes back! wasserstein GANs are not optimal transport? In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL `https://openreview.net/forum?id=VtEEpi-dGlt`.

Arthur Leclaire and Julien Rabin. A stochastic multi-layer algorithm for semi-discrete optimal transport with applications to texture synthesis and style transfer. *Journal of Mathematical Imaging and Vision*, 63 (2):282–308, 2021.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Na Lei, Kehua Su, Li Cui, Shing-Tung Yau, and Xianfeng David Gu. A geometric view of optimal transportation and generative model. *Computer Aided Geometric Design*, 68:1–21, 2019.

Oscar Leong. *Learned Generative Priors for Imaging Inverse Problems*. PhD thesis, Rice University, 2021.

Dong Liu, Minh Thành Vu, Saikat Chatterjee, and Lars K Rasmussen. Entropy-regularized optimal transport generative models. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3532–3536. IEEE, 2019.

Anton Mallasto, Jes Frellsen, Wouter Boomsma, and Aasa Feragen. (q, p)-wasserstein gans: Comparing ground metrics for wasserstein gans. *arXiv preprint arXiv:1902.03642*, 2019a.

Anton Mallasto, Guido Montúfar, and Augusto Gerolin. How well do wgans estimate the wasserstein metric? *arXiv preprint arXiv:1910.03875*, 2019b.

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 2437–2445, 2020.

Q. Mérigot. A multiscale approach to optimal transport. *Computer Graphics Forum*, 30(5):1583–1592, 2011.

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.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D Lee. On the convergence and robustness of training gans with regularized optimal transport. In *Advances in Neural Information Processing Systems*, pp. 7091–7101, 2018.

F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 2015.

Vivien Seguy, Bharath Bhushan Damodaran, Remi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. In *ICLR 2018-International Conference on Learning Representations*, pp. 1–15, 2018.

Fahad Shamshad and Ali Ahmed. Compressed sensing-based robust phase retrieval via deep generative priors. *IEEE Sensors Journal*, 21(2):2286–2298, 2020.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Jan Stanczuk, Christian Etmann, Lisa Maria Kreusser, and Carola-Bibiane Schönlieb. Wasserstein gans work because they fail (to approximate the wasserstein distance). *arXiv preprint arXiv:2103.01678*, 2021.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

A Notations

Here is the list of the main notations introduced in the paper.

- $\mathcal{X}, \mathcal{Y}$ compact subsets of $\mathbb{R}^d$
- If $\mathcal{Y}$ is finite, we denote by J its cardinality.
- $\mathscr{C}(\mathcal{X})$ set of real-valued continuous functions over $\mathcal{X}$
- Θ open subset of $\mathbb{R}^q$
- $\theta \in \Theta$ vector containing the parameters of the generative model
- For a function $f : \Theta \to \mathbb{R}$, we denote by $\nabla f(\theta)$ the gradient of f at point θ.
  For a fixed θ_0, in order to specify the variable in (7) we write it as $\nabla f(\theta_0) = \nabla_\theta(f(\theta))_{|\theta=\theta_0}$.
- $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ continuous cost function (ground cost)
- L Lipschitz constant of the cost function c that appears in Definition 1
- $\nabla_x c$ gradient of c with respect to its first variable
  For $x_0 \in \mathcal{X}$, $\nabla_x c(x_0, y) \in \mathbb{R}^d$ is the gradient of $x \mapsto c(x, y)$ at point x_0
- μ probability measure on $\mathcal{X}$ (source measure)
- ν probability measure on $\mathcal{Y}$ (target measure)
- $\mu \otimes \nu$ product of the measures μ, ν
- π probability measure on $\mathcal{X} \times \mathcal{Y}$ (transport plan)

- $T : \mathcal{X} \to \mathcal{Y}$ measurable map (transport map)
- $T\sharp\mu$ pushforward measure on $\mathcal{Y}$ (defined by $T\sharp\mu(B) = \mu(T^{-1}(B))$)
- $\lambda \geq 0$ entropic regularization parameter
- $W_\lambda(\mu, \nu)$ regularized optimal transport cost between μ, ν for the ground cost c, defined by (8)
- $W(\mu, \nu) = W_0(\mu, \nu)$ standard optimal transport cost for the ground cost c (without regularization)
- When $\mathcal{X} = \mathcal{Y}$ and the cost symmetric, the Sinkhorn divergence is defined by

  $$S_\lambda(\mu, \nu) = W_\lambda(\mu, \nu) - \frac{1}{2}\Big(W_\lambda(\mu, \mu) + W_\lambda(\nu, \nu)\Big).$$

- $\varphi : \mathcal{X} \to \mathbb{R}$, $\psi : \mathcal{Y} \to \mathbb{R}$ variables of the dual optimal transport problem
- c-transform $\psi^c = \psi^{c,0}$ and regularized c-transform $\psi^{c,\lambda}(x)$ defined by

  $$\forall x \in \mathcal{X}, \quad \psi^{c,\lambda}(x) = \operatorname*{softmin}_{y \in \mathcal{Y}} c(x, y) - \psi(y)$$

  with

  $$\operatorname*{softmin}_{y \in \mathcal{Y}} u(y) = \begin{cases} \min\limits_{y \in \mathcal{Y}} u(y) & \text{if } \lambda = 0, \\ -\lambda \log \displaystyle\int e^{-\frac{u(y)}{\lambda}} d\nu(y) & \text{if } \lambda > 0. \end{cases}$$

- semi-dual formulation of regularized optimal transport

  $$W_\lambda(\mu, \nu) = \sup_{\psi \in \mathscr{C}(\mathcal{Y})} \int \psi^{c,\lambda}(x) d\mu(x) + \int \psi(y) d\nu(y).$$

- δ_θ Dirac mass located at θ; it is the probability distribution of the constant random variable θ
- $\mathcal{Z}$ probability space
- Z random variable in $\mathcal{Z}$, random input of the generator with probability distribution ζ
- $g : \Theta \times \mathcal{Z} \to \mathbb{R}^d$
- $D_\theta g$ partial differential of g with respect to its first variable
  For $\theta_0 \in \Theta$, $D_\theta g(\theta_0, z)$ is the differential of $\theta \mapsto g(\theta, z)$ at point θ_0; it can be represented with a matrix in $\mathbb{R}^{d \times q}$. The operator norm of $D_\theta g(\theta_0, z)$ is denoted by $\|D_\theta g(\theta_0, z)\|$.
- g_θ refers to $g(\theta, \cdot)$; g_θ is the generator
- $\mu_\theta = g_\theta\sharp\zeta$ generative distribution, i.e. distribution of the random variable $g_\theta(Z)$
- Hypothesis (G_θ), Hypothesis (G_Θ): regularity hypotheses on the generator (see Definition 3)
- Cost function for WGAN learning with regularized optimal transport cost $h_\lambda(\theta) = W_\lambda(\mu_\theta, \nu)$
- $I : \mathscr{C}(\mathcal{X}) \times \Theta \to \mathbb{R}$ defined by $I(\varphi, \theta) = \displaystyle\int_{\mathcal{X}} \varphi d\mu_\theta, \quad (\varphi \in \mathscr{C}(\mathcal{X}),\ \theta \in \Theta)$
- $F_\lambda : \mathscr{C}(\mathcal{Y}) \times \Theta \to \mathbb{R}$ defined by $F_\lambda(\psi, \theta) = I(\psi^{c,\lambda}, \theta) = \displaystyle\int_{\mathcal{X}} \psi^{c,\lambda} d\mu_\theta = \mathbb{E}[\psi^{c,\lambda}(g_\theta(Z))]$
- $H_\lambda : \mathscr{C}(\mathcal{Y}) \times \Theta \to \mathbb{R}$ defined by $H_\lambda(\psi, \theta) = F_\lambda(\psi, \theta) + \displaystyle\int_{\mathcal{Y}} \psi d\nu = \int_{\mathcal{X}} \psi^{c,\lambda} d\mu_\theta + \int_{\mathcal{Y}} \psi d\nu$
  H_λ allows to write

  $$W_\lambda(\mu_\theta, \nu) = h_\lambda(\theta) = \max_{\psi \in \mathscr{C}(\mathcal{Y})} H_\lambda(\psi, \theta)$$

- For $\lambda = 0$, we write h, W, H, F instead of h_0, W_0, H_0, F_0
- $\mathcal{F}_\lambda : \mathscr{C}(\mathcal{X}) \times \mathscr{C}(\mathcal{X}) \times \Theta$ is defined by

$$\mathcal{F}_\lambda(\chi, \eta, \theta) = \int_{\mathcal{X}} \chi d\mu_\theta + \int_{\mathcal{X}} \eta d\mu_\theta - \lambda \int_{\mathcal{X}\times\mathcal{X}} e^{\frac{\chi(x)+\eta(y)-c(x,y)}{\lambda}} d\mu_\theta(x) d\mu_\theta(y) + \lambda.$$

- $K \in L^1(\mathcal{Z})$: $K(Z)$ is the random upper bound that appears in the regularity hypothesis on g
- For $\psi \in \mathbb{R}^J$, $\mathtt{L}_\psi(y)$ is the Laguerre cell of y (see Definition 34)
- $(\mathtt{L}_\psi(y))_{y\in\mathcal{Y}}$ is the Laguerre diagram associated to $\psi \in \mathbb{R}^J$ (see Fig. 2)
- A_ψ is the Laguerre interface associated with the Laguerre diagram of $\psi \in \mathbb{R}^J$ (see Fig. 2)
- $[y_1, y_2]$ is the segment connecting the points y_1, y_2
- $f(\psi, \theta, Z) = \psi^c(g(\theta, Z))$ so that $F(\psi, \theta) = \mathbb{E}[f(\psi, \theta, Z)]$
- $f_\lambda(\psi, \theta, Z) = \psi^{c,\lambda}(g(\theta, Z))$ so that $F_\lambda(\psi, \theta) = \mathbb{E}[f_\lambda(\psi, \theta, Z)]$
- U open subset of $\mathbb{R}^d$
- $\mathcal{D}(U)$ the space of test functions i.e. $\mathscr{C}^\infty$ functions with compact support on U
- $\mathcal{D}'(U)$ the space of distributions on U
- $T \in \mathcal{D}'(\Theta \times U)$ a distribution on $\Theta \times U$
- $\dfrac{\partial T}{\partial \theta_i}$ partial derivative of T w.r.t. to θ_i in the sense of distributions
- $\partial_\psi f$ super-gradient of the concave function $\psi \mapsto f(\psi)$
- $\mathcal{D}(\psi, \theta) = \partial_\psi H_\lambda(\psi_0, \theta)$: super-gradient of the concave function $\psi \mapsto H_\lambda(\psi, \theta)$ at point (ψ_0, θ)

B Continuity of c-transforms

In this section, we will recall some well-known facts about regularized c-transforms. For that, we need a modulus of continuity of the cost function, that is, the smallest function ω such that

$$\forall x, x' \in \mathcal{X},\ \forall y, y' \in \mathcal{Y}, \quad |c(x,y) - c(x',y')| \leqslant \omega\left(\|x - x'\| + \|y - y'\|\right). \tag{103}$$

Since c is continuous on the compact $\mathcal{X} \times \mathcal{Y}$, it is uniformly continuous, thus $\lim_{\delta\to 0} \omega(\delta) = 0$.

Lemma 6 ((Santambrogio, 2015; Feydy et al., 2019))**.** *For $\lambda \geqslant 0$, any c-transform $\psi^{c,\lambda}$ has a modulus of continuity that is bounded by the modulus of continuity of the cost function.*

Proof. If $u \leqslant v$ holds pointwise, then $\operatorname{softmin} u \leqslant \operatorname{softmin} v$ pointwise. Also, for a constant $k \in \mathbb{R}$, $\operatorname{softmin}(k + u) = k + \operatorname{softmin}(u)$. But, from the definition of ω, we have

$$c(x, y) - \psi(y) \leqslant \omega(\|x - x'\|) + c(x', y) - \psi(y). \tag{104}$$

By taking the soft-min, we thus obtain

$$\psi^{c,\lambda}(x) \leqslant \omega(\|x - x'\|) + \psi^{c,\lambda}(x'), \tag{105}$$

and by symmetry, this leads to

$$|\psi^{c,\lambda}(x) - \psi^{c,\lambda}(x')| \leqslant \omega(\|x - x'\|). \tag{106}$$

□

Lemma 7. *For $\lambda \geqslant 0$, and any $\psi, \chi \in \mathscr{C}(\mathcal{Y})$, $\|\psi^{c,\lambda} - \chi^{c,\lambda}\|_\infty \leqslant \|\psi - \chi\|_\infty$.*

In other words, the map $\psi \mapsto \psi^{c,\lambda}$ is 1-Lipschitz for the uniform norm.

Proof. Applying again the monotonicity of the softmin operation to the inequality

$$c(x,y) - \psi(y) \leqslant \|\psi - \chi\|_\infty + c(x,y) - \chi(y) \tag{107}$$

we get $\psi^{c,\lambda}(x) \leqslant \|\psi - \chi\|_\infty + \chi^{c,\lambda}(x)$, which gives the desired result by a symmetry argument. ☐

The following lemma states that the optimal dual potentials vary continuously with respect to the input measure as soon as they are unique up to additive constants.

Lemma 8 ((Feydy et al., 2019))**.** *Assume that $\mathcal{X}, \mathcal{Y}$ are compact, and that c is continuous. Let us fix $x_0 \in \mathcal{X}$. Assume that ν is fixed, and that μ_n converges weak-$\star$ in $\mathcal{M}_+^1(\mathcal{X})$ to a measure μ. Assume furthermore that the Kantorovich potentials associated with $W_\lambda(\mu,\nu)$ are unique up to an additive constant (which is always the case if $\lambda > 0$). For each n, let φ_n be a c-concave Kantorovich potential for $W_\lambda(\mu_n,\nu)$ such that $\varphi_n(x_0) = 0$ and let φ be the (necessarily c-concave) Kantorovich potential for $W_\lambda(\mu,\nu)$ such that $\varphi(x_0) = 0$.*

Then $\varphi_n \to \varphi$ uniformly on $\mathcal{X}$.

Proof. By compactness of $\mathcal{X} \times \mathcal{Y}$, c has a bounded modulus of continuity $\omega(\delta)$ which tends to zero when $\delta \to 0$. By (106), we obtain that the functions $|\varphi_n|$ are bounded by $\sup_{x\in\mathcal{X}} \omega(\|x - x_0\|) < \infty$. Besides, Lemma 6 also shows that the functions φ_n are uniformly equicontinuous on $\mathcal{X}$. Therefore, Arzela-Ascoli theorem ensures that the family $\{\varphi_n, n \in \mathbb{N}\}$ is relatively compact in $\mathscr{C}(\mathcal{X})$.

Now, assume, by contradiction, that (φ_n) does not tend to φ in $\mathscr{C}(\mathcal{X})$. Then there would exist $\varepsilon > 0$ and a subsequence $(\varphi_{r(n)})$ such that

$$\|\varphi_{r(n)} - \varphi\|_\infty > \varepsilon \quad \forall n. \tag{108}$$

By relative compactness, one can then extract a subsequence $(\varphi_{r(s(n))})$ which converges in $\mathscr{C}(\mathcal{X})$ to a function $\tilde{\varphi}$. Using the monotonicity of soft-min, this implies that $(\varphi_{r(s(n))}^{c,\lambda})$ also converges in $\mathscr{C}(\mathcal{X})$ to $\tilde{\varphi}^{c,\lambda}$. Thus

$$W(\mu_n,\nu) = \int \varphi_n d\mu + \int \varphi_n^{c,\lambda} d\nu \xrightarrow[n\to\infty]{} \int \tilde{\varphi} d\mu + \int \tilde{\varphi}^{c,\lambda} d\nu. \tag{109}$$

Finally, we also have $W(\mu_n,\nu) \to W(\mu,\nu)$ since $\mu_n \overset{*}{\rightharpoonup} \mu$. Thus

$$W(\mu,\nu) = \int \tilde{\varphi} d\mu + \int \tilde{\varphi}^{c,\lambda} d\nu \tag{110}$$

and $\tilde{\varphi}$ is a Kantorovich potential for $W(\mu,\nu)$ and $\tilde{\varphi}(x_0) = \lim \varphi_n(x_0) = 0$. Using the uniqueness assumption, we get $\tilde{\varphi} = \varphi$ which contradicts (108). ☐

C Technical Results

We first recall the proof of the envelope theorem (Oyama & Takenawa, 2018, Prop. A.1).

Theorem 7 (Envelope theorem Oyama & Takenawa (2018))**.** *Let X be a topological space. Let A be an open set of a normed vector space E. Let $f : X \times A \to \mathbb{R}$ be a function and let us denote*

$$\forall a \in A, \quad v(a) = \sup_{x\in X} f(x,a). \tag{111}$$

Let $s : A \to X$ be such that for all $a \in A$, $v(a) = f(s(a),a)$. Let $\alpha \in A$ be a point such that

- *s is continuous at α,*
- *the partial differential $D_a f$ of f with respect to a exists in a neighborhood of $(s(\alpha),\alpha)$, and is continuous at $(s(\alpha),\alpha)$.*

Then v is differentiable at α and $D_a v(\alpha) = D_a f(s(\alpha), \alpha)$.

Proof. Let $\xi = s(\alpha)$ and let $\varepsilon > 0$. The second hypothesis gives an open neighborhood $U \times V$ of (ξ, α) in $X \times A$ such that for any $(x, a) \in U \times V$, $f(x, \cdot)$ is differentiable at a and such that the partial differential $(x, a) \mapsto D_a f(x, a)$ is continuous on $U \times V$. By continuity of s, $s^{-1}(U) \cap V$ is an open neighborhood of α and thus it exists $\eta > 0$ such that $\|h\| < \eta$ implies $\alpha + h \in V$ and $s(\alpha + h) \in U$.

By definition of v, we have for any $h \in E$ such that $\|h\| < \eta$,

$$f(\xi, \alpha + h) - f(\xi, \alpha) \leqslant v(\alpha + h) - v(\alpha) \leqslant f(s(\alpha + h), \alpha + h) - f(s(\alpha + h), \alpha). \tag{112}$$

On the one hand, by definition of $D_a f(\xi, \alpha)$, there exists $\eta_1 \in (0, \eta)$ such that $\|h\| < \eta_1$ implies

$$|f(\xi, \alpha + h) - f(\xi, \alpha) - D_a f(\xi, \alpha)h| \leqslant \varepsilon\|h\|. \tag{113}$$

On the other hand, for $\|h\| < \eta$, $t \in [0, 1] \mapsto f(s(\alpha + h), \alpha + th)$ is differentiable on $[0, 1]$ and therefore, there exists $\theta_h \in (0, 1)$ such that

$$f(s(\alpha + h), \alpha + h) - f(s(\alpha + h), \alpha) = D_a f(s(\alpha + h), \alpha + \theta_h h)h. \tag{114}$$

By continuity of $D_a f$, there is an open neighborhood $\bar{U} \times \bar{V} \subset U \times V$ such that

$$\forall (x, a) \in U \times V, \quad \|D_a f(x, a) - D_a f(\xi, \alpha)\| \leqslant \varepsilon, \tag{115}$$

where $\| \cdot \|$ denotes the dual norm. Again, by continuity of s, $s^{-1}(\bar{U}) \cap \bar{V}$ is an open neighborhood of α and thus, there exists $\eta_2 \in (0, \eta)$ such that $\|h\| < \eta_2$ implies $\alpha + h \in s^{-1}(\bar{U}) \cap \bar{V}$. Therefore, for $\|h\| < \eta_2$, $(s(\alpha + h), \alpha + \theta_h h) \in \bar{U} \times \bar{V}$ and thus

$$|f(s(\alpha + h), \alpha + h) - f(s(\alpha + h), \alpha) - D_a f(\xi, \alpha)h| \tag{116}$$

$$\leqslant \|D_a f(s(\alpha + h), \alpha + \theta_h h) - D_a f(\xi, \alpha)\|\|h\| \leqslant \varepsilon\|h\|. \tag{117}$$

Finally, for $\|h\| < \min(\eta_1, \eta_2)$ we get

$$|v(\alpha + h) - v(\alpha) - D_a f(\xi, \alpha)h| \leqslant \varepsilon\|h\|, \tag{118}$$

which proves that v is differentiable at α and $D_a v(\alpha) = D_a f(\xi, \alpha)$. ☐

The next proposition gives the support of a push-forward distribution.

Proposition 3. *Let $Q = [-1, 1]^s$, and $g : Q \to \mathbb{R}^d$ continuous. Let Z be a random variable with uniform distribution ζ on Q and let $\mu = g\sharp\zeta$ be the distribution of $g(Z)$. Then the support of μ is exactly $g(Q)$.*

Proof. Since g is continuous, $g(Q)$ is compact and in particular closed. Thus $U = g(Q)^c$ is open, and one has that

$$\mu_\theta(U) = \mathbb{P}(g(Z) \in U) = \zeta(g^{-1}(U)) = 0, \tag{119}$$

because $g^{-1}(U)$ does not intersect Q. This proves that $\mathrm{Supp}(\mu) \subset g_\theta(Q)$. Now, if V is an open set such that $\mu(V) = 0$, then $\mathbb{P}(Z \in g_\theta^{-1}(V)) = 0$, which gives $g_\theta^{-1}(V) \cap Q = \varnothing$ because $g_\theta^{-1}(V)$ is open. It follows that $V \subset g_\theta(Q)^c$, which proves that $\mathrm{Supp}(\mu)$ is exactly $g_\theta(Q)$. ☐

D Detailed Settings used for Experiments on generator learning

D.1 Detailed Settings for the experiments with synthetic datasets

The following paragraph gathers the parameters and network architectures used in the experiments of Section 6.2.

- $N = 50$ iterations on θ with SGD algorithm (with constant learning rate $\eta = 0.2$), except for the MLP generator which is optimized with Adam (with constant learning rate 0.05)
- $K = 100$ iterations of the inner loop with ASGD algorithm (see learning rates below) with cold start initialization (unless otherwise specified)
- The cost $c(x, y)$ is the quadratic cost $\|x - y\|^2$.
- For the generator we consider various architectures g_θ and input noise Z:
  - a Dirac generator $g_\theta^{\text{dirac}}(z) = \theta$ parameterized by $\theta \in \mathbb{R}^2$,
  - a Disk generator with fixed radius $r > 0$,

    $$g_\theta^{\text{disk}}(z) = \theta + r\sqrt{z_1}(\cos(2\pi z_2), \sin(\pi z_2))$$

    with noise Z uniform on $[0,1]^2$, parameterized by $\theta \in \mathbb{R}^2$,
  - an ellipse generator (denoted by 1-ellipse)

    $$g_\theta^{\text{ell}}(z) = t + \sqrt{z_1}(a\cos(2\pi z_2), b\sin(\pi z_2))$$

    with noise Z uniform on $[0,1]^2$, parameterized by $\theta = (t, a, b) \in \mathbb{R}^2 \times \mathbb{R} \times \mathbb{R}$
  - a 2-mixture of ellipse generators (denoted by 2-ellipse)

    $$g_\theta^{\text{mixell}}(z) = \begin{cases} g_{\theta_1}^{\text{ell}}(z_1, z_2) & \text{if} \quad z_3 < \frac{1}{2} \\ g_{\theta_2}^{\text{ell}}(z_1, z_2) & \text{otherwise} \end{cases}$$

    with noise Z uniform on $[0,1]^3$, parameterized by $\theta = (\theta_1, \theta_2) \in (\mathbb{R}^2 \times \mathbb{R} \times \mathbb{R})^2$
  - a multilayer perceptron (MLP) with input noise Z uniform on $[0,1]^{\mathsf{din}}$, nhid fully-connected layers with dimension dhid and activation function $\mathsf{Elu}(1)$. In the experiments shown in Section 6.2.2, the layers are set by $\mathsf{din} = 5$, $\mathsf{dhid} = 10$, $\mathsf{nhid} = 3$. For the experiments in Appendix E, the parameters are indicated accordingly.
- All batches of Z are made of $b = 100$ samples, except for the Dirac generator where 1 sample suffices.
- The dual variable ψ is modeled by a vector $\psi \in \mathbb{R}^J$.
- The dual variable ψ is optimized with the ASGD algorithm written in (99). The step size strategy is $\frac{\gamma}{k^{0.5}}$ with learning rate $\gamma > 0$ (values indicated in the text).

D.2 Detailed Settings for the MNIST experiment

The following paragraph gathers the parameters and network architectures used in the experiments of Section 6.3.

- $N = 3000$ iterations on θ with ADAM algorithm Kingma & Ba (2015) with learning rate 0.001
- $K = 10$ iterations of the inner loop with ASGD algorithm (see learning rates below) with warm start initialization
- The cost $c(x, y)$ is the quadratic cost $\alpha^{-1}\|x - y\|^2$ normalized by $\alpha = \frac{1}{J}\sum_{y \in \mathcal{Y}} \|y\|^2$
- For the generator g_θ we consider two different architectures:
  - a multilayer perceptron (MLP) with four fully-connected layers; the number of channels for the successive hidden layers is $256, 512, 1024$ with activation functions $\mathsf{LeakyReLU}(0.2)$.
  - a Deep Convolutional Adversarial Network (DCGAN) (Radford et al., 2015) adapted for the dimension 28×28 of MNIST images, with four deconvolution layers; the number of channels for the successive hidden layers is $256, 128, 64$.

- The input of these generators is a random variable Z following the uniform distribution on $[-1, 1]^{100}$ (the choice of dimensionality 100 is commonly encountered when fitting a generative network to the MNIST database). All batches of Z are made of $b = 200$ samples.
- The dual variable ψ is either directly modeled by a vector $\psi \in \mathbb{R}^J$, or parameterized by a multilayer perceptron with four fully-connected layers; the number of channels for the successive hidden layers is $512, 256, 128$. These two different settings are respectively referred to as SDOT (for semi-dual OT) and SDOTNN (for semi-dual OT with neural network).
- The dual variable ψ is optimized with the Pytorch implementation of ASGD (`torch.asgd`), which has parameters `lr` and `alpha`. The step size strategy for the ASGD inner loop has been chosen in the following manner:
  - for the parameterization SDOT, $\texttt{lr} = 5, \texttt{alpha} = 0.5$,
  - for the parameterization SDOTNN, $\texttt{lr} = 0.1, \texttt{alpha} = 0.8$.

  The other parameters are fixed to default values.

E Additional experiments on 6-point dataset

In this appendix we provide additional results of generative model learning on the 6-point dataset.

In Fig. 11, for a fixed large MLP architecture, we vary the regularization parameter λ.

In Fig. 12, we fix the regularization parameter $\lambda = 0$ and we vary the configuration of the MLP architecture.

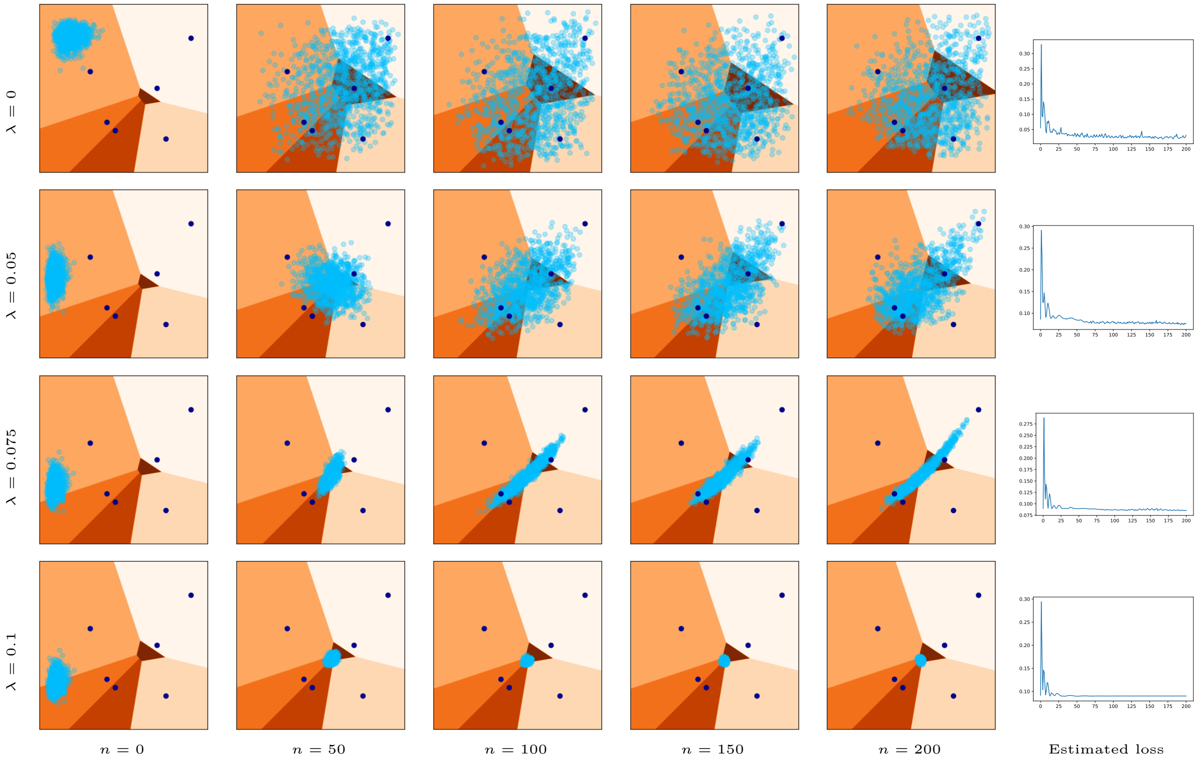


Figure 11: **Generator learning for 6-point dataset $\mathcal{Y}$ with MLP generator.** The parameters of the MLP are here $\mathsf{din} = 100$, $\mathsf{dhid} = 256$, $\mathsf{nhid} = 3$. In the first columns, along the number of iterations n, we display $\mathcal{Y}$ (dark blue) and the current position of μ_θ (by showing sampled points in light blue) The Laguerre diagram associated to the current value of $\underline{\psi}_\theta$ is displayed in the background with the partition indicated in colors. The last column shows the evolution of the estimated loss $H_\lambda(\underline{\psi}_\theta, \theta)$ along the iterates. On the left we indicate the regularization parameter λ.

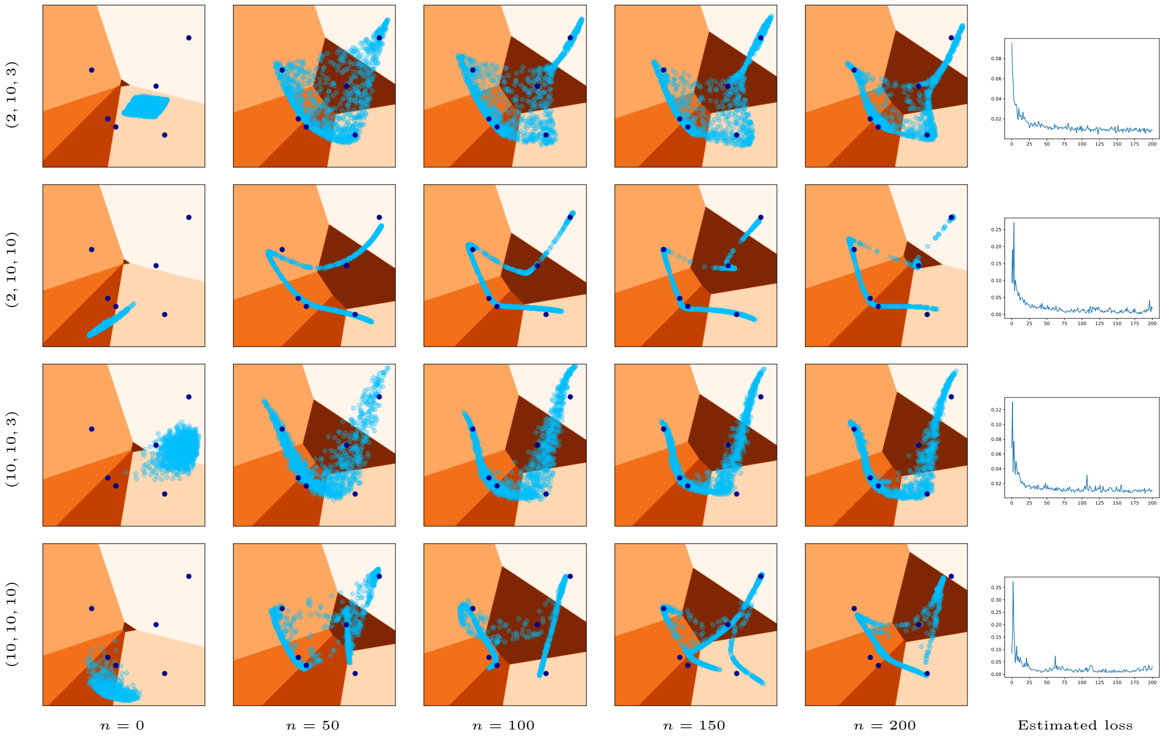


Figure 12: **Generator learning for 6-point dataset $\mathcal{Y}$ with MLP generator.** On the left we indicate the tuple of parameters $(\mathsf{din}, \mathsf{dhid}, \mathsf{nhid})$. In the first columns, along the number of iterations n, we display $\mathcal{Y}$ (dark blue) and the current position of μ_θ (by showing sampled points in light blue) The Laguerre diagram associated to the current value of $\underline{\psi}_\theta$ is displayed in the background with the partition indicated in colors. The last column shows the evolution of the estimated loss $H_\lambda(\underline{\psi}_\theta, \theta)$ along the iterates. The regularization parameter is set here to $\lambda = 0$.