Regularized Optimal Transport is Ground Cost Adversarial

François-Pierre Paty [1] Marco Cuturi [2 1]

Abstract

Regularizing the optimal transport (OT) problem has proven crucial for OT theory to impact the field of machine learning. For instance, it is known that regularizing OT problems with entropy leads to faster computations and better differentiation using the Sinkhorn algorithm, as well as better sample complexity bounds than classic OT. In this work we depart from this practical perspective and propose a new interpretation of regularization as a robust mechanism, and show using Fenchel duality that any convex regularization of OT can be interpreted as ground cost adversarial. This incidentally gives access to a robust dissimilarity measure on the ground space, which can in turn be used in other applications. We propose algorithms to compute this robust cost, and illustrate the interest of this approach empirically.

1. Introduction

Optimal transport (OT) has become a generic tool in machine learning, with applications in various domains such as supervised machine learning (Frogner et al., 2015; Abadeh et al., 2015; Courty et al., 2016), graphics (Solomon et al., 2015; Bonneel et al., 2016), imaging (Rabin & Papadakis, 2015; Cuturi & Peyré, 2016), generative models (Arjovsky et al., 2017; Salimans et al., 2018), biology (Hashimoto et al., 2016; Schiebinger et al., 2019) or NLP (Grave et al., 2019; Alaux et al., 2019). The key to using OT in these applications lies in the different forms of regularization of the original OT problem, as introduced in references (Villani, 2009; Santambrogio, 2015). Adding a small convex regularization to the classical linear cost not only helps on the algorithmic side, by convexifying the objective and allowing for faster solvers, but also introduces a regularity trade-off that prevents from overfitting on data measures.

Regularizing OT Although entropy-regularized OT is the most studied regularization of OT, due to its algorithmic advantages (Cuturi, 2013), several other convex regularizations of the transport plan have been proposed in the community: quadratically-regularized OT (Essid & Solomon, 2017), OT with capacity constraints (Korman & McCann, 2015), Group-Lasso regularized OT (Courty et al., 2016), OT with Laplacian regularization (Flamary et al., 2014), Tsallis Regularized OT (Muzellec et al., 2017), among others. On the other hand, regularizing the dual Kantorovich problem was shown in (Liero et al., 2018) to be equivalent to unbalanced OT, that is optimal transport with relaxed marginal constraints.

Understanding why regularization helps The question of understanding why regularizing OT proves critical has triggered several approaches. A compelling reason is statistical: Although classical OT suffers from the curse of dimensionality, as its empirical version converges at a rate of order $(1/n)^{1/d}$ (Dudley, 1969; Fournier & Guillin, 2015; Weed et al., 2019), regularized OT and more precisely Sinkhorn divergences have a sample complexity of $O(1/\sqrt{n})$ (Genevay et al., 2019; Mena & Niles-Weed, 2019). Entropic OT was also shown to perform maximum likelihood estimation in the Gaussian deconvolution model (Rigollet & Weed, 2018). Taking another approach, (Dessein et al., 2018; Blondel et al., 2018) have considered general classes of convex regularizations and characterized them from a more geometrical perspective.

Robustness Recently, several papers (Genevay et al., 2018; Flamary et al., 2018; Deshpande et al., 2019; Kolouri et al., 2019; Niles-Weed & Rigollet, 2019; Paty & Cuturi, 2019) have proposed to maximize OT with respect to the ground cost, which can in turn be interpreted in light of ground metric learning (Cuturi & Avis, 2014). This approach can also be viewed as an instance of robust optimization (Ben-Tal & Nemirovski, 1998; Ben-Tal et al., 2009; Bertsimas et al., 2011): instead of considering a data-dependent, hence unstable minimization problem $\min_x f_{\hat{\theta}}(x)$ where $\hat{\theta}$ represents the data, the robust optimization literature adversarially chooses the parameters θ in a neighborhood of the data: $\max_{\theta\in\Theta} \min_x f(x)$. Continuing along these lines, we make a connection between *regularizing* and *maximizing* OT.

[1] CREST / ENSAE Paris, Institut Polytechnique de Paris [2] Google Brain. Correspondence to: François-Pierre Paty <francois.pierre.paty@ensae.fr>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

Contributions Our main goal is to provide a novel interpretation of regularized optimal transport in terms of ground cost robustness: regularizing OT amounts to maximizing **un**regularized OT with respect to the ground cost. Our contributions are:

1. We show that any convex regularization of the transport plan corresponds to ground-cost robustness (§ 3);
2. We reinterpret classical regularizations of OT in the ground-cost adversarial setting (§ 4);
3. We prove, under some technical assumption, a duality theorem for regularized OT, which we use to show that under the same assumption, there exists an optimal adversarial ground-cost that is separable (§ 5);
4. We extend ground-cost robustness to the case of more than two measures (§ 6);
5. We propose algorithms to solve the above-mentioned problems (§7) and illustrate them on data (§ 8).

2. Background on Optimal Transport and Notations

Let $\mathcal{X}$ be a compact Hausdorff space, and define $\mathscr{P}(\mathcal{X})$ the set of Borel probability measures over $\mathcal{X}$. We write $\mathcal{C}(\mathcal{X})$ for the set of continuous functions from $\mathcal{X}$ to $\mathbb{R}$, endowed with the supremum norm. For $\phi, \psi \in \mathcal{C}(\mathcal{X})$, we write $\phi \oplus \psi \in \mathcal{C}(\mathcal{X}^2)$ for the function $\phi \oplus \psi : (x, y) \mapsto \phi(x) + \psi(y)$.

For $n \in \mathbb{N}$, we write $[\![n]\!] = \{1, ..., n\}$. All vectors will be denoted with **bold** symbols. For a Boolean assertion A, we write $\iota(A)$ for its indicator function $\iota(A) = 0$ if A is true and $\iota(A) = +\infty$ otherwise.

Kantorovich Formulation of OT For $\mu, \nu \in \mathscr{P}(\mathcal{X})$, we write $\Pi(\mu, \nu)$ for the set of couplings

$$\begin{aligned}\Pi(\mu, \nu) = \{\pi \in \mathscr{P}(\mathcal{X}^2) \text{ s.t.} \forall A, B \subset \mathcal{X} \text{ Borel},\\ \pi(A \times \mathcal{X}) = \mu(A), \pi(\mathcal{X} \times B) = \nu(B)\}.\end{aligned}$$

For a real-valued continuous function $c \in \mathcal{C}(\mathcal{X}^2)$, the optimal transport cost between μ and ν is defined as

$$\mathscr{T}_c(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}^2} c(x, y)\, d\pi(x, y). \tag{1}$$

Since c is continuous and $\mathcal{X}$ is compact, the infimum in (1) is attained, see Theorem 1.4 in (Santambrogio, 2015). Problem (1) admits the following dual formulation, see Proposition 1.11 and Theorem 1.39 in (Santambrogio, 2015):

$$\mathscr{T}_c(\mu, \nu) = \max_{\substack{\phi,\psi \in \mathcal{C}(\mathcal{X}) \\ \phi \oplus \psi \leq c}} \int \phi\, d\mu + \int \psi\, d\nu. \tag{2}$$

Space of Measures Since $\mathcal{X}$ is compact, the dual space of $\mathcal{C}(\mathcal{X}^2)$ is the set $\mathscr{M}(\mathcal{X}^2)$ of Borel finite signed measures over $\mathcal{X}^2$. For $F : \mathscr{M}(\mathcal{X}^2) \to \mathbb{R}$, we recall that F is Fréchet-differentiable at π if there exists $\nabla F(\pi) \in \mathcal{C}(\mathcal{X}^2)$ such that for any $h \in \mathscr{M}(\mathcal{X}^2)$, as $t \to 0$

$$F(\pi + th) = F(\pi) + t \int \nabla F(\pi)\, dh + o(t).$$

Similarly, $G : \mathcal{C}(\mathcal{X}^2) \to \mathbb{R}$ is Fréchet-differentiable at c if there exists $\nabla G(c) \in \mathscr{M}(\mathcal{X}^2)$ such that for any $h \in \mathcal{C}(\mathcal{X}^2)$, as $t \to 0$

$$G(c + th) = G(c) + t \int h\, d\nabla G(c) + o(t).$$

Legendre–Fenchel Transformation For any functional $F : \mathscr{M}(\mathcal{X}^2) \to \mathbb{R} \cup \{+\infty\}$, we can define its convex conjugate $F^* : \mathcal{C}(\mathcal{X}^2) \to \mathbb{R} \cup \{+\infty\}$ and biconjugate $F^{**} : \mathscr{M}(\mathcal{X}^2) \to \mathbb{R} \cup \{+\infty\}$ as

$$F^*(c) := \sup_{\pi \in \mathscr{M}(\mathcal{X}^2)} \int c\, d\pi - F(\pi),$$

$$F^{**}(\pi) := \sup_{c \in \mathcal{C}(\mathcal{X}^2)} \int c\, d\pi - F^*(c).$$

F^* is always lower semi-continuous (lsc) and convex as the supremum of continuous linear functions.

Specific notations For $F : \mathscr{M}(\mathcal{X}^2) \to \mathbb{R} \cup \{+\infty\}$, we write $\operatorname{dom}(F) = \{\pi \in \mathscr{M}(\mathcal{X}^2) \mid F(\pi) < +\infty\}$ for its domain and will say that F is proper if $\operatorname{dom}(F) \neq \emptyset$.

We denote by $\mathscr{F}$ the set of proper lsc convex functions $F : \mathscr{M}(\mathcal{X}^2) \to \mathbb{R} \cup \{+\infty\}$, and for $\mu, \nu \in \mathscr{P}(\mathcal{X})$, we define the set $\mathscr{F}(\mu, \nu)$ of lsc convex functions that are proper on $\Pi(\mu, \nu)$:

$$\mathscr{F}(\mu, \nu) = \{F \in \mathscr{F} \mid \exists \pi \in \Pi(\mu, \nu), F(\pi) < +\infty\}.$$

3. Ground Cost Adversarial Optimal Transport

3.1. Definition

Instead of considering the classical *linear* formulation of optimal transport (1), we consider in this paper the following more general *nonlinear* convex formulation:

Definition 1. *Let $F \in \mathscr{F}$. For $\mu, \nu \in \mathscr{P}(\mathcal{X})$, we define:*

$$\mathscr{W}_F(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} F(\pi). \tag{3}$$

When $F(\pi) = \int c\, d\pi$, problem (3) corresponds to the classical optimal transport problem defined in (1) and $\mathscr{W}_F = \mathscr{T}_c$.

Lemma 1. *The infimum in (3) is attained. Moreover, if $F \in \mathscr{F}(\mu,\nu)$, $\mathscr{W}_F(\mu,\nu) < +\infty$.*

Proof. We can apply Weierstrass's theorem since $\Pi(\mu,\nu)$ is compact and F is lsc by definition. For $F \in \mathscr{F}(\mu,\nu)$, there exists $\pi_0 \in \Pi(\mu,\nu)$ such that $F(\pi_0) < +\infty$, so $\mathscr{W}_F(\mu,\nu) \leq F(\pi_0) < +\infty$. □

The main result of this paper is the following interpretation of problem (3) as a ground-cost adversarial OT problem:

Theorem 1. *For $\mu,\nu \in \mathscr{P}(\mathcal{X})$ and $F \in \mathscr{F}(\mu,\nu)$, minimizing F over $\Pi(\mu,\nu)$ is equivalent to the following convex problem:*

$$\mathscr{W}_F(\mu,\nu) = \sup_{c\in\mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - F^*(c). \tag{4}$$

Proof. Since F is proper, lsc and convex, Fenchel-Moreau theorem ensures that it is equal to its convex biconjugate F^{**}, so:

$$\begin{aligned}\min_{\pi\in\Pi(\mu,\nu)} F(\pi) &= \min_{\pi\in\Pi(\mu,\nu)} F^{**}(\pi)\\ &= \min_{\pi\in\Pi(\mu,\nu)} \sup_{c\in\mathcal{C}(\mathcal{X}^2)} \int c\, d\pi - F^*(c).\end{aligned}$$

Define the objective $l(\pi, c) := \int c\, d\pi - F^*(c)$. Since F^* is lsc as the convex conjugate of F, for any $\pi \in \Pi(\mu,\nu)$, $l(\pi,\cdot)$ is usc. It is also concave as the sum of concave functions. Likewise, for any $c \in \mathcal{C}(\mathcal{X}^2)$, $l(\cdot, c)$ is continuous and convex (in fact linear). Since $\Pi(\mu,\nu)$ and $\mathcal{C}(\mathcal{X}^2)$ are convex, and $\Pi(\mu,\nu)$ is compact, we can use Sion's minimax theorem to swap the min and the sup:

$$\min_{\pi\in\Pi(\mu,\nu)} F(\pi) = \sup_{c\in\mathcal{C}(\mathcal{X}^2)} \min_{\pi\in\Pi(\mu,\nu)} \int c\, d\pi - F^*(c).$$

Finally, $c \mapsto \mathscr{T}_c(\mu,\nu) - F^*(c)$ is concave since F^* is convex and $c \mapsto \mathscr{T}_c(\mu,\nu)$ is concave as the minimum of linear functionals. □

Remark 1. *Note that the inequality*

$$\mathscr{W}_F(\mu,\nu) \geq \sup_{c\in\mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - F^*(c)$$

is in fact verified for any $F : \mathscr{M}(\mathcal{X}^2) \to \mathbb{R} \cup \{+\infty\}$ since $F \geq F^{**}$ is always verified.

The supremum in equation (4) is not necessarily attained. Under some regularity assumption on F, we show that the supremum is attained and relate the optimal couplings and the optimal ground costs:

Proposition 1. *Let $\mu,\nu \in \mathscr{P}(\mathcal{X})$ and $F \in \mathscr{F}(\mu,\nu)$. Suppose that F is Fréchet-differentiable on $\Pi(\mu,\nu)$. Then the supremum in (4) is attained at $c_\star = \nabla F(\pi_\star)$ where $\pi_\star$ is any minimizer of (3). Conversely, suppose F^* is Fréchet-differentiable everywhere. If $c_\star$ is the unique maximizer in (4), then $\pi_\star = \nabla F^*(c_\star)$ is a minimizer of (3).*

See a proof in appendix. In section 5, we will further characterize $c_\star$ for a certain class of functions $F \in \mathscr{F}$.

One interesting particular case of Theorem 1 is when the convex cost $\pi \mapsto F(\pi)$ is a convex regularization of the classical linear optimal transport:

Corollary 1. *Let $c_0 \in \mathcal{C}(\mathcal{X}^2)$, $\mu,\nu \in \mathscr{P}(\mathcal{X})$. Let $\varepsilon > 0$ and $R \in \mathscr{F}(\mu,\nu)$. Then:*

$$\begin{aligned}&\min_{\pi\in\Pi(\mu,\nu)} \int c_0\, d\pi + \varepsilon R(\pi)\\ &= \sup_{c\in\mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - \varepsilon R^*\left(\frac{c - c_0}{\varepsilon}\right).\end{aligned} \tag{5}$$

Proof. We apply theorem 1 with $F(\pi) = \int c_0\, d\pi + \varepsilon R(\pi)$, for which we only need to compute the convex conjugate:

$$\begin{aligned}F^*(c) &= \sup_{\pi\in\mathscr{M}(\mathcal{X}^2)} \int c - c_0\, d\pi - \varepsilon R(\pi)\\ &= \varepsilon \sup_{\pi\in\mathscr{M}(\mathcal{X}^2)} \int \frac{c - c_0}{\varepsilon}\, d\pi - R(\pi)\\ &= \varepsilon R^*\left(\frac{c - c_0}{\varepsilon}\right).\end{aligned}$$

□

Corollary 1 shows that the ground cost c_0 in regularized optimal transport acts as a prior on the adversarial ground cost. Indeed, in equation (5) the penalization term $\varepsilon R^*\left(\frac{c-c_0}{\varepsilon}\right)$ forces any optimal adversarial ground cost to be "close" to c_0, the closeness being measured in terms of the convex conjugate of the regularization: R^*.

Remark 2. *We can also consider the minimization of a proper usc concave function F over $\Pi(\mu,\nu)$. Since $-F \in \mathscr{F}$, by reusing the argument of the proof of Theorem 1 (see a proof in appendix):*

$$\inf_{\pi\in\Pi(\mu,\nu)} F(\pi) = \inf_{c\in\mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) + (-F)^*(-c).$$

Minimizing a concave function of the transport plan $\pi \in \Pi(\mu,\nu)$, or equivalently maximizing a convex function of π, amounts to finding a ground cost $c \in \mathcal{C}(\mathcal{X}^2)$ that minimizes the transport cost between μ and ν plus a convex penalization on c. Note that this is not a convex problem since the objective is the sum of a concave and a convex functions. When μ and ν are discrete measures, $\Pi(\mu,\nu)$ is a finite-dimensional compact polytope so one of its extreme points has to be a minimizer of F.

In the ground cost maximization problem, the maximization is carried out on any continuous function c on $\mathcal{X}^2$, and in particular we do not impose that c takes only nonnegative values. In other words, an optimal adversarial ground cost may take negative values, which prevents us from directly interpreting optimal adversarial ground costs as suitable dissimilarity measures over $\mathcal{X}$. In the following subsection, we impose that $c \geq 0$ in the adversarial problem when the space $\mathcal{X}$ is discrete and prove an analogue of Corollary 1.

3.2. Discrete Separable Case

In this subsection, we will focus on the discrete case where the space $\mathcal{X} = [\![n]\!]$ for some $n \in \mathbb{N}$. A probability measure $\mu \in \mathscr{P}(\mathcal{X})$ is then a histogram of size n that we will represent by a vector $\boldsymbol{\mu} \in \mathbb{R}_+^n$ such that $\sum_{i=1}^n \boldsymbol{\mu}_i = 1$. Cost functions $c \in \mathcal{C}(\mathcal{X}^2)$ and transport plans $\pi \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ are now matrices $\mathbf{c}, \boldsymbol{\pi} \in \mathbb{R}^{n\times n}$.

We focus on regularization functions R that are separable, *i.e.* of the form

$$R(\boldsymbol{\pi}) = \sum_{i=1}^n \sum_{j=1}^n R_{ij}(\boldsymbol{\pi}_{ij})$$

for some differentiable convex proper lsc $R_{ij} : \mathbb{R} \to \mathbb{R}$.

In applications, it is natural to constrain the adversarial ground cost $\mathbf{c} \in \mathbb{R}^{n\times n}$ to take nonnegative entries. Adding this constraint on the adversarial cost corresponds to linearizing "at short range" the regularization R for "small transport values".

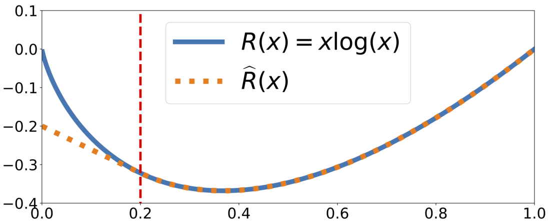


Figure 1. The entropy regularization $R(x) = x\log(x)$ and its linearized version $\widehat{R}(x)$ for small transport values.

Proposition 2. *Let $\varepsilon > 0$. For $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathscr{P}(\mathcal{X})$, it holds:*

$$\begin{aligned}&\sup_{\mathbf{c}\in\mathbb{R}_+^{n\times n}} \mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu}, \boldsymbol{\nu}) - \varepsilon \sum_{ij} R_{ij}^* \left(\frac{\mathbf{c}_{ij} - \mathbf{c}_{0ij}}{\varepsilon}\right) \\ &= \min_{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{c}_0, \boldsymbol{\pi}\rangle + \varepsilon \sum_{ij} \widehat{R}_{ij}(\boldsymbol{\pi}_{ij}) \qquad (6)\end{aligned}$$

where $\widehat{R}_{ij} : \mathbb{R} \to \mathbb{R}$ is the continuous convex function defined as

$$\widehat{R}_{ij}(x) := \begin{cases} R_{ij}(x) & \textit{if } x \geq {R_{ij}^*}' \left(-\frac{\mathbf{c}_{0\,ij}}{\varepsilon}\right) \\ \frac{-\mathbf{c}_{0\,ij}}{\varepsilon} x - R_{ij}^* \left(-\frac{\mathbf{c}_{0\,ij}}{\varepsilon}\right) & \textit{otherwise.}\end{cases}$$

Moreover, if R_{ij} is of class C^1, then $\widehat{R}_{ij}$ is also C^1.

4. Examples

4.1. Ground Cost Adversarial Interpretation of Classical OT Regularizations

As presented in the introduction, several convex regularizations R have been proposed in the literature. We give the ground cost adversarial counterpart for some of them: two examples in the continuous setting, and four p-norm based regularizations in the discrete case.

Example 1 (Entropic Regularization)**.** *Let $\mu, \nu \in \mathscr{P}(\mathcal{X})$. For $\pi \in \Pi(\mu, \nu)$, we define its relative entropy as $\mathrm{KL}(\pi\|\mu\otimes\nu) = \int \log \frac{d\pi}{d\mu\otimes\nu} d\pi$. Then for $c_0 \in \mathcal{C}(\mathcal{X}^2)$ and $\varepsilon > 0$, it holds:*

$$\begin{aligned}&\min_{\pi\in\Pi(\mu,\nu)} \int c_0 \, d\pi + \varepsilon \, \mathrm{KL}(\pi\|\mu\otimes\nu) \\ &= \sup_{c\in\mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu, \nu) - \varepsilon \int \exp\left(\frac{c - c_0}{\varepsilon}\right) d\mu\otimes\nu + \varepsilon.\end{aligned}$$

Proof. For $\pi \in \mathscr{M}(\mathcal{X}^2)$, let

$$R(\pi) = \begin{cases} \int \log \frac{d\pi}{d\mu\otimes\nu} d\pi - \int d\pi + 1 & \text{if } \pi \ll \mu\otimes\nu \\ +\infty & \text{otherwise.}\end{cases}$$

R is convex, and using proposition 7 in (Feydy et al., 2019),

$$R^*(c) = \int e^c - 1 \, d\mu\otimes\nu.$$

Applying corollary 1 concludes the proof. □

Another case of interest is the so-called Subspace Robust Wasserstein distance recently proposed by (Paty & Cuturi, 2019). Here, the set of adversarial metrics is parameterized by a finite-dimensional parameter Ω, which allows to recover an adversarial metric defined on the whole space even when the measures are finitely supported.

Example 2 (Subspace Robust Wasserstein)**.** *Let $d \in \mathbb{N}$, $k \in [\![d]\!]$ and $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$ with a finite second-order moment. For $\pi \in \Pi(\mu, \nu)$, define $V_\pi = \int (x-y)(x-y)^\top d\pi(x, y)$ and $\lambda_1(V_\pi) \geq \ldots \geq \lambda_d(V_\pi)$ its ordered eigenvalues.*

Then $F : \pi \mapsto \sum_{l=1}^k \lambda_l(V_\pi)$ is convex, and

$$\mathcal{S}_k(\mu, \nu) := \min_{\pi\in\Pi(\mu,\nu)} \sum_{l=1}^k \lambda_l(V_\pi) = \max_{\substack{0\preceq\Omega\preceq I \\ \mathrm{Tr}(\Omega)=k}} \mathscr{T}_{d_\Omega^2}(\mu, \nu)$$

where $d_\Omega^2(x, y) = (x-y)^\top \Omega (x-y)$ is the squared Mahalanobis distance.

Proof. See Theorem 1 in (Paty & Cuturi, 2019). Note that in this case, $\mathcal{X} = \mathbb{R}^d$ is not compact. This is not a problem since $F^* \equiv +\infty$ outside a compact set, *i.e.* the set on metrics on which the maximization takes place is compact. Indeed, one can show that:

$$F^*(c) = \iota(\exists 0 \preceq \Omega \preceq I \text{ with } \mathrm{Tr}(\Omega) = k \text{ s.t. } c = d_\Omega^2). \quad \square$$

Let us now consider p-norm based examples, which will subsume quadratically-regularized ($p = 2$) OT studied in (Essid & Solomon, 2017; Lorenz et al., 2019), capacity-constrained ($p = +\infty$) OT proposed by (Korman & McCann, 2015) and Tsallis regularized ($p < 0$) OT introduced by (Muzellec et al., 2017).

For a matrix $\mathbf{w} \in \mathbb{R}_+^{n\times n}$ with $\sum_{ij} \mathbf{w}_{ij} = n^2$ and $\boldsymbol{\pi} \in \mathbb{R}^{n\times n}$, we denote by $\|\boldsymbol{\pi}\|_{\mathbf{w},p}^p = \sum_{ij} \mathbf{w}_{ij} |\boldsymbol{\pi}_{ij}|^p$ the $\mathbf{w}$-weighted (powered) p-norm of $\boldsymbol{\pi}$. We also write $1/\mathbf{w}$ for the matrix defined by $(1/\mathbf{w})_{ij} = 1/\mathbf{w}_{ij}$. In the following, except otherwise mentioned, we take $p, q \in [1, +\infty]$ such that $1/p + 1/q = 1$, $\mathbf{c}_0 \in \mathbb{R}^{n\times n}$, $\varepsilon > 0$.

Example 3 ($\|\cdot\|_{\mathbf{w},p}^p$ Regularization)**.**

$$\min_{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{c}_0, \boldsymbol{\pi}\rangle + \varepsilon\frac{1}{p}\|\boldsymbol{\pi}\|_{\mathbf{w},p}^p$$
$$= \sup_{\mathbf{c}\in\mathbb{R}^{n\times n}} \mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\nu}) - \varepsilon\frac{1}{q}\left\|\frac{\mathbf{c}-\mathbf{c}_0}{\varepsilon}\right\|_{1/\mathbf{w}^{q-1},q}^q .$$

In particular when $p = 2$ and $\mathbf{w} = 1$, this corresponds to quadratically-regularized OT studied in (Essid & Solomon, 2017; Lorenz et al., 2019).

We give the details of the (straightforward) computations in the appendix.

Example 4 ($\|\cdot\|_{\mathbf{w},p}$ Penalization)**.**

$$\min_{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{c}_0, \boldsymbol{\pi}\rangle + \varepsilon\|\boldsymbol{\pi}\|_{\mathbf{w},p} = \sup_{\substack{\mathbf{c}\in\mathbb{R}^{n\times n}\\ \|\mathbf{c}-\mathbf{c}_0\|_{1/\mathbf{w},q}\le\varepsilon}} \mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\nu}).$$

Proof. We apply Corollary 1 with $R : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ defined as $R(\boldsymbol{\pi}) = \|\boldsymbol{\pi}\|_{\mathbf{w},p}$, for which we need to compute its convex conjugate. We know that the dual of $\|\cdot\|_p$ is $\iota(\|\cdot\|_q \le 1)$, and using classical results about convex conjugates, $\|\cdot\|_{\mathbf{w},p}^* = \iota(\|\cdot\|_{1/\mathbf{w},q} \le 1)$. □

Example 5 ($\|\cdot\|_{\mathbf{w},p}$ Regularization)**.**

$$\min_{\substack{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})\\ \|\boldsymbol{\pi}\|_{\mathbf{w},p}\le\varepsilon}} \langle \mathbf{c}_0, \boldsymbol{\pi}\rangle = \sup_{\mathbf{c}\in\mathbb{R}^{n\times n}} \mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\nu}) - \varepsilon\|\mathbf{c}-\mathbf{c}_0\|_{1/\mathbf{w},q}.$$

In particular when $p = +\infty$ and $\mathbf{w} = 1$, this coincides with capacity-constrained OT proposed by (Korman & McCann, 2015).

Proof. We apply Corollary 1 with $R : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ defined as $R(\boldsymbol{\pi}) = \iota(\|\boldsymbol{\pi}\|_{\mathbf{w},p} \le 1)$, for which we need to compute its convex conjugate. We know that the dual of $\iota(\|\cdot\|_p \le 1)$ is $\|\cdot\|_q$, and using classical results about convex conjugates, $\iota(\|\cdot\|_{\mathbf{w},p} \le 1)^* = \|\cdot\|_{1/\mathbf{w},q}$. □

Example 6 (Tsallis Regularization)**.** *For $q \in (0,1)$, the Tsallis regularized OT problem (Muzellec et al., 2017)*

$$\min_{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \langle \mathbf{c}_0, \boldsymbol{\pi}\rangle - \varepsilon\frac{1}{1-q}\sum_{ij}\left(\boldsymbol{\pi}_{ij}^q - \boldsymbol{\pi}_{ij}\right)$$

is equivalent to

$$\sup_{\substack{\mathbf{c}\in\mathbb{R}^{n\times n}\\ \mathbf{c}\le\mathbf{c}_0}} \mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\nu}) - \varepsilon^{\frac{1}{1-q}}(-p)^{-p}\left\|\frac{1}{\mathbf{c}_0-\mathbf{c}}\right\|_{-p}^{-p} + \frac{\varepsilon}{1-q}$$

where $p < 0$ is such that $1/p + 1/q = 1$.

We give the details of the computations in appendix.

4.2. A Link With the Matching Literature in Economics

Maximizing the OT problem with respect to the ground cost has been proposed in the matching literature in economics as a way to recover a ground cost when only a matching is observed, see *e.g.* (Dupuy & Galichon, 2014; Galichon & Salanié, 2015; Dupuy et al., 2016). In this subsection, we reinterpret their methods by showing that they are equivalent to some regularized OT problems. In other words, instead of interpreting a regularization problem as a robust OT problem as in subsection 4.1, we go the other way around and show that this practical OT maximization problem corresponds to a regularized OT problem.

Practitioners observe two probability measures $\mu, \nu \in \mathscr{P}(\mathcal{X})$ (*e.g.* features from a group of men and a group of women) and a matching $\pi_0 \in \Pi(\mu,\nu)$ (*e.g.* dating or marriage data). Under the assumption that the matching is optimal for some criteria, we can determine these by finding a ground cost $c_\star \in \mathcal{C}(\mathcal{X}^2)$ such that the matching π_0 is an optimal transport plan for the cost c. Then $c_\star(x,y)$ can be interpreted as the unwillingness for two people with characteristics x and y to be matched.

As shown in Theorem 3 in (Galichon & Salanié, 2015),

$$\sup_{c\in\mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - \int c\, d\pi_0 = \iota\left(\pi_0 \in \Pi(\mu,\nu)\right) \qquad (7)$$

and if $\pi_0 \in \Pi(\mu,\nu)$, the supremum is attained at any $c_\star \in \mathcal{C}(\mathcal{X}^2)$ such that π_0 is an optimal transport plan for the cost $c_\star$. Indeed, the first order condition for the maximization problem and the envelope theorem give the result.

In practice, economists are more interested in discovering which features explain the most the observed matching π_0. To this end, they choose a parametric model for the cost c, for example a Mahalanobis model $c \in \left\{d_\Omega^2 : (x,y) \mapsto (x-y)^\top\Omega(x-y) \,|\, \Omega \succeq 0, \|\Omega\| \le 1\right\}$. More generally, we can rewrite problem (7) as

$$\sup_{c\in\mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - \int c\, d\pi_0 - R^*(c) \qquad (8)$$

where $R \in \mathscr{F}$ is a lsc convex functional, *e.g.* $R^*(c) = \iota\left(\exists \Omega \succeq 0, \|\Omega\| \leq 1, c = d_\Omega^2\right)$ for the Mahalanobis model.

Using Theorem 1, we can then reinterpret problem (8):

$$\begin{aligned}&\sup_{c \in \mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - \int c \, d\pi_0 - R^*(c)\\ &= \min_{\pi \in \Pi(\mu,\nu)} R(\pi - \pi_0)\end{aligned}$$

where we have used the fact that $R^{**} = R$. Solving equation (8) amounts to finding a matching $\pi \in \Pi(\mu,\nu)$ that is close to the observed matching π_0, as measured by R.

5. Characterization of the Adversarial Cost and Duality

Theorem 1 shows that regularizing OT is equivalent to maximizing unregularized OT with respect to the ground cost. This gives access to a robustly computed ground-cost $c_\star$. In this section, we first prove a duality theorem for problem (3) that we use to further characterize $c_\star$. We will first need a technical assumption on F:

Definition 2. *Let $F \in \mathscr{F}$. We will say that F is* separably $*$-increasing *if for any $\phi, \psi \in \mathcal{C}(\mathcal{X})$ and any $c \in \mathcal{C}(\mathcal{X}^2)$:*

$$\phi \oplus \psi \leq c \Rightarrow F^*(\phi \oplus \psi) \leq F^*(c). \tag{9}$$

In particular if F^* is increasing, F is separably $*$-increasing.

This definition, albeit not always verified *e.g.* in the discrete separable case of Proposition 2 and in the SRW case of Example 2, is indeed verified in various cases of interest, *e.g.* for the entropic or $\|\cdot\|_{\mathbf{w},p}^p$ regularizations:

Example 7. *For $\mu, \nu \in \mathscr{P}(\mathcal{X})$, $c_0 \in \mathcal{C}(\mathcal{X}^2)$ and $\varepsilon > 0$, the entropy-regularized OT function*

$$F : \pi \mapsto \int c_0 \, d\pi + \varepsilon \, \mathrm{KL}(\pi \| \mu \otimes \nu)$$

is separably $*$-increasing.

Proof. As in the proof of example 1,

$$F^*(c) = \varepsilon \int \exp\left(\frac{c - c_0}{\varepsilon}\right) - 1 \, d\mu \otimes \nu$$

which verifies condition (9) as an increasing functional. □

Example 8. *In the discrete setting $\mathcal{X} = [\![n]\!]$, let $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathscr{P}(\mathcal{X})$, $\mathbf{c}_0 \in \mathbb{R}^{n \times n}$, $\mathbf{w} \in \mathbb{R}_+^{n \times n}$ summing to n^2. Take $p > 1$ and $\varepsilon > 0$. With $\varphi_p(x) = x^p$ if $x \geq 0$ and $\varphi_p(x) = +\infty$ if $x < 0$, the $\|\cdot\|_{\mathbf{w},p}^p$-regularized OT function*

$$F : \boldsymbol{\pi} \mapsto \langle \mathbf{c}_0, \boldsymbol{\pi} \rangle + \varepsilon \sum_{ij} \mathbf{w}_{ij} \varphi_p(\boldsymbol{\pi}_{ij})$$

is separably $*$-increasing.

Proof. Note that minimizing F over $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) \subset \mathbb{R}_+^{n \times n}$ is equivalent to minimizing $\widetilde{F} : \boldsymbol{\pi} \mapsto \langle \mathbf{c}_0, \boldsymbol{\pi} \rangle + \varepsilon \sum_{ij} \mathbf{w}_{ij} |\boldsymbol{\pi}_{ij}|^p$. One can show that, with $q > 1$ such that $1/p + 1/q = 1$ and $(x)_+ := \max\{0, x\}$:

$$F^*(\mathbf{c}) = \varepsilon \frac{1}{q} \left\| \frac{(\mathbf{c} - \mathbf{c}_0)_+}{\varepsilon} \right\|_{1/\mathbf{w}^{q-1}, q}^q$$

which clearly verifies condition (9). □

When F is separably $*$-increasing, we can easily prove a duality theorem for problem (3):

Theorem 2 ($\mathscr{W}_F$ duality)**.** *Let $\mu, \nu \in \mathscr{P}(\mathcal{X})$ and $F \in \mathscr{F}(\mu, \nu)$ a separably $*$-increasing function. Then:*

$$\mathscr{W}_F(\mu,\nu) = \max_{\phi, \psi \in \mathcal{C}(\mathcal{X})} \int \phi d\mu + \int \psi d\nu - F^*(\phi \oplus \psi). \tag{10}$$

Proof. Using Theorem 1 and Kantorovich duality (2):

$$\begin{aligned}\mathscr{W}_F(\mu,\nu) &= \sup_{c \in \mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - F^*(c)\\ &= \sup_{c \in \mathcal{C}(\mathcal{X}^2)} \max_{\substack{\phi,\psi \in \mathcal{C}(\mathcal{X})\\ \phi \oplus \psi \leq c}} \int \phi \, d\mu + \int \psi \, d\nu - F^*(c)\\ &= \sup_{c \in \mathcal{C}(\mathcal{X}^2)} \max_{\phi,\psi \in \mathcal{C}(\mathcal{X})} \int \phi \, d\mu + \int \psi \, d\nu - F^*(c)\\ &\qquad - \iota(\phi \oplus \psi \leq c)\\ &= \max_{\phi,\psi \in \mathcal{C}(\mathcal{X})} \int \phi \, d\mu + \int \psi \, d\nu\\ &\qquad + \sup_{c \in \mathcal{C}(\mathcal{X}^2)} -F^*(c) - \iota(\phi \oplus \psi \leq c)\\ &= \max_{\phi,\psi \in \mathcal{C}(\mathcal{X})} \int \phi \, d\mu + \int \psi \, d\nu - \inf_{\substack{c \in \mathcal{C}(\mathcal{X}^2)\\ \phi \oplus \psi \leq c}} F^*(c).\end{aligned}$$

Since F is separably $*$-increasing, for any $\phi, \psi \in \mathcal{C}(\mathcal{X})$,

$$\inf_{\substack{c \in \mathcal{C}(\mathcal{X}^2)\\ \phi \oplus \psi \leq c}} F^*(c) = F^*(\phi \oplus \psi),$$

which shows the desired duality result. □

Theorem 2 subsumes the already known duality results for entropy-regularized OT and quadratically-regularized OT. It also enables us to characterize of the optimal adversarial ground cost when the convex objective $F \in \mathscr{F}$ is separably $*$-increasing:

Corollary 2. *If $\phi_\star, \psi_\star$ are optimal solutions in (10), the cost $\phi_\star \oplus \psi_\star \in \mathcal{C}(\mathcal{X}^2)$ is an optimal adversarial cost in (4).*

Proof. For $\phi, \psi \in \mathcal{C}(\mathcal{X})$, note that

$$\mathscr{T}_{\phi \oplus \psi}(\mu,\nu) = \int \phi \, d\mu + \int \psi \, d\nu.$$

Then using $\mathscr{W}_F$ duality:

$$\begin{aligned}\mathscr{W}_F(\mu,\nu) &= \max_{\phi,\psi\in\mathcal{C}(\mathcal{X})} \int \phi\, d\mu + \int \psi\, d\nu - F^*(\phi\oplus\psi)\\ &= \max_{\phi,\psi\in\mathcal{C}(\mathcal{X})} \mathscr{T}_{\phi\oplus\psi}(\mu,\nu) - F^*(\phi\oplus\psi)\\ &\leq \sup_{c\in\mathcal{C}(\mathcal{X}^2)} \mathscr{T}_c(\mu,\nu) - F^*(c)\\ &= \mathscr{W}_F(\mu,\nu)\end{aligned}$$

where we have used Theorem 1 in the last line. This shows that the inequality is in fact an equality, so if $\phi_\star, \psi_\star$ are optimal dual potentials in (10), $\phi_\star\oplus\psi_\star$ is an optimal adversarial cost in (4). □

Corollary 2 is quite striking. Indeed, in the regularized formulation of Corollary 1, any optimal ground cost $c_\star$ in equation (5) should be close (in R^* sense) to the prior cost c_0 because of the penalization term $\varepsilon R^*\left(\frac{c-c_0}{\varepsilon}\right)$. But under the assumption that F is separably $*$-increasing, we have just shown that regardless of c_0, there exists an optimal adversarial ground cost that is separable.

6. Adversarial Ground-Cost for Several Measures

For two measures $\mu,\nu \in \mathscr{P}(\mathcal{X})$ and a separably $*$-increasing function $F \in \mathscr{F}(\mu,\nu)$, corollary 2 shows that there exists an optimal adversarial ground cost $c_\star$ that is separable. This separability, which is verified *e.g.* in the entropic or quadratic case, means that the OT problem for $c_\star$ is degenerate in the sense that any transport plan is optimal for the cost $c_\star$. From a metric learning point of view, $c_\star$ is not a suitable dissimilarity measure on $\mathcal{X}$. But why limit ourselves to two measures? If we observe $N\in\mathbb{N}$ measures $\mu_1,\dots,\mu_N \in \mathscr{P}(\mathcal{X})$, we could look for a ground cost $c\in\mathcal{C}(\mathcal{X}^2)$ that is adversarial to all the pairs:

$$\sup_{c\in\mathcal{C}(\mathcal{X}^2)} \sum_{i\neq j} \mathscr{T}_c(\mu_i,\mu_j) - F^*(c)$$

for some convex regularization $F^* : \mathcal{C}(\mathcal{X}^2)\to\mathbb{R}\cup\{+\infty\}$. We will specifically focus on the case where we observe a sequence of measures $\mu_{1:T} := \mu_1,\dots,\mu_T \in \mathscr{P}(\mathcal{X}), T\geq 2$. When we observe such time-dependent data, we can look for a sequence of adversarial costs $c_{1:T-1} := c_1,\dots,c_{T-1} \in \mathcal{C}(\mathcal{X}^2)$ which is globally adversarial:

Definition 3. *For $D : \mathcal{C}(\mathcal{X}^2)\times\mathcal{C}(\mathcal{X}^2)\to\mathbb{R}\cup\{+\infty\}$ and $F_t \in \mathscr{F}(\mu_t,\mu_{t+1})$, $t\in[\![T-1]\!]$, we define:*

$$\begin{aligned}\mathcal{W}_{D,F}(\mu_{1:T}) := \sup_{c_{1:T-1}} \sum_{t=1}^{T-1} &\mathscr{T}_{c_t}(\mu_t,\mu_{t+1}) \qquad (11)\\ &- D(c_t,c_{t+1}) - F_t^*(c_t)\end{aligned}$$

with the convention $D(c_{T-1},c_T)=0$.

In problem (11), D acts as a time-regularization by forcing the adversarial sequence of ground-costs to vary “continuously” with time.

Taking inspiration from the Subspace Robust Wasserstein (SRW) distance, we propose as a particular case of definition 3 a generalization of SRW to the case of a sequence of measures $\mu_1,\dots,\mu_T$, $T\geq 2$:

Definition 4. *Let $d\in\mathbb{N}$ and $k\in[\![d]\!]$. Define $\mathcal{R}_k = \left\{\Omega\in\mathbb{R}^{d\times d}\,|\,0\preceq\Omega\preceq I, \mathrm{Tr}(\Omega)=k\right\}$. We define the* sequential SRW *between $\mu_1,\dots,\mu_T\in\mathscr{P}(\mathbb{R}^d)$ as:*

$$\begin{aligned}\mathcal{TS}_{k,\eta}(\mu_{1:T}) := \sup_{\Omega_1,\dots,\Omega_{T-1}\in\mathcal{R}_k} \sum_{t=1}^{T-1} &\mathscr{T}_{d^2_{\Omega_t}}(\mu_t,\mu_{t+1}) \quad (12)\\ &- \eta\mathfrak{B}^2(\Omega_t,\Omega_{t+1})\end{aligned}$$

where $\mathfrak{B}^2(A,B) = \mathrm{Tr}(A+B-2(A^{\frac{1}{2}}BA^{\frac{1}{2}})^{\frac{1}{2}})$ is the squared Bures metric (Bures, 1969; Bhatia et al., 2018) on the SDP cone.

Note that problem (12) is convex. If $T=2$, the sequential SRW is equal to the usual SRW distance: $\mathcal{TS}_{k,\eta}(\mu_1,\mu_2) = \mathcal{S}_k(\mu_1,\mu_2)$.

7. Algorithms

From now on, we only consider the discrete case $\mathcal{X}=[\![n]\!]$.

7.1. Projected (Sub)gradient Ascent Solves Nonnegative Adversarial Cost OT

In the setting of subsection 3.2, we propose to run a projected subgradient ascent on the ground cost $\mathbf{c}\in\mathbb{R}_+^{n\times n}$ to solve problem (6). Note that in this case, $\widehat{F}(\boldsymbol{\pi}) := \langle\mathbf{c}_0,\boldsymbol{\pi}\rangle + \varepsilon\sum_{ij}\widehat{R}^*_{ij}\left(\frac{\mathbf{c}_{ij}-\mathbf{c}_{0\,ij}}{\varepsilon}\right)$ is **not** separably $*$-increasing, so we can hope that the optimal adversarial ground cost will not be separable.

At each iteration of the ascent, we need to compute a subgradient of $g:\mathbf{c}\mapsto\mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\nu}) - \varepsilon R^*\left(\frac{\mathbf{c}-\mathbf{c}_0}{\varepsilon}\right)$ given by Danskin’s theorem:

$$\begin{aligned}&\partial g(\mathbf{c}) =\\ &\mathrm{conv}\left\{\boldsymbol{\pi}_\star - \nabla R^*\left(\frac{\mathbf{c}-\mathbf{c}_0}{\varepsilon}\right)\,\middle|\, \boldsymbol{\pi}_\star\in\operatorname*{arg\,min}_{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})}\langle\mathbf{c},\boldsymbol{\pi}\rangle\right\}.\end{aligned}$$

Although projected subgradient ascent does converge, having access to gradients instead of subgradients, hence regularity, helps the convergence. We therefore propose to replace $\mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu},\boldsymbol{\nu})$ by its entropy-regularized version

$$\mathscr{S}^\eta_{\mathbf{c}}(\mu,\nu) = \min_{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})}\langle\mathbf{c},\boldsymbol{\pi}\rangle + \eta\sum_{ij}\boldsymbol{\pi}_{ij}(\log\boldsymbol{\pi}_{ij}-1)$$

Algorithm 1 Projected *(sub)*Gradient Ascent for Nonnegative Adversarial Cost

```
Input: Histograms $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}^n$, learning rate lr
Initialize $\mathbf{c} \in \mathbb{R}_+^{n \times n}$
for $i = 0$ to MAXITER do
  $\boldsymbol{\pi}_\star \leftarrow \mathrm{OT}(\boldsymbol{\mu}, \boldsymbol{\nu}, \mathrm{cost} = \mathbf{c})$
  $\mathbf{c} \leftarrow \mathrm{Proj}_{\mathbb{R}_+^{n \times n}} \left[\mathbf{c} + \mathrm{lr}\boldsymbol{\pi}_\star - \mathrm{lr}\nabla R^* \left(\frac{\mathbf{c} - \mathbf{c}_0}{\varepsilon}\right)\right]$
end for
```

in the definition of the obective g. Then g is differentiable, because there exists a unique solution $\boldsymbol{\pi}_\star$ in the entropic case (hence $\partial g(\mathbf{c})$ is a singleton). This will also speed up the computations of the gradient at each iteration using Sinkhorn's algorithm. We can interpret this addition of a small entropy term in the adversarial cost formulation as a further regularization of the primal:

Corollary 3. *Using the same notations as in Theorem 1, for* $\eta \geq 0$*:*

$$\sup_{\mathbf{c} \in \mathbb{R}^{n \times n}} \mathscr{S}_{\mathbf{c}}^\eta(\boldsymbol{\mu}, \boldsymbol{\nu}) - F^*(\mathbf{c})$$
$$= \min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} F(\boldsymbol{\pi}) + \eta \sum_{ij} \boldsymbol{\pi}_{ij}(\log \boldsymbol{\pi}_{ij} - 1).$$

Proof. Let $R(\boldsymbol{\pi}) := \sum_{ij} \boldsymbol{\pi}_{ij}(\log \boldsymbol{\pi}_{ij} - 1)$. Then:

$$\begin{aligned}
&\sup_{\mathbf{c} \in \mathbb{R}^{n \times n}} \mathscr{S}_{\mathbf{c}}^\eta(\boldsymbol{\mu}, \boldsymbol{\nu}) - F^*(\mathbf{c}) \\
&\quad = \sup_{\mathbf{c} \in \mathbb{R}^{n \times n}} \min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \langle \boldsymbol{\pi}, \mathbf{c} \rangle + \eta R(\boldsymbol{\pi}) - F^*(\mathbf{c}) \\
&\quad = \min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \eta R(\boldsymbol{\pi}) + \sup_{\mathbf{c} \in \mathbb{R}^{n \times n}} \langle \boldsymbol{\pi}, \mathbf{c} \rangle - F^*(\mathbf{c}) \\
&\quad = \min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \eta R(\boldsymbol{\pi}) + F(\boldsymbol{\pi})
\end{aligned}$$

where we have used Sion's minimax theorem as in the proof of Theorem 1 to swap the min and the sup, and used as well the fact that $F = F^{**}$ given by Fenchel-Moreau theorem. □

7.2. Sinkhorn-like Algorithm for $*$-increasing $F \in \mathscr{F}$

If the function $F \in \mathscr{F}$ is separably $*$-increasing, we can directly write the optimality conditions for the concave dual problem (10):

$$\boldsymbol{\mu} = \nabla F^*(\boldsymbol{\phi}_\star \oplus \boldsymbol{\psi}_\star)\mathbb{1} \tag{13}$$
$$\boldsymbol{\nu} = \nabla F^*(\boldsymbol{\phi}_\star \oplus \boldsymbol{\psi}_\star)^\top \mathbb{1} \tag{14}$$

where $\mathbb{1}$ is the vector of all ones. We can then alternate between fixing $\boldsymbol{\psi}$ and solving for $\boldsymbol{\phi}$ in (13) and fixing $\boldsymbol{\phi}$ and solving for $\boldsymbol{\psi}$ in (14). In the case of entropy-regularized OT, this is equivalent to Sinkhorn's algorithm. In quadratically-regularized OT, this is equivalent to the alternate minimization proposed by (Blondel et al., 2018). We give the detailed derivation of these facts in the appendix.

7.3. Coordinate Ascent for Sequential SRW

Problem (12) is a globally convex problem of $\Omega_1, \ldots, \Omega_{T-1}$. We propose to run a randomized coordinate ascent on the concave objective, *i.e.* to select $\tau \in [\![T-1]\!]$ randomly at each iteration and doing a gradient step for Ω_τ. We need to compute a subgradient of the objective $h : \Omega_\tau \mapsto \sum_{t=1}^{T-1} \mathscr{T}_{d^2_{\Omega_t}}(\mu_t, \mu_{t+1}) - \eta \mathfrak{B}^2(\Omega_t, \Omega_{t+1})$, given by:

$$\begin{aligned}\nabla h(\Omega_\tau) = V(\boldsymbol{\pi}_{\tau\star}) &- \eta \partial_1 \mathfrak{B}^2(\Omega_\tau, \Omega_{\tau+1}) \\ &- \eta \partial_2 \mathfrak{B}^2(\Omega_{\tau-1}, \Omega_\tau)\end{aligned} \tag{15}$$

where $\boldsymbol{\pi} \mapsto V(\boldsymbol{\pi})$ is defined in Example 2, $\boldsymbol{\pi}_{\tau\star} \in \mathbb{R}^{n \times n}$ is any optimal transport plan between $\mu_\tau, \mu_{\tau+1}$ for cost $d^2_{\Omega_\tau}$, and $\partial_1 \mathfrak{B}^2, \partial_2 \mathfrak{B}^2$ are the gradients of the squared Bures metric with respect to the first and second arguments, computed *e.g.* in (Muzellec & Cuturi, 2018).

Algorithm 2 Randomized (Block) Coordinate Ascent for sequential SRW

```
Input: Measures $\mu_1, \ldots, \mu_T \in \mathscr{P}(\mathbb{R}^d)$, dimension $k$, learning rate lr
Initialize $\Omega_1, \ldots, \Omega_{T-1} \in \mathbb{R}^{d \times d}$
for $i = 0$ to MAXITER do
  Draw $\tau \in [\![T-1]\!]$
  $\pi_{\tau\star} \leftarrow \mathrm{OT}(\mu_\tau, \mu_{\tau+1}, \mathrm{cost} = d^2_{\Omega_\tau})$
  $\Omega_\tau \leftarrow \mathrm{Proj}_{\mathcal{R}_k} [\Omega_\tau + \mathrm{lr}\nabla h(\Omega_\tau)]$ using (15)
end for
```

8. Experiments

8.1. Linearized Entropy-Regularized OT

We consider the entropy-regularized OT problem in the discrete setting:

$$\mathscr{S}_{\mathbf{c}_0}^\varepsilon(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \langle \mathbf{c}_0, \boldsymbol{\pi} \rangle + \varepsilon R(\boldsymbol{\pi})$$

where $\mathbf{c}_0 \in \mathbb{R}^{n \times n}$ and $R : \boldsymbol{\pi} \mapsto \sum_{ij} \boldsymbol{\pi}_{ij}(\log \boldsymbol{\pi}_{ij} - 1)$. Since R is separable, we can constrain the associated adversarial cost to be nonnegative by linearizing the entropic regularization. By proposition 2, this amounts to solve

$$\begin{aligned}&\sup_{\mathbf{c} \in \mathbb{R}_+^{n \times n}} \mathscr{T}_{\mathbf{c}}(\boldsymbol{\mu}, \boldsymbol{\nu}) - \varepsilon \sum_{ij} \exp\left(\frac{\mathbf{c}_{ij} - \mathbf{c}_{0ij}}{\varepsilon}\right) \\ &= \min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \langle \mathbf{c}_0, \boldsymbol{\pi} \rangle + \varepsilon \sum_{ij} \widehat{R}_{ij}(\boldsymbol{\pi}_{ij})\end{aligned} \tag{16}$$

where $\widehat{R}_{ij} : \mathbb{R} \to \mathbb{R}$ is defined as

$$\widehat{R}_{ij}(x) := \begin{cases} x(\log x - 1) & \text{if } x \geq \exp\left(-\frac{\mathbf{c}_{0ij}}{\varepsilon}\right) \\ \frac{-\mathbf{c}_{0ij}}{\varepsilon} x - \exp\left(-\frac{\mathbf{c}_{0ij}}{\varepsilon}\right) & \text{otherwise.} \end{cases}$$

We first consider $N = 100$ couples of measures $(\boldsymbol{\mu}_i, \boldsymbol{\nu}_i)$ in dimension $d = 1000$, each measure being a uniform measure on $n = 100$ samples from a Gaussian distribution with covariance matrix drawn from a Wishart distribution with $k = d$ degrees of freedom. For each couple, we run Algorithm 1 to solve problem (16). This gives an adversarial cost $\mathbf{c}_\star^\varepsilon$. We plot in Figure 2 the mean value of $\left|\widehat{W}_\varepsilon - \mathscr{T}_{\|\cdot\|^2}(\boldsymbol{\mu}_i, \boldsymbol{\nu}_i)\right|$ depending on ε, for $\widehat{W}_\varepsilon$ equal to $\mathscr{T}_{\mathbf{c}_\star^\varepsilon}(\boldsymbol{\mu}_i, \boldsymbol{\nu}_i)$, $\mathscr{S}^\varepsilon(\boldsymbol{\mu}_i, \boldsymbol{\nu}_i)$ and the value of (16). For small values of ε, all three values converge to the real Wasserstein distance. For large ε, Sinkhorn stabilizes to the MMD (Genevay et al., 2016) while the robust cost goes to 0 (for the adversarial cost goes to 0).

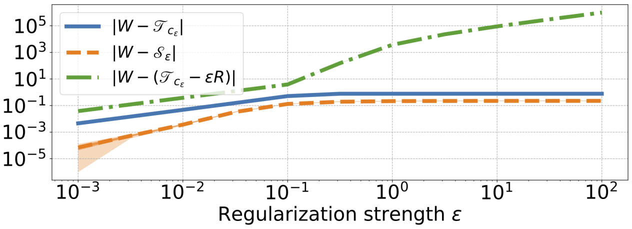
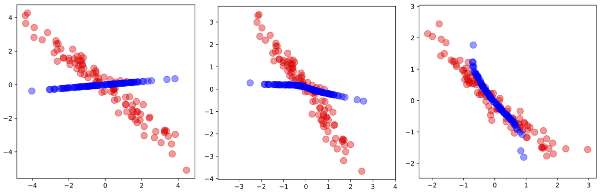


Figure 2. Mean value (over 100 runs) of the difference between the classical (2-Wasserstein) OT cost W and Sinkhorn $\mathscr{S}^\varepsilon$ (orange dashed), OT cost with adversarial nonnegative cost $\mathscr{T}_{c_\varepsilon}$ (blue line) and the value of problem (16) $\mathscr{T}_{c_\varepsilon} - \varepsilon R$ (green dot-dashed) depending on ε. The shaded areas represent the min-max, 10%-90% and 25%-75% percentiles, and appear negligeable except for numerical errors.

In Figure 3, we visualize the effect of the regularization ε on the ground cost $\mathbf{c}_\star^\varepsilon$ itself, for measures $\boldsymbol{\mu}, \boldsymbol{\nu}$ plotted in Figure 3a. We use multidimensional scaling on the adversarial cost matrix $\mathbf{c}_\star^\varepsilon$ (with distances between points from the same measures unchanged) to recover points in $\mathbb{R}^2$. For large values of ε, the adversarial cost goes to 0, which corresponds in the primal to a fully diffusive transport plan $\boldsymbol{\pi} = \boldsymbol{\mu}\boldsymbol{\nu}^\top$.



(a) Original Points (b) $\varepsilon = 0.01$ (c) $\varepsilon = 2$

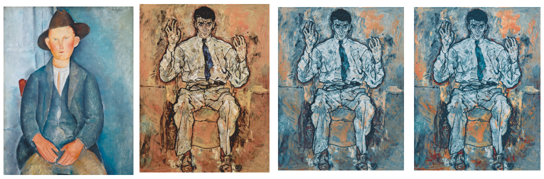
Figure 3. Effect of the regularization strength on the metric: as ε grows, the associated adversarial cost shrinks the distances.

8.2. Learning a Metric on the Color Space

We consider 20 measures $(\boldsymbol{\mu}_i)_{i=1,\dots,10}, (\boldsymbol{\nu}_j)_{j=1,\dots,10}$ on the red-green-blue color space identified with $\mathcal{X} = [0,1]^3$. Each measure is a point cloud corresponding to the colors used in a painting, divided into two types: ten portraits by Modigliani $(\boldsymbol{\mu}_i, i \in M)$ and ten by Schiele $(\boldsymbol{\nu}_j, j \in S)$. As in SRW and sequential SRW formulations, we learn a metric $c_\Omega \in \mathcal{C}(\mathcal{X}^2)$ parameterized by a matrix $0 \preceq \Omega \preceq I$ such that $\operatorname{Tr}\Omega = 1$ that best separates the Modiglianis and the Schieles:

$$\Omega_\star \in \underset{\Omega \in \mathcal{R}_1}{\arg\max} \sum_{i \in M} \sum_{j \in S} \mathscr{T}_{d_\Omega^2}(\boldsymbol{\mu}_i, \boldsymbol{\nu}_j).$$

We compute $\Omega_\star$ using projected SGD. We then use this “one-dimensional” metric $d_{\Omega_\star}^2$ as a ground metric for OT-based color transfer (Rabin et al., 2014): an optimal transport plan $\boldsymbol{\pi}$ between two color palettes $\boldsymbol{\mu}_i, \boldsymbol{\nu}_j$ gives a way to transfer colors from one painting to the other. Visually, transferring the colors using the classical quadratic cost $\|\cdot\|^2$ or the adversarially-learnt one-dimensional metric $d_{\Omega_\star}^2$ makes no major difference, showing that when regularized, OT can extract sufficient information from lower dimensional representations.



(a) Modigliani (b) Schiele (c) One-dimensional (d) Three-dimensional

Figure 4. Color transfer, best zoomed in. *(a)* and *(b)*: Original paintings. *(c)*: Schiele's painting with Modigliani's colors, using the learn adversarial one-dimensional metric $d_{\Omega_\star}^2$. *(d)*: Schiele's painting with Modigliani's colors, using the Euclidean metric $\|\cdot\|^2$.

9. Conclusion

In this paper, we have shown that any convex regularization of optimal transport can be recast as a ground cost adversarial problem. Under some technical assumption on the regularization, we proved a duality theorem for regularized OT, which we use to characterize the optimal ground cost as a separate function of its two arguments. In order to overcome this degeneration, we proposed to constrain the robust ground-cost to take non-negative values. We also proposed a framework to learn an adversarial sequence of ground costs which is adversarial to a time-varying sequence of measures. Future work includes learning a continuous adversarial cost c_θ parameterized by a neural network, under some regularity constraints (*e.g.* c_θ is Lipschitz). On the application side, learning low-dimensional representations of time-evolving data could be applied in biology as a refinement of the methodology of (Schiebinger et al., 2019).

Acknowledgements We acknowledge the support of a "Chaire d'excellence de l'IDEX Paris Saclay". We would like to thank Boris Muzellec and Théo Lacombe for fruitful discussions and relevant remarks.

References

Abadeh, S. S., Esfahani, P. M. M., and Kuhn, D. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pp. 1576–1584, 2015.

Alaux, J., Grave, E., Cuturi, M., and Joulin, A. Unsupervised hyper-alignment for multilingual word embeddings. In *International Conference on Learning Representations*, 2019.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. *Proceedings of the 34th International Conference on Machine Learning*, 70:214–223, 2017.

Ben-Tal, A. and Nemirovski, A. Robust convex optimization. *Mathematics of operations research*, 23(4):769–805, 1998.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. *Robust optimization*, volume 28. Princeton University Press, 2009.

Bertsimas, D., Brown, D. B., and Caramanis, C. Theory and applications of robust optimization. *SIAM review*, 53(3): 464–501, 2011.

Bhatia, R., Jain, T., and Lim, Y. On the bures-wasserstein distance between positive definite matrices. *Expositiones Mathematicae, to appear*, 2018.

Blondel, M., Seguy, V., and Rolet, A. Smooth and sparse optimal transport. In Storkey, A. and Perez-Cruz, F. (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 880–889, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR. URL http://proceedings.mlr.press/v84/blondel18a.html.

Bonneel, N., Peyré, G., and Cuturi, M. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4):71:1–71:10, 2016.

Bures, D. An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w^*-algebras. *Transactions of the American Mathematical Society*, 135:199–212, 1969.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9): 1853–1865, 2016.

Cuturi, M. Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pp. 2292–2300, 2013.

Cuturi, M. and Avis, D. Ground metric learning. *Journal of Machine Learning Research*, 15:533–564, 2014.

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

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D., and Schwing, A. G. Max-sliced wasserstein distance and its use for gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10648–10656, 2019.

Dessein, A., Papadakis, N., and Rouas, J.-L. Regularized optimal transport and the rot mover's distance. *The Journal of Machine Learning Research*, 19(1):590–642, 2018.

Dudley, R. M. The speed of mean Glivenko-Cantelli convergence. *Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Dupuy, A. and Galichon, A. Personality traits and the marriage market. *Journal of Political Economy*, 122(6):1271–1319, 2014.

Dupuy, A., Galichon, A., and Sun, Y. Estimating matching affinity matrix under low-rank constraints. *Arxiv:1612.09585*, 2016.

Essid, M. and Solomon, J. Quadratically-regularized optimal transport on graphs. *arXiv preprint arXiv:1704.08200*, 2017.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouve, A., and Peyré, G. Interpolating between optimal transport and mmd using sinkhorn divergences. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2681–2690. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/feydy19a.html.

Flamary, R., Courty, N., Rakotomamonjy, A., and Tuia, D. Optimal transport with laplacian regularization. In *NIPS 2014, Workshop on Optimal Transport and Machine Learning*, 2014.

Flamary, R., Cuturi, M., Courty, N., and Rakotomamonjy, A. Wasserstein discriminant analysis. *Machine Learning*, 107(12):1923–1945, 2018.

Fournier, N. and Guillin, A. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pp. 2053–2061, 2015.

Galichon, A. and Salanié, B. Cupid's invisible hand: Social surplus and identification in matching models. *Available at SSRN 1804623*, 2015.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pp. 3440–3448, 2016.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617, 2018.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. Sample complexity of sinkhorn divergences. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1574–1583. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/genevay19a.html.

Grave, E., Joulin, A., and Berthet, Q. Unsupervised alignment of embeddings with wasserstein procrustes. 2019.

Hashimoto, T., Gifford, D., and Jaakkola, T. Learning population-level diffusions with generative RNNs. In *International Conference on Machine Learning*, pp. 2417–2426, 2016.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. Generalized sliced wasserstein distances. In *Advances in Neural Information Processing Systems*, pp. 261–272, 2019.

Korman, J. and McCann, R. Optimal transportation with capacity constraints. *Transactions of the American Mathematical Society*, 367(3):1501–1521, 2015.

Liero, M., Mielke, A., and Savaré, G. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

Lorenz, D. A., Manns, P., and Meyer, C. Quadratically regularized optimal transport. *arXiv preprint arXiv:1903.01112*, 2019.

Mena, G. and Niles-Weed, J. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, pp. 4543–4553, 2019.

Muzellec, B. and Cuturi, M. Generalizing point embeddings using the wasserstein space of elliptical distributions. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 10258–10269. Curran Associates, Inc., 2018.

Muzellec, B., Nock, R., Patrini, G., and Nielsen, F. Tsallis regularized optimal transport and ecological inference. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Niles-Weed, J. and Rigollet, P. Estimation of wasserstein distances in the spiked transport model. *arXiv preprint arXiv:1909.07513*, 2019.

Paty, F.-P. and Cuturi, M. Subspace robust Wasserstein distances. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5072–5081, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/paty19a.html.

Rabin, J. and Papadakis, N. Convex color image segmentation with optimal transport distances. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 256–269. Springer, 2015.

Rabin, J., Ferradans, S., and Papadakis, N. Adaptive color transfer with relaxed optimal transport. In *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 4852–4856. IEEE, 2014.

Rigollet, P. and Weed, J. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Mathematique*, 356(11-12):1228–1235, 2018.

Salimans, T., Zhang, H., Radford, A., and Metaxas, D. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkQkBnJAb.

Santambrogio, F. *Optimal transport for applied mathematicians*. Birkhauser, 2015.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional Wasserstein distances: efficient optimal transportation on geometric domains. *ACM Transactions on Graphics*, 34 (4):66:1–66:11, 2015.

Villani, C. *Optimal Transport: Old and New*, volume 338. Springer Verlag, 2009.

Weed, J., Bach, F., et al. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.