Regularized Optimal Transport

Tao Li[*] TAOLI@NYU.EDU
Yuan Ni[*] YN754@NYU.EDU
Michael Stanley[*] MHS592@NYU.EDU
Nikhil Supekar[*] NS4486@NYU.EDU

Abstract

Optimal transport (OT) provides a geometrically meaningful framework to compare probability distributions. However, the practical impact of OT is rather limited because of its computational burden and statistical concerns. As a remedy, regularized OT is proposed, consisting of original OT and a regularization term, which turns out to be helpful both in theory and practice. In this paper, we aim to explain why regularized optimal transport helps both computationally and statistically. We first point out that thanks to the dual formulation of regularized OT, the computation of OT is essentially a stochastic programming, which can be efficiently solved by first-order methods, which further paves the way of training generative models. In addition to its computational advantages, regularized OT also breaks the curse of dimensionality when estimating the Wasserstein distance from samples.

1. Introduction

Optimal Transport (OT) refers to the problem of optimizing cost of transportation and allocation of resources. The problem was first formalized by the french mathematician Gaspard Monge in 1781 who was looking for the most economical way of moving soil from one place to another. The effectiveness of the solution is determined by the minimality of the integral of the distances the soil is moved (point-to-point). This is referred to as Monge's variational formulation. The point-to-point mapping is referred to as a transport plan. Intuitively, we can consider moving soil from one place to another as comparing two piles of soils which ultimately resembles comparison of two probability distributions. The amount of work taken for moving soil from one place to another can be compared with the cost associated with transforming distributions into one another, measuring the dissimilarity between distributions. Such measurement is of great significance in modern data science and machine learning applications.

Probabilistic language provides a natural means of formalizing notions of uncertainty in high-dimensional statistical contexts, and optimal transport as we have briefly discussed above, shows promise for making geometry work in probabilistic regime. Therefore, it is not surprising that in recent years, optimal transport has been applied to a wide range of problems in machine learning, computer graphics, image processing, and document retrieval. Bonneel et al. (2011) solve a mass transport problem by applying partial transport to perform interpolation between pairs of values in computer graphics applications. Solomon et al. (2014) apply the theory of optimal transportation to introduce a novel method for computing the earth mover's distance and apply it to problems in graphics and geometry processing. de Goes et al. (2011) propose a robust 2D shape reconstruction algorithm which devises a fine-to-coarse scheme using optimal transport to approximate a defect-laden point set by a simplical complex. Kusner et al. (2015) propose the Word Mover's Distance that measures the dissimilarity between two text documents which can be cast as an instance of the Earth Mover's Distance based on Optimal Transport.

[*] Authors are listed in alphabetic order and contribute equally to this manuscript

However, the direct application of traditional optimal transport suffers from a few drawbacks. First, optimal transport is shown to be computationally expensive. Pele and Werman (2009) show that, without the use of embeddings, the cost of computing optimal transport scales in at least $O(d^3 \log(d))$. Dobri ́c and Yukich (1995) derived a rate of convergence for transportation cost in high dimensions and showed that vanilla optimal transport applied to sampled data is cursed by dimensionality. Both computational and statistical concerns limits the efficacy of optimal transport for the ever larger data of modern applications and these concerns suggest the need for regularization in order to scale optimal transport to high-dimensional problems.

Regularized Optimal Transport Compared with vanilla optimal transport. the regularized has appealing advantages. First, it yields a more reasonable optimal transport model for specific problems, enforcing certain properties of the transport plan in question, such as sparsity (Blondel et al., 2018). Second, these regularizers, often convex, serve the purpose of convex relaxation, which makes the computation of the transport plan more tractable and more efficient. Also, adding suitable regularizers helps overcome the curse of dimensionality coming from sampling noise and make the Wasserstein distance estimated from high-dimensional data more statistically stable.

Regularized optimal transport has found success in many recent applications. With **efficient computation** of regularized optimal transport, Gramfort et al. (2015) use entropic smoothing leading to a smooth convex optimization problem applied to functional neuroimaging data, functional MRI and magnetoencephalography (MEG) source estimates. Abraham et al. (2015) explore Tomographic reconstruction to yield an estimate of a system from a finite number of projections. Sanjabi et al. (2018) use the regularized Wasserstein distance for computational efficiency of gradients which allows first order optimization techniques with convergence guarantees in Generative Adverserial Networks (GANs). Leveraging **statistical stability** brought by regularization, Genevay et al. (2019b) rewrite the k-means algorithm as an optimal transport task for learning a low-dimensional embedding that can better reflect the geometry of the data and derive a differentiable loss function that can be optimized via SGD. Cuturi et al. (2019) recover differentiable operators for sorting and ranking which can be integrated into end-to-end, automatically differentiable deep learning pipelines.

Main focus Given both theoretical and empirical successes of regularized optimal transport from both computational and statistical perspectives, there is a call for a better understanding of how and when regularization helps. The main purpose of this survey paper is to answer these questions by analyzing regularizers across two dimensions. One is about **computation** or optimization, which means that we treat the optimal transport problem as an optimization, where the regularizer mainly contributes to reducing computational complexity and imposing extra constraints on the solution, e.g., sparsity. The second is **statistical**, where we are interested in how regularizers impact statistical stability or robustness when dealing with high-dimensional data. Throughout the survey paper, we shall mainly revolve around the family of entropy regularization, which is one of the most popular regularizers in machine learning. Our theoretical analysis will be accompanied by illustrative examples from the application side of optimal transport.

The rest of the paper is organized as follows. Section 2 introduces the mathematical background of optimal transport and its regularization. Section 3 includes the computational advantages of regularized optimal transport and especially, Wasserstein generative adversarial networks is discussed for illustrating those advantages brought by regularization. In addition to the computational aspect, the statistical aspect of regularized optimal transport is detailed in section 4. Finally, we summarize our findings in the conclusion.

2. Background

In this section, we shall lay a solid mathematical foundation for our further discussion on optimal transport and its regularization. To be specific, we will first introduce Kantorovich formulation of optimal transport and its dual formulation. Then, we shall move to regularized optimal transport, where the regularizers can be drawn from the family of φ-divergences. Among these divergences, we are particularly interested in the one that leads to entropy regularization, which is the main focus of our discussions on computational and statistical aspect of regularized optimal transport in the following sections.

2.1 Optimal Transport

To illustrate optimal transport, consider α, the source measure, denoting a pile of soil and β, the target measure, denoting a hole to be filled up. Both measures are assumed to be probability measures over spaces $\mathcal{X}, \mathcal{Y}$ respectively. Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ denote the cost function denoting the cost of transporting unit mass of soil from $x \in \mathcal{X}$ to $y \in \mathcal{Y}$. The optimal transport problem concerns transporting α to β while minimizing c. More formally, if we denote $\mathcal{M}_+(\mathcal{X})$ the set of positive Radon probability measures on $\mathcal{X}$, given $\alpha \in \mathcal{M}_+(\mathcal{X})$ and $\beta \in \mathcal{M}_+(\mathcal{Y})$, the Monge formulation of OT is as below,

$$\begin{aligned} \min_{T} \quad & M(T) = \int_{\mathcal{X}} c(x, T(x)) d\alpha(x) \\ \text{subject to } & \beta = T_{\#}\alpha \end{aligned}$$

where $T : \mathcal{X} \to \mathcal{Y}$ is a measurable map and $T\#\alpha$ is called the push-forward of α by T on $\mathcal{Y}$, defined as

$$\forall B \subseteq \mathcal{Y}, \quad T_{\#}\alpha(B) = \alpha\left(T^{-1}(B)\right).$$

Monge's problem is difficult due to the non-linear constraint and remained unsolved until progress made more than a century later by Russian Mathematician Leonid Vitaliyevich Kantarovich. Kantarovich provided a natural relaxation to the problem by allowing mass to be split amongst multiple locations. To formalize this, consider a measure $\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$ and $d\pi(x, y)$ as the amount of mass transferred from x to y. Kantarovich's Optimal Transport Problem can now be formulated as: given $\alpha \in \mathcal{M}_+(\mathcal{X})$ and $\beta \in \mathcal{M}_+(\mathcal{Y})$

$$W(\alpha, \beta) = \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y), \tag{OT}$$

where $\Pi(\alpha, \beta)$ is composed of probability distributions over the product space $\mathcal{X} \times \mathcal{Y}$ with fixed marginals α, β:

$$\Pi(\alpha, \beta) := \left\{\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y}); P_{1\#}\pi = \alpha, P_{2\#}\pi = \beta\right\},$$

where $P_{1\#}\pi(P_{2\#}\pi)$ s the marginal distribution of π for the first (second) variable.

eq. (OT) represents the primal form of the Kantorovich formulation of standard optimal transport. The dual formulation is

$$W(\alpha, \beta) = \sup_{(u,v) \in U(c)} \int_{\mathcal{X}} u(x) d\alpha(x) + \int_{\mathcal{Y}} v(y) \beta(y) \tag{3}$$

where $U(c)$ is the following set constrained by the cost function c

$$U(c) := \{(u, v) \in C(\mathcal{X}) \times C(\mathcal{Y}) | u(x) + v(y) \leq c(x, y), \forall (x, y) \in \mathcal{X} \times \mathcal{Y}\}, \tag{4}$$

where $C(\mathcal{X})$ denotes the set of continuous functions on $\mathcal{X}$. Using the shippers problem as intuition, $u(x)$ can be thought of as the "loading cost" and $v(y)$ as the "unloading cost." The constraint in the dual formulation of standard OT makes it difficult to compute.

2.2 Regularized Optimal Transport

We now introduce regularized optimal transport, which regularizes the original problem by penalizing it with the $\varphi-$ divergence. We first briefly review the $\varphi-$ divergence and then move to regularized optimal transport.

φ-divergences φ-divergences make up a family of functions that measure the difference between probability distributions. For two distributions denoted $\alpha \in \mathcal{M}_+(\mathcal{X})$ and $\beta \in \mathcal{M}_+(\mathcal{Y})$, φ-divergences compare $\frac{d\alpha}{d\beta}$ to 1. φ-divergence is 0 when $\frac{d\alpha}{d\beta} = 1$ and grows larger as $\frac{d\alpha}{d\beta}$ diverges from 1, which occurs when α and β differ greatly. Formally, φ-divergence is defined as:

$$D_\varphi(\alpha|\beta) := \int_{\mathcal{X}} \varphi(\frac{d\alpha}{d\beta}(x))d\beta(x),$$

where $\varphi(x)$ is a convex, lower semi-continuous function, and $\varphi(1) = 0$. The choice of the $\varphi(x)$ determines the specific divergence, as shown in table 1 for a few common divergences. φ-divergences have many applications, and in OT they can be used as regularization terms.

Divergence	$\varphi(x)$
Kullback-Leibler(KL)	$x\log(x)$
Total Variation	$\frac{1}{2}\lvert x-1\rvert$
Jensen-Shannon(JS)	$(x+1)\log(\frac{2}{x+1}) + x\log(x)$

Table 1: Example φ-divergences and their corresponding φ functions.

Regularized Optimal Transport As originally proposed in Cuturi (2013), regularized optimal transport consists of penalizing standard optimal transport with the φ-divergence of the transport plan $(\pi(x,y))$ with respect to the product measure $(\alpha \otimes \beta)$. The primal formulation is:

$$\begin{aligned} W_\varepsilon^\varphi(\alpha,\beta) &:= \min_{\pi\in\Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y) + \varepsilon D_\varphi(\pi(x,y)|\alpha(x)\otimes\beta(y)) \\ &= \min_{\pi\in\Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y) + \epsilon \int_{\mathcal{X}\times\mathcal{Y}} \varphi(\frac{d\pi(x,y)}{d\alpha(x)d\beta(y)})d\alpha(x)d\beta(y) \end{aligned} \tag{5}$$

As the transport plan deviates more from the product measure, the regularization term increases. Hence, transport plans that are more similar to the product measure are preferred.

It is also noted that strong duality holds for regularized OT, as proven in Genevay (2019), with the dual formulation:

$$W_\varepsilon^\varphi(\alpha,\beta) = \sup_{u,v\in C(\mathcal{X})\times C(\mathcal{Y})} \int_{\mathcal{X}} u(x)d\alpha(x) + \int_{\mathcal{Y}} v(y)d\beta(y) - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \varphi^*(\frac{u(x)+v(y)-c(x,y)}{\epsilon})d\alpha(x)d\beta(y) \tag{6}$$

where φ^* is the Legendre transform of φ:

$$\varphi^*(p) := \sup_w wp - \varphi(w) \tag{7}$$

Notably, the constraints of the dual formulation of standard OT in eq. (4) have been replaced with a smooth regularization term (the last term in eq. (6). The existence of maximizers of eq. (6) is not guaranteed in general, but is guaranteed for particular choices of $\varphi(x)$ such as $\varphi(x) = x\log(x) - x + 1$, which is the choice of interest here. Regularizing OT with this $\phi(x)$ results in entropy-regularized OT, the primal form of which can be defined as:

$$W_\varepsilon(\alpha, \beta) := \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y) + \epsilon D_{KL}(\pi|\alpha \otimes \beta) \tag{ROT}$$

where D_{KL} is the relative entropy (or KL divergence) of the transport plan $\pi(x, y)$ with respect to the product measure $\alpha \otimes \beta$:

$$D_{KL}(\pi|\alpha \otimes \beta) := \int_{\mathcal{X}\times\mathcal{Y}} (\log(\frac{d\pi(x, y)}{d\alpha(x)d\beta(y)}) - 1)d\pi(x, y) + 1$$

The dual formulation is, as expected, eq. (6) with the specified φ:

$$\begin{aligned} W_\varepsilon(\alpha, \beta) = \max_{u\in C(\mathcal{X}), v\in C(\mathcal{Y})} & \int_{\mathcal{X}} u(x)d\alpha(x) + \int_{\mathcal{Y}} v(y)d\beta(y) \\ & - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}} d\alpha(x)d\beta(y) + \varepsilon \end{aligned} \tag{Dual ROT}$$

eq. (Dual ROT)[1] is a result of simplifying the Legendre transform for $\varphi(x) = x\log(x) - x + 1$:

$$\varphi^*(p) := \sup_w wp - w\log(w) + w - 1 = e^p + 1$$

The dual formulation of entropy regularized OT is concave in both u and v. Genevay (2019) introduces and derives a semi-dual formulation of entropy regularized OT, which is of the form:

$$W_\varepsilon = \max_{v\in C(\mathcal{Y})} \int_{\mathcal{X}} v^{c,\epsilon}(x)d\alpha(x) + \int_{\mathcal{Y}} v(y)d\beta(y), \tag{Semi-Dual ROT}$$

where $v^{c,\varepsilon}$ is the c, ε-transform, defined as:

$$v^{c,\varepsilon} := -\epsilon \log(\int_{\mathcal{Y}} e^{\frac{v(y)-c(x,y)}{\varepsilon}} d\beta(y)).$$

3. Computational Optimal Transport with Regularization

As we have mentioned in the introduction, the computation of OT for discrete distributions involves large-scale linear programming, which can become prohibitive computationally in some cases and hence limits the practical impact of OT. What's more, linear programming based methods can not be applied when dealing with continuous densities. Even though there are known methods called semi-discrete solvers (Aurenhammer et al., 1998) for computing the distance between a discrete distribution and a continuous one, they are restricted to the Euclidean squared cost, and can only be implemented in low dimensions. As the solvers were originally designed for assignment problems, they can hardly applied in high-dimensional statistical contexts. Therefore, it is of great significance to develop computational methods that can cope with three possible settings:

1. discrete OT: compare a discrete with another discrete measure;

[1] In our following discussion, we often omit the constant ε in the dual formulation for simplicity.

2. semi-discrete OT: compare a discrete with a continuous measure;

3. continuous OT: compare a continuous with another continuous measure.

In this section, we shall present how to kill the above three "birds" with one "stone": entropy regularization. This regularization approach relies on two main ideas: 1) when using the dual formulation, the OT problem can be viewed as a stochastic optimization; 2) for such an optimization problem, entropy regularization results in a smooth concave objective function. Therefore, computational methods in stochastic optimization (such as stochastic gradient descent) can carried over to OT problems, as we shall see more clearly in the following.

3.1 Stochastic Optimization Formulation

Recall that the dual formulation of the entropy regularized OT in eq. (Dual ROT) is

$$W_\varepsilon(\alpha, \beta) = \max_{u \in \mathcal{C}(\mathcal{X}), v \in \mathcal{C}(\mathcal{Y})} F_\varepsilon(u, v),$$

where

$$F_\varepsilon(u, v) = \int_{\mathcal{X}} u(x) \mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(y) \mathrm{d}\beta(y) - \varepsilon \int_{\mathcal{X} \times \mathcal{Y}} \exp\left(\frac{u(x) + v(y) - c(x, y)}{\varepsilon}\right) \mathrm{d}\alpha(x) \mathrm{d}\beta(y).$$

If we define

$$f_\varepsilon(x, y, u, v) := u(x) + v(y) - \varepsilon \exp\left(\frac{u(x) + v(y) - c(x, y)}{\varepsilon}\right),$$

then we can rewrite $F_\varepsilon(u, v)$ as an expectation, i.e.,

$$F_\varepsilon(u, v) = \mathbb{E}_{X \sim \alpha, Y \sim \beta}[f_\varepsilon(X, Y, u, v)].$$

Hence, the dual problem can be rewritten as a stochastic optimization.

$$\max_{u \in \mathcal{C}(\mathcal{X}), v \in \mathcal{C}(\mathcal{Y})} \mathbb{E}_{X \sim \alpha, Y \sim \beta}[f_\varepsilon(X, Y, u, v)]. \tag{8}$$

It is noted that the dual cannot be cast as an unconstrained expectation maximization problem when $\varepsilon = 0$, which indicates the significance of adding such regularization: by introducing entropy regularization, we replace the hard constraint by the soft constraint.

For the semi-dual formulation in eq. (Semi-Dual ROT), we can derive corresponding stochastic optimization problem. From equation, we know that

$$W_\varepsilon(\alpha, \beta) = \max_{v \in \mathcal{C}(\mathcal{Y})} H_\varepsilon(v),$$

where

$$H_\varepsilon(v) = \int_{\mathcal{X}} v^{c,\varepsilon}(x) \mathrm{d}\alpha(x) + \int_{\mathcal{Y}} v(y) \mathrm{d}\beta(y)$$
$$v^{c,\varepsilon}(x) = -\varepsilon \log\left(\int_{\mathcal{Y}} \exp\left(\frac{v(y) - c(x, y)}{\varepsilon}\right) \mathrm{d}\beta(y)\right).$$

Let $h_\varepsilon(x, v) = \int_{\mathcal{Y}} v(y) \mathrm{d}\beta(y) + v^{c,\varepsilon}(x)$, then the semi-dual problem is equivalent to the following optimziation

$$\max_{v \in \mathcal{C}(\mathcal{Y})} \mathbb{E}_X \left[h_\varepsilon(X, v)\right] \tag{9}$$

We note that though eq. (8) is written in the form of stochastic programming, it becomes a finite-dimensional convex programming problem, i.e., the objective function becomes concave, when the distribution is a finite sum of dirac measures. In this case, we can have a much simpler computational scheme, such as Sinkhorn algorithm (Cuturi, 2013) as we show below. On the other hand, when the distribution is continuous, though direct computation is impossible, we can leverage sampling and/or kernel methods as a workaround.

3.2 Discrete and Semi-discrete Optimal Transport

For discrete OT, both α and β are discrete measures, i.e., $\alpha = \sum_{i=1}^{I} \alpha_i \delta_{x_i}$ and $\beta = \sum_{j=1}^{J} \beta_j \delta_{y_j}$, where $\{x_i\}_{i=1}^{I} \subset \mathcal{X}$ and $\{y_j\}_{j=1}^{J} \subset \mathcal{Y}$. Equivalently, we can view each discrete measure as a vector from some simplex, that is $\alpha \in \Sigma_I$ and $\beta \in \Sigma_J$, where Σ_I, Σ_J are simplexes in $\mathbb{R}^I$ and $\mathbb{R}^J$ respectively. These discrete measures may come from the evaluation of continuous densities on a grid, which serves as a discretization (Liu et al., 2018; Bosc, 2010; Mérigot, 2011) or empirical measure based on samples for some machine learning applications.

Apparently, as $\alpha \in \Sigma_I, \beta \in \Sigma_J$ are both of finite dimension, the optimization problem eq. (8) also becomes a finite dimensional optimization problem as

$$\max_{u \in \mathbb{R}^I, v \in \mathbb{R}^J} g(u, v) := \langle u, \alpha \rangle + \langle v, \beta \rangle + \varepsilon \sum_{i,j} \exp\left(\frac{u_i + v_j - c_{i,j}}{\varepsilon}\right) \alpha_i \beta_j. \tag{10}$$

We note that the function $g(u, v)$ is concave and continuously differentiable, hence it admits a maximizer, which be can obtained by performing alternating maximization.

$$\begin{aligned} u^{k+1} &= \arg\max_u g(u, v^k), \\ v^{k+1} &= \arg\max_v g(u^{k+1}, v) \end{aligned} \tag{11}$$

If we let $a_i^k = e^{u_i^k/\varepsilon}, b_i^k = e^{v_i^k/\varepsilon}$ and denote $e^{-c/\varepsilon}$ the matrix whose entry is $(e^{-c/\varepsilon})_{ij} = e^{-c_{ij}/\varepsilon}$, then the alternating maximization eq. (11) becomes

$$\begin{aligned} a^{k+1} &= 1 \oslash e^{-c/\varepsilon}(b^k \otimes \beta), \\ b^{k+1} &= 1 \oslash e^{-c/\varepsilon}(a^{k+1} \otimes \alpha), \end{aligned} \tag{SK}$$

where $\oslash$ is the pointwise division and $\otimes$ is the pointwise multiplication. We note that this alternating maximization scheme eq. (SK) is often referred to as Sinkhorn Algorithm, which is the state of the art for computing discrete optimal transport (Cuturi, 2013) The algorithm introduced above relies on the the dual formulation, which involves matrix-vector multiplication at each iteration with complexity $O(N), N = \max(I, J)$. On the other hand, this computational issue can be alleviated by applying online stochastic optimization methods to the stochastic optimization problem derived from the semi-dual formulation, as pointed out in (Genevay et al., 2016).

Recall that for the semi-dual formulation, the equivalent optimization problem is given by eq. (9), where the objective function under discrete measures α, β becomes

$$\mathbb{E}_X[h_\varepsilon(X, v)] = \sum_{i \in I} h_\varepsilon(x_i, v) \alpha_i, \tag{12}$$

where

$$h_\varepsilon(x, v) = \sum_{j \in J} v_j \beta_j - \varepsilon \log\left(\sum_{j \in J} \exp\left(\frac{v_j - c(x, y_j)}{\varepsilon}\right)\right) - \varepsilon.$$

For such a stochastic optimization problem, when we apply stochastic gradient descent (SGD), we first initialize $v = \mathbf{0}_J$ and at $k-$th iteration, we draw a sample x_k from the distribution α, then we compute the following gradient as a proxy for the full gradient

$$\nabla_v h_\varepsilon(x_k, v) = \beta - \chi^{\varepsilon}_{(c(x,y_\ell)-\mathbf{v}_\ell)_\ell}$$

where $\left(\chi^{\varepsilon}_{r}\right)_j = e^{-\frac{r_j}{\varepsilon}} \beta_j \left(\sum_\ell e^{-\frac{r_\ell}{\varepsilon}} \beta_\ell \right)^{-1}$. Hence, SGD gives the following updating rule

$$v_{k+1} = v_k + \eta_k \nabla_v h_\varepsilon(x_k, v_k), \tag{SGD}$$

where η_k is the step size, decaying fast enough to zero in order to ensure that the "noise" created by using the proxy $\nabla_v h_\varepsilon(x_k, v)$ is canceled in the limit. One of such step sizes can be $\eta_k = \frac{\eta_0}{1+k/k_0}$, where k_0 is roughly the number of iterations serving as a warmup phase. One can prove by using stochastic approximation (Peyré and Cuturi, 2019) that $|H_\varepsilon(v^*) - H_\varepsilon(v_k)| = O(1/\sqrt{k})$.

Since SGD is slow because of the fast decay of η_k toward zero, Schmidt et al. (2017) propose to apply stochastic average gradient descent (SAG), in order to improve the convergence speed. The key is to find a better proxy for the full gradient. Different from SGD, SAG keeps in memory the gradients it has computed in the past and updates the stored gradient every time the particular sample, based on which the gradient was computed previously, is sampled again. The key of SAG is that it applies an update in the direction of the average of all gradients stored so far, which improves the convergence rate to $|H_\varepsilon(v^*) - H_\varepsilon(v_k)| = O(1/k)$. The updating rule of SAG is

$$v_{k+1} = v_k + \frac{\eta_k}{I} \sum_{i=1}^{I} g_i, \tag{SAG}$$

where $g_i = \nabla_v h_\varepsilon(x_i, v_{k_i})$ if x_i has been sampled most recently at k_i-th iteration, otherwise $g_i = 0$. Obviously, compared with eq. (SGD) dealing with only one gradient at a time, eq. (SAG) has to store the gradients for each of the I points. This expense can be mitigated by adopting mini-batches instead of individual points.

In addition to SGD and SAG, another popular first-order method that can be applied to speedup the convergence is stochastic gradient descent with averaging (SGA)(Bach, 2014). Similar to SAG, this method improves the convergence speed by averaging the past iterates, yet it only takes one gradient per iteration. To be specific, SGA first runs a standard SGD on auxiliary variable $\tilde{v}$

$$\tilde{v}_{k+1} = \tilde{v}_k + \eta_k \nabla_v h_\varepsilon(x_k, \tilde{v}_k),$$

where x_k is drawn from α and then outputs the averaged vector $v_{k+1} = \frac{1}{k} \sum_{i=1}^{k} \tilde{v}_k$. To avoid explicitly storing all the iterates, we can adopt a running average as follows

$$\begin{aligned} \tilde{v}_{k+1} &= \tilde{v}_k + \eta_k \nabla_v h_\varepsilon(x_k, \tilde{v}_k), \\ v_{k+1} &= \frac{1}{k+1} \tilde{v}_{k+1} + \frac{k}{k+1} v_k. \end{aligned} \tag{SGA}$$

In this case, a typical choice of the step size is of the form $\eta_k = \frac{\eta_0}{1+\sqrt{k/k_0}}$, where k_0 is defined in the same way as in SGD. We note that this decays to 0 in a much slower pace than that in SGD. The convergence rate is $O(1/\sqrt{k})$ and though it is of the same order as SGD, Bach (2014) proves that the constants involved in the big O are smaller than SGD, since in contrast to SGD, SGA is adaptive to the local strong convexity of the functional.

Our discussions above revolve around discrete OT, and in fact, we see no difficulty in carrying these deductions over to semi-discrete problems, as the two enjoy the same stochastic optimization formulation, as we shall see more clearly in the following.

We assume that α is now an arbitrary measure that may not be discrete and $\beta = \sum_{j=1}^{J} \beta_j \delta_{y_j}$. Naturally, one can consider approximating α by an empirical measure $\hat{\alpha}_N := \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$, where $\{x_i\}$ are iid samples from α. However, this introduces bias as the discretized problem is different from the original one. In fact, In this case, we can resort to the semi-dual problem, which is still a finite-dimensional maximization problem, written as

$$W_\varepsilon(\alpha, \beta) = \max_{v \in \mathbb{R}^J} \mathbb{E}_{X \sim \alpha}[h_\varepsilon(X, v)].$$

Though we can no longer write the expectation as a finite sum as we do in eq. (12), we assume we can draw samples from α, which is a common setting in machine learning applications. By this assumption, we see that our previous algorithms such as SGD and SAG also apply here, as all of them only rely on sampling when performing the update.

3.3 Continuous Optimal Transport

We note that so far the optimal transport problems we have discussed can all be converted to finite-dimensional stochastic optimization as at least one of the measures is discrete. Apparently, when neither α or β is discrete, the problem eq. (9) is infinite-dimensional, which cannot be solved by simply gradient methods. Intuitively, we can find some finite-dimensional approximations which can be either parametric or non-parametric, as we shall see in the following.

We start with the non-parametric approach proposed in (Dieuleveut and Bach, 2016). The key is that we approximate the continuous functions $u \in \mathcal{C}(\mathcal{X}), v \in \mathcal{C}$ in the dual formulation by linear combination of kernel functions in reproducing kernel hilbert spaces (RKHS) and then searching potentials u, v is equivalent to searching the linear combination coefficients, which becomes a finite-dimensional problem. Specifically, we consider two RKHS $\mathcal{H}$ and $\mathcal{G}$ defined on $\mathcal{X}$ and $\mathcal{Y}$ with kernels κ and ℓ. Recall that if $u \in \mathcal{H}$ then $u(x) = \langle u, \kappa(\cdot, x)\rangle_{\mathcal{H}}$, then we rewrite $f_\varepsilon(x, y, u, v)$ as

$$f_\varepsilon(x, y, u, v) = \langle u, \kappa(x, \cdot)\rangle_{\mathcal{H}} + \langle v, \ell(y, \cdot)\rangle_{\mathcal{G}} - \varepsilon \exp\left(\frac{\langle u, \kappa(x, \cdot)\rangle_{\mathcal{H}} + \langle v, \ell(y, \cdot)\rangle_{\mathcal{G}} - c(x, y)}{\varepsilon}\right)$$

By the reproducing property of $\mathcal{H}$ and $\mathcal{G}$, we can compute derivatives of f_ε as

$$\begin{aligned} \partial_u f_\varepsilon(x, y, u, v) &= \kappa(\cdot, x)\left(1 - \exp\left(\frac{u(x) + v(y) - c(x, y)}{\varepsilon}\right)\right) \\ \partial_u f_\varepsilon(x, y, u, v) &= \ell(\cdot, y)\left(1 - \exp\left(\frac{u(x) + v(y) - c(x, y)}{\varepsilon}\right)\right). \end{aligned} \tag{13}$$

Hence, if we consider the stochastic gradient descent

$$(u_k, v_k) = (u_{k-1}, v_{k-1}) + \frac{C}{\sqrt{k}} \nabla f_\varepsilon(x_k, y_k, u_{k-1}, v_{k-1}),$$

where x_k, y_k are iid samples from $\alpha \otimes \beta$, we obtian the following recursion formula for iterates (u_k, v_k) as we substitute the gradient ∇f_ε with eq. (13)

$$\begin{aligned} (u_k, v_k) &= \sum_{i=1}^{k} \alpha_i (\kappa(\cdot, x_i), \ell(\cdot, y_i)), \\ \alpha_i &= P_{B_r}\left(\frac{C}{\sqrt{i}}\left(1 - \exp\left(\frac{u_{i-1}(x_i) - v_{i-1}(y_i) - c(x_i, y_i)}{\varepsilon}\right)\right)\right), \end{aligned}$$

where P_{B_r} is the projection on the centered ball of radius r. Such projection places a bound on the iterates and ensures convergence as shown in (Dieuleveut and Bach, 2016).

We note that though using kernel methods reduces infinite-dimensional problem to finite-dimensional one, there is no free lunch: the main computation cost arises from $u_{k-1}(x_k) = \sum_{i=1}^{k-1} \alpha_i \kappa(x_k, x_i)$ and $v_{k-1}(y_k) = \sum_{i=1}^{k-1} \alpha_i \ell(y_k, y_i)$, which leads to complexity $O(k^2)$. In words, as the recursive procedure proceeds, the running time complexity for non-parametric approach is growing quadratically with respect to the the number of iterations.

In order to alleviate this issue, Seguy et al. (2017) restrict that infinite-dimensional optimization problem over a space of continuous functions to a much smaller subset, such as that spanned by deep neural networks. Simply put, we approximate u, v by u_ϕ and v_ψ respectively, where u_ϕ, v_ψ are parametrized by two independent neural networks with ϕ, ψ being the parameters. At each iteration, we first draw a batch of samples from α, i.e., $(x_1, \ldots, x_p) \sim \alpha$ and another batch from β, i.e., $(y_1, \ldots, y_p) \sim \beta$, then we update the parameters as follows

$$\begin{cases} \phi \leftarrow \phi + \gamma \sum_{ij} \nabla_\phi u(x_i) + \partial_u f_\varepsilon(x_i, y_j, u_\phi, v_\psi) \nabla_\phi u(x_i) \\ \\ \psi \leftarrow \psi + \gamma \sum_{ij} \nabla_\psi v(y_j) + \partial_v f_\varepsilon(x_i, y_j, u_\phi, v_\psi) \nabla_\psi v(y_j). \end{cases}$$

Compared with the non-parametric one where the complexity is $O(k^2)$ for $k-$th iteration, this parametric one only requires $O(p^2)$ per iteration. Hence, this approach can provide an effective way to compute a proxy for the Wasserstein distance in high-dimensional scenarios, though this approach leads to a nonconvex finite-dimensional optimization problem without theoretical guarantees.

So far we have discussed efficient computational schemes for solving computational optimal transport problems. Furthermore, efficient computation brings optimal transport to a wider audience in machine learning research, where regularized OT improves the performance of some machine learning models, such as generative adversarial networks (GAN), as we shall present in the subsequent.

3.4 Regularized Optimal Transport in GAN

Generative Adversarial Networks (GANs) have emerged as one of the most practical methods for learning data distributions. GANs train a generative model to capture the data distribution, and a discriminative model that tries to distinguish between the samples from training data and the ones generated by the generative model. Wasserstein GANs are a GAN formulation based on the Earth Mover's Distance, a special case of the Wasserstein distance. More formally, assume that we have two distributions α and β defined over $\mathcal{X}$ and $\mathcal{Y}$ respectively[2], where the two are subsets of $\mathbb{R}^d$. Let Π be the family of couplings of α and β. Recall that given a cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, the optimal transport cost between α and β is $W(\alpha, \beta)$ as defined in eq. (OT)

The goal of generative modeling is to learn a mapping $G_{\theta^*} \in \{G_\theta, \theta \in \Theta\}$ such that the cost $W(G_{\theta^*}(\alpha), \beta)$ is minimized. In other words, we seek an optimal θ that minimizes the transportation cost.

$$\min_{\theta \in \Theta} h_0(\theta) = W(G_\theta(\alpha), \beta) = \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}} \int_{\mathcal{Y}} \pi(x, y) c(G_\theta(x), y) \, \mathrm{d}x \mathrm{d}y \tag{WGAN}$$

[2] In machine learning practice, we often assume $\mathcal{X}$ and $\mathcal{Y}$ have finite supports, as we often deal with finite data points such as images.

The primal problem is computationally challenging. Instead, the dual formulation of the generative problem is easier to parameterize. This dual formulation is called the Wasserstein GAN. Let u and v be the dual functions parameterized by ϕ and ψ respectively.

$$\begin{aligned} &\min_{\theta} \max_{\phi,\psi} \quad \mathbb{E}_{x\sim\alpha} u_{\phi}\left(G_{\theta}(x)\right) - \mathbb{E}_{y\sim\beta} v_{\psi}(y) \\ &\text{subject to } u_{\phi}\left(G_{\theta}(x)\right) - v_{\psi}(y) \leq c\left(G_{\theta}(x), y\right), \forall (x,y) \end{aligned} \tag{14}$$

Wasserstein GANs are popular due to training stability over the standard GANs. In fact, when we use the euclidean norm as the cost function, we can have a much simpler dual formulation, known as Kantorovich-Rubinstein formulation (Villani, 2009), which we shall detail when we argue why regularization improves Wasserstein GAN. For now, we first introduce what regularzied Wasserstein GAN is and how it works.

Apparently, minimizing the Wasserstein distance is computationally challenging due to the non-convex and non-smooth nature of the objective. Regularized optimal transport applied to Wasserstein GANs has been shown to be computationally efficient by application of first order optimization methods for optimizing the objective. Simply put, we replace $W(G_\theta(\alpha), \beta)$ with the entropy regularized version $W_\varepsilon(G_\theta(\alpha), \beta)$ defined in eq. (ROT). When adopting the dual formulation in eq. (Semi-Dual ROT), we obtain a similar problem as we do in eq. (14) but without any constraints

$$\min_{\theta\in\Theta} h_\varepsilon(\theta), \tag{15}$$

where $h_\varepsilon(\theta)$ is defined as

$$h_\varepsilon(\theta) = \max_{u\in\mathcal{C}(\mathcal{X}), v\in\mathcal{C}(\mathcal{Y})} \mathbb{E}_{x\sim\alpha}[u(G_\theta(x))] - \mathbb{E}_{y\sim\beta}[v(y)] - \mathbb{E}_{x,y\sim\alpha\times\beta}\left[\exp\left(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\right)\right] \tag{16}$$

As introduced, GANs train two models, a generator to capture the data distribution, i.e., G_θ and the associated training problem eq. (15) and a discriminator to distinguish between samples and generator output, i.e., eq. (16). The dual formulation of the regularized optimal transport for GANs models the discriminators as dual functions. Therefore, solving the dual formulation yields the discriminator parameters. The dual is an unconstrained concave maximization problem and can be solved so, however parametric methods such as neural networks are also applicable. For the generator problem, consider $h_\varepsilon(\theta)$ to be the generative model objective. It has been shown in (Sanjabi et al., 2018) that h_ε is L-Lipschitz smooth, for some $L \in \mathbb{R}$. The smoothness ensures that small changes in ε result in small changes in the optimal transport plan. More importantly, dual solver can generate approximate gradients for h_ε. The approximated gradients $\nabla_\theta h_\varepsilon(\theta)$ can be used to produce algorithms with guaranteed convergence to approximate solutions to the GAN generator and can be solved by first order methods such as Stochastic Gradient Descent (SGD).

In practice, larger values of ε introduce bias into the Wasserstein distance. However, Sanjabi et al. (2018) point out that the solution isn't as trivial as reducing ε since it changes the optimal discriminator parameters, changes the convergence guarantees and reduces the Lipschitz smoothness of h_ε. Another proposed solution is to use the Sinkhorn loss defined in eq. (17) to get an objective that is meaningful for larger values of ε.

$$D^{\varepsilon}_{SK}(\alpha,\beta) := 2W_\varepsilon(\alpha,\beta) - W_\varepsilon(\alpha,\alpha) - W_\varepsilon(\beta,\beta). \tag{17}$$

As highlighted in (Sanjabi et al., 2018), in addition to the computational benefits, by incorporating the regularization, the algorithm only requires solving the discriminator to approximate optimality with theoretical guarantees. For the rest of the section, we try to answer the following question

Our arguments mainly revolve around the following key facts:

1. The entropy regularization renders $W_\varepsilon(\alpha, \beta)$ Fréchet-differentiable with respect to α, β, which makes the generative objective smooth
2. The entropy regularization, by placing soft constraints, makes the dual problem an unconstrained maximization, where the dual solutions are within the class of continuous functions. This make the discriminator problem "solvable" when using neural networks.

To further develop our points, we briefly review the vanilla Wassertein GAN and discuss issues in the training process, with which we compare the regularized version and explain its advantages.

In generative models, one key step is to find a discrepancy function measuring dissimilarity between the data distribution and the learned distribution. As argued in (Arjovsky et al., 2017), different from KL or JS divergence, wassersetin distance yields a weaker topology on the space of probability measures. Hence, if we use wasserstein distance as the discrepancy function, i.e., the loss function, it is continuous with respect to the learning parameters, which improves the stability of learning, avoids mode collapse and provides helpful learning curves for debugging and hyperparameter searches. Moreover, it is differentiable almost everywhere given the mapping G_θ is locally Lipschitz, providing useful gradient for learning the model. However, we argue that this vanilla Wasserstein GAN is not learnable or solvable when adopting neural networks as the function approximator, as shown below.

Following our previous notations, and let $c(x, y) = \|x - y\|$ on $\mathbb{R}^d$, then by Kantorovich-Rubinstein duality (Villani, 2009),

$$W(\beta, G_\theta(\alpha)) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x\sim\beta}[f(x)] - \mathbb{E}_{x\sim G_\theta(\alpha)}[f(x)], \tag{18}$$

where $\{f | \|f\|_L \leq 1\}$ denotes all 1-Lipschits functions. In (Arjovsky et al., 2017), it has been shown that

$$\nabla_\theta W(\beta, G_\theta(\alpha)) = -\mathbb{E}_{z\sim\alpha}[\nabla_\theta f(G_\theta(z))], \tag{19}$$

where f is the solution to eq. (18) and we note such a gradient can be estimated via sampling methods (Gulrajani et al., 2017), as it is written as an expectation. Though such process is principled as argued in (Arjovsky et al., 2017), approximating the solution f using neural networks turns out to be messy. Arjovsky et al. (2017) proposed the following substitute for seeking the solution, which relies on a family of functions $\{f_w\}_{w\in\mathcal{W}}$ parameterized by neural networks.

$$\max_{w\in\mathcal{W}} \mathbb{E}_{x\sim\beta}[f_w(x)] - \mathbb{E}_{z\in\alpha}[f_w(G_\theta(z))], \tag{20}$$

where $\mathcal{W}$ denotes a compact set, ensuring that f_w is of Lipschitz. In practice, in order to parameters lie in a compact set, weights clipping (Arjovsky et al., 2017) and other remedies, such as gradient penalty (Gulrajani et al., 2017) have been proposed, which alleviate the stability issue in GAN training to some extent.

However, such techniques are rather engineering practices than a mathematical workaround. In fact, when we solve for eq. (20), we are essentially solving a problem different from eq. (18), as $\{f_w\}_{w\in\mathcal{W}}$ is merely a subset of Lipschitz functions. What's worse, as pointed out in (Bousquet et al., 2017), no matter how accurately we solve for eq. (20), it may fail to provide a good estimate of the gradient, when we plug it in eq. (19). We denote the solution to eq. (18) by f^* and the $\epsilon-$optimal one by f^ϵ, i.e.,

$$\begin{aligned}&|W_{f^*}(\alpha,\beta) - W_{f^\epsilon}(\alpha,\beta)| < \epsilon,\\ &W_f(\alpha,\beta) := \mathbb{E}_{x\sim\beta}[f(x)] - \mathbb{E}_{x\sim G_\theta(\alpha)}[f(x)]\end{aligned}$$

It has been shown in (Bousquet et al., 2017) that there exists a constant $C > 0$ such that for any $\epsilon > 0$, one can always find α, β such that $\cos(\nabla_\theta W_{f^*}, \nabla_\theta W_{f^\epsilon}) \leq 1 - C$. Therefore, we claim that for Wasserstein GAN, the discriminator problem, i.e., finding the dual solution is not "solvable" for neural networks.

From our discussions above, we can see that though Wasserstein GAN enjoys many theoretical advantages over other divergences, its training process is quite challenging. Thanks to regularization, we can turn to regularized wasserstein distance for a better training process with theoretical guarantees. First of all, by adding entropy regularization, the dual problem is an unconstrained convex programming, though infinite dimensional. Moreover, the dual solutions (u, v) vary continuously with the input measures (α, β), which further implies the differentiability of W_ε with respect to α, β. The technical details underneath are beyond the scope of this paper, and we refer reader to (Feydy et al., 2018). Hence, for the training process, we obtain a smooth loss function $W_\varepsilon(\beta, G_\theta(\alpha))$ to which we can apply some stochastic programming solvers, e.g. stochastic gradient descent.

Besides, we note that when using the dual formulation eq. (Dual ROT), the solutions are within the class of continuous functions, which can be approximated by neural networks, according to the universal approximation theorem . In other words, instead of solving for a substitute eq. (20) in vanilla Wasserstein GAN, we are indeed solving for the true discriminator problem, which is solvable for neural networks at lease theoretically. Even though there is bound to exists some approximation error, it has been shown that such approximations still provides helpful information for estimating the gradient when solving the generator problem (Sanjabi et al., 2018), as we have seen above.

4. Statistical

Besides the computational benefits of ROT, the statistical behaviours of the ROT are also worth mentioning. A good statistical understanding of this problem is critical because we often use an approximated version of the Wasserstein distance in machine learning. In section 4.1, we would compare the rate of convergence on a continuous set for OT and ROT. A discussion of the different limiting behaviours of the OT and ROT for distributions supported on finite spaces will be shown in section 4.2.

In many real applications as we have mentioned in the introduction, the distribution of α (β or both) is estimated from data by its empirical distribution. Suppose we have a sample of iid. random variables $X_1, \cdots, X_n \sim \alpha$ (or/also $Y_1, \cdots, Y_m \sim \beta$). We estimate α (β or both) by its empirical version

$$\hat{\alpha}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \quad \hat{\beta}_m = \frac{1}{m}\sum_{i=1}^{m} \delta_{Y_i}$$

Recall the definition of p-Wasserstein distance. Let $(\mathcal{X}, d)$ be a complete metric space with $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$. Let $\mathcal{P}_p(\mathcal{X})$ be the set of Borel probability measures supported on $\mathcal{X}$ with finite moment of order $p \geq 1$. The p-Wasserstein distance between two measures α and β in $\mathcal{P}_p(\mathcal{X})$ is defined by

$$W_p(\alpha,\beta) = (\inf_{\pi\in\Pi(\alpha,\beta)} \int_{\mathcal{X}^2} d^p(x,y)d\pi(x,y))^{1/p}$$

We can also define $W_{p,\epsilon}$ as the p-Wasserstein distance for the ROT with a regularization term according to eq. (ROT).

4.1 Rate of Convergence

We surveyed some results on the rate of convergence on compact set $\mathcal{X} \in \mathcal{R}^n$ and compare the rate for OT and ROT. For more general metric spaces and refinements to the rates, see (Dudley, 1969), (Alfonsi, 2013) and (Fournier and Guillin, 2015).

Theorem 1 *[Dudley (1969)] For $\mathcal{X} \in \mathbb{R}^d$ and measure supported on bounded domain. For $d \geq 3, 1 \leq q < \infty$*

$$\mathbb{E}(|W_p(\hat{\alpha_n}, \hat{\beta_m}) - W_p(\alpha_n, \beta_m)|) = O(n^{-1/d})$$

From the above theorem, we see the quality of the approximation scales like $O(n^{-1/d})$. In order to achieve the same estimation quality, we'd need to have the sample size scale exponentially. This would be a nightmare when using OT in the context of high dimensional data, and it is often referred to as the curse of dimensionality. We have seen in Section 3 how entropy regularization helps with efficient computation, the following theorem proved in a recent paper by Genevay et al. (2019a) gives insights on how it also helps break the curse.

Theorem 2 *[Genevay et al. (2019a)]*

Consider the the Sinkhorn divergence between two measures $\alpha \in \mathcal{X}, \beta \in \mathcal{Y}$, where $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^d$ are bounded. Also, require the cost c be L-Lipschitz and $\mathcal{C}^\infty$.

$$\mathbb{E}(|W_\varepsilon(\hat{\alpha_n}, \hat{\beta_m}) - W_\varepsilon(\alpha_n, \beta_m)|) = O(\frac{e^{\frac{\kappa}{\beta}}}{\sqrt{n}}(1 + \frac{1}{\epsilon^{\lfloor d/2 \rfloor}})),$$

where $\kappa = 2L|\mathcal{X}| + \|c\|_\infty$

In particular, this theorem gives the following asymptotic behavior.

$$\mathbb{E}(|W_\varepsilon(\hat{\alpha_n}, \hat{\beta_m}) - W_\epsilon(\alpha_n, \beta_m)|) \longrightarrow O(\frac{e^{\frac{\kappa}{\beta}}}{\sqrt{n}} \frac{1}{\varepsilon^{\lfloor d/2 \rfloor}}) \quad as \quad \varepsilon \to 0 \tag{21}$$

$$\mathbb{E}(|W_\varepsilon(\hat{\alpha_n}, \hat{\beta_m}) - W_\varepsilon(\alpha_n, \beta_m)|) \longrightarrow O(\frac{1}{\sqrt{n}}) \quad as \quad \varepsilon \to \infty \tag{22}$$

From the previous asymptotic behavior, we observe the following. First, as $\varepsilon \to \infty$, the convergence rate in eq. (22) does not depend on d or ϵ anymore . Therefore, when ε is large, increasing ε will not significantly improve convergence rate. Secondly, when ϵ is small, we see that the dependence on ε is crucial. Furthermore, we observe the influence of ε on the convergence rate increases with the dimension d as shown in eq. (21), but much more mildly compared with $O(n^{-1/d})$ in theorem 1.

4.2 Limiting behaviours on Finite Sets

Understanding the limiting behaviours of the empirical Wasserstein distance is also critical, especially to statistical inference (e.g. confidence intervals for sample based Wasserstein distances). However, its applications in a statistically rigorous way is severely hindered by a lack of inferential tools. A central limit theorem was established in (Barrio and Loubes, 2019) for empirical transportation cost where data are sampled from absolutely continuous measures on $\mathbb{R}^d$. However, these results in the continuous setting lead to test statistics whose numerical implementation become prohibitive for $d \geq 2$. Sommerfeld and Munk (2018) argue that this issue can be solved by restricting the distributions on finite sets and they give a distributional limit of the empirical Wasserstein distance. The restriction is not only an approximation to the truth, but rather a sufficient setting for many real world applications. Following their work, a central limit theorem for the entropic regularized optimal transport was given by Bigot et al. (2019) in parallel to Klatt et al. (2019). In this subsection, we briefly compare the main limiting behaviors for the OT and ROT on finite sets.

Theorem 3 *[Sommerfeld and Munk (2018)]*

Under the null hypothesis $r = s$, the empirical OT distance has the following limiting behaviour

$$(\frac{nm}{m+n})^{\frac{1}{2p}} W_p(\hat{\alpha}_n, \hat{\beta}_m) \xrightarrow{D} \{\underset{u \in \Phi^*}{max} \langle G, u \rangle\}^{1/p}, \quad as \quad n, m \to \infty,$$

where Φ^* is the set of dual solutions to the OT. G is a centered N-d Gaussian distribution whose covariance matrix depends on r.

We see that the this theorem applies to all probability measures with finite support, regardless of the dimension of the underlying space. Furthermore, the scaling rate depends solely on p, which contrasts the bounds for a continuous case where a strong dependence on the dimensionality was exhibited.

The central limit theorem for empirical regularized optimal transport on finite spaces was given by Klatt et al. (2019) in parallel with Bigot et al. (2019). Since the work by Klatt et al. (2019) applies to regularizer of Legendre type (sufficiently smooth), which includes the entropy regularization, we includes the main results of the paper here.

Theorem 4 *[Klatt et al. (2019)]*

Let $\alpha, \beta \in \Sigma_N$ on the finite metric space $(\mathcal{X}, d)$ and if we denote $\xrightarrow{D}$ the convergence in distribution, then we have

1. *One Sample: α is estimated by its empirical version*

$$\sqrt{n}\{W_{p,\varepsilon}(\hat{\alpha}_n, \beta) - W_{p,\varepsilon}(\alpha, \beta)\} \xrightarrow{D} \mathcal{N}_1(0, \sigma^2_{p,\varepsilon}(\alpha|\beta)), \, as \, n \to \infty$$

2. *Two Samples: α, β both estimated by their empirical version*

$$\sqrt{\frac{nm}{n+m}}\{W_{p,\varepsilon}(\hat{\alpha}_n, \hat{\beta}_m) - W_{p,\varepsilon}(\alpha, \beta)\} \xrightarrow{D} \mathcal{N}_1(0, \sigma^2_{p,\varepsilon}(\alpha, \beta)), \, as \, m, n \to \infty$$

This theorem is valid for a general class of regularizers, for a detailed definition see (Klatt et al., 2019). In particular, it is valid for Boltzmann-Shannon entropy, Burg entropy and Fermi-Dirac entropy.

We observe the following difference and relationships to the non-regularized case. First, the theorem does not differentiate between the case for the hypothesis $\alpha = \beta$ and $\alpha \neq \beta$, which is different from the OT case. Secondly, the limiting behaviour of the ROT ($\varepsilon = 0$) is asymptotic to a Gaussian distribution, which is different from the behaviour of the non-regularized OT case (see theorem 3). Furthermore, if we apply this theorem to the negative Boltzmann-Shannon entropy, we obtain the empirical Sinkhorn divergence, where empirical measures are the inputs for eq. (17). Since we know that the Sinkhorn divergence approximates the OT distance exponentially fast as $\varepsilon \to 0$. This different limit behaviour shed light onto approximating the non-regularized OT distance by the Sinkhorn divergence as $\varepsilon \to 0$. In particular, when we let ε tends to 0 at an appropriate rate depending on the sample size, $\varepsilon = o(1/\log(\sqrt{n}))$, we recover the distributional limit given by theorem 3 in the setting of vanilla optimal transport. Finally, a real world application of the benefit of using ROT dependent test statistics was shown in (Klatt et al., 2019) for the study of protein interaction networks, which are intended to allow more statistical inference from the outcome of the optimal transport plan.

5. Conclusion

We have discussed in this paper that regularization is indeed a game changer for optimal transport and related applications, from both computational and statistical perspectives. In particular, we focus on entropy regularized optimal transport and its numerous advantages over vanilla optimal transport, as it has been widely applied in modern machine learning practice, such as generative models. Thanks to the smoothing effect of the regularization, we obtain an unconstrained maximization problem as the dual formulation, where the objective is concave. Such formulation provides a stochastic programming viewpoint for computational optimal transport and improves the training process in Wasserstein GAN, as it renders the training loss smooth with respect to the generator and makes the discriminator solvable for neural networks. On the other hand, from the statistical perspective, existing results for the vanilla OT supported on R^d lead to test statistics whose numerical implementation become prohibitive for high dimensional empirical measures. With the help of entropic regularized OT, it is of great help to propose test statistics based on fast Sinkhorn divergences for statistical inference, e.g. confidence intervals for sample based Wasserstein distance.

References

I. Abraham, R. Abraham, M. Bergounioux, and G. Carlier. Tomographic reconstruction from a few views: a multi-marginal optimal transport approach]. 2015.

Aurélien Alfonsi. Strong order one convergence of a drift implicit euler scheme: Application to the cir process. *Statistics & Probability Letters*, 83(2):602 – 607, 2013. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2012.10.034. URL `http://www.sciencedirect.com/science/article/pii/S0167715212004063`.

Martin Arjovsky, Soumith Chintala, and Leon Bottou. Wasserstein Generative Adversarial Networks. volume 70 of *Proceedings of Machine Learning Research*, pages 214—223, International Convention Centre, Sydney, Australia, 2017. PMLR. URL `http://proceedings.mlr.press/v70/arjovsky17a.html`.

F. Aurenhammer, F. Hoffmann, and B. Aronov. Minkowski-Type Theorems and Least-Squares Clustering. *Algorithmica*, 20(1):61–76, 1998. ISSN 0178-4617. doi: 10.1007/pl00009187.

Francis Bach. Adaptivity of Averaged Stochastic Gradient Descent to Local Strong Convexity for Logistic Regression. *J. Mach. Learn. Res.*, 15(1):595–627, 2014. ISSN 1532-4435.

Eustasio Del Barrio and Jean-Michel Loubes. Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47(2):926–951, 2019. doi: 10.1214/18-aop1275.

Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications, 2019. URL `https://projecteuclid.org/euclid.ejs/1576119711`.

Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and Sparse Optimal Transport. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 880—889, Playa Blanca, Lanzarote, Canary Islands, 2018. PMLR. URL `http://proceedings.mlr.press/v84/blondel18a.html`.

Nicolas Bonneel, Michiel van de Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using lagrangian mass transport. *ACM Trans. Graph*, 158:1–12, 2011.

Damien Bosc. Numerical approximation of optimal transport maps. *SSRN Electronic Journal, DOI*, 10, 2010.

Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, Carl-Johann Simon-Gabriel, and Bernhard Schoelkopf. From optimal transport to generative modeling: the VEGAN cookbook. *arXiv*, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26, pages 2292–2300. Curran Associates, Inc., 2013. URL `https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf`.

Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. Differentiable Ranks and Sorting using Optimal Transport. 2019.

Fernando de Goes, David Cohen-Steiner, Pierre Alliez, and Mathieu Desbrun. An optimal transport approach to robust reconstruction and simplification of 2d shapes. *Computer Graphics Forum*, 30:1593–1602, 2011.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large stepsizes. *The Annals of Statistics*, 44(4):1363–1399, 2016. ISSN 0090-5364. doi: 10.1214/15-aos1391.

V. Dobri´c and J.E. Yukich. Asymptotics for transportation cost in high dimensions. *J. Theoret. Probab.*, 8(1):97–118, 1995.

R. M. Dudley. The speed of mean glivenko-cantelli convergence. *Ann. Math. Statist.*, 40(1):40–50, 02 1969. doi: 10.1214/aoms/1177697802. URL `https://doi.org/10.1214/aoms/1177697802`.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. *arXiv*, 2018.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707, August 2015. URL `https://hal.archives-ouvertes.fr/hal-00915365`.

Aude Genevay. Entropy-regularized optimal transport for machine learning. *PhD Thesis*, 2019.

Aude Genevay, Marco Cuturi, Gabriel Peyr\'e, and Francis Bach. Stochastic Optimization for Large-Scale Optimal Transport. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3440–3448, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1574–1583. PMLR, 16–18 Apr 2019a. URL `http://proceedings.mlr.press/v89/genevay19a.html`.

Aude Genevay, Gabriel Dulac-Arnold, and Jean-Philippe Vert. Differentiable Deep Clustering with Cluster SizeConstraints. 2019b.

Alexandre Gramfort, Gabriel Peyr ́e, and Marco Cuturi. Fast optimal transport averaging of neuroimaging data. *In Information Processing in Medical Imaging*, pages 261–272, 2015.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved Training of Wasserstein GANs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5769–5779, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Marcel Klatt, Carla Tameling, and Axel Munk. Empirical regularized optimal transport: Statistical theory and applications, 2019.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. pages 957–966, 2015.

Jialin Liu, Wotao Yin, Wuchen Li, and Yat Tin Chow. Multilevel Optimal Transport: a Fast Approximation of Wasserstein-1 distances. 2018.

Quentin Mérigot. A Multiscale Approach to Optimal Transport. *Computer Graphics Forum*, 30(5): 1583–1592, 2011. ISSN 1467-8659. doi: 10.1111/j.1467-8659.2011.02032.x.

Offir Pele and Michael Werman. Fast and robust earth mover's distances. *IEEE 12th International Conference on Computer Vision*, pages 406–467, 2009.

Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–206, 2019. ISSN 1935-8237. doi: 10.1561/2200000073.

Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D. Lee. On the Convergence and Robustness of Training GANs with Regularized Optimal Transport. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 7091–7101, Red Hook, NY, USA, 2018. Curran Associates Inc.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017. ISSN 0025-5610. doi: 10.1007/s10107-016-1030-6.

Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-Scale Optimal Transport and Mapping Estimation. *arXiv*, 2017.

Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher. Earth mover's distances on discrete surfaces. *ACM Transactions on Graphics*, 33(4):1–12, 2014.

Max Sommerfeld and Axel Munk. Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society Series B*, 80(1):219–238, January 2018. doi: 10.1111/rssb.12236. URL `https://ideas.repec.org/a/bla/jorssb/v80y2018i1p219-238.html`.

Cédric Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, Heidelberg, 2009. ISBN 9783540710493. doi: 10.1007/978-3-540-71050-9.