Optimal transport with f-divergence regularization and generalized Sinkhorn algorithm

Dávid Terjék*
Alfréd Rényi Institute of Mathematics

Diego González-Sánchez*
Alfréd Rényi Institute of Mathematics

Abstract

Entropic regularization provides a generalization of the original optimal transport problem. It introduces a penalty term defined by the Kullback-Leibler divergence, making the problem more tractable via the celebrated Sinkhorn algorithm. Replacing the Kullback-Leibler divergence with a general f-divergence leads to a natural generalization. The case of divergences defined by superlinear functions was recently studied by Di Marino and Gerolin. Using convex analysis, we extend the theory developed so far to include all f-divergences defined by functions of Legendre type, and prove that under some mild conditions, strong duality holds, optimums in both the primal and dual problems are attained, the generalization of the c-transform is well-defined, and we give sufficient conditions for the generalized Sinkhorn algorithm to converge to an optimal solution. We propose a practical algorithm for computing an approximate solution of the optimal transport problem with f-divergence regularization via the generalized Sinkhorn algorithm. Finally, we present experimental results on synthetic 2-dimensional data, demonstrating the effects of using different f-divergences for regularization, which influences convergence speed, numerical stability and sparsity of the optimal coupling.

*Both authors should be equally credited for this work.

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

1 INTRODUCTION

Since its inception in the 18th century with the work of Gaspard Monge, the theory of optimal transport (Villani, 2008) has found its applications in many areas such as physics, economics and statistics. Among other developments, the optimal transport problem led L. V. Kantorovich to develop his duality theory (Kantorovich, 1940) and to pioneer the field of linear programming (Kantorovich, 1939) for practical solutions during World War II. This theory has been applied successfully in computer vision in tasks such as image retrieval (Rubner et al., 1997). However, computing the optimal transport involved solving a linear program which was computationally too costly to apply it to machine learning. Cuturi showed that slightly modifying the original optimal transport problem by introducing a regularization term one can compute the (regularized) optimal transport cost using the Sinkhorn algorithm (Sinkhorn and Knopp, 1967) in significantly less time (Cuturi, 2013). In recent years, this generalization of the optimal transport problem called entropy-regularized optimal transport (Peyré and Cuturi, 2019) has become a popular tool in the machine learning community (Feydy et al., 2019; Lorenz and Mahler, 2020; Di Marino and Gerolin, 2020b).

1.1 Our contributions

Let μ and ν be Borel probability measures defined on compact metric spaces X and Y respectively. Let D_ϕ be an f-divergence defined by a convex and lower semicontinuous function $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ such that $\phi(1) = 0$. Let $c : X \times Y \to \mathbb{R}$ be a Lipschitz continuous cost function and $\epsilon > 0$ a constant. We are interested in the optimal transport problem with f-divergence regularization (or Primal Problem) defined as

$$\underset{\epsilon}{\mathrm{OT}}(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left\{ \int c d\pi + \epsilon D_\phi(\pi \| \mu \otimes \nu) \right\} \quad (1)$$

where $\Pi(\mu,\nu)$ is the set of Borel couplings of μ and ν. The corresponding Dual Problem is then

$$\sup_{f\oplus g\leq c+\epsilon\phi'(\infty)} \left\{ \int f\oplus g d\mu\otimes\nu - \epsilon\int \phi_+^*\circ\frac{1}{\epsilon}(f\oplus g-c)d\mu\otimes\nu \right\}, \quad (2)$$

where the potentials f and g are assumed to be real-valued Lipschitz functions on X and Y, respectively, $\phi'(\infty)=\lim_{x\to\infty}\frac{\phi(x)}{x}$, and ϕ_+^* is the convex conjugate of $\phi_+=\phi+\iota_{\mathbb{R}_+}$.

In this paper we prove that if ϕ is of Legendre type then the Primal and Dual Problems have equal optimums. Furthermore, there exists optimal couplings for (1) and optimal potentials for (2). This generalizes the work of Di Marino and Gerolin (2020b), which develops the theory for superlinear ϕ, i.e. for $\phi'(\infty)=\infty$. We also prove that the singular part (which is always 0 for superlinear ϕ) of an optimal coupling is supported on a c-cyclically monotone set (Villani, 2008, Definition 5.1) (see Theorem 3).

In order to prove these results, we also generalize the (c,ϵ,ϕ)-transform (Di Marino and Gerolin, 2020b, Definition 3.1) so that it also works for non-superlinear ϕ, i.e. for $\phi'(\infty)<\infty$. This turned out to be a non-trivial task. Moreover, an interesting phenomenon occurs in the case of non-superlinear divergences as the corresponding (c,ϵ,ϕ)-transform sometimes collapses to (almost) the c-transform (Villani, 2008, Definition 5.2) (see Proposition 15 in Appendix B). This shows a more explicit connection between the classical theory of optimal transport and the regularized versions.

We show that a generalized version of the Sinkhorn algorithm (also denoted IPFP sequences (Di Marino and Gerolin, 2020b)) converge to an optimal solution even in the non-superlinear case under mild assumptions (see Definition 5 and Theorem 6). Finally, we propose a practical algorithm for computing an approximate solution of the optimal transport problem with f-divergence regularization using the generalized Sinkhorn algorithm.

We demonstrate the method on synthetic 2-dimensional point clouds. Our results indicate that for practical implementations the χ^2 divergence can compete with the Kullback-Leibler divergence of classical entropy-regularized OT. The corresponding algorithm is slightly slower but gives sparse optimal couplings. Thus, it could be useful in any task where we can make use of this sparsity, see Appendix D.3.

1.2 Related work

Since the breakthrough of Cuturi (2013), the area of entropy-regularized optimal transport has grown quickly (Peyré and Cuturi, 2019; Santambrogio, 2015). Some of them have focused on studying the case of the Kullback-Leibler divergence and Γ-convergence to the unregularized problem (Clason et al., 2019). Others have focused on generalizing the regularization to tackle linear programming problems (Benamou et al., 2015). There are results on Γ-convergence for the squared Euclidean cost and a proof of convergence of the discrete entropic smoothing of the Wasserstein gradient flow (Carlier et al., 2017). We can also find a theoretical proof together with practical experiments of the usefulness of Sinkhorn divergences, which remove the bias introduced to the optimal coupling by the regularization term (Feydy et al., 2019). Other types of generalizations have also been proposed (Roberts et al., 2017).

But the work that motivated the most our results (and which is clearly closest to this paper) is Di Marino and Gerolin (2020b). In this paper we find general results on strong duality and convergence of Sinkhorn iterations in the superlinear case ($\phi'(\infty)=\infty$). Indeed, our initial motivation was to understand the difficulties that arise in the non-superlinear case as most of the popular f-divergences used nowadays are non-superlinear (Agrawal and Horel, 2020, Table 1), whereas in many places in the literature this assumption seems necessary (see (Carlier et al., 2017, Assumption 3.1) and (Lorenz and Mahler, 2020, Section 4)). Thus, we decided to follow the same structure as Di Marino and Gerolin (2020b) in the theoretical section of our paper, generalizing the proofs and concepts present in their work. In addition, we wanted to give rigorous proofs in the context of Lipschitz functions, that, as we explain in the paper, model better the case of neural networks.

To conclude this section, we would like to highlight Dessein et al. (2018) where we find results on regularized optimal transport in finite spaces with Bregman divergences, which intersect with the set of f-divergences only at the Kullback-Leibler divergence. And Muzellec et al. (2017) where the case of Tsallis entropies (which are a subset of f-divergences) for the discrete case is covered. Other works focusing on finite spaces are Genevay et al. (2016); Altschuler et al. (2017); Blondel et al. (2018); Luise et al. (2018, 2019). Closer to our work are also Ferradans et al. (2014); Rakotomamonjy et al. (2015); Cuturi and Peyré (2016); Lorenz et al. (2019); Di Marino and Gerolin (2020a); Kurose et al. (2021); Eckstein and Nutz (2021); Lin et al. (2019).

2 BACKGROUND

2.1 Notation

We denote the extended reals by $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, the nonnegative reals by $\mathbb{R}_+$, and the extended nonnegative reals by $\overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \infty$. The indicator of a set A is denoted by ι_A with $\iota_A(x) = 0$ if $x \in A$ and $\iota_A(x) = \infty$ otherwise. We denote by $\operatorname{int} A$ the interior of a set A inside a topological space. Absolute continuity and singularity of measures will be denoted by $\ll$ and $\perp$ respectively. The Radon-Nikodym derivative of a measure μ with respect to a nonnegative measure ν such that $\mu \ll \nu$ is denoted by $\frac{d\mu}{d\nu}$ and the support of a measure μ by $\operatorname{supp}(\mu)$. The product of measures μ, ν is denoted by $\mu \otimes \nu$ and the set of measures having μ and ν as marginals by $\Pi(\mu, \nu)$. The set of probability measures on a measurable space X is denoted by $P(X)$. For functions $f : X \to \mathbb{R}$ and $g : Y \to \mathbb{R}$, the tensor sum $f \oplus g : X \times Y \to \mathbb{R}$ is defined as $f \oplus g(x, y) = f(x) + g(y)$. Given a convex function $\phi : \mathbb{R} \to \overline{\mathbb{R}}$, its effective domain $\operatorname{dom} \phi \subset \mathbb{R}$ is defined as $\operatorname{dom} \phi = \{s \in \mathbb{R} : \phi(s) < \infty\}$ and the convex conjugate $\phi^* : \mathbb{R} \to \overline{\mathbb{R}}$ as $\phi^*(t) = \sup_{s \in \mathbb{R}} \{st - \phi(s)\}$. Such a function ϕ is proper if $\operatorname{dom} \phi \neq \emptyset$ and $\phi > -\infty$.

2.2 f-divergences

Given a proper, convex and lower semicontinuous function[1] $\phi : \mathbb{R} \to \overline{\mathbb{R}}$, a measure μ and a nonnegative measure ν on a measurable space X, the f-divergence of μ from ν is defined (Csiszár, 1963; Ali and Silvey, 1966; Csiszár, 1967; Csiszár et al., 1999; Borwein and Lewis, 1993; Agrawal and Horel, 2020) as

$$D_\phi(\mu \| \nu) = \int \phi \circ \frac{d\mu_c}{d\nu} d\nu + \phi'(\infty)\mu_s^+(X) - \phi'(-\infty)\mu_s^-(X).$$

Here, $\mu_c \ll \nu, \mu_s \perp \nu$ are the absolutely continuous and singular parts of the Lebesgue decomposition of μ with respect to ν and $\mu_s^+, \mu_s^- \geq 0$ is the Jordan decomposition of the singular part. By definition $\phi'(\pm\infty) = \lim_{x \to \pm\infty} \frac{\phi(x)}{x} \in \overline{\mathbb{R}}$. Restricting to nonnegative measures can be done by using $\phi_+ = \phi + \iota_{\mathbb{R}_+}$ in place of ϕ, inducing $D_{\phi_+}(\mu \| \nu) = D_\phi(\mu \| \nu)$ if $\mu \geq 0$ and ∞ otherwise.

A subset of f-divergences including the Kullback-Leibler, reverse Kullback-Leibler, χ^2, reverse χ^2, squared Hellinger, Jensen-Shannon, Jeffreys and triangular discrimination divergences, but excluding the total variation, consists of those defined by functions ϕ of Legendre type. A proper, convex and lower semicontinuous function $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ is said to be of Legendre type (Borwein and Lewis, 1993, Definition 2.5) if it is strictly convex on $\operatorname{dom} \phi$ and differentiable on $\operatorname{int} \operatorname{dom} \phi$ with $\lim_{s \to \inf \operatorname{dom} \phi} \phi'(s) = -\infty$ if $\inf \operatorname{dom} \phi > -\infty$ and $\lim_{s \to \sup \operatorname{dom} \phi} \phi'(s) = \infty$ if $\sup \operatorname{dom} \phi < \infty$.

[1] Originally, f is used in place of ϕ (hence the name), but we reserve the symbol f for other functions.

2.3 Entropy-regularized optimal transport

Let $\mu \in P(X)$ and $\nu \in P(Y)$ be probability measures defined on spaces X and Y and let D_ϕ be an f-divergence. The generalized entropy regularized optimal transport problem with cost function $c : X \times Y \to \overline{\mathbb{R}}$ and regularization coefficient $\epsilon \in \mathbb{R}_+$ is defined in (1), and the corresponding dual problem[2] is defined in (2) (Di Marino and Gerolin, 2020b). Research in this area deals with the problem of finding suitable conditions under which strong duality holds, i.e. (2) equals (1). In some cases of interest, there are known sufficient conditions ensuring that the infimum and the supremum are achieved by optimal primal and dual variables, and characterizations of such optimal variables have been developed as well.

The case $\epsilon = 0$ with any ϕ reduces to the original, unregularized optimal transport problem, the duality theory of which is named after its most prominent contributor L. V. Kantorovich (Villani, 2008, Theorem 5.10). In this case, one has that there exists a closed, c-cyclically monotone set $C \subset X \times Y$ such that any optimal primal variable π is supported on C. A set $C \subset X \times Y$ is called c-cyclically monotone (Villani, 2008, Definition 5.1) if for any subset $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset C$ for $n \in \mathbb{N}$, one has $\sum_{i=1}^n c(x_i, y_i) \leq \sum_{i=1}^{n-1} c(x_i, y_{i+1}) + c(x_n, y_1)$. This means that an optimal π only assigns mass to pairs $(x_1, y_1), (x_2, y_2)$ such that one can not get lower transport cost by rerouting π to assign mass to $(x_1, y_2), (x_2, y_1)$ instead.

The case $\epsilon > 0$ with $\phi = x \log(x) - x + 1$, corresponding to the Kullback-Leibler divergence, became a popular tool in machine learning due to its better computational performance over the unregularized case. Cuturi proved that the Sinkhorn algorithm can be used in this case to obtain the optimal variables in a significantly smaller timeframe compared to the unregularized case (Cuturi, 2013). The price of efficiency is the optimal coupling being biased, an issue that has been investigated and remedied (Feydy et al., 2019). For more references on the state of the art see Section 1.2.

[2] The constraint $f \oplus g \leq c + \epsilon\phi'(\infty)$ is absent if ϕ is superlinear.

3 OPTIMAL TRANSPORT WITH f-DIVERGENCE REGULARIZATION

3.1 (c, ϵ, ϕ)-transform and f-Kantorovich duality

In this paper we study the problem of regularized optimal transport under the assumptions that the underlying spaces X and Y are compact metric spaces, and the cost function $c : X \times Y \to \mathbb{R}$ and the potentials $f : X \to \mathbb{R}$, $g : Y \to \mathbb{R}$ are Lipschitz. The reason we have chosen this family of functions is that for most applications the costs involved satisfy this hypothesis. Also, for deep learning applications, any function represented by a neural network is a Lipschitz function, and if one aims to implement the potentials by neural networks such as in a GAN setting, it makes sense to develop the theory of regularized optimal transport on Lipschitz functions.

Remark 1. The results presented in this paper can be also applied for Polish spaces X and Y as long as the measures $\mu \in P(X)$ and $\nu \in P(Y)$ are compactly supported. Furthermore, we can always assume that both μ and ν are of full support. To see this, note that if $\pi \in \Pi(\mu, \nu)$ then $\text{supp}(\pi) \subset \text{supp}(\mu) \times \text{supp}(\nu)$. Thus, for many problems (such as the ones we deal with in this paper), given compactly supported measures μ and ν on Polish spaces X and Y respectively, we can assume that $\text{supp}(\mu) = X$ and $\text{supp}(\nu) = Y$. If this is not the case, we can always restrict ourselves to the support, apply all the results that we are going to present to $\text{supp}(\mu)$ and $\text{supp}(\nu)$ and then go back to the original spaces. Given a measure defined in $\text{supp}(\mu) \times \text{supp}(\nu)$ it is trivial how to define a measure on $X \times Y$, and any function $f \in \text{Lip}(\text{supp}(\mu))$ or $g \in \text{Lip}(\text{supp}(\nu))$ can be extended to a Lipschitz function on X or Y with the same Lipschitz norm, respectively (Cobzaş et al., 2019, Theorem 4.1.1).

Our first main result concerns the generalization of the c-transform. Recall the classical problem of optimal transport $\text{OT}(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \{\int c \, d\pi\}$ for a cost function $c : X \times Y \to \mathbb{R}$ (Villani, 2008). For the sake of simplicity we assume that c is continuous and X and Y are compact metric spaces. It is trivial that for any pair of continuous functions $f : X \to \mathbb{R}$ and $g : Y \to \mathbb{R}$, if $f \oplus g \leq c$ then $\int f \, d\mu + \int g \, d\nu \leq \int c \, d\pi$ for any $\pi \in \Pi(\mu, \nu)$. A classical result of Kantorovich shows that in fact the supremum of $\int f \, d\mu + \int g \, d\nu$ over all functions $f \oplus g \leq c$ equals the infimum of $\int c \, d\pi$ over all couplings (Villani, 2008).

Let us now think about this problem in the following way, if $f \oplus g \leq c$ for any pair of functions (f, g) we can "improve" the value of $\int f \, d\mu + \int g \, d\nu$ by replacing g with $\inf_{x \in X}\{c(x, y) - f(x)\} := f^c(y)$. The latter function is called the c-transform of f (Villani, 2008). Clearly $f \oplus f^c \leq c$. Similarly, we could replace f with $(f^c)^c$, defined analogously. The values that we will obtain in the dual problem will never decrease, i.e. $\int f \, d\mu + \int g \, d\nu \leq \int f \, d\mu + \int f^c \, d\nu \leq \int (f^c)^c \, d\mu + \int f^c \, d\nu \leq \cdots$. Unfortunately, after repeating this process we will see that we get stuck (Villani, 2008, Proposition 5.8) and in general we will not reach the value $\text{OT}(\mu, \nu)$. The great advantage of regularized optimal transport is that if we replace $\text{OT}(\mu, \nu)$ by $\text{OT}_\epsilon(\mu, \nu)$ defined in (1), at the cost of introducing a bias, the analogue of the previous argument will in fact converge (under certain conditions) to $\text{OT}_\epsilon(\mu, \nu)$.

The analogue of the c-transform for the problem $\text{OT}_\epsilon(\mu, \nu)$ was introduced by Di Marino and Gerolin (2020b) for superlinear divergences ($\phi'(\infty) = \infty$) (Di Marino and Gerolin, 2020b, Definition 3.1). We generalize that definition to the case of any f-divergence defined by ϕ of Legendre type.

Definition 2 ((c, ϵ, ϕ)-transform)**.** Let $c \in \text{Lip}(X \times Y)$, $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ a proper, convex and lower semicontinuous function of Legendre type with $\phi(1) = 0$, $\epsilon > 0$, $\mu \in P(X)$ and $\nu \in P(Y)$ with full supports. We define the (c, ϵ, ϕ)-transform $f^{(c,\epsilon,\phi)} \in \text{Lip}(Y)$ of $f \in \text{Lip}(X)$ as follows:

$$f^{(c,\epsilon,\phi)}(y) := \underset{\gamma \leq f^c(y) + \epsilon\phi'(\infty)}{\arg\max} \left\{ \frac{1}{\epsilon}\gamma - \int \phi_+^* \left(\frac{1}{\epsilon}(f(x) + \gamma - c(x, y)) \right) \, d\mu(x) \right\}.$$

See Proposition 15 in Appendix B for properties of the (c, ϵ, ϕ)-transform. Let us now check why this definition is the natural generalization of the c-transform. Let (f, g) be a pair of potentials such that $f \oplus g \leq c + \epsilon\phi'(\infty)$. It follows from the convex conjugate of $D_{\phi_+}(\cdot \| \mu \otimes \nu)$ (Borwein and Lewis, 1993; Agrawal and Horel, 2020) and the Young-Fenchel inequality[3] that $\int f d\mu + \int g d\nu - \epsilon \int \phi_+^* \circ \frac{1}{\epsilon}(f \oplus g - c) d\mu \otimes \nu \leq \int c d\pi + \epsilon D_\phi(\pi \| \mu \otimes \nu)$. Looking at the left hand side of the inequality, notice that if we try to adjust the value of $g(y)$ **pointwise** at any fixed point y in such a way that $g(y) - \epsilon \int \phi_+^*(\frac{1}{\epsilon}(f(x) + g(y) - c(x, y)) d\mu(x)$ is maximized we obtain precisely the (c, ϵ, ϕ)-transform of f. Hence, we have an analogous inequality as before, $\int f d\mu + \int g d\nu - \epsilon \int \phi_+^* \circ \frac{1}{\epsilon}(f \oplus g - c) d\mu \otimes \nu \leq \int f d\mu + \int f^{(c,\epsilon,\phi)} d\nu - \epsilon \int \phi_+^* \circ \frac{1}{\epsilon}(f \oplus f^{(c,\epsilon,\phi)} - c) d\mu \otimes \nu$.

The similarities do not end here, we encourage the reader to compare Proposition 15 with Proposition 13 where we have stated many properties of the (c, ϵ, ϕ)-

[3] As any coupling π is by definition positive, we can replace ϕ by $\phi_+ = \phi + \iota_{\mathbb{R} \geq 0}$ and $D_\phi(\pi \| \mu \otimes \nu) = D_{\phi_+}(\pi \| \mu \otimes \nu)$.

and c-transforms, respectively. For now, let us mention how we can compute the value of the (c, ϵ, ϕ)-transform. The following result is (i) of Proposition 15:

$f^{(c,\epsilon,\phi)}(y)$ is well-defined for all $y \in Y$ implicitly by $\int_X {\phi_+^*}' \circ \frac{1}{\epsilon}(f + f^{(c,\epsilon,\phi)}(y) - c(\cdot, y))d\mu = 1$ if there exists such a number $f^{(c,\epsilon,\phi)}(y) \in \mathbb{R}$ or explicitly as $f^{(c,\epsilon,\phi)}(y) = \min_{x\in X}\{\epsilon\phi'(\infty) + c(x, y) - f(x)\} = f^c(y) + \epsilon\phi'(\infty)$ otherwise.

This shows precisely why if $\phi'(\infty) = \infty$ this definition reduces to solving the implicit equation

$$\int_X {\phi_+^*}' \circ \frac{1}{\epsilon}(f + \gamma - c(\cdot, y))d\mu = 1 \tag{3}$$

for γ. However, if this is not the case, there may be cases where the (c, ϵ, ϕ)-transform is just the c-transform plus $\epsilon\phi'(\infty)$. Indeed, this behaviour can happen as we can see in Example 17. Analogously to the c-subdifferential (Villani, 2008, Definition 5.2), the (c, ϵ, ϕ)-subdifferential of $f \in \mathrm{Lip}(X)$ defined as $\partial_{(c,\epsilon,\phi)}f = \{(x, y) \in X \times Y : f(x) + f^{(c,\epsilon,\phi)}(y) = c(x, y) + \epsilon\phi'(\infty)\}$ is a closed, c-cyclically monotone set (see Proposition 19). Analogous results hold for the (c, ϵ, ϕ)-transform, $g^{(c,\epsilon,\phi)} \in \mathrm{Lip}(X)$, of $g \in \mathrm{Lip}(Y)$.

We can now state one of the main results of this paper, generalizing the Kantorovich duality of optimal transport (see Di Marino and Gerolin (2020b, Section 3.2) and Villani (2008, Theorem 5.10)).

Theorem 3 (f-Kantorovich duality)**.** *Let $\mu \in P(X)$ and $\nu \in P(Y)$ be probability measures of full support on compact metric spaces (X, d_X) and (Y, d_Y). Let $c \in \mathrm{Lip}(X \times Y)$, $0 < \epsilon \in \mathbb{R}$ be a regularization coefficient and $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ a proper, convex and lower semicontinuous function of Legendre type. Then one has*

$$\begin{aligned}
&\min_{\pi\in\Pi(\mu,\nu)} \left\{ \int c d\pi + \epsilon D_\phi(\pi\|\mu\otimes\nu) \right\} \\
&= \max_{\substack{f\in\mathrm{Lip}(X), g\in\mathrm{Lip}(Y) \\ f\oplus g\le c+\epsilon\phi'(\infty)}} \left\{ \int f \oplus g d\mu\otimes\nu \right. \\
&\left. -\epsilon \int \phi_+^* \circ \frac{1}{\epsilon}(f \oplus g - c)d\mu\otimes\nu \right\} \\
&= \max_{f\in\mathrm{Lip}(X)} \left\{ \int f \oplus f^{(c,\epsilon,\phi)} d\mu\otimes\nu \right. \\
&\left. -\epsilon \int \phi_+^* \circ \frac{1}{\epsilon}(f \oplus f^{(c,\epsilon,\phi)} - c)d\mu\otimes\nu \right\} \\
&= \max_{g\in\mathrm{Lip}(Y)} \left\{ \int g^{(c,\epsilon,\phi)} \oplus g d\mu\otimes\nu \right. \\
&\left. -\epsilon \int \phi_+^* \circ \frac{1}{\epsilon}(g^{(c,\epsilon,\phi)} \oplus g - c)d\mu\otimes\nu \right\},
\end{aligned}$$

i.e., strong duality holds and optimums in both the Primal and Dual Problems are attained. The absolutely continuous part (with respect to $\mu\otimes\nu$) π_c of any optimal coupling π is unique with its density given by

$$\frac{d\pi_c}{d\mu\otimes\nu} = {\phi_+^*}' \circ \frac{1}{\epsilon}(f \oplus g - c), \tag{4}$$

where $(f, g) \in \mathrm{Lip}(X) \times \mathrm{Lip}(Y)$ are any pair of optimal potentials. Optimal potentials (f, g) are such that $f \oplus g$ is unique almost everywhere with respect to π_c, and $f^{(c,\epsilon,\phi)} = g$ and $g^{(c,\epsilon,\phi)} = f$ always hold. Moreover, there exists a closed, c-cyclically monotone set C, which can be taken to be the intersection of the (c, ϵ, ϕ)-subdifferentials $\partial_{(c,\epsilon,\phi)}f = \partial_{(c,\epsilon,\phi)}g = \{(x, y) \in X \times Y : f(x)+g(y) = c(x, y)+\epsilon\phi'(\infty)\}$ of all optimal couplings (f, g), such that the singular part (with respect to $\mu\otimes\nu$) π_s of any optimal coupling π is supported on C, i.e., $\mathrm{supp}(\pi_s) \subset C$.

See Theorem 18 and Proposition 20 in Appendix B for the proof. As a sketch, the proof of this result consists of two main parts. The first one is proving that both the Primal (1) and Dual (2) problems have the same optimum. The proof of this fact follows from convex analytic tools (Zalinescu, 2002, Theorem 2.6.1(v)). To prove attainment in the Primal problem we can use a standard functional analytic argument. To prove attainment in the Dual Problem we make use of the properties of the (c, ϵ, ϕ)-transform given by Proposition 15. Then, uniqueness properties of the optimal couplings and potentials follow from the characterization of the subdifferentials of f-divergences (Borwein and Lewis, 1993, Theorem 2.10) and the Young-Fenchel inequality.

3.2 Generalized Sinkhorn algorithm

The goal of this section is to prove that under certain conditions, given any starting pair of potentials (f, g) if we start replacing g with $f^{(c,\epsilon,\phi)}$, then f with $(f^{(c,\epsilon,\phi)})^{(c,\epsilon,\phi)}$ and so on, we are able to recover a pair of optimal potentials of the Dual Problem and an optimal coupling for the Primal Problem. This process is called the generalized Sinkhorn algorithm. A single Sinkhorn iteration is defined as follows. Note that this definition yields a generalization of IPFP sequences (Di Marino and Gerolin, 2020b, Section 4) but with a stabilizing factor that will be helpful both in theory to prove convergence and in practice to prevent overflow.

Definition 4 (Sinkhorn operator)**.** Let X and Y be compact metric spaces and $\mu \in P(X)$, $\nu \in P(Y)$ be Borel probability measures of full support. Let also $c \in \mathrm{Lip}(X \times Y)$, $0 < \epsilon \in \mathbb{R}$ be a regularization coefficient and $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ a proper, convex and lower semicontinuous function of Legendre type. Fix any point $y_0 \in Y$. Given a pair $(f, g) \in \mathrm{Lip}(X) \times \mathrm{Lip}(Y)$ we define the operator $\mathcal{F}^{(c,\epsilon,\phi)} : \mathrm{Lip}(X)\times\mathrm{Lip}(Y) \to \mathrm{Lip}(X)\times\mathrm{Lip}(Y)$

as[4]

$$\mathcal{F}^{(c,\epsilon,\phi)}(f,g) := \\ ((f^{(c,\epsilon,\phi)} - f^{(c,\epsilon,\phi)}(y_0))^{(c,\epsilon,\phi)}, f^{(c,\epsilon,\phi)} - f^{(c,\epsilon,\phi)}(y_0)).$$

The most important properties of this operator are that if $(f',g') = \mathcal{F}^{(c,\epsilon,\phi)}(f,g)$ then $\|f'\|_L, \|g'\|_L, \|f'\|_\infty$ and $\|g'\|_\infty$ are uniformly bounded in terms of the diameters of X and Y, $\|c\|_L$, $\|c\|_\infty$ and ϵ. Also, this operator is continuous in the product topology generated by $\|\cdot\|_\infty$ on $\mathrm{Lip}(X) \times \mathrm{Lip}(Y)$. See Proposition 22 for more details.

We have seen before that iterating the c-transform in the classical optimal transport problem usually does not converge to a pair of optimal potentials. However, we know that using the Kullback-Leibler divergence for regularization we get convergence of the Sinkhorn algorithm to optimal potentials (Cuturi, 2013). As we saw before, as soon as $\phi'(\infty) < \infty$ the (c,ϵ,ϕ)-transform can collapse to almost the usual c-transform, in which case convergence is not guaranteed. Therefore, we introduce a mild condition that ensures that even in this case, the (c,ϵ,ϕ)-transform never collapses to the c-transform plus $\epsilon\phi'(\infty)$. This condition on the other hand is general enough to be able to include many examples and different f-divergences, and ensures that the (c,ϵ,ϕ)-subdifferentials are always empty. This implies that any optimal coupling is absolutely continuous with respect to $\mu \otimes \nu$, so that the optimal coupling is actually unique, as in the case $\phi'(\infty) = \infty$.

Definition 5 (Good triple)**.** Let X be a compact metric space and μ a Borel probability measure on X. Let ϕ be proper, convex and lower semicontinuous function of Legendre type and suppose that $\phi'(\infty) < \infty$. Let $C > 0$ be a constant. We say that (X,μ,ϕ) is a *good triple* with respect to C if for all $x_0 \in X$ one has

$$\lim_{\delta \downarrow 0} \int_X {\phi_+^*}'(\phi'(\infty) - Cd(x_0,x) - \delta)\, d\mu(x) > 1.$$

This condition can be trivially verified if X is a discrete space and the measure μ has full support. But more generally it applies to other functions ϕ even in general compact metric spaces. For example, if $X = [0,1]$ with the usual Lebesgue measure and the Euclidean distance then it is easy to check by hand that if ϕ is the function defining either the Jensen-Shannon, the squared Hellinger or the reverse Kullback-Leibler divergence then (X,μ,ϕ) is a good triple with respect to any fixed constant C.

With this definition we can now state the main result of this section. Note that this result generalizes Di Marino and Gerolin (2020b, Theorem 4.1) to some cases where the divergence is not superlinear and it is adapted to the context of Lipschitz functions.

Theorem 6 (Convergence of generalized Sinkhorn algorithm)**.** *Let X and Y be compact metric spaces and $\mu \in P(X)$, $\nu \in P(Y)$ be Borel probability measures of full support. Let also $c \in \mathrm{Lip}(X \times Y)$, $0 < \epsilon \in \mathbb{R}$ be a regularization coefficient and $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ a proper, convex and lower semicontinuous function of Legendre type. Suppose that either $\phi'(\infty) = \infty$ or (X,μ,ϕ) and (Y,ν,ϕ) are good triples with respect to $2\|c\|_L/\epsilon$. Take any pair $(f_0,g_0) \in \mathrm{Lip}(X) \times \mathrm{Lip}(Y)$ and define inductively $(f_n,g_n) := \mathcal{F}^{(c,\epsilon,\phi)}(f_{n-1},g_{n-1})$ for $n \geq 1$. Let us also define the dual functional for any pair of functions $(f,g) \in \mathrm{Lip}(X) \times \mathrm{Lip}(Y)$ as*

$$D_\epsilon(f,g) := \int f \oplus g - \epsilon\phi_+^* \circ \left(\frac{1}{\epsilon}(f \oplus g - c)\right) d\mu \otimes \nu.$$

Then one has $D_\epsilon(f_n,g_n) \to \mathrm{OT}_\epsilon(\mu,\nu)$ as $n \to \infty$, and $f_n \oplus g_n \to \tilde{f} \oplus \tilde{g}$ in $L^\infty(\pi)$ as well with π being the unique optimal coupling and $(\tilde{f},\tilde{g})$ any pair of optimal potentials. Moreover, π can be recovered as $\pi = {\phi_+^*}' \circ \frac{1}{\epsilon}(\tilde{f} \oplus \tilde{g} - c) \cdot \mu \otimes \nu$.

4 EXPERIMENTS

4.1 Practical implementation

For measures with finite supports, if $\mathrm{supp}(\mu) = \{x_1,\dots,x_k\} = X$, the potential reduces to a finite-dimensional vector $f \in \mathbb{R}^k$ as $f_i = f(x_i)$ (and similarly for ν and g). In this case, the equation (3) defining the values of (c,ϵ,ϕ)-transforms can always be solved (Terjék, 2021) via Newton's method in parallel[5], which is included here as Algorithm 1. For $\phi'(\infty) < \infty$, initial values are chosen to be just below the boundary value by some parameter $\delta > 0$. For $\phi'(\infty) = \infty$, initial values are chosen to be $\gamma_i = \log\langle e^{h_{\cdot,i}}, \xi\rangle$, which is exactly the closed-form solution of $\gamma_{\phi,\xi}(h)$ for ϕ corresponding to the Kullback-Leibler divergence. Theoretically, as we are minimizing a convex function any initial value will eventually converge using Newton's method. We tried several initializations and this one seemed to give the best performance and that is why we have used it. Since we are running n parallel Newton's method iterations, we set the stopping criterion to be the mean of the squared Newton steps falling below a tolerance parameter τ.

We propose a practical implementation of the generalized Sinkhorn algorithm in Algorithm 2. In both algorithms, vectors are understood as row vectors, and

[4]Technically this operator depends as well on y_0, but as this is fixed and arbitrary, we decided not to include it explicitly.

[5]Note that we solve for $-\frac{1}{\epsilon}\gamma$ for better stability.

Algorithm 1 Calculate $\gamma_{\phi,\xi}(h)$

Input:
$h \in M_{m\times n}(\mathbb{R})$, $\xi \in \mathbb{R}^m$, $\phi : \mathbb{R} \to \overline{\mathbb{R}}$, $0 < \delta, \tau \in \mathbb{R}$
Output:
$\gamma_{\phi,\xi}(h) \in \mathbb{R}^n$

if $\phi'(\infty) < \infty$ **then**
  $\gamma_i = \max(h_{\cdot,i}) - \phi'(\infty) + \delta.$
else
  $\gamma_i = \log\langle e^{h_{\cdot,i}}, \xi\rangle$
end if
repeat
  $s_i = \frac{-\langle(\phi_+^*)'(h_{\cdot,i}-\gamma),\xi\rangle+1}{\langle(\phi_+^*)''(h_{\cdot,i}-\gamma),\xi\rangle}$
  $\gamma = \gamma - s$
until $\frac{1}{n}\sum_{i=1}^n s_i^2 < \tau$

Algorithm 2 Generalized Sinkhorn algorithm for computing optimal potentials f, g and optimal coupling π

Input:
$\mu \in \mathbb{R}^k$, $\mathrm{supp}(\mu) = \{x_1, \ldots, x_k\} \subset X$,
$\nu \in \mathbb{R}^l$, $\mathrm{supp}(\nu) = \{y_1, \ldots, y_l\} \subset Y$,
$c : X \times Y \to \mathbb{R}$, $\phi : \mathbb{R} \to \overline{\mathbb{R}}$, $0 < \epsilon, \tau \in \mathbb{R}$
Output:
$f \in \mathbb{R}^k$, $g \in \mathbb{R}^l$, $\pi \in M_{k\times l}(\mathbb{R})$

$C_{i,j} = c(x_i, y_i)$
$f_i = 0$
repeat
  $f_{prev} = f$
  $g = -\epsilon\gamma_{\phi,\mu}\left(\frac{1}{\epsilon}(f \otimes \mathbf{1}^l - C)\right)$
  $g = g - g_1$
  $f = -\epsilon\gamma_{\phi,\nu}\left(\frac{1}{\epsilon}(g \otimes \mathbf{1}^k - C^*)\right)$
until $\|f - f_{prev}\|_\infty < \tau$
$\pi_{i,j} = {\phi_+^*}'(\frac{1}{\epsilon}(f_i + g_j - C_{i,j}))\mu_i\nu_j$

statements containing indices i and/or j are to be executed for each index value in parallel. The vectors $\mathbf{1}^l$ and $\mathbf{1}^k$ represent column vectors of dimension l and k with all their coordinates equal to 1, and thus their tensor products with row vectors of dimension k and l give matrices of dimension $k \times l$ and $l \times k$, respectively. We denote the adjoint of $C \in M_{k\times l}(\mathbb{R})$ by $C^* \in M_{l\times k}(\mathbb{R})$. Since our convergence results are in terms of the infinity norm, we set the stopping criterion to be the infinity norm of the difference of the potantials falling below a given tolerance parameter τ.

4.2 Experimental setup

To demonstrate the feasibility of the approach, we apply the algorithm to synthetic 2-dimensional data obtained from `https://github.com/jeanfeydy/global-divergences`, the official codebase of Feydy et al. (2019)[6]. The data consists of 4 pairs of densities on $\mathbb{R}^2$, nicknamed "crescents", "densities", "moons" and "slopes". The task with each pair is to compute the regularized optimal transport problem between measures obtained by sampling a set of points independently from each density. Using different f-divergences and ϵs influences many aspects of the task, which are detailed below. In all examples, the cost function is $c(x, y) = \frac{1}{2}\|x - y\|_2^2$, i.e. half of the squared Euclidean distance on the plane. We consider classical f-divergences defined by ϕ of Legendre type, specifically the Kullback-Leibler, reverse Kullback-Leibler, χ^2 (or Neyman χ^2), reverse χ^2 (or Pearson χ^2), squared Hellinger, Jensen-Shannon, Jeffreys and triangular discrimination (or Vincze-Le Cam) divergences. The corresponding functions needed for the algorithms (such as ϕ_+^* and its first and second derivatives) are collected in Appendix C.

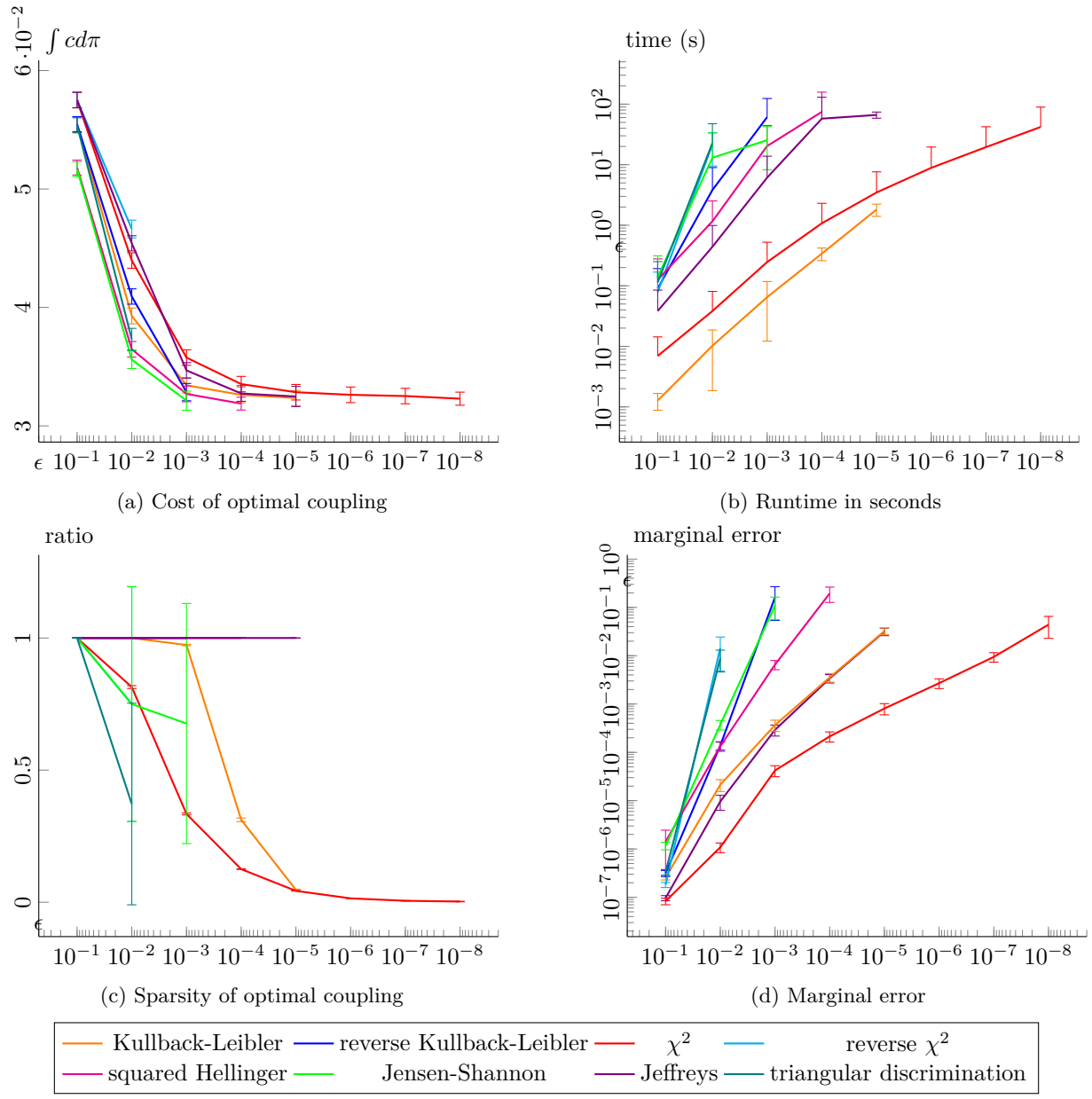
The source code to reproduce the experimental results can be found at `https://github.com/renyi-ai/optimal-transport-with-f-divergence-regularization-and-generalized-sinkhorn-algorithm`. In order to make the experiments more robust, for each one of the four different densities, each f-divergence and a range of ϵs we run the experiments with four different point cloud sizes (500, 1000, 2000 and 5000) and five different random seeds (which determine the point clouds sampled from the densities). In Appendix D we include a detailed account of the data, the hyperparameters and the experimental results. All experiments were run on NVIDIA A100 40GB SXM GPUs.

4.3 Cost of optimal coupling and convergence speed

Entropic regularization introduces a tradeoff between convergence speed of the Sinkhorn algorithm and bias in the optimal coupling. Increasing ϵ leads to faster convergence, but pushes the optimal coupling further away from the coupling which is optimal in the unregularized problem. In Figure 1a and Figure 1b we can see, depending on ϵ, the cost of the coupling obtained (i.e., $\int cd\pi$) as well as the time needed to compute it (in seconds). The values presented correspond to the "crescents" density pair, with means and standard deviations computed over all random seeds and pointcloud sizes[7].

[6] The data is used according to its terms of use, which can be found following the link above.

[7] Different densities lead to markedly different ranges of costs of optimal couplings, which is why we did not average over them. Results for the other 3 pairs of densities can be found in Appendix D. The size of the point clouds sampled from the densities determines the memory requirements but seems to have little effect on convergence speed, which is why we averaged over this hyperparameter.



(a) Cost of optimal coupling

(b) Runtime in seconds

(c) Sparsity of optimal coupling

(d) Marginal error

Figure 1: Experimental results

We eliminated[8] the data of couplings with a marginal error $\sum_{j=1}^{l} |\langle \pi_{\cdot,j}, \mathbf{1}^k \rangle - \nu_j|$ greater than 0.2. In theory the optimal coupling should be (as its name says) a coupling, but for small ϵ the couplings obtained had marginals that differ greatly from its theoretical values μ and ν. In these cases, one needs to set a lower tolerance parameter τ for Algorithm 2 in order to obtain a coupling with negligible marginal error. In Appendix D we include a more detailed account of this issue, and an additional experiment that shows the visual manifestation of the bias by pushing forward one of the point clouds through the transportation map defined by the gradient of $\int cd\pi$.

From the data it seems that the Kullback-Leibler and χ^2 divergences lead to faster convergence compared to the others by a large margin. It is not surprising that Kullback-Leibler is the fastest, since $\gamma_{\phi,\xi}(h)$ is available in closed form for in this case, and no Newton's method iterations are needed. On the other hand, χ^2 allows choosing ϵ from a much larger interval.

[8] We compute the marginal error with respect to ν, because the order of (c, ϵ, ϕ)-transforms makes the marginal error with respect to μ vanish.

4.4 Sparsity of optimal coupling and marginal error

Let us first informally discuss why some divergences lead naturally to sparse optimal solutions. Recall from (3) that when computing the (c, ϵ, ϕ)-transform we are solving the problem of finding some γ_y such that $\int {\phi_+^*}' \circ \frac{1}{\epsilon}(f + \gamma_y - c(\cdot, y))d\mu = 1$. If ϵ starts to decrease, the argument of ${\phi_+^*}'$ increases. Thus, for the integral to be 1 (guaranteed by the theory) and since ${\phi_+^*}'$ is monotonic, $(f + \gamma_y - c(\cdot, y))$ needs to have small values. When ${\phi_+^*}'$ has 0 in its range, such as for the χ^2 and triangular discrimination divergences, this will typically force ${\phi_+^*}'((f + \gamma_y - c(\cdot, y)))$ to be 0. Therefore, by equation (4), the density matrix of the optimal coupling will be sparse.

For the χ^2 divergence, one has ${\phi_+^*}'((-\infty, -2]) = 0$, and for the triangular discrimination divergence, one has ${\phi_+^*}'((-\infty, -3]) = 0$. For all other divergences considered, one always has ${\phi_+^*}' > 0$. Since the density of the optimal coupling is obtained as the image of ${\phi_+^*}'$, this leads to sparse couplings in the former case. In the latter, the optimal couplings are strictly positive. When using $\int cd\pi$ as a loss function, sparsity in the coupling π leads to sparsity in the gradient tensors. This can be useful in practical scenarios as most automatic differentiation engines contain implementations of subroutines tailored for sparse tensors, which can be used in these cases to increase efficiency. An example is when a practitioner uses $\int cd\pi$ as the loss function with μ being a pointcloud output by a neural network and ν being a ground truth point cloud. In this case, if the coupling π is not sparse, any point of μ receives backpropagated gradients from most of the points of ν, whereas if π is sparse, then it only receives gradients from a few of them.

Quantitatively, sparsity in terms of the quotient of positive elements to all elements in the optimal couplings and marginal errors obtained in the experiments are visualized in Figure 1c and Figure 1d. As expected, the χ^2 and triangular discrimination divergences naturally lead to sparse couplings. A consequence of limited machine precision is that the couplings will be empirically sparse even in other cases, notably for the Kullback-Leibler divergence, which reaches the same sparsity as χ^2 at $\epsilon = 10^{-5}$. However, for this ϵ, the marginal error in the Kullback-Leibler case is 40 times larger than for χ^2. For greater values of ϵ (and therefore shorter running time), couplings obtained using χ^2 are more sparse by a large margin.

4.5 Conclusions

The classical setup using the Kullback-Leiber divergence is the fastest to compute and gives low costs in general terms. However, the χ^2 divergence, albeit being marginally slower, can obtain a similar cost but with a much more sparse coupling for values of ϵ corresponding to shorter running times. As we discussed above, the optimal coupling can be used to compute the gradient tensor with respect to the cost and thus a sparse tensor could lead to benefits using subroutines tailored for sparse tensors present in most automatic differentiation engines. Other f-divergences do not seem to induce practical benefits from this limited set of experiments (intended to showcase the feasibility of the generalized Sinkhorn algorithm), but may turn out to be useful in other scenarios.

5 LIMITATIONS

From the theoretical side, the main limitation of our paper is the assumption that the cost function is Lipschitz. We explained in the corresponding section the reasons why we decided to work in this setup. A more general theory may be able to include lower semicontinuous costs, but we did not pursue this in the present work. The Legendre type assumption on ϕ excludes the total variation divergence, but it is necessary in order to have a well-defined (c, ϵ, ϕ)-transform. Another limitation in our work is that while we believe that the Sinkhorn algorithm may fail to converge to optimal variables if no condition like the Good Triple is assumed, we did not present an explicit example of this behavior. Finally, we did not study the theoretical complexity of the generalized Sinkhorn algorithm, Γ-convergence of OT_ϵ to OT_0, explicit formulas of the subdifferential of $\mathrm{OT}_\epsilon(\cdot, \nu)$, nor the generalization of Sinkhorn divergences. We leave these for future projects.

On the practical part, we believe that the implementation of Newton's method could be optimized for each ϕ. For the tolerances, there should be at least a heuristic way of choosing them in terms of ϵ and ϕ in order to have the marginal conditions satisfied at convergence.

Acknowledgements

Dávid Terjék is supported by the Hungarian National Excellence Grant 2018-1.2.1-NKP-00008 and by the Hungarian Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program. Diego González-Sánchez is supported by projects KPP 133921 and Momentum (Lendület) 30003.

The authors would like to thank Mihály Weiner from the Department of Mathematical Analysis at Budapest University of Technology and Economics for his help and in particular for proposing Example 17, as well as the anonymous reviewers for their useful comments.

References

Agrawal, R. and Horel, T. (2020). Optimal bounds between f-divergences and integral probability metrics. *CoRR*, abs/2006.05973.

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B*, 28(1):131–142.

Altschuler, J., Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms foroptimal transport via sinkhorn iteration. *Advances in Neural Information Processing Systems*, pages 1964–1974.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. (2015). Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138.

Blondel, M., Seguy, V., and Rolet, A. (2018). Smooth and sparse optimal transport. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, PMLR 84:880–889.

Borwein, J. M. and Lewis, A. S. (1993). Partially-finite programming in l_1 and the existence of maximum entropy estimates. *SIAM J. Optim.*, 3(2):248–267.

Carlier, G., Duval, V., Peyré, G., and Schmitze, B. (2017). Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418.

Clason, C., Lorenz, D. A., Mahler, H., and Wirth, B. (2019). Entropic regularization of continuous optimal transport problems. *Preprint*.

Cobzaş, Ş., Miculescu, R., and Nicolae, A. (2019). *Lipschitz Functions*. Lecture Notes in Mathematics. Springer International Publishing.

Csiszár, I. (1963). Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizität von markoffschen ketten. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 8(1–2):85–108.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318.

Csiszár, I., Gamboa, F., and Gassiat, E. (1999). MEM pixel correlated solutions for generalized moment and interpolation problems. *IEEE Trans. Inf. Theory*, 45(7):2253–2270.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In Burges, C. J. C., Bottou, L., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2292–2300.

Cuturi, M. and Peyré, G. (2016). A smoothed dual approach for variational wasserstein problem. *SIAM J. Imaging Sci.*, 9(1):320–343.

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

Di Marino, S. and Gerolin, A. (2020a). An optimal transport approach for the schrödinger bridge problem and convergence of sinkhorn algorithm. *J. Sci. Comput.*, 85(2):27.

Di Marino, S. and Gerolin, A. (2020b). Optimal transport losses and sinkhorn algorithm with general convex regularization.

Eckstein, S. and Nutz, M. (2021). Quantitative stability of regularized optimal transport and convergence of sinkhorn's algorithm. abs/2110.06798.

Ferradans, S., Papadakis, N., Peyré, G., and Aujol, J.-F. (2014). Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882.

Feydy, J., Séjourné, T., Vialard, F., Amari, S., Trouvé, A., and Peyré, G. (2019). Interpolating between optimal transport and MMD using sinkhorn divergences. In Chaudhuri, K. and Sugiyama, M., editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 2681–2690. PMLR.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. (2016). Stochastic optimization for large-scaleoptimal transport. *Advances in Neural Information Processing Systems*, pages 3440–3448.

Kantorovich, L. (1939). Mathematical methods in the organization and planning of production. *Leningrad Univ.*

Kantorovich, L. (1940). On an effective method of solving certain classes ofextremal problems. *Dokl. Akad. Nauk. USSR*, 28:212–215.

Kurose, T., Yoshizawa, S., and Amari, S. (2021). Optimal transportation plans with escort entropy regularization. *Info. Geo.*

Lin, T., Ho, N., and Jordan, M. I. (2019). On the efficiency of sinkhorn and greenkhorn and their acceleration for optimal transport. volume abs/1906.01437.

Lorenz, D. A. and Mahler, H. (2020). Orlicz-space regularization for optimal transport and algorithms for quadratic regularization. *Preprint*.

Lorenz, D. A., Manns, P., and Meyer, C. (2019). Quadratically regularized optimal transport. *Preprint*.

Luise, G., Rudi, A., Pontil, M., and Cilibert, C. (2018). Differential properties ofsinkhorn approximation for learning with wasserstein distance. *Advances in Neural Information Processing Systems*, pages 5859–5870.

Luise, G., Salzo, S., Pontil, M., and Cilibert, C. (2019). Sinkhorn barycenters with free support via frank-wolfe algorithm. *Advances in Neural Information Processing Systems*, pages 9318–9329.

Muzellec, B., Nock, R., Patrini, G., and Nielsen, F. (2017). Tsallis regularized optimal transport and ecological inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI'17, pages 2387–2393.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport. *Found. Trends Mach. Learn.*, 11(5-6):355–607.

Rakotomamonjy, A., Flamary, R., and Courty, N. (2015). Generalized conditional gradient: analysis of convergence and applications. *LITIS Lagrange IRISA HAL Id: hal-01217870*.

Roberts, L., Razoumov, L., Su, L., and Wang, Y. (2017). Gini-regularized optimal transport with an application to spatio-temporal forecasting. *CoRR*, abs/1712.02512.

Rubner, Y., Guibas, L., and Tomasi, C. (1997). The earth movers distance, multi-dimensional scaling, and color-based image retrieval. *Proceedings of the ARPA ImageUnderstanding Workshop*, pages 661–668.

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians*. Birkhäuser Basel.

Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math*, 21(2):343–348.

Terjék, D. (2021). Moreau-Yosida f-divergences. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10214–10224. PMLR.

Villani, C. (2008). *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.

Zalinescu, C. (2002). *Convex Analysis in General Vector Spaces*. World Scientific.

Supplementary Material: Optimal transport with f-divergence regularization and generalized Sinkhorn algorithm

A Mathematical background

A.1 Functional analysis

We are going to recite some of the most important results that we need for our paper. We refer the reader to standard reference in the area (Cobzaş et al., 2019) for a detailed account on them.

In this paper, X and Y will denote compact metric spaces except if noted otherwise. We will be interested in studying the set of Lipschitz functions on these sets. We say that a function $f : X \to \mathbb{R}$ is Lipschitz if there exists $K \geq 0$ such that $|f(x) - f(x')| \leq K d_X(x, x')$ for all $x, x' \in X$ where d_X is a metric on X. We will denote the set of Lipschitz functions on X by $\mathrm{Lip}(X)$. The optimal K such that the above holds will be the Lipschitz constant of f, $\|f\|_L = \sup_{x \neq x'} \left\{ \frac{|f(x)-f(x')|}{d_X(x,x')} \right\}$. Clearly this definition depends on the metric d_X (resp. d_Y) chosen on X (resp. Y) but for our purposes we will fix some metric on X (resp. Y) and all the Lipschitz constants will be relative to it. Furthermore, in the product space $X \times Y$ we will assume that we have a metric $d_{X\times Y}$ such that $d_{X\times Y}((x, y), (x, y')) = d_Y(y, y')$ (and similarly fixing y instead of x). For example, it can be assumed for the rest of the paper that $d_{X\times Y}((x, y), (x', y')) = \max\{d_X(x, x'), d_Y(y, y')\}$.

As X and Y are compact spaces, it will always make sense to talk also about the $\|\cdot\|_\infty$ norm of a Lipschitz function on X or Y (and it will always be finite). Thus, we define $\|f\|_\infty := \sup_{x\in X}\{|f(x)|\}$. We can combine the Lipschitz constant with the supremum norm to create the following norm on $\mathrm{Lip}(X)$ (resp. $\mathrm{Lip}(Y)$), $\|f\|_{\max} := \max\{\|f\|_L, \|f\|_\infty\}$.

For any compact metric space (X, d_X) we will denote by $\mathcal{B}(X)$ the Borel σ-algebra on X. A measure on X is a function $\mu : \mathcal{B}(X) \to \mathbb{R}$ such that $\mu(\emptyset) = 0$ and for pairwise disjoint elements $(A_i \in \mathcal{B}(X))_{i\geq 1}$ we have $\mu(\cup_{i=1}^\infty A_i) = \sum_{i=1}^\infty \mu(A_i)$. The variation of a measure μ is defined by:

$$|\mu|(A) := \sup_{A=\cup_{i=1}^n B_i} \sum_{i=1}^n |\mu(B_i)|$$

where the B_i are pairwise disjoint and the supremum is taken over all possible partitions. The total variation of μ is then defined as $\|\mu\| := |\mu|(X)$. We will denote by $\mathcal{M}(X)$ the set of Borel measures on X with finite total variation. Similarly, $\mathcal{M}_+(X)$ will denote the subset of measures $\mu \in \mathcal{M}(X)$ such that $\mu \geq 0$ and $\mathcal{M}(X, \xi)$ for some $\xi \in \mathbb{R}$ will denote the set of measures such that $\mu(X) = \xi$. Finally $P(X)$ will denote the set of probability measures on X, i.e., $\mathcal{M}_+(X) \cap \mathcal{M}(X, 1)$.

Given two measures $\mu, \nu \in \mathcal{M}(X)$ we will say that μ is absolutely continuous with respect to ν and denote it by $\mu \ll \nu$ if for all $A \in \mathcal{B}(X)$ if $\nu(A) = 0$ then $\mu(A) = 0$. In this case, we will denote the Radon-Nikodym derivative as $\frac{d\mu}{d\nu} \in L^1(\nu)$. We will say that two measures $\mu, \nu \in \mathcal{M}(X)$ are singular to each other when there exists $A \in \mathcal{B}(X)$ such that $|\mu|(A) = 0$ and $|\nu|(X \setminus A) = 0$. Given a measure $\mu \in \mathcal{M}(X)$ for a compact metric space X we define its support as $\mathrm{supp}(\mu) := X \setminus (\cup_{\{U\subset X \text{ open and } |\mu|(U)=0\}} U)$. Note that this is always a closed and compact subset of X.

For any $\mu \in \mathcal{M}(X, 0)$ we are interested in defining

$$\|\mu\|_{KR} := \sup\left\{ \int f \, d\mu : \|f\|_L \leq 1 \right\}.$$

With this, we can define the Hanin norm, which will be central in this paper. Given $\mu \in \mathcal{M}(X)$ we define

$$\|\mu\|_H := \inf_{\nu \in \mathcal{M}(X,0)} \{\|\nu\|_{KR} + \|\mu - \nu\|\}.$$

The importance of this norm relies on the following theorem:

Theorem. *Let (X, d_X) be a compact metric space. Then*

$$(\mathrm{Lip}(X), \|\cdot\|_{\max}) \simeq (\mathcal{M}(X), \|\cdot\|_H)^*$$

and furthermore there is an isometric linear isomorphism given by the mapping that sends $f \in \mathrm{Lip}(X)$ to the functional $\mu \mapsto \int f \, d\mu$.

Remark 7. For the rest of the paper, the normed spaces of measures $\mathcal{M}(X), \mathcal{M}(Y)$ and $\mathcal{M}(X \times Y)$ will be assumed to have the Hanin norm and the normed spaces $\mathrm{Lip}(X), \mathrm{Lip}(Y)$ and $\mathrm{Lip}(X \times Y)$ the max norm unless stated otherwise (as for example in (vi) of Proposition 15 where the $\|\cdot\|_\infty$-norm is used).

Given X and Y compact metric spaces and $\pi \in \mathcal{M}(X \times Y)$ let $p_1 : X \times Y \to X$ be the map $(x, y) \mapsto x$. We denote by $p_1^*(\pi) \in \mathcal{M}(X)$ the pushforward measure of π, i.e. $p_1^*(\pi)(A) := \pi(p_1^{-1}(A))$ for any $A \in \mathcal{B}(X)$. We do an analogous definition with with $p_2 : X \times Y \to Y$ $(x, y) \mapsto y$. If we let now $\mu \in P(X)$ and $\nu \in P(Y)$ we can define $\Pi(\mu, \nu) := \{\pi \in \mathcal{M}_+(X \times Y) : p_1^*(\pi) = \mu, p_2^*(\pi) = \nu\}$.

A.2 Convex analysis (Zalinescu, 2002)

Given a topological vector space X, denote its topological dual by X^*, i.e. the set of real-valued continuous linear maps on X, which is a topological vector space itself, and the canonical pairing by $\langle \cdot, \cdot \rangle : X \times X^* \to \mathbb{R}$, which is the continuous bilinear map $(x, x^*) \to \langle x, x^* \rangle = x^*(x)$. Given a function $f : X \to \overline{\mathbb{R}}$, the set $\mathrm{dom}\, f = \{x \in X : f(x) < \infty\}$ is the effective domain of f. A function f is proper if $\mathrm{dom}\, f \neq \emptyset$ and $f(x) > -\infty$ for all $x \in X$, otherwise it is improper. For a convex function $f : X \to \overline{\mathbb{R}}$, its convex conjugate is $f^* : X^* \to \overline{\mathbb{R}}$ defined by $f^*(x^*) = \sup_{x \in X}\{\langle x, x^* \rangle - f(x)\}$, and its subdifferential at $x \in X$ is the set $\partial f(x) = \{x^* \in X^* \mid \forall \hat{x} \in X : \langle \hat{x} - x, x^* \rangle \leq f(\hat{x}) - f(x)\}$, singleton if and only if f is Gateaux differentiable at x.

Remark 8. It should not be confused the pushforward operators p_1^* and p_2^* with the convex conjugate of a convex function (represented also with the symbol *). We believe that this will make no confusion as the only pushforward of a measure will be represented as p_1^* and p_2^*. All the rest of * are convex conjugates.

B Complete proofs

Let us start with an easy result that shows that the pushforward operation of a measure is continuous between the spaces of measures in the Hanin norm:

Proposition 9. *Let X and Y be compact metric spaces. The map $p_1^* : \mathcal{M}(X \times Y) \to \mathcal{M}(X)$ is linear and continuous.*

Proof. Let $p_1 : X \times Y \to X$ be the projection to the first coordinate, $(x, y) \mapsto x$. As $p_1^*(\pi) = \pi \circ p_1^{-1}$ and thus it follows directly that this operator is linear. To see that it is continuous it is enough to check that it is bounded. Let $\pi \in \mathcal{M}(X \times Y)$ be such that $\|\pi\|_H \leq 1$. Let $\nu \in \mathcal{M}(X \times Y, 0)$ be such that $\|\nu\|_{KR} + \|\pi - \nu\| \leq 1 + \epsilon$ for some $\epsilon > 0$. We have to see that $\|p_1^*(\pi)\|_H$ is bounded.

It suffices to see that $\|\nu \circ p_1^{-1}\|_{KR} + \|\pi \circ p_1^{-1} - \nu \circ p_1^{-1}\|$ is bounded by some constant (independent of ν). Clearly we have that $\nu \circ p_1^{-1} \in \mathcal{M}(X, 0)$ as $\nu \circ p_1^{-1}(X) = \nu(X \times Y) = 0$. By definition

$$\|\nu \circ p_1^{-1}\|_{KR} = \sup \left\{ \int f \circ p_1 \, d\nu \ : \ \|f\|_L \leq 1 \right\}.$$

Consider $X \times Y$ with the maximun distance, $d_{X \times Y} = \max(d_X, d_Y)$. It is easy to see that $f \circ p_1 \in \mathrm{Lip}(X \times Y)$ as $|(f \circ p_1)(x, y) - (f \circ p_1)(x', y')| = |f(x) - f(x')| \leq \|f\|_L d_X(x, x') \leq \|f\|_L d_{X \times Y}((x, y), (x'y'))$. If $\|f\|_L \leq 1$ then $\|f \circ p_1\|_L \leq 1$ as well. Therefore $\|\nu \circ p_1^{-1}\|_{KR} \leq \|\nu\|_{KR}$ (as essentially we are taking the supremum over a larger set).

For the total variation part, note that given a partition $A_1, \ldots, A_m$ of X, this automatically gives us a partition of $X \times Y$ induced by p_1^{-1}, namely $A_1 \times Y, \ldots, A_m \times Y$. Therefore $\|\pi \circ p_1^{-1} - \nu \circ p_1^{-1}\| \leq \|\pi - \nu\|$. Thus $\|p_1^*(\pi)\|_H \leq 1+\epsilon$ for all positive ϵ and therefore $\|p_1^*(\pi)\|_H \leq \|\pi\|_H$ and the functional is continuous. □

Clearly a similar argument shows that p_2^* is linear and continuous.

Proposition 10. *If a proper, convex and lower semicontinuous function $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ is of Legendre type, then $\phi_+ = \phi + \iota_{\mathbb{R}_+}$ is strictly convex and differentiable on $\operatorname{dom}\phi_+ = \operatorname{dom}\phi \cap \mathbb{R}_+$, ϕ_+^* is strictly convex and differentiable on $\operatorname{dom}\phi_+^*$, and $(\phi_+')^{-1} = \phi_+^{*\prime}$ on the set $\{t \in \mathbb{R} : \phi_+^{*\prime}(t) > 0\}$.*

Proof. If $\operatorname{dom}\phi \subset \mathbb{R}_+$, the proposition is immediate. Assume the contrary, so that $\phi(0) \in \mathbb{R}$. By definition, for $t \in \mathbb{R}$

$$\phi_+^*(t) = \sup_{s \in \mathbb{R}} \{st - \phi_+(s)\} = \sup_{s \in \mathbb{R}_+} \{st - \phi(s)\}, \tag{5}$$

which is a strictly concave constrained maximization problem. The first derivative test gives

$$t - \phi'(s) = 0, \tag{6}$$

giving the optimum

$$s = \phi^{*\prime}(t) \tag{7}$$

(Borwein and Lewis, 1993, Lemma 2.6). If $\phi^{*\prime}(t) \geq 0$, the constraint is satisfied and the optimum is valid. Otherwise, since the problem is strictly concave, the optimum is going to be $s = 0$, giving

$$\phi_+^*(t) = \begin{cases} \phi^{*\prime}(t)t - \phi(\phi^{*\prime}(t)) = \phi^*(t) & \text{if } \phi^{*\prime}(t) \geq 0, \\ -\phi(0) & \text{otherwise.} \end{cases} \tag{8}$$

Differentiating gives

$$\phi_+^{*\prime}(t) = \begin{cases} \phi^{*\prime}(t) & \text{if } \phi^{*\prime}(t) \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

proving the proposition. □

In the following, we denote by $\langle \mu, f \rangle$ the integral $\int f d\mu$ of $f \in \operatorname{Lip}(X)$ and $\mu \in \mathcal{M}(X)$, since it is exactly the dual pairing for the duality of $(\mathcal{M}(X), \|.\|_H)$ and $(\operatorname{Lip}(X), \|.\|_{\max})$ (and similarly for the spaces Y and $X \times Y$). The mapping $(\mu \to D_\phi(\mu\|\nu))$ is denoted $I_{\phi,\nu}$, the theory of which can be found in the literature (Agrawal and Horel, 2020; Borwein and Lewis, 1993; Terjék, 2021). We begin with a technical result that will help us prove strong duality.

Proposition 11. *Let X and Y be compact metric spaces and $c \in \operatorname{Lip}(X \times Y)$. Let also $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ be a proper, convex and lower semicontinuous function of Legendre type with $\phi(1) = 0$. Let $\epsilon > 0$, $\mu \in P(X)$ and $\nu \in P(Y)$, and define the map $T_{c,\phi,\epsilon,\mu,\nu} : \mathcal{M}(X \times Y) \to \overline{\mathbb{R}}$ as*

$$T_{c,\phi,\epsilon,\mu,\nu}(\pi) = \langle \pi, c \rangle + \epsilon I_{\phi_+,\mu\otimes\nu}(\pi) + \iota_{\{(\mu,\nu)\}}(p_1^*(\pi), p_2^*(\pi)).$$

Then this map is proper, convex, and lower semicontinuous, and its conjugate $T^*_{c,\phi,\epsilon,\mu,\nu} : \operatorname{Lip}(X \times Y) \to \overline{\mathbb{R}}$ is

$$T^*_{c,\phi,\epsilon,\mu,\nu}(\varphi) = \inf_{(f,g) \in \operatorname{Lip}(X) \times \operatorname{Lip}(Y)} \left\{ \epsilon I^*_{\phi_+,\mu\otimes\nu}\left(\frac{1}{\epsilon}(\varphi - c - f \oplus g)\right) + \langle \mu, f \rangle + \langle \nu, g \rangle \right\}.$$

Proof. First we define $\Phi_\varphi : \mathcal{M}(X \times Y) \times (\mathcal{M}(X) \times \mathcal{M}(Y)) \to \overline{\mathbb{R}}$ as

$$\Phi_\varphi(\pi, (\xi, \rho)) = \iota_{\{(\mu,\nu)\}}(p_1^*(\pi) - \xi, p_2^*(\pi) - \rho) + \langle \pi, c - \varphi \rangle + \epsilon I_{\phi_+,\mu\otimes\nu}(\pi).$$

for a function $\varphi \in \operatorname{Lip}(X \times Y)$. Now note that

$$T^*_{c,\phi,\epsilon,\mu,\nu}(\varphi) = \sup_{\pi\in\mathcal{M}(X\times Y)} \{\langle \pi, \varphi\rangle - \epsilon I_{\phi_+,\mu\otimes\nu}(\pi) - \iota_{\{(\mu,\nu)\}}(p_1^*(\pi), p_2^*(\pi)) - \langle \pi, c\rangle\}$$
$$= \sup_{\pi\in\mathcal{M}(X\times Y)} \{-\Phi_\varphi(\pi, 0, 0)\} = -\inf_{\pi\in\mathcal{M}(X\times Y)} \{\Phi_\varphi(\pi, 0, 0)\}\,.$$

Further, suppose that the following convex optimization problem can be solved:

$$-\inf_{\pi\in\mathcal{M}(X\times Y)} \{\Phi_\varphi(\pi, 0, 0)\} = \inf_{(f,g)\in \mathrm{Lip}(X)\times \mathrm{Lip}(Y)} \{\Phi^*_\varphi(0, (-f, -g))\} \tag{10}$$

Then this would imply that

$$T^*_{c,\phi,\epsilon,\mu,\nu}(\varphi) = \inf_{(f,g)\in \mathrm{Lip}(X)\times \mathrm{Lip}(Y)} \{\Phi^*_\varphi(0, (-f, -g))\}.$$

Now, by definition of $\Phi^*_\varphi(0, (-f, -g))$ we have that

$$\Phi^*_\varphi(0, (-f, -g)) = \sup_{\pi,\xi,\rho} \{\langle \xi, -f\rangle + \langle \rho, -g\rangle + \langle \pi, \varphi\rangle - \langle \pi, c\rangle$$
$$- \iota_{\{(\mu,\nu)\}}(p_1^*(\pi) - \xi, p_2^*(\pi) - \rho) - \epsilon I_{\phi_+,\mu\otimes\nu}(\pi)\}.$$

Changing the variables $\eta := p_1^*(\pi) - \xi$ and $\tau := p_2^*(\pi) - \rho$ the previous equation equals:

$$\sup_{\pi,\eta,\tau} \{\langle \varphi - f \oplus g, \pi\rangle + \langle f, \eta\rangle + \langle g, \tau\rangle - \langle \pi, c\rangle - \iota_{\{(\mu,\nu)\}}(\eta, \tau) - \epsilon I_{\phi_+,\mu\otimes\nu}(\pi)\}.$$

It is clear that without loss of generality we can assume that $\eta = \mu$ and $\tau = \nu$ (as otherwise the value inside the supremum is $-\infty$). Hence

$$\Phi^*_\varphi(0, (-f, -g)) = \sup_{\pi} \{\langle \varphi - f \oplus g, \pi\rangle - \epsilon I_{\phi_+,\mu\otimes\nu}(\pi) - \langle \pi, c\rangle\} + \langle f, \mu\rangle + \langle g, \nu\rangle$$
$$= (\epsilon I_{\phi_+,\mu\otimes\nu})^*(\varphi - f \oplus g - c) + \langle f, \mu\rangle + \langle g, \nu\rangle.$$

And this will conclude the proof since

$$T^*_{c,\phi,\epsilon,\mu,\nu}(\varphi) = \inf_{(f,g)\in \mathrm{Lip}(X)\times \mathrm{Lip}(Y)} \{(\epsilon I_{\phi_+,\mu\otimes\nu})^*(\varphi - f \oplus g - c) + \langle f, \mu\rangle + \langle g, \nu\rangle\}$$

and $(\epsilon I_{\phi_+,\mu\otimes\nu})^* = \epsilon I^*_{\phi_+,\mu\otimes\nu}(\frac{1}{\epsilon}\cdot)$ (Zalinescu, 2002, Theorem 2.3.1(v)).

Thus, it only remains to check that (10) holds. To do so, we know that we have strong duality if the marginal function

$$h_\varphi(\xi, \rho) := \inf_{\pi\in\mathcal{M}(X\times Y)} \Phi_\varphi(\pi, \xi, \rho)$$

is lower semicontinuous at the origin and $h_\varphi(0, 0) \in \mathbb{R}$ (Zalinescu, 2002, Theorem 2.6.1(v)). First note that taking $\pi = \mu \otimes \nu$ it is easy to see that the infumum is not ∞. To see that it is not equal to $-\infty$, note that (Agrawal and Horel, 2020, Paragraph before Remark 4.1.4)

$$I_{\phi_+,\mu\otimes\nu}(\pi) \geq 0$$

for any $\pi \in \mathcal{M}(X \times Y)$. If we take now any π such that $\Phi_\varphi(\pi, 0, 0) \neq \infty$ it is clear that $p_1^*(\pi) = \mu$ and π is a positive measure. Thus, we have the bound $\Phi_\varphi(\pi, 0, 0) \geq -\|c - \varphi\|_\infty$ for all $\pi \in \mathcal{M}(X \times Y)$. Hence, we have that $h_\varphi(0, 0) \in \mathbb{R}$.

To prove lower semicontinuity at the origin we have to prove that given $(\xi_n, \rho_n) \in \mathcal{M}(X) \times \mathcal{M}(Y)$ with $(\xi_n, \rho_n) \to (0, 0)$ as $n \to \infty$ then we have that $h_\varphi(0, 0) \leq \liminf_{n\to\infty} h_\varphi(\xi_n, \rho_n)$. Note that for n large enough

we can assume without loss of generality $\max(\|\xi_n\|_H, \|\rho_n\|_H) \leq 1$. Note also that if $\|\xi\|_H \leq 1$ then $\Phi_\varphi(\pi, \xi, \rho)$ is bounded from below. This is because in order to have a value different from ∞ we must have that $\pi \geq 0$ (otherwise $I_{\phi_+, \mu\otimes\nu}(\pi) = \infty$) and also $p_1^*(\pi) = \xi + \mu$. In particular, the total variation of π can be bounded as follows

$$\|\pi\| = \pi(X \times Y) = p_1^*(\pi)(X) = (\xi + \mu)(X) \leq \|\xi\|_H + 1 \leq 2,$$

where in the first equality we have used that $\pi \geq 0$ and for the first inequality we have used Cobzaş et al. (2019, Proposition 8.5.2(ii)). Thus, we have that

$$\Phi_\varphi(\pi, \xi, \rho) \geq \langle c - \varphi, \pi \rangle \geq -\|\varphi - c\|_\infty \|\pi\| \geq -2\|\varphi - c\|_\infty$$

whenever $\|\xi\|_H \leq 1$.

Hence, if $\max(\|\xi_n\|_H, \|\rho_n\|_H) \leq 1$ then $h_\varphi(\xi_n, \rho_n) > -\infty$. It is clear that if $\liminf_{n\to\infty} h_\varphi(\xi_n, \rho_n) = \infty$ then the lower semicontinuity of this sequence is verified. Hence, passing through a subsequence if necessary we can assume that $h_\varphi(\xi_n, \rho_n)$ are all finite and that $\liminf_{n\to\infty} h_\varphi(\xi_n, \rho_n) = \lim_{n\to\infty} h_\varphi(\xi_n, \rho_n)$. Now, for each n let $\pi_n \in \mathcal{M}(X \times Y)$ be such that $|\Phi_\varphi(\pi_n, \xi_n, \rho_n) - h_\varphi(\xi_n, \rho_n)| < 1/n$. Note that without loss of generality we can assume that $\pi_n \geq 0$ (using the same arguments as we used in the previous paragraph). In particular, we have that $\|\pi_n\| \leq 2$ for all n large enough (so that $\|\xi_n\|_H \leq 1$). By Cobzaş et al. (2019, Remark 8.5.9) we have that the set $\{\pi \in \mathcal{M}(X \times Y) : \|\pi\| \leq 2\}$ is compact in the Hanin norm and therefore there exists a convergent subsequence $\pi_n \to \pi$ (that abusing the notation we denote just by n). Hence

$$\begin{aligned}\liminf_{n\to\infty} h_\varphi(\xi_n, \rho_n) &\geq \liminf_{n\to\infty} \Phi_\varphi(\pi_n, \xi_n, \rho_n) - 1/n = \liminf_{n\to\infty} \Phi_\varphi(\pi_n, \xi_n, \rho_n) \\ &\geq \Phi_\varphi(\pi, 0, 0) \geq \inf_{\pi \in \mathcal{M}(X\times Y)} \Phi_\varphi(\pi, 0, 0) = h_\varphi(0, 0).\end{aligned}$$

Where we have used that Φ_φ is lower semicontinuous. To prove this, we just have to prove that it is the sum of lower semicontinuous functions. Clearly $\langle \cdot, c - \varphi \rangle$ is continuous, the indicator function is also lower semi-continuous, and $I_{\phi, \mu\otimes\nu}$ is lower semi-continuous in the Hanin norm (Terjék, 2021, Proposition 7). Using that clearly $(\pi_n, \xi_n, \rho_n) \to (\pi, 0, 0)$ as $n \to \infty$ the result follows. The map $T_{c,\phi,\epsilon,\mu,\nu}$ is easily seen to be proper, convex and lower semicontinuous. □

Let us recall the definition of c-transform (Villani, 2008, Definition 5.2).

Definition 12 (c-transform)**.** Let $f \in \mathrm{Lip}(X)$ and $c \in \mathrm{Lip}(X \times Y)$ for some compact metric spaces X and Y. We define the c-transform of f as follows:

$$f^c(y) := \inf_{x\in X} \{c(x, y) - f(x)\}.$$

Proposition 13. *Let $f \in \mathrm{Lip}(X)$ and $c \in \mathrm{Lip}(X \times Y)$ for some compact metric spaces X and Y. Then the c-transform of f has the following properties:*

(i) If $g \in \mathrm{Lip}(Y)$ is such that $f \oplus g \leq c$ then $g \leq f^c$.

(ii) $f \oplus f^c \leq c$.

(iii) $f^c \in \mathrm{Lip}(Y)$ and $\|f^c\|_L \leq \|c\|_L$.

(iv) $\|f^c\|_\infty \leq \|f\|_\infty + \|c\|_\infty$.

Proof. Most of the properties follow immediatelly from the definitions. For *(iii)*, note that given $y, y' \in Y$ we have

$$\begin{aligned}f^c(y) - f^c(y') &= \inf_{x\in X} \{c(x, y) - f(x)\} - \inf_{x\in X} \{c(x, y') - f(x)\} \\ &\leq \inf_{x\in X} \{c(x, y') - f(x) + \|c\|_L d_Y(y, y')\} - \inf_{x\in X} \{c(x, y) - f(x)\} \\ &\leq \|c\|_L d_Y(y, y'),\end{aligned}$$

where we have assumed that $d_{X\times Y}((x, y), (x, y')) = d_Y(y, y')$. Swapping the roles of y and y' we have the other inequality and therefore $|f^c(y) - f^c(y')| \leq \|c\|_L d_Y(y, y')$. □

The natural generalization of the c-transform to the regularized optimal transport problem is the following.

Definition 14 ((c,ϵ,ϕ)-transform)**.** Let $c \in \mathrm{Lip}(X \times Y)$, $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ a proper, convex and lower semicontinuous function of Legendre type with $\phi(1) = 0$, $\epsilon > 0$, $\mu \in P(X)$ and $\nu \in P(Y)$. We define the (c,ϵ,ϕ)-transform of f as follows:

$$f^{(c,\epsilon,\phi)}(y) := \arg\max_{\gamma \in \mathbb{R}} \left\{ \frac{1}{\epsilon}\gamma - \int \phi_+^* \left(\frac{1}{\epsilon}(f(x) + \gamma - c(x,y)) \right) \, d\mu(x) \right\}.$$

Note that in this definition we can assume that $\gamma \leq f^c(y) + \epsilon\phi'(\infty)$ as otherwise it is clear that the function inside the arg max is going to be $-\infty$ (Borwein and Lewis, 1993, Lemma 2.1).

Let us now prove some properties of the (c,ϵ,ϕ)-transform:

Proposition 15. *Let (X, d_X) and (Y, d_Y) be compact metric spaces. Let also $\mu \in P(X)$ be of full support, i.e. $\mathrm{supp}(\mu) = X$, $c \in \mathrm{Lip}(X \times Y)$ a cost function, $0 < \epsilon \in \mathbb{R}$ a regularization coefficient and $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ a proper, convex and lower semicontinuous function of Legendre type. Then one has that for any $f \in \mathrm{Lip}(X)$:*

(i) $f^{(c,\epsilon,\phi)}(y)$ *is well-defined for all $y \in Y$ implicitly by $\int_X {\phi_+^*}' \circ \frac{1}{\epsilon}(f + f^{(c,\epsilon,\phi)}(y) - c(\cdot, y))d\mu = 1$ if there exists such number $f^{(c,\epsilon,\phi)}(y)$ or explicitly as $f^{(c,\epsilon,\phi)}(y) = \min_{x \in X}\{\epsilon\phi'(\infty) + c(x,y) - f(x)\} = f^c(y) + \epsilon\phi'(\infty)$ otherwise.*

(ii) $f(x) + f^{(c,\epsilon,\phi)}(y) \leq c(x,y) + \epsilon\phi'(\infty)$ *for all $x \in X$ and $y \in Y$.*

(iii) $\|f^{(c,\epsilon,\phi)}\|_L \leq \|c\|_L$.

(iv) $\|f^{(c,\epsilon,\phi)}\|_\infty \leq \|f\|_\infty + \|c\|_\infty$ *if $\phi'(\infty) = \infty$ and $\|f^{(c,\epsilon,\phi)}\|_\infty \leq \|f\|_\infty + \|c\|_\infty + \epsilon\phi'(\infty)$ otherwise.*

(v) For any $a \in \mathbb{R}$ we have $(f + a)^{(c,\epsilon,\phi)} = f^{(c,\epsilon,\phi)} - a$.

(vi) The map from $\mathrm{Lip}(X) \to \mathrm{Lip}(Y)$ that sends a function to its (c,ϵ,ϕ)-transform is 1-Lipschitz with respect to the $\|\cdot\|_\infty$-norm.

Clearly analogous properties hold if we consider the (c,ϵ,ϕ)-transform defined as

$$g^{(c,\epsilon,\phi)}(x) := \arg\max_{\gamma \in \mathbb{R}} \left\{ \frac{1}{\epsilon}\gamma - \int \phi_+^* \left(\frac{1}{\epsilon}(\gamma + g(y) - c(x,y)) \right) \, d\nu(y) \right\}.$$

of a function $g \in \mathrm{Lip}(Y)$.

Proof. Let us start proving (i). Fix any $y \in Y$ and consider the following Primal Problem

$$\inf_{\xi \in \mathcal{M}(X,1)} \left\{ I_{\phi_+,\mu}(\xi) - \int \frac{1}{\epsilon}(f - c(\cdot, y)) \, d\xi \right\} \tag{11}$$

and the corresponding Dual Problem

$$\sup_{\gamma \in \mathbb{R}} \left\{ \frac{1}{\epsilon}\gamma - \int \phi_+^* \left(\frac{1}{\epsilon}(f + \gamma - c(\cdot, y)) \right) \, d\mu \right\}. \tag{12}$$

First, let us verify that the Primal Constraint Qualifications (Primal CQ) and the Dual Constraint Qualifications (Dual CQ) (Borwein and Lewis, 1993, p. 254 and p. 255) are satisfied. To verify the Primal CQ just note that taking $\frac{d\xi}{d\mu} = 1$ this condition holds (i.e. $\xi = \mu$). For the Dual CQ, note that if γ is such that $\gamma < f^c(y) + \epsilon\phi'(\infty)$ then this condition holds as well (as $\phi'_+(-\infty) = -\infty$).

Thus, we get that both the Primal and Dual Problems have (in principle non-necessarily unique) optimal solutions $\widehat{\xi}$ and $\widehat{\gamma}$ respectively (Borwein and Lewis, 1993, Theorem 4.1 (i), (ii) and (iii)). Furthermore, if we decompose $\widehat{\xi} = \frac{d\widehat{\xi}_c}{d\mu}\mu + (\widehat{\xi}_s)_+ - (\widehat{\xi}_s)_-$ where $\widehat{\xi}_c$ is the absolutely continuous part with respect to μ and $(\widehat{\xi}_s)_+$

and $(\widehat{\xi}_s)_-$ is the Jordan decomposition of the singular part we have that $\frac{d\widehat{\xi}_c}{d\mu}$ is uniquely defined μ-a.e. and $\frac{d\widehat{\xi}_c}{d\mu} = {\phi_+^*}'(\frac{1}{\epsilon}(f+\widehat{\gamma}-c(\cdot,y)))$.

Now suppose that we have two optimal $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ and that the absolutely continuous part is nonzero. By uniqueness of the absolutely continuous part we have that ${\phi_+^*}'(\frac{1}{\epsilon}(f+\widehat{\gamma}_1-c(\cdot,y))) = {\phi_+^*}'(\frac{1}{\epsilon}(f+\widehat{\gamma}_2-c(\cdot,y)))$ μ-a.e. First note that $\int_X {\phi_+^*}'(\frac{1}{\epsilon}(f+\widehat{\gamma}_1-c(\cdot,y)))d\mu > 0$ as otherwise the absolutely continuous part would be 0. Thus, there exists an open set $U_y \subset X$ of positive measure such that ${\phi_+^*}'(\frac{1}{\epsilon}(f(x)+\widehat{\gamma}_1-c(x,y))) > 0$ for all $x \in U_y$. Without loss of generality we can assume that $\widehat{\gamma}_2 \geq \widehat{\gamma}_1$ (otherwise swap the roles of $\widehat{\gamma}_2$ and $\widehat{\gamma}_1$) so in particular this inequality holds as well for every $x \in U_y$ replacing $\widehat{\gamma}_1$ with $\widehat{\gamma}_2$. By Proposition 10, ${\phi_+^*}'$ is invertible in U_y and thus we have that $\frac{1}{\epsilon}(f+\widehat{\gamma}_1-c(\cdot,y)) = \frac{1}{\epsilon}(f+\widehat{\gamma}_2-c(\cdot,y))$ for U_y-a.e. (if we want to be very precise, this would be with the restriction of μ to U_y) and this clearly shows that $\widehat{\gamma}_1 = \widehat{\gamma}_2$. In particular, this unique value is precisely $f^{(c,\epsilon,\phi)}(y)$. For simplicity and smoothness of the notation we will denote $f^{(c,\epsilon,\phi)}(y) = \gamma_y$.

We have that $\text{supp}((\widehat{\xi}_s)_-) \subset \{\frac{1}{\epsilon}(f+\gamma_y-c(\cdot,y)) = \phi'_+(-\infty) = -\infty\} = \emptyset$ (Borwein and Lewis, 1993, Corollary 3.6) and, in particular $(\widehat{\xi}_s)_- = 0$, and that $\text{supp}((\widehat{\xi}_s)_+) \subset \{\frac{1}{\epsilon}(f+\gamma_y-c(\cdot,y)) = \phi'_+(\infty)\}$. As ϕ is of Legendre type ${\phi_+^*}'$ is always nonnegative and increasing by Proposition 10. In particular $\widehat{\xi}$ is a probability measure. If $\phi'(\infty) = \infty$ or $\int_X {\phi_+^*}'(\frac{1}{\epsilon}(f+\widehat{\gamma}_y-c(\cdot,y))) = 1$ then we have no singular part but if $\int \frac{d\widehat{\xi}_c}{d\mu}\, d\mu < 1$ then there must exists some $x \in X$ such that $f(x)+\gamma_y-c(x,y) = \epsilon\phi'(\infty)$. If we assume that (ii) of this proposition holds and $f(x)+f^{(c,\epsilon,\phi)}(y) \leq c(x,y)+\epsilon\phi'(\infty)$ for all $x \in X$ and $y \in Y$, we have that in this case $f^{(c,\epsilon,\phi)}(y) = \gamma_y = f^c(y)+\epsilon\phi'(\infty)$, in particular uniqueness holds even if the absolutely continuous part is 0.

Let us prove (ii) now. First assume that we have not proved the uniqueness part of (i) yet. Suppose by contradiction that $f(x_0)+\gamma_{y_0} > c(x_0,y_0)+\epsilon\phi'(\infty)$ for some $x_0 \in X$ and $y_0 \in Y$ where $\gamma_{y_0} = \hat{\gamma}_{y_0}$ is an optimal solution of the Dual Problem. Let us now define $U_{y_0} := \{x \in X : f(x)+\gamma_{y_0} > c(x,y_0)+\epsilon\phi'(\infty)\}$ which by hypothesis is a non-empty open set. Now we use the assumption that $\text{supp}(\mu) = X$ to see that in this case $\frac{1}{\epsilon}\gamma_{y_0} - \int_X \phi_+^*(\frac{1}{\epsilon}(f+\gamma_{y_0}-c))\, d\mu$ equals

$$\frac{1}{\epsilon}\gamma_{y_0} - \int_{X\setminus U_{y_0}} \phi_+^*\left(\frac{1}{\epsilon}(f+\gamma_{y_0}-c)\right)\, d\mu - \int_{U_{y_0}} \phi_+^*\left(\frac{1}{\epsilon}(f+\gamma_{y_0}-c)\right)\, d\mu.$$

But $\mu(U_{y_0}) > 0$ and ϕ_+^* equals ∞ in that set. As the other part of the integral is always bounded from below, we get that $\frac{1}{\epsilon}\gamma_{y_0} - \int_X \phi_+^*(\frac{1}{\epsilon}(f+\gamma_{y_0}-c))\, d\mu = -\infty$ but this is impossible as we know that the Dual Problem has a solution strictly larger than $-\infty$. Hence, the uniqueness part of (i) holds and therefore as γ_y is by definition $f^{(c,\epsilon,\phi)}(y)$ we conclude (ii).

We continue now by proving (iii). Let us define $\beta^y(\gamma) := \int_X {\phi_+^*}'(\frac{1}{\epsilon}(f+\gamma-c(\cdot,y)))\, d\mu$ for any $\gamma \in \mathbb{R}$. As we saw before, either $\beta^y(\gamma_y) = 1$ or $\beta^y(\gamma_y) < 1$ and $\gamma_y = f^c(y)+\epsilon\phi'(\infty)$. To prove that $f^{(c,\epsilon,\phi)}$ defined pointwise by γ_y is Lipschitz, given $y, y' \in Y$ first suppose that $\beta^{y'}(\gamma_{y'}) \geq \beta^y(\gamma_y)$. Then, as ${\phi_+^*}'$ is an increasing function so is $\beta^y(\gamma)$ as a function of γ. Thus

$$\begin{aligned}\beta^{y'}(\gamma_{y'}) \geq \beta^y(\gamma_y) &\geq \int_X {\phi_+^*}'\left(\frac{1}{\epsilon}(f(x)+\gamma-c(x,y')-\|c\|_L d_Y(y,y'))\right)\, d\mu \\ &= \beta^{y'}(\gamma_y - \|c\|_L d_Y(y,y')).\end{aligned}$$

Hence, $\gamma_y - \gamma_{y'} \leq \|c\|_L d_Y(y,y')$.

If $\beta^{y'}(\gamma_{y'}) = \beta^y(\gamma_y) = 1$ then we are done, as we can repeat the above argument switching the roles of y and y'. Similarly, if $\beta^{y'}(\gamma_{y'}) < 1$ and $\beta^y(\gamma_y) < 1$, using the fact that in this case the transform is just a translate of the regular c-transform we get the result. The only case left is what happens if (say) $\beta^{y'}(\gamma_{y'}) = 1$ and $\beta^y(\gamma_y) < 1$. By the previous argument we already know that $\gamma_y - \gamma_{y'} \leq \|c\|_L d_Y(y,y')$. For the other inequality, note that as $f^c(y') - f^c(y) \leq \|c\|_L d_Y(y,y')$ but we also know that $\gamma_y = f^c(y)+\epsilon\phi'(\infty)$ and $\gamma_{y'} \leq f^c(y')+\epsilon\phi'(\infty)$. Plugging this into the previous inequality the result follows.

Let us now prove (iv). Again we have to divide into two cases. Given $y \in Y$, if $\beta^y(\gamma_y) = 1$ using that $\phi(1) = 0$ we know that ${\phi_+^*}'(0) = 1$ and as ${\phi_+^*}'$ is increasing and nonnegative we have that $\sup_{x\in X}\{\frac{1}{\epsilon}(f+\gamma_y-c)\} \geq 0$

(as otherwise $\beta^y(\gamma_y)$ would be strictly smaller than 1). From this it is easy to see that $\gamma_y \geq -\|f\|_\infty - \|c\|_\infty$. An analogous argument shows that $\gamma_y \leq \|f\|_\infty + \|c\|_\infty$. If $\gamma_y = f^c(y) + \epsilon\phi'(\infty)$ then we use the bound $\|f^c\|_\infty \leq \|f\|_\infty + \|c\|_\infty$ and the result follows.

Part (v) follows directly from the definitions.

To prove the last part, let $f_1, f_2 \in \mathrm{Lip}(X)$. We want to prove that if $\|f_1 - f_2\|_\infty \leq L$ then $\|f_1^{(c,\epsilon,\phi)} - f_2^{(c,\epsilon,\phi)}\| \leq L$. We have to consider 3 different cases. Fix any $y \in Y$. First assume that both $f_i^{(c,\epsilon,\phi)}(y)$ for $i = 1, 2$ are calculated by the formula $f_i^{(c,\epsilon,\phi)}(y) = \min_{x\in X}\{\epsilon\phi'(\infty) + c(x, y) - f_i(x)\}$. If we use that $-f_1(x) \geq -f_2(x) - L$ for all $x \in X$ we have that $f_1^{(c,\epsilon,\phi)}(y) \geq \min_{x\in X}\{\epsilon\phi'(\infty) + c(x, y) - f_2(x) - L\} = f_2^{(c,\epsilon,\phi)}(y) - L$. Using the inequality $-f_1(x) \leq -f_2(x) + L$ we obtain the converse inequality and we are done in this case.

Next, assume that for $i = 1, 2$, the value of $f_i^{(c,\epsilon,\phi)}(y)$ is given implicitly as the unique value such that $\int_X {\phi_+^*}' \circ \frac{1}{\epsilon}(f_i + f_i^{(c,\epsilon,\phi)}(y) - c(\cdot, y))d\mu = 1$. Then we would have that for example $1 = \int_X {\phi_+^*}' \circ \frac{1}{\epsilon}(f_1 + f_1^{(c,\epsilon,\phi)}(y) - c(\cdot, y))d\mu \leq \int_X {\phi_+^*}' \circ \frac{1}{\epsilon}(f_2 + L + f_1^{(c,\epsilon,\phi)}(y) - c(\cdot, y))d\mu$. As the function β^y as we defined it before is increasing, we must have[9] that by definition $f_2^{(c,\epsilon,\phi)}(y) \leq f_1^{(c,\epsilon,\phi)}(y) + L$. By an analogous argument but using that $f_1(x) \geq f_2(x) - L$ for all $x \in X$ we have the opposite inequality.

Finally, in the mixed case when (say) $f_1^{(c,\epsilon,\phi)}(y)$ is given explicitly and $f_2^{(c,\epsilon,\phi)}(y)$ is implicit, we have to combine the previous arguments to conclude our result. On the one hand, $f_1^{(c,\epsilon,\phi)}(y) \geq \min_{x\in X}\{\epsilon\phi'(\infty) + c(x, y) - f_2(x) - L\} \geq f_2^{(c,\epsilon,\phi)}(y) - L$ (as we always have the inequality $f_2^{(c,\epsilon,\phi)}(y) \leq f_2^c(y) + \epsilon\phi'(\infty)$. For the other inequality note that $1 \geq \int_X {\phi_+^*}' \circ \frac{1}{\epsilon}(f_1 + f_1^{(c,\epsilon,\phi)}(y) - c(\cdot, y))d\mu$ always (because $\widehat{\xi}$ is always a probability measure). Then we use the inequality $f_1(x) \geq f_2(x) - L$ which give us at the end that $1 \geq \int_X {\phi_+^*}' \circ \frac{1}{\epsilon}(f_2 - L + f_1^{(c,\epsilon,\phi)}(y) - c(\cdot, y))d\mu$. Similarly as before, this implies that $f_2^{(c,\epsilon,\phi)}(y) \geq f_1^{(c,\epsilon,\phi)}(y) - L$. □

Remark 16. Note that in some cases the (c, ϵ, ϕ)-transform collapses to "almost" the c-transform.

Example 17. *Consider the following example. Let $X = Y = [0, 1]$ with the measure $d\mu = 2x\,dx$ (where dx is the usual Lebesgue measure). Let also $\epsilon = 1$, the cost function $c(x, y) = 3x - 1$ and $f(x) = 0$. Let also D_ϕ be the reverse Kullback-Leibler divergence (see Section C.2). Then $f^{(c,\epsilon,\phi)}(y) = f^c(y) + \epsilon\phi'(\infty) = 0$ for all $y \in Y$. To prove this, note that we just have to compute $\int \phi_+^*(\frac{1}{\epsilon}(f + \gamma - c)d\mu = \int_0^1 \frac{2x}{3x-\gamma}dx = \frac{2}{3} + \frac{2}{9}\gamma(\log(3 - \gamma) - \log(-\gamma))$. From here it is easy to check that there is no $\gamma \leq 0$ such that the previous integral equals 1. Thus, the (c, ϵ, ϕ)-transform of f collapses to $f^c(y) + \epsilon\phi'(\infty)$ for all $y \in Y$.*

Theorem 18. *Let $\mu \in P(X)$ and $\nu \in P(Y)$ be probability measures of full support on compact metric spaces (X, d_X) and (Y, d_Y). Let $c \in \mathrm{Lip}(X \times Y)$, $0 < \epsilon \in \mathbb{R}$ be a regularization coefficient and $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ a proper, convex and lower semicontinuous function of Legendre type. Then one has*

$$\begin{aligned}
&\min_{\pi\in\Pi(\mu,\nu)}\{\langle\pi, c\rangle + \epsilon D_\phi(\pi\|\mu\otimes\nu)\}\\
&= \max_{\substack{f\in\mathrm{Lip}(X), g\in\mathrm{Lip}(Y)\\ f\oplus g\leq c+\epsilon\phi'(\infty)}}\{\langle\mu\otimes\nu, f\oplus g\rangle - \epsilon\langle\mu\otimes\nu, \phi_+^*\circ\frac{1}{\epsilon}(f\oplus g - c)\rangle\}\\
&= \max_{f\in\mathrm{Lip}(X)}\{\langle\mu\otimes\nu, f\oplus f^{(c,\epsilon,\phi)}\rangle - \epsilon\langle\mu\otimes\nu, \phi_+^*\circ\frac{1}{\epsilon}(f\oplus f^{(c,\epsilon,\phi)} - c)\rangle\}\\
&= \max_{g\in\mathrm{Lip}(Y)}\{\langle\mu\otimes\nu, g^{(c,\epsilon,\phi)}\oplus g\rangle - \epsilon\langle\mu\otimes\nu, \phi_+^*\circ\frac{1}{\epsilon}(g^{(c,\epsilon,\phi)}\oplus g - c)\rangle\},
\end{aligned}$$

and $\pi_* \in \Pi(\mu, \nu)$ is optimal in the primal problem if and only if there exists $(f_*, g_*) \in \mathrm{Lip}(X) \times \mathrm{Lip}(Y)$ such that

$$\frac{1}{\epsilon}(f_* \oplus g_* - c) \leq \phi'(\infty), \tag{13}$$

$$\frac{d\pi_c}{d\mu\otimes\nu} = {\phi_+^*}' \circ \frac{1}{\epsilon}(f_* \oplus g_* - c) \tag{14}$$

[9] Note that in principle these integrals are only well-defined if the argument of ${\phi_+^*}'$ is less or equal than $\phi'(\infty)$. However, we can assume that the value of ${\phi_+^*}'$ is ∞ for values larger than $\phi'(\infty)$ as this will be consistent with the definition of the (c, ϵ, ϕ)-transform given in (i).

and

$$\operatorname{supp}(\pi_s) \subset \{(x,y) \in X \times Y : \frac{1}{\epsilon}(f_*(x) + g_*(y) - c(x,y)) = \phi'(\infty)\} \tag{15}$$

hold. In this case, (f_*, g_*) are a pair of optimal potentials in the dual problem.

Proof. Since $\inf_{x\in X}\{f(x)\} = -f^*(0)$ for any proper, convex and lower semicontinuous function f, by Proposition 11 one has

$$\begin{aligned}\inf_{\pi\in\mathcal{M}(X\times Y)} &\{\langle \pi, c\rangle + \epsilon I_{\phi_+,\mu\otimes\nu}(\pi) + \iota_{\{(\mu,\nu)\}}(\pi(\cdot\times Y), \pi(X\times\cdot))\} \\ &= -\inf_{(f,g)\in \mathrm{Lip}(X)\times\mathrm{Lip}(Y)} \left\{\epsilon I^*_{\phi_+,\mu\otimes\nu}\left(\frac{1}{\epsilon}(-c - f\oplus g)\right) + \langle\mu, f\rangle + \langle\nu, g\rangle\right\},\end{aligned}$$

or equivalently

$$\sup_{(f,g)\in \mathrm{Lip}(X)\times\mathrm{Lip}(Y)} \left\{\langle \mu\otimes\nu, f\oplus g\rangle - \epsilon I^*_{\phi_+,\mu\otimes\nu}\left(\frac{1}{\epsilon}(f\oplus g - c)\right)\right\}.$$

Since $I^*_{\phi,\mu\otimes\nu}(\varphi) = \infty$ unless $\varphi(X) \subseteq [\phi'(-\infty), \phi'(\infty)]$ (Terjék, 2021, Proposition 7), $\phi'_+(-\infty) = -\infty$ and $\phi'_+(\infty) = \phi'(\infty)$, one has the constraint $\frac{1}{\epsilon}(f\oplus g - c) \le \phi'(\infty)$, leading to

$$\sup_{f\oplus g\le c+\epsilon\phi'(\infty)} \left\{\langle \mu\otimes\nu, f\oplus g\rangle - \epsilon\left\langle \mu\otimes\nu, \phi^*_+\circ\frac{1}{\epsilon}(f\oplus g - c)\right\rangle\right\}.$$

By definition of the (c,ϵ,ϕ)-transform, it is clear that

$$g - \epsilon\int \phi^*_+\circ\frac{1}{\epsilon}(f\oplus g - c)\, d\mu \le f^{(c,\epsilon,\phi)} - \epsilon\int \phi^*_+\circ\frac{1}{\epsilon}(f\oplus f^{(c,\epsilon,\phi)} - c)\, d\mu \tag{16}$$

for every $y\in Y$. Thus we can always replace g by $f^{(c,\epsilon,\phi)}$. A similar argument shows that we can always replace f by $g^{(c,\epsilon,\phi)}$.

Let us now check that both the supremum and the infimum are attained. Let us start with the infimum. Let $\pi_n \in \Pi(\mu,\nu)$ be such that $\langle\pi_n, c\rangle + \epsilon I_{\phi,\mu\otimes\nu}(\pi_n) \to \inf_{\pi\in\Pi(\mu,\nu)}\{\langle\pi, c\rangle + \epsilon I_{\phi,\mu\otimes\nu}(\pi)\}$ as $n\to\infty$. As the set of probability measures is a compact set in the Hanin norm (Cobzaş et al., 2019, Theorem 8.4.25(3),Theorem 8.5.7) and any coupling is a probability measure, we can assume that there is a convergent subsequence (that abusing the notation we denote by π_n) such that $\pi_n \to \pi^*$ in the Hanin norm. Moreover, as p_1^* and p_2^* are continuous functions we know that $\pi^*\in\Pi(\mu,\nu)$. And finally note that as the function $\langle\cdot, c\rangle + \epsilon I_{\phi,\mu\otimes\nu}(\cdot)$ is lower semicontinuous we have that $\langle\pi^*, c\rangle + \epsilon I_{\phi,\mu\otimes\nu}(\pi^*) \le \lim_{n\to\infty}\langle\pi_n, c\rangle + \epsilon I_{\phi,\mu\otimes\nu}(\pi_n) = \inf_{\pi\in\Pi(\mu,\nu)}\{\langle\pi, c\rangle + \epsilon I_{\phi,\mu\otimes\nu}(\pi)\}$, so that the maximum is achieved by π^*.

As for the supremum, we want to prove that

$$S := \sup_{f\oplus g\le c+\epsilon\phi'(\infty)} \left\{\langle \mu\otimes\nu, f\oplus g\rangle - \epsilon\left\langle \mu\otimes\nu, \phi^*_+\circ\frac{1}{\epsilon}(f\oplus g - c)\right\rangle\right\}$$

is attained for some pair of functions $(f,g)\in\mathrm{Lip}(X)\times\mathrm{Lip}(Y)$. Let $(f_n, g_n)\in\mathrm{Lip}(X)\times\mathrm{Lip}(Y)$ be a sequence of functions such that $f_n\oplus g_n \le c + \epsilon\phi'(\infty)$ and $\left|S - \langle\mu, f_n\rangle - \langle\nu, g_n\rangle + \epsilon\left\langle\mu\otimes\nu, \phi^*_+\circ\frac{1}{\epsilon}(f_n\oplus g_n - c)\right\rangle\right| \le 1/n$. First note that by (16) we can replace g_n by $f_n^{(c,\epsilon,\phi)}$ and we are still at most $1/n$ away from S. As Y is compact and metric, it has finite diameter, $\operatorname{diam}(Y) = \sup_{y,y'\in Y} d_Y(y,y') < \infty$. By (iii) of Proposition 15, the Lipschitz constant of $f_n^{(c,\epsilon,\phi)}$ is bounded by $\|c\|_L$ for all $n\ge 0$. Moreover, note that we can replace the pair $(f_n, f_n^{(c,\epsilon,\phi)})$ by $(f_n + a, f_n^{(c,\epsilon,\phi)} - a)$ for any constant $a\in\mathbb{R}$. Thus, taking $a = f_n^{(c,\epsilon,\phi)}(y_0)$ for some $y_0\in Y$ we have that $f_n^{(c,\epsilon,\phi)} - f_n^{(c,\epsilon,\phi)}(y_0)$ is a function with Lipschitz constant at most $\|c\|_L$ and $|f_n^{(c,\epsilon,\phi)} - f_n^{(c,\epsilon,\phi)}(y_0)| \le \|c\|_L d_Y(y,y') \le \|c\|_L \operatorname{diam}(Y)$.

Now, again we use (16) and instead of the pair $(f_n + f_n^{(c,\epsilon,\phi)}(y_0), f_n^{(c,\epsilon,\phi)} - f_n^{(c,\epsilon,\phi)}(y_0))$ we take $((f_n^{(c,\epsilon,\phi)} - f_n^{(c,\epsilon,\phi)}(y_0))^{(c,\epsilon,\phi)}, f_n^{(c,\epsilon,\phi)} - f_n^{(c,\epsilon,\phi)}(y_0))$. By Proposition 15 we know that the Lipschitz constant of $(f_n^{(c,\epsilon,\phi)} - f_n^{(c,\epsilon,\phi)}(y_0))^{(c,\epsilon,\phi)}$ is at most $\|c\|_L$ and that $\|(f_n^{(c,\epsilon,\phi)} - f_n^{(c,\epsilon,\phi)}(y_0))^{(c,\epsilon,\phi)}\|_\infty \le \|c\|_\infty + \|c\|_L \operatorname{diam}(Y)$. Thus,

if we denote by $h_n := f_n^{(c,\epsilon,\phi)} - f_n^{(c,\epsilon,\phi)}(y_0)$ we have that $|S - \langle \mu, h_n^{(c,\epsilon,\phi)} \rangle - \langle \nu, h_n \rangle| \leq 1/n$ and $\|h_n\|_{\max} = \max\{\|h_n\|_\infty, \|h_n\|_L\} \leq (\operatorname{diam}(Y)+1)\|c\|_L$. Similarly we get that $\|h_n^{(c,\epsilon,\phi)}\|_{\max} = \max\{\|h_n^{(c,\epsilon,\phi)}\|_\infty, \|h_n^{(c,\epsilon,\phi)}\|_L\} \leq (\operatorname{diam}(Y)+1)\|c\|_L + \|c\|_\infty$. The key fact now is that these constants do not depend on n, and therefore, as $(\operatorname{Lip}(X), \|\cdot\|_{\max})$ is the dual of a normed space (namely $(\mathcal{M}(X), \|\cdot\|_H)$), by the Banach-Alaoglu theorem we know that the unit ball is compact in the weak* topology. Thus, we can assume (passing to a subsequence if necessary) that $h_n \to h$ and $h_n^{(c,\epsilon,\phi)} \to h'$ in the weak* topology. Using the fact that $I^*_{\phi_+,\mu\otimes\nu}$ is weak* lower semicontinuous (Zalinescu, 2002, Theorem 2.3.1) this implies that

$$S = \langle \mu \otimes \nu, h' \oplus h \rangle - \epsilon \left\langle \mu \otimes \nu, \phi_+^* \circ \frac{1}{\epsilon}(h' \oplus h - c) \right\rangle$$

and similarly changing h' by $h^{(c,\epsilon,\phi)}$ or h by $h'^{(c,\epsilon,\phi)}$. Note that $h'(x)+h(y) \leq c(x,y)+\epsilon\phi'(\infty)$ for all $(x,y) \in X \times Y$ as otherwise the right hand side of the previous equation will be $-\infty$.

If π is optimal and (f,g) are optimal potentials then

$$\langle \pi, c \rangle + \epsilon I_{\phi_+,\mu\otimes\nu}(\pi) = \langle \pi, f \oplus g \rangle - \epsilon I^*_{\phi_+,\mu\otimes\nu}\left(\frac{1}{\epsilon}(f \oplus g - c)\right),$$

or equivalently

$$\left\langle \pi, \frac{1}{\epsilon}(f \oplus g - c) \right\rangle = I_{\phi_+,\mu\otimes\nu}(\pi) + I^*_{\phi_+,\mu\otimes\nu}\left(\frac{1}{\epsilon}(f \oplus g - c)\right).$$

The optimality conditions then follow Borwein and Lewis (1993, Theorem 2.10). □

We can say even a little more about the structure of the optimal potentials and coupling. A set $C \subset X \times Y$ is called c-cyclically monotone (Villani, 2008, Definition 5.1) if for any subset $\{(x_1,y_1),\dots,(x_n,y_n)\} \subset C$ for $n \in \mathbb{N}$, one has

$$\sum_{i=1}^n c(x_i,y_i) \leq \sum_{i=1}^{n-1} c(x_i,y_{i+1}) + c(x_n,y_1). \tag{17}$$

Proposition 19. *The (c,ϵ,ϕ)-subdifferential of $f \in \operatorname{Lip}(X)$ defined as*

$$\partial_{(c,\epsilon,\phi)} f = \{(x,y) \in X \times Y : f(x) + f^{(c,\epsilon,\phi)}(y) = c(x,y) + \epsilon\phi'(\infty)\}, \tag{18}$$

and the (c,ϵ,ϕ)-subdifferential of $g \in \operatorname{Lip}(Y)$ defined as

$$\partial_{(c,\epsilon,\phi)} g = \{(x,y) \in X \times Y : g(y) + g^{(c,\epsilon,\phi)}(x) = c(x,y) + \epsilon\phi'(\infty)\} \tag{19}$$

are both closed, c-cyclically monotone sets.

Proof. If $\phi'(\infty) = \infty$, then $\partial_{(c,\epsilon,\phi)} f = \partial_{(c,\epsilon,\phi)} g = \emptyset$, so the statement is vacuously true. Now assume that $\phi'(\infty) \in \mathbb{R}$. Being the level sets of Lipschitz continuous functions implies that both sets are closed. Let $\{(x_1,y_1),\dots,(x_n,y_n)\} \subset \partial_{(c,\epsilon,\phi)} f$, so that one has

$$\sum_{i=1}^n c(x_i,y_i) = \sum_{i=1}^n [f(x_i) + f^{(c,\epsilon,\phi)}(y_i) - \epsilon\phi'(\infty)]. \tag{20}$$

On the other hand, one always has $c(x_i,y_j) + \epsilon\phi'(\infty) \geq f(x_i) + f^{(c,\epsilon,\phi)}(y_j)$, implying that

$$\begin{aligned}\sum_{i=1}^{n-1} c(x_i,y_{i+1}) + c(x_n,y_1) &\geq \sum_{i=1}^{n-1} [f(x_i) + f^{(c,\epsilon,\phi)}(y_{i+1}) - \epsilon\phi'(\infty)] + f(x_n) + f^{(c,\epsilon,\phi)}(y_1) - \epsilon\phi'(\infty) \\ &= \sum_{i=1}^n [f(x_i) + f^{(c,\epsilon,\phi)}(y_i) - \epsilon\phi'(\infty)]. \end{aligned} \tag{21}$$

The last two equations imply the proposition for $\partial_{(c,\epsilon,\phi)} f$, and a symmetric argument clearly works for $\partial_{(c,\epsilon,\phi)} g$. □

Proposition 20. *Let $\mu \in P(X)$ and $\nu \in P(Y)$ be probability measures of full support on compact metric spaces (X, d_X) and (Y, d_Y). Let $c \in \mathrm{Lip}(X \times Y)$, $0 < \epsilon \in \mathbb{R}$ be a regularization coefficient and $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ a proper, convex and lower semicontinuous function of Legendre type. Let $\pi' \in \Pi(\mu, \nu)$ be an optimal coupling for the primal problem. Let π'_c be its absolutely continuous part with respect to $\mu \otimes \nu$ and π'_s the singular part. Let also $(f', g') \in \mathrm{Lip}(X) \times \mathrm{Lip}(Y)$ be a pair of optimal potentials. Then $\frac{d\pi'_c}{d\mu \otimes \nu}$ is unique for any optimal coupling. If $(\tilde{f}, \tilde{g}) \in \mathrm{Lip}(X) \times \mathrm{Lip}(Y)$ are also optimal potentials then $f' \oplus g' = \tilde{f} \oplus \tilde{g}$ π'_c-a.e.. If ${\phi^*_+}'$ is invertible in $(-\infty, \phi'(\infty))$ then any optimal potential equals $(f' + a, g' - a)$ for some $a \in \mathbb{R}$. Finally, the support of π_s lies in the intersection of the (c, ϵ, ϕ)-subdifferentials of all optimal dual variables.*

Proof. Let $\pi^1, \pi^2 \in \Pi(\mu, \nu)$ and $(f^1, g^1), (f^2, g^2) \in \mathrm{Lip}(X) \times \mathrm{Lip}(Y)$ be optimal primal and dual variables. If $g^j = {f^j}^{(c,\epsilon,\phi)}$ and $f^j = {g^j}^{(c,\epsilon,\phi)}$ would not hold for $j \in \{1, 2\}$, one could replace g^j with ${f^j}^{(c,\epsilon,\phi)}$ to increase the value of the dual problem, contradicting optimality of (g^j, f^j). By optimality, one has

$$\int c d\pi^i + \epsilon I_{\phi_+, \mu\otimes\nu}(\pi^i) = \int f^j \oplus g^j d\mu \otimes \nu - \epsilon I^*_{\phi_+, \mu\otimes\nu}\left(\frac{1}{\epsilon}(f^j \oplus g^j - c)\right) \tag{22}$$

for $i, j \in \{1, 2\}$. Since $\pi^i \in \Pi(\mu, \nu)$, one has $\int f^j \oplus g^j d\mu \otimes \nu = \int f^j \oplus g^j d\pi^i$, so we can rearrange as

$$I_{\phi_+, \mu\otimes\nu}(\pi^i) + I^*_{\phi_+, \mu\otimes\nu}\left(\frac{1}{\epsilon}(f^j \oplus g^j - c)\right) = \int \frac{1}{\epsilon}(f^j \oplus g^j - c) d\pi^i. \tag{23}$$

By Borwein and Lewis (1993, Theorem 2.10), since $\phi'_+(-\infty) = -\infty$, this holds if and only if

$$\frac{1}{\epsilon}(f^j \oplus g^j - c) \leq \phi'(\infty), \tag{24}$$

$$\frac{d\pi^i_c}{d\mu \otimes \nu} = {\phi^*_+}' \circ \frac{1}{\epsilon}(f^j \oplus g^j - c) \ \mu \otimes \nu\text{-a.e.} \tag{25}$$

and

$$\mathrm{supp}(\pi^i_s) \subset \left\{(x, y) \in X \times Y : \frac{1}{\epsilon}(f^j(x) + g^j(y) - c(x, y)) = \phi'(\infty)\right\}, \tag{26}$$

where one has $\left\{(x, y) \in X \times Y : \frac{1}{\epsilon}(f^j(x) + g^j(y) - c(x, y)) = \phi'(\infty)\right\} = \partial_{(c,\epsilon,\phi)} f^j = \partial_{(c,\epsilon,\phi)} g^j$. As (25) holds for fixed j and $i = 1, 2$ we have that the absolutely continuous part of any optimal coupling is unique. If we let $C := \{(x, y) \in X \times Y : {\phi^*_+}'(\frac{1}{\epsilon}(f^1 \oplus g^1 - c)) = {\phi^*_+}'(\frac{1}{\epsilon}(f^2 \oplus g^2 - c))\}$ we know that $\mu \otimes \nu(C) = 1$. As ${\phi^*_+}' \geq 0$ and it is invertible in the points where ${\phi^*_+}' > 0$ by Proposition 10 if $P := \{(x, y) \in X \times Y : {\phi^*_+}'(\frac{1}{\epsilon}(f^1 \oplus g^1 - c)) > 0\}$ we know that for all $(x, y) \in C \cap P$ we have $f^1 \oplus g^1 = f^2 \oplus g^2$. But clearly $\pi^1_c(C \cap P) = \pi^1_c(X \times Y)$.

Furthermore, if ${\phi^*_+}'$ is invertible in its domain from the same equation we deduce that $f^1 \oplus g^1 = f^2 \oplus g^2$ $\mu \otimes \nu$-a.e. As μ and ν have full support so do $\mu \otimes \nu$, and as f^j and g^j for $j = 1, 2$ are continuous functions, then $f^1 \oplus g^1 = f^2 \oplus g^2$ must hold for every $(x, y) \in X \times Y$[10]. The last part of the proposition follows from (26) for a fixed i and any $j = 1, 2$. □

Definition 21 (Sinkhorn operator)**.** Let X and Y be compact metric spaces and $\mu \in P(X)$, $\nu \in P(Y)$ be Borel probability measures of full support. Let also $c \in \mathrm{Lip}(X \times Y)$, $0 < \epsilon \in \mathbb{R}$ be a regularization coefficient and $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ a proper, convex and lower semicontinuous function of Legendre type. Fix any point $y_0 \in Y$. Given a pair $(f, g) \in \mathrm{Lip}(X) \times \mathrm{Lip}(Y)$ we define the operator $\mathcal{F}^{(c,\epsilon,\phi)} : \mathrm{Lip}(X) \times \mathrm{Lip}(Y) \to \mathrm{Lip}(X) \times \mathrm{Lip}(Y)$ as

$$\mathcal{F}^{(c,\epsilon,\phi)}(f, g) := ((f^{(c,\epsilon,\phi)} - f^{(c,\epsilon,\phi)}(y_0))^{(c,\epsilon,\phi)}, f^{(c,\epsilon,\phi)} - f^{(c,\epsilon,\phi)}(y_0))$$

Technically this operator depends also on the point y_0 but as it is not very important which point it is, we decided not to put it in the definition of Sinkhorn iteration. This operator has the following very nice property:

[10] Here we use an standard continuity argument. If for some $(x_0, y_0) \in X \times Y$ we have $f^1(x_0) + g^1(y_0) \neq f^2(x_0) + g^2(y_0)$ then this will hold in an open neighborhood of (x_0, y_0). But this will contradict the fact that $f^1 \oplus g^1 = f^2 \oplus g^2$ $\mu \otimes \nu$-a.e. as any open set has positive measure if the measure has full support.

Proposition 22. *Let X and Y be compact metric spaces and $\mu \in P(X)$, $\nu \in P(Y)$ be Borel probability measures of full support. Let also $c \in \mathrm{Lip}(X \times Y)$, $0 < \epsilon \in \mathbb{R}$ be a regularization coefficient and $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ a proper, convex and lower semicontinuous function of Legendre type. Fix any point $y_0 \in Y$. Then for any $(f, g) \in \mathrm{Lip}(X) \times \mathrm{Lip}(Y)$ we have $\|\mathcal{F}^{(c,\epsilon,\phi)}(f, g)\|_{\max} \leq K$ where K depends only on the diameters of X and Y, ϵ and on $\|c\|_{\max}$. Moreover, $\mathcal{F}^{(c,\epsilon,\phi)}$ is continuous with respect to the $\|\cdot\|_\infty$-norm*[11].

Proof. By (iii) of Proposition 15 we have that the Lipschitz constants of $f^{(c,\epsilon,\phi)}$ and $(f^{(c,\epsilon,\phi)} - f^{(c,\epsilon,\phi)}(y_0))^{(c,\epsilon,\phi)}$ are uniformly bounded by $\|c\|_L$. As clearly $f^{(c,\epsilon,\phi)} - f^{(c,\epsilon,\phi)}(y_0)$ is a function that attains the value 0 we have that $\|f^{(c,\epsilon,\phi)} - f^{(c,\epsilon,\phi)}(y_0)\|_\infty \leq \|f^{(c,\epsilon,\phi)}\|_L \sup_{y,y' \in Y} d_Y(y, y') \leq \|c\|_L \operatorname{diam}(Y)$ where the diameter of Y is finite because Y is compact. Using (iv) of Proposition 15 we have that $\|(f^{(c,\epsilon,\phi)} - f^{(c,\epsilon,\phi)}(y_0))^{(c,\epsilon,\phi)}\|_\infty \leq \|(f^{(c,\epsilon,\phi)} - f^{(c,\epsilon,\phi)}(y_0))\|_\infty + \|c\|_\infty + \epsilon\phi'(\infty)\chi_{\mathbb{R}}(\phi'(\infty))$ (where the last summand vanishes if $\phi'(\infty) = \infty$). The last part of the proposition follows easily from (vi) of Proposition 15. □

Definition 23 (Good triple)**.** Let X be a compact metric space and μ a Borel probability measure on X. Let ϕ be a proper, convex and lower semicontinuous function of Legendre type and suppose that $\phi'(\infty) < \infty$. Let also $C > 0$ be a constant. We say that (X, μ, ϕ) is a *good triple* with respect to C if for all $x_0 \in X$

$$\lim_{\delta \downarrow 0} \int_X {\phi_+^*}'(\phi'(\infty) - Cd(x_0, x) - \delta)\, d\mu(x) > 1.$$

As we said in the main body of the paper, this condition is the one which ultimately will allow us to prevent the (c, ϵ, ϕ)-transform to collapse to the c-transform plus $\epsilon\phi'(\infty)$. More specifically, in the next proposition we will see how $\max_{x \in X, y \in Y}\{\frac{1}{\epsilon}(f(x) + f^{(c,\epsilon,\phi)}(y) - c(x, y))\}$ is separated from the critical value $\phi'(\infty)$ assuming this condition.

Proposition 24. *Let X and Y be compact metric spaces and $\mu \in P(X)$, $\nu \in P(Y)$ be Borel probability measures of full support. Let also $c \in \mathrm{Lip}(X \times Y)$, $0 < \epsilon \in \mathbb{R}$ be a regularization coefficient and $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ a proper, convex and lower semicontinuous function of Legendre type. Suppose that (X, μ, ϕ) is a good triple with respect to $2\|c\|_L/\epsilon$. Then for any $f \in \mathrm{Lip}(X)$ with $\|f\|_L \leq \|c\|_L$ we have that $\max_{x \in X, y \in Y}\{\frac{1}{\epsilon}(f(x) + f^{(c,\epsilon,\phi)}(y) - c(x, y))\} \leq \phi'(\infty) - \tau$ for some positive constant $\tau > 0$.*

Proof. Fix any $y \in Y$ and let $x_y \in X$ be such that the maximum of $\{\frac{1}{\epsilon}(f(x) + f^{(c,\epsilon,\phi)}(y) - c(x, y))\}$ with respect to $x \in X$ is attained at x_y (it always exists because X is compact and the functional continuous). Then

$$\left|\frac{1}{\epsilon}(f(x) + f^{(c,\epsilon,\phi)}(y) - c(x, y)) - \frac{1}{\epsilon}(f(x_y) + f^{(c,\epsilon,\phi)}(y) - c(x_y, y))\right| \leq \frac{2\|c\|_L}{\epsilon} d_X(x, x_y).$$

As ${\phi_+^*}'$ is increasing we have that

$$\int_X {\phi_+^*}'\left(\frac{1}{\epsilon}(f(x) + f^{(c,\epsilon,\phi)}(y) - c(x, y))\right) d\mu \geq \int_X {\phi_+^*}'\left(\frac{1}{\epsilon}(f(x_y) + f^{(c,\epsilon,\phi)}(y) - c(x_y, y)) - \frac{2\|c\|_L}{\epsilon} d_X(x, x_y)\right) d\mu.$$

Relabeling $\frac{1}{\epsilon}(f(x_y) + f^{(c,\epsilon,\phi)}(y) - c(x_y, y))$ as $\phi'(\infty) - \delta$ we see that this is a contradiction if δ is too small because the left hand side has to integrate to a value at most 1 by (the proof of) Proposition 15. By the definition of good triple we see that this is independent from the point $y \in Y$ so it holds for all of them. □

Let us now state our main final result, which we will prove in several steps:

Theorem 25. *Let X and Y be compact metric spaces and $\mu \in P(X)$, $\nu \in P(Y)$ be Borel probability measures of full support. Let also $c \in \mathrm{Lip}(X \times Y)$, $0 < \epsilon \in \mathbb{R}$ be a regularization coefficient and $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ a proper, convex and lower semicontinuous function of Legendre type. Suppose that (X, μ, ϕ) and (Y, ν, ϕ) are a good*

[11] In the space $\mathrm{Lip}(X) \times \mathrm{Lip}(Y)$ we define the $\|\cdot\|_\infty$ norm as $\|(f, g)\|_\infty := \max(\|f\|_\infty, \|g\|_\infty)$ for any $(f, g) \in \mathrm{Lip}(X) \times \mathrm{Lip}(Y)$.

triples with respect to $2\|c\|_L/\epsilon$. *Take any pair* $(f,g) \in \mathrm{Lip}(X) \times \mathrm{Lip}(Y)$ *and define inductively* $(f_0, g_0) := (f, g)$ *and* $(f_n, g_n) := \mathcal{F}^{(c,\epsilon,\phi)}(f_{n-1}, g_{n-1})$ *for* $n \geq 1$. *Let us also define the dual functional for any pair of functions* $(f,g) \in \mathrm{Lip}(X) \times \mathrm{Lip}(Y)$*:*

$$D_\epsilon(f,g) := \langle \mu, f\rangle + \langle \nu, g\rangle - \epsilon I^*_{\phi_+, \mu\otimes\nu}\left(\frac{1}{\epsilon}(f \oplus g - c)\right).$$

The optimal primal problem $\mathrm{OT}_\epsilon(\mu,\nu)$ *is defined as*

$$\underset{\epsilon}{\mathrm{OT}}(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \{\langle \pi, c\rangle + \epsilon I_{\phi_+,\mu\otimes\nu}(\pi)\}.$$

Then $D_\epsilon(f_n, g_n) \to \mathrm{OT}_\epsilon(\mu,\nu)$ *as* $n \to \infty$. *Also, there exists a unique optimal coupling* $\tilde{\pi}$ *that attains the infimum in* $\mathrm{OT}_\epsilon(\mu,\nu)$ *and if* $(\tilde{f}, \tilde{g})$ *are optimal potentials for the dual problem we have that* $f_n \oplus g_n \to \tilde{f} \oplus \tilde{g}$ *in* $L^\infty(\pi)$.

Proof. By Proposition 22 as soon as $n \geq 1$ we have that all the functions (f_n, g_n) will have $\|\cdot\|_{\max}$ norm bounded uniformly in terms of c, ϵ and the diameters of X and Y. Therefore we will assume from now on that all the functions in this sequence have this property. We can apply the Arzelà-Ascoli theorem (Cobzaş et al., 2019, Theorem 8.4.11) (as the sum norm defines the same topology as the max norm, clearly $\|\cdot\|_{\max} \leq \|\cdot\|_{\mathrm{sum}} \leq 2\|\cdot\|_{\max}$). Thus we have that $(f_{n_k}, g_{n_k}) \to (\tilde{f}, \tilde{g})$ as $k \to \infty$ in the $\|\cdot\|_\infty$ norm for some subsequence n_k.

In particular, as all the elements f_n and g_n have Lipschitz norm bounded by $\|c\|_L$, so do $\tilde{f}$ and $\tilde{g}$. The Sinkhorn operator is continuous with the $\|\cdot\|_\infty$-norm by Proposition 22. By Proposition 24 and the definition of the pair (f_n, g_n) we know that $\frac{1}{\epsilon}(f_n \oplus g_n - c)$ has its image in $(-\infty, \phi'(\infty) - \tau]$ and therefore (Borwein and Lewis, 1993, Theorem 2.7) the operator $D_\epsilon(\cdot,\cdot)$ is continuous in the set where (f_n, g_n) lives.

Thus, we have that $D_\epsilon(f_{n_k}, g_{n_k}) \to D_\epsilon(\tilde{f}, \tilde{g})$ and $D_\epsilon(\mathcal{F}^{(c,\epsilon,\phi)}(f_{n_k}, g_{n_k})) \to D_\epsilon(\mathcal{F}^{(c,\epsilon,\phi)}(\tilde{f}, \tilde{g})))$. Furthermore, by definition of the sequence (f_n, g_n) we have that $D_\epsilon(f_{n_k}, g_{n_k})) \leq D_\epsilon(\mathcal{F}^{(c,\epsilon,\phi)}(f_{n_k}, g_{n_k})) \leq D_\epsilon(f_{n_{k+1}}, g_{n_{k+1}}))$. Hence, $D_\epsilon(\tilde{f}, \tilde{g}) = D_\epsilon(\mathcal{F}^{(c,\epsilon,\phi)}(\tilde{f}, \tilde{g}))$. In particular we have that $D_\epsilon(\tilde{f}, \tilde{g})) \leq D_\epsilon(\tilde{f}, \tilde{f}^{(c,\epsilon,\phi)}) = D_\epsilon(\tilde{f} + a, \tilde{f}^{(c,\epsilon,\phi)} - a)) \leq D_\epsilon((\tilde{f}^{(c,\epsilon,\phi)} - a)^{(c,\epsilon,\phi)}, \tilde{f}^{(c,\epsilon,\phi)} - a)) = D_\epsilon(\tilde{f}, \tilde{g})$ for some constant $a \in \mathbb{R}$. Hence, all previous inequalities are equalities. By (v) of Proposition 15 we have that $D_\epsilon(\tilde{f}, \tilde{g})) = D_\epsilon(\tilde{f}, \tilde{f}^{(c,\epsilon,\phi)})$ and $D_\epsilon(\tilde{f}, \tilde{f}^{(c,\epsilon,\phi)})) = D_\epsilon((\tilde{f}^{(c,\epsilon,\phi)})^{(c,\epsilon,\phi)}, \tilde{f}^{(c,\epsilon,\phi)}))$. Now we need the following lemma:

Lemma 26. *With the same hypothesis as in Theorem 25 let* $f \in \mathrm{Lip}(X)$ *and* $g \in \mathrm{Lip}(Y)$ *be any functions. Suppose that* $D_\epsilon(f,g)) = D_\epsilon(f, f^{(c,\epsilon,\phi)})$. *Then* $g = f^{(c,\epsilon,\phi)}$ *for every* $y \in Y$.

Proof of lemma: Assume that equality fails at some $y_0 \in Y$. For any $h \in \mathrm{Lip}(Y)$ and $y \in Y$ let us define

$$H_h(y) := h(y) - \epsilon \int_X \phi_+^*(f + h(y) - c(\cdot, y))d\mu.$$

By hypothesis we also have that $\int H_g(y)d\nu = \int H_{f^{(c,\epsilon,\phi)}}(y)d\nu$. Note that by definition of the (c,ϵ,ϕ)-transform we have that $H_g(y) \leq H_{f^{(c,\epsilon,\phi)}}(y)$ for all $y \in Y$. Thus, those functions are equal ν-a.e. If for some $y_0 \in Y$ we have $g(y_0) \neq f^{(c,\epsilon,\phi)}(y_0)$ then $H_g(y_0) < H_{f^{(c,\epsilon,\phi)}}(y_0)$. Let us denote that positive difference as $e := H_{f^{(c,\epsilon,\phi)}}(y_0) - H_g(y_0)$.

Now note first that $H_g(y)$ is upper semi-continuous in y (Zalinescu, 2002, Theorem 2.3.1) (we apply this result for general Lipschitz functions in $\mathrm{Lip}(X)$ and then note that we are just specializing that result to the concrete family of functions $\frac{1}{\epsilon}(f(x) + g(y) - c(x,y))$ indexed by $y \in Y$). For the part of $H_{f^{(c,\epsilon,\phi)}}$ we need to prove continuity instead of just upper semicontinuity. To do so, recall that by Proposition 24 we know that the functions $\frac{1}{\epsilon}(f(x) + f^{(c,\epsilon,\phi)}(y) - c(x,y)) \in \mathrm{Lip}(X)$ (this is a family of functions indexed by $y \in Y$) have their image strictly inside the range $(-\infty, \phi'(\infty))$. Thus we have that for $y \in Y$, $H_{f^{(c,\epsilon,\phi)}}$ is continuous (Borwein and Lewis, 1993, Theorem 2.7).

Hence, by upper semicontinuity of $H_g(y)$ there exists some $\delta > 0$ such that if $d_Y(y, y_0) < \delta$ then $H_g(y) < H_g(y_0) + e/3$. Similarly, by continuity of $H_{f^{(c,\epsilon,\phi)}}(y)$ there exists $\delta' > 0$ such that if $d_Y(y, y_0) < \delta'$ then $H_{f^{(c,\epsilon,\phi)}}(y) \geq H_{f^{(c,\epsilon,\phi)}}(y_0) - e/3$. Therefore if $d_Y(y, y_0) < \min(\delta, \delta')$ then $H_g(y) < H_{f^{(c,\epsilon,\phi)}}(y)$ and this is a contradiction with the fact that ν has full support. In particular, we have found an open set $\{y \in Y : d_Y(y, y_0) < \min(\delta, \delta')\}$ (which has positive measure as ν has full support) such that $H_g(y) < H_{f^{(c,\epsilon,\phi)}}(y)$ and this is a contradiction with the fact that these two functions are equal ν-a.e. □

Note that under similar hypothesis an analogous argument shows that if $D_\epsilon(f,g) = D_\epsilon(g^{(c,\epsilon,\phi)},g)$ then $g^{(c,\epsilon,\phi)} = f$.

This lemma implies that the potentials $(\tilde{f},\tilde{g})$ that we have found satisfy that they are (c,ϵ,ϕ)-transforms of each other. We claim that the measure

$$\tilde{\pi} := \phi_+^{*\prime}(\tilde{f} \oplus \tilde{g} - c)\mu \otimes \nu$$

is an optimal solution to the primal problem, i.e. $D_\epsilon(\tilde{f},\tilde{g}) = \mathrm{OT}_\epsilon(\mu,\nu)$. First, note that by Theorem 18 if we manage to prove that $\tilde{\pi}$ is in $\Pi(\mu,\nu)$ then we would be done (as this clearly satisfy all the remaining conditions to be optimal). As $\phi_+^{*\prime}$ is nonnegative so is $\tilde{\pi}$.

Let us see that $(p_2^*)(\tilde{\pi}) = \nu$ (the other projection follows analogously). Given any $A \subset Y$ measurable we have that

$$(p_2^*)(\tilde{\pi})(A) = \int_A \int_X \phi_+^{*\prime}(\tilde{f}(x) + \tilde{g}(y) - c(x,y))d\mu(x)d\nu(y)$$

$$= \int_A \int_X \phi_+^{*\prime}(\tilde{f}(x) + \tilde{f}^{(c,\epsilon,\phi)}(y) - c(x,y))d\mu(x)d\nu(y).$$

By hypothesis, recall that (X,μ,ϕ) is a good triple with respect to $2\|c\|_L/\epsilon$ and therefore by Proposition 24 and (i) of Proposition 15 we know that for every $y \in Y$ we have $\int_X \phi_+^{*\prime}(\tilde{f} + \tilde{f}^{(c,\epsilon,\phi)}(y) - c(\cdot,y))d\mu = 1$. Thus, the integral above reduces to $\int_A d\nu = \nu(A)$ and the result follows.

This proves that for the subsequence n_k we have that $D_\epsilon(f_{n_k},g_{n_k}) \to \mathrm{OT}_\epsilon(\mu,\nu)$. However, it is easy to extend this result to the full sequence using the fact that $D_\epsilon(f_n,g_n)$ is an increasing sequence for all $n \geq 1$. Just note that given $\delta > 0$ we know that there exists $K(\delta)$ such that if $k \geq K(\delta)$ we have that $|D_\epsilon(f_{n_k},g_{n_k}) - \mathrm{OT}_\epsilon(\mu,\nu)| < \delta$. Thus, for all $n \geq n_{K(\delta)}$ we know that $|D_\epsilon(f_n,g_n) - \mathrm{OT}_\epsilon(\mu,\nu)| < \delta$ which proves convergence.

By Proposition 20 we know that the absolutely continuous part of an optimal solution is unique. As in our case we know that this defines directly a probability measure, we know that the solution to the primal problem $\mathrm{OT}_\epsilon(\mu,\nu)$ is unique (the singular part must be just 0). To prove the last part of the theorem, let $(f',g') \in \mathrm{Lip}(X) \times \mathrm{Lip}(Y)$ be a pair of optimal potentials and suppose by contradiction that $f_n \oplus g_n \not\to f' \oplus g'$ in $L^\infty(\tilde{\pi})$ as $n \to \infty$. This means that there exists a subsequence n_i and a positive $\delta > 0$ such that $\|f_{n_i} \oplus g_{n_i} - f' \oplus g'\|_{L^\infty(\tilde{\pi})} \geq \delta$ for all $i \geq 1$. We can now repeat the same arguments as before using the Arzelà-Ascoli theorem to prove that there exists a subsubsequence n_{i_j} for $j \geq 1$ such that $(f_{n_{i_j}},g_{n_{i_j}}) \to (f'',g'')$ as $j \to \infty$ in the $\|\cdot\|_\infty$ norm. In particular, $f_{n_{i_j}} \oplus g_{n_{i_j}} \to f'' \oplus g''$ in $L^\infty(\tilde{\pi})$. By Proposition 20 we have that $f' \oplus g' = f'' \oplus g''$ π-a.e. which is clearly a contradiction. □

C Functions related to f-divergences

C.1 Kullback-Leibler divergence

$$\phi_+(x) = \begin{cases} x\log(x) - x + 1 \text{ if } x \geq 0, \\ \infty \text{ otherwise.} \end{cases} \tag{27}$$

$$\partial\phi_+(x) = \begin{cases} \{\log(x)\} \text{ if } x > 0, \\ \emptyset \text{ otherwise.} \end{cases} \tag{28}$$

$$\phi'(\infty) = \infty. \tag{29}$$

$$\phi_+^*(x) = e^x - 1. \tag{30}$$

$$\phi_+^{*\prime}(x) = e^x. \tag{31}$$

$$\phi_+^{*\prime\prime}(x) = e^x. \tag{32}$$

C.2 Reverse Kullback-Leibler divergence

$$\phi_+(x) = \begin{cases} x - 1 - \log(x) \text{ if } x \geq 0, \\ \infty \text{ otherwise.} \end{cases} \tag{33}$$

$$\partial\phi_+(x) = \begin{cases} \left\{\frac{x-1}{x}\right\} \text{ if } x > 0, \\ \emptyset \text{ otherwise.} \end{cases} \tag{34}$$

$$\phi'(\infty) = 1. \tag{35}$$

$$\phi_+^*(x) = \begin{cases} -\log(1-x) \text{ if } x \leq 1, \\ \infty \text{ otherwise.} \end{cases} \tag{36}$$

$$\phi_+^{*\prime}(x) = \frac{1}{1-x}. \tag{37}$$

$$\phi_+^{*\prime\prime}(x) = \frac{1}{(1-x)^2}. \tag{38}$$

C.3 χ^2 divergence

$$\phi_+(x) = \begin{cases} (x-1)^2 \text{ if } x \geq 0, \\ \infty \text{ otherwise.} \end{cases} \tag{39}$$

$$\partial\phi_+(x) = \begin{cases} \{2x-2\} \text{ if } x \geq 0, \\ \emptyset \text{ otherwise.} \end{cases} \tag{40}$$

$$\phi'(\infty) = \infty. \tag{41}$$

$$\phi_+^*(x) = \begin{cases} \frac{1}{4}x^2 + x \text{ if } x \geq -2, \\ -1 \text{ otherwise.} \end{cases} \tag{42}$$

$$\phi_+^{*\prime}(x) = \begin{cases} \frac{1}{2}x + 1 \text{ if } x \geq -2, \\ 0 \text{ otherwise.} \end{cases} \tag{43}$$

$$\phi_+^{*\prime\prime}(x) = \begin{cases} \frac{1}{2} \text{ if } x \geq -2, \\ 0 \text{ otherwise.} \end{cases} \tag{44}$$

C.4 Reverse χ^2 divergence

$$\phi_+(x) = \begin{cases} \frac{1}{x} + x - 2 \text{ if } x \geq 0, \\ \infty \text{ otherwise.} \end{cases} \tag{45}$$

$$\partial\phi_+(x) = \begin{cases} \left\{1 - \frac{1}{x^2}\right\} \text{ if } x > 0, \\ \emptyset \text{ otherwise.} \end{cases} \tag{46}$$

$$\phi'(\infty) = 1. \tag{47}$$

$$\phi_+^*(x) = \begin{cases} 2 - 2\sqrt{1-x} \text{ if } x \leq 1, \\ \infty \text{ otherwise.} \end{cases} \tag{48}$$

$$\phi_+^{*\prime}(x) = \frac{1}{\sqrt{1-x}}. \tag{49}$$

$$\phi_+^{*\prime\prime}(x) = \frac{1}{2\sqrt{1-x}^3}. \tag{50}$$

C.5 Squared Hellinger divergence

$$\phi_+(x) = \begin{cases} (\sqrt{x}-1)^2 \text{ if } x \geq 0, \\ \infty \text{ otherwise.} \end{cases} \tag{51}$$

$$\partial\phi_+(x) = \begin{cases} \left\{1 - \frac{1}{\sqrt{x}}\right\} \text{ if } x > 0, \\ \emptyset \text{ otherwise.} \end{cases} \tag{52}$$

$$\phi'(\infty) = 1. \tag{53}$$

$$\phi_+^*(x) = \begin{cases} \frac{x}{1-x} \text{ if } x \leq 1, \\ \infty \text{ otherwise.} \end{cases} \tag{54}$$

$$\phi_+^{*\prime}(x) = \frac{1}{(1-x)^2}. \tag{55}$$

$$\phi_+^{*\prime\prime}(x) = \frac{2}{(1-x)^3}. \tag{56}$$

C.6 Jensen-Shannon divergence

$$\phi_+(x) = \begin{cases} x\log(x) - (x+1)\log(\frac{x+1}{2}) \text{ if } x \geq 0, \\ \infty \text{ otherwise.} \end{cases} \tag{57}$$

$$\partial\phi_+(x) = \begin{cases} \{\log(x) - \log(x+1) + \log(2)\} \text{ if } x > 0, \\ \emptyset \text{ otherwise.} \end{cases} \tag{58}$$

$$\phi'(\infty) = \log(2). \tag{59}$$

$$\phi_+^*(x) = \begin{cases} -\log(2-e^x) \text{ if } x \leq \log(2), \\ \infty \text{ otherwise.} \end{cases} \tag{60}$$

$$\phi_+^{*\prime}(x) = \frac{1}{2e^{-x}-1}. \tag{61}$$

$$\phi_+^{*\prime\prime}(x) = \frac{2e^x}{(e^x-2)^2}. \tag{62}$$

C.7 Jeffreys divergence

$$\phi_+(x) = \begin{cases} (x-1)\log(x) \text{ if } x \geq 0, \\ \infty \text{ otherwise.} \end{cases} \tag{63}$$

$$\partial\phi_+(x) = \begin{cases} \left\{\log(x) - \frac{1}{x} + 1\right\} \text{ if } x > 0, \\ \emptyset \text{ otherwise.} \end{cases} \tag{64}$$

$$\phi'(\infty) = \infty. \tag{65}$$

$$\phi_+^*(x) = x + W(e^{1-x}) + \frac{1}{W(e^{1-x})} - 2. \tag{66}$$

$$\phi_+^{*\prime}(x) = \frac{1}{W(e^{1-x})}. \tag{67}$$

$$\phi_+^{*\prime\prime}(x) = \frac{1}{W(e^{1-x})} - \frac{1}{W(e^{1-x})+1}. \tag{68}$$

Here W denotes the principal branch of the Lambert W function, also called the product logarithm, defined implicitly by the relation $W(x)e^{W(x)} = x$. This can be computed by Newton's method and differentiated implicitly. For stability, since we only need the value of $W(e^{1-x})$, we compute $W(e^{1-x})$ explicitly instead of composing the Lambert W function with e^{1-x}.

C.8 Triangular discrimination divergence

$$\phi_+(x) = \begin{cases} \frac{(x-1)^2}{x+1} \text{ if } x \geq 0, \\ \infty \text{ otherwise.} \end{cases} \tag{69}$$

$$\partial\phi_+(x) = \begin{cases} \left\{ \frac{(x-1)(x+3)}{(x+1)^2} \right\} \text{ if } x \geq 0, \\ \emptyset \text{ otherwise.} \end{cases} \tag{70}$$

$$\phi'(\infty) = 1. \tag{71}$$

$$\phi_+^*(x) = \begin{cases} -1 \text{ if } x < -3, \\ (\sqrt{1-x}-1)(\sqrt{1-x}-3) \text{ if } -3 \leq x \leq 1, \\ \infty \text{ otherwise.} \end{cases} \tag{72}$$

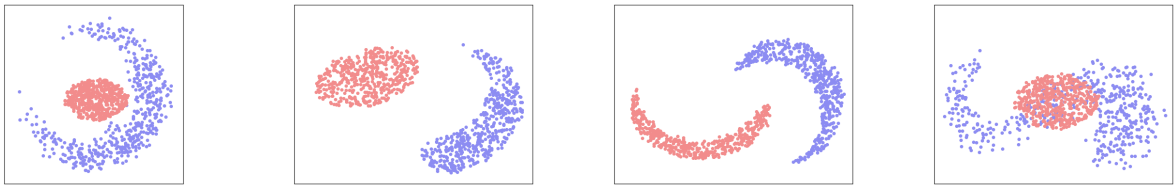
$$\phi_+^{*\prime}(x) = \begin{cases} 0 \text{ if } x < -3, \\ \frac{2}{\sqrt{1-x}} - 1 \text{ if } -3 \leq x \leq 1. \end{cases} \tag{73}$$

$$\phi_+^{*\prime\prime}(x) = \begin{cases} 0 \text{ if } x < -3, \\ \frac{1}{(\sqrt{1-x})^3} \text{ if } -3 \leq x \leq 1. \end{cases} \tag{74}$$

D Experimental results

D.1 Experimental setup

As we explained in the main paper, we tested our algorithm on synthetic 2-dimensional data obtained from the codebase of Feydy et al. (2019). These data consists of 4 pairs of densities in the 2-dimensional space named "crescents", "densities", "moons" and "slopes". For each of these density pairs, and for point cloud sizes in $\{500, 1000, 2000, 5000\}$, we sample 5 different point clouds fixing the random seed in $\{0, 1, 2, 3, 4\}$. Thus, we have in total $4 \times 4 \times 5 = 80$ different pairs of point clouds that we are going to use in our experiments. An example of such pointclouds for each density pair can be seen in Figure 2. Then, for each of the 8 divergences considered (Kullback-Leibler, reverse Kullback-Leibler, χ^2, reverse χ^2, squared Hellinger, Jensen-Shannon, Jeffreys and triangular discrimination) we tried different ϵ regularization coefficients ranging from 0.1 to 10^{-8}.



(a) "crescents" (b) "densities" (c) "moons" (d) "slopes"

Figure 2: Example of generated point cloud.

D.2 Cost of the optimal coupling, convergence speed, sparsity and marginal error

We can find in Figures 3, 4 and 5 the plots of the costs vs running time and the plots of sparsity vs marginal error corresponding to the datasets of "moons", "densities" and "slopes" respectively. Similarly as before, we eliminated the values with marginal error greater than 0.2.

D.3 Gradients through the optimal coupling

As we remarked in the main paper, the algorithm that we present can be used as the loss function between point clouds defined by empirical measures in automatic differentiation engines. In order to do so, one has to compute the gradient with respect to the points in the supports of the measures. An obvious solution is to backpropagate through the Sinkhorn iterations, which is computationally demanding. It is possible to do so via the optimal potentials f, g by generalizing the "graph surgery" method of Feydy et al. (2019) and the gradient formula of Di Marino and Gerolin (2020b, Proposition 3.7). Instead, we propose to do so via the optimal coupling π. Detaching π from the computational graph and calculating $\int c d\pi = \sum_{i,j} C_{i,j}\pi_{i,j}$ leads to a scalar loss which depends on the points $\{x_i\}$ and $\{y_j\}$ only through the cost function c.

An intrinsic feature of entropic regularization is that introduces a tradeoff between convergence speed of the Sinkhorn algorithm and bias in the optimal coupling (i.e., the coupling obtained minimizes $\int c\, d\pi + \epsilon D_f(\pi \| \mu \otimes \nu)$ instead of the original $\int c\, d\pi$ for $\pi \in \Pi(\mu, \nu)$). Increasing ϵ leads to faster convergence, but pushes the optimal coupling further away from the coupling which is optimal in the unregularized problem. Using different f-divergences for regularization leads to different biases. Since the range of values of these divergences can be quite different, there is no point in comparing the induced biases with equal ϵs. To make a fair comparison, we tuned the value of ϵ for each task-divergence setting in order for the Sinkhorn algorithm to converge in 200 iterations with a tolerance of $\tau = 10^{-6}$.

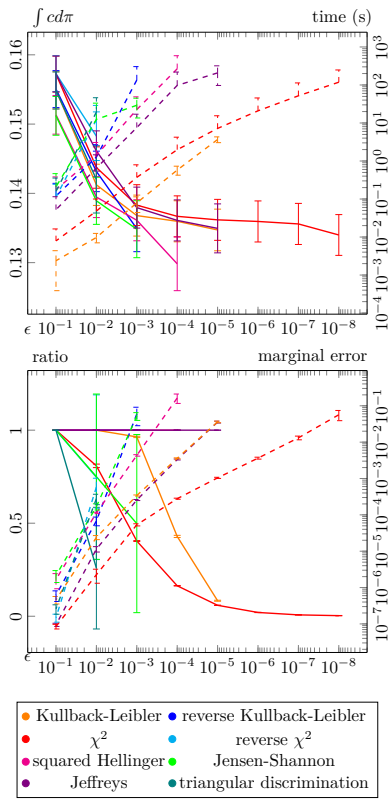


Figure 3: Dataset: "moons". Above: cost of optimal coupling (solid line) and runtime in seconds (dashed line). Below: ratio of positive elements to all elements in optimal coupling (solid line) and marginal error (dashed line).

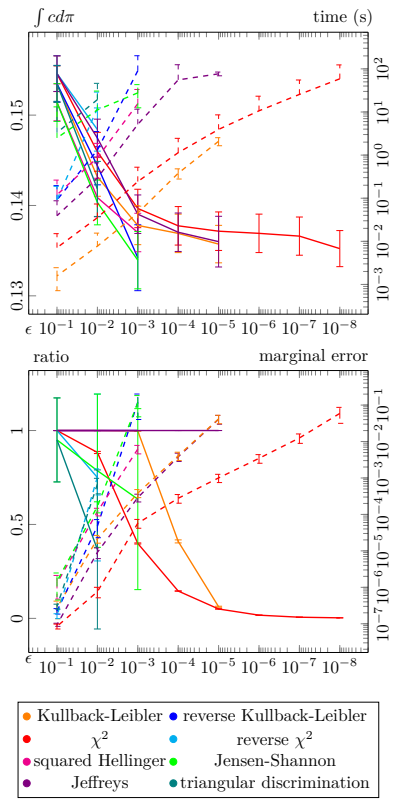


Figure 4: Dataset: "densities". Above: cost of optimal coupling (solid line) and runtime in seconds (dashed line). Below: ratio of positive elements to all elements in optimal coupling (solid line) and marginal error (dashed line).

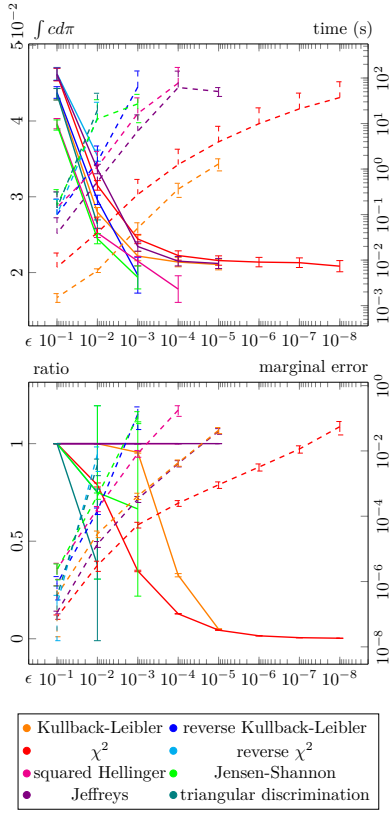
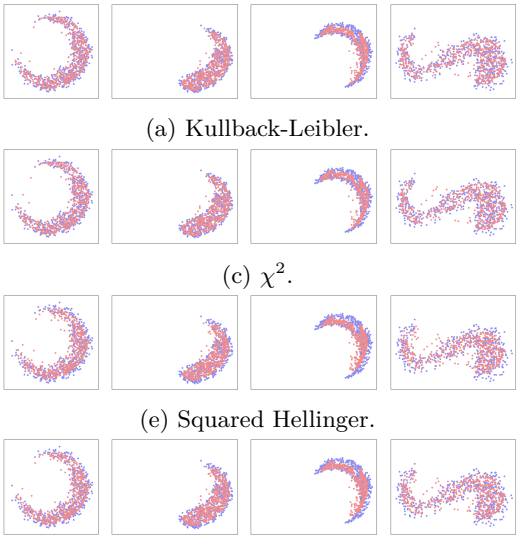
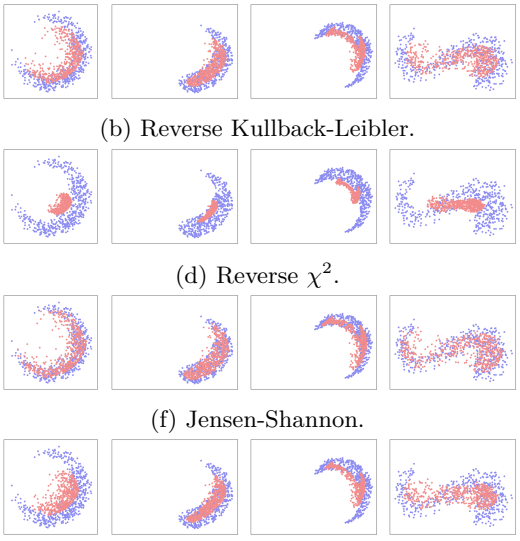


Figure 5: Dataset: "slopes". Above: cost of optimal coupling (solid line) and runtime in seconds (dashed line). Below: ratio of positive elements to all elements in optimal coupling (solid line) and marginal error (dashed line).



(a) Kullback-Leibler.



(b) Reverse Kullback-Leibler.

(c) χ^2.

(d) Reverse χ^2.

(e) Squared Hellinger.

(f) Jensen-Shannon.

(g) Jeffreys.

(h) Triangular discrimination.

Figure 6: Bias of optimal couplings when ϵ is tuned to reach tolerance of 10^{-6} in 200 Sinkhorn iterations.

Upon convergence, we backpropagated the resulting loss $\int c d\pi$ as described above, and took 1 gradient descent step on the points belonging to the support of μ with a learning rate of 1. If the coupling π were unbiased, this procedure should transport the red pointcloud μ exactly onto the blue one ν. The results are visualized in Figure 6. The tradeoff leads to a visually similar, small amount of bias in the case of the Kullback-Leibler, χ^2, squared Hellinger, Jensen-Shannon and Jeffreys divergences. On the other hand, the bias is more pronounced for the reverse Kullback-Leibler, reverse χ^2 and triangular discrimination divergences. The bias can be reduced in all cases by decreasing ϵ, at a price of slower convergence speed. For other application scenarios, practitioners might benefit from evaluating all considered f-divergences, since the biases in other tasks could differ.

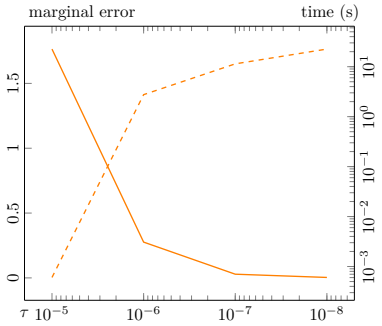


Figure 7: Marginal error (solid line) can be decreased by decreasing the tolerance parameter τ at the cost of increased running time (dashed line). This plots was done using the Kullback-Leibler divergece, the "crescents" dataset, a point cloud size of 500 and 0 as the random seed.

D.4 Fixing marginal errors

Depending of the hyperparemeters, the resulting coupling can have a large marginal error. There are algorithms in the literature that correct an approximate coupling to an exact one such as Altschuler et al. (2017, Algorithm 2), which we have included in the source code of the experiments. However, the final cost of the resulting coupling was worse in general than the one without this rounding step and also the sparsity of the coupling disappears. Hence, we decided not to include this rounding step in our algorithm. Without such a rounding step, practitioners should set a lower tolerance parameter τ to decrease the marginal error of the resulting coupling. We ran several experiments to see this effect with different divergences, random seeds, and point cloud sizes, and the results were very similar. Thus, we decided to include just one of them as an example in Figure 7.