



HAL
open science

Equivalence of optimal transport problems to regularization on the family of f-divergences

Maxime Nicaise, Yaiza Bermudez, Samir M. Perlaza

► To cite this version:

Maxime Nicaise, Yaiza Bermudez, Samir M. Perlaza. Equivalence of optimal transport problems to regularization on the family of f-divergences. ISIT 2026 - IEEE International Symposium on Information Theory, Jun 2026, Guangzhou, China. ⟨hal-05591964⟩

HAL Id: hal-05591964

<https://inria.hal.science/hal-05591964v1>

Submitted on 15 Apr 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

Equivalence of optimal transport problems to regularization on the family of f -divergences

Maxime Nicaise*, Yaiza Bermudez*, and Samir M. Perlaza*^{†§}

Emails: maxime.nicaise@student-cs.fr, {yaiza.bermudez, samir.perlaza}@inria.fr

*Centre Inria d'Université Côte d'Azur, INRIA, Sophia Antipolis, France

[†]Laboratoire GAATI, Université de la Polynésie française, Fa'a'a, French Polynesia

[§]ECE Dept. Princeton University, Princeton, 08544 NJ, USA

Abstract—This work establishes that an optimal transport (OT) problem regularized by a given f -divergence admits the same solution as another OT problem regularized by a different g -divergence, under an appropriate transformation of the cost function. This structural equivalence between OT problems regularized by distinct divergences, in the sense of sharing the same unique minimizer, is demonstrated within the framework of Polish spaces with bounded cost functions.

Index Terms—Optimal Transport, Optimization, Regularization.

I. INTRODUCTION

Optimal transport (OT) is established as a fundamental tool for comparing probability distributions, with applications ranging from machine learning to computer graphics [1]–[6]. The choice of a suitable discrepancy measure between distributions is essential for tasks such as regression, classification, and generative modeling. In this respect, OT-based distances offer a significant advantage over classical divergences, such as the Kullback–Leibler divergence or other f -divergences, as demonstrated in [2] and [4]. Indeed, the topology induced through Wasserstein distances is weaker and better adapted to comparing measures with non-overlapping supports or those supported on lower-dimensional sets. See [2] and [7]. This property makes OT particularly attractive for modern data analysis.

Despite these appealing properties, from a computational perspective, solving the Kantorovich formulation of optimal transport involves tackling a large-scale linear program. This task often becomes intractable for high-dimensional or large-sample problems, as shown in [8]. To mitigate the computational complexity of the original problem and improve the sample complexity of the estimators, regularization is commonly employed. This ranges from classic entropic regularization to structured [3] or quadratic regularizers [9]. In particular, classic entropic regularization leads to the celebrated Sinkhorn algorithm [10]. However, more recent developments highlight the benefits of broadening this framework to general

f -divergences [11], [12]. These schemes yield desirable properties such as sparsity or robustness to outliers. Parallel to these advances in optimal transport, similar variational problems are extensively studied within the field of statistical learning theory, particularly within the statistical empirical risk minimization (ERM) framework [13]–[16]. Specifically, theoretical equivalence between variational problems regularized with two different f -divergences is established in [16], [17] and [18]. Therein, it is proved that an ERM problem regularized by a specific f -divergence is reformulable using a different divergence and a suitably transformed cost function, yielding the exact same minimizer.

This work establishes an equivalence between optimal transport problems with f -divergence regularizations, which is analogous to [16, Theorem 2] in the context of ERM with f -divergence regularizations. More specifically, in this work, it is proved that two optimal transport problems regularized by different divergences admit the same unique optimal coupling, up to a suitable transformation of the cost function. This structural equivalence between OT problems is demonstrated within the framework of Polish spaces with bounded cost functions. Under these general assumptions, the existence of optimal potentials for the dual problem and the uniqueness of the optimal coupling for the primal problem are rigorously established. Furthermore, a closed-form expression for the optimal coupling is obtained.

A. Notations and preliminaries

Given a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, its effective domain is denoted by $\text{dom}(f) \triangleq \{x \in \mathcal{X} : f(x) < +\infty\}$; and its range is denoted by $\text{Ran}(f) \triangleq \{f(x) : x \in \mathcal{X}\}$. Given a set \mathcal{A} , its interior is denoted by $\text{int}(\mathcal{A})$. Given two measures μ and ν , the notation $\mu \ll \nu$ signifies that μ is absolutely continuous with respect to ν . In the following, \mathcal{X} and \mathcal{Y} are Polish spaces. The set of probability measures over \mathcal{X} equipped with the σ -field of Borel sets is denoted by $\Delta(\mathcal{X})$. The set of p -integrable functions with respect to P_X is denoted by $L^p(\mathcal{X}, P_X)$. Given two probability measures $P_X \in \Delta(\mathcal{X})$ and $P_Y \in \Delta(\mathcal{Y})$, the notation $\Pi(P_X, P_Y)$ denotes the set of probability measures on $\mathcal{X} \times \mathcal{Y}$ with marginals P_X and P_Y in $\Delta(\mathcal{X})$ and $\Delta(\mathcal{Y})$, respectively. The probability measures in $\Pi(P_X, P_Y)$ are called couplings. The product measure of $P_X \in \Delta(\mathcal{X})$ and $P_Y \in \Delta(\mathcal{Y})$ in $\Delta(\mathcal{X} \times \mathcal{Y})$ is denoted by $P_X P_Y$. Given

This work is supported in part by the European Commission through the H2020-MSCA-RISE-2019 project 872172; the French National Agency for Research (ANR) through the Project ANR-21-CE25-0013 and the project ANR-22-PEFT-0010 of the France 2030 program PEPR Réseaux du Futur; and in part by the Agence de l'innovation de défense (AID) through the project UK-FR 2024352.

two functions $f : \mathcal{X} \rightarrow \mathbb{R}$ and $g : \mathcal{X} \rightarrow \mathbb{R}$, the notation $f \stackrel{a.s.}{\underset{P_X}{\approx}} g$ means that f and g are almost surely equal with respect to P_X . Similarly, the notation $f \stackrel{a.s.}{\underset{P_X}{\geq}} g$ means that f is greater than g almost surely with respect to P_X . Given \mathcal{A} and \mathcal{B} , two subsets of \mathcal{X} , the notation $\mathcal{A} \stackrel{a.s.}{\underset{P_X}{\subset}} \mathcal{B}$ means that $P_X(\mathcal{A} \setminus \mathcal{B}) = 0$. Let \mathcal{E} be a Banach space and \mathcal{E}' be its topological dual space. Given $x \in \mathcal{E}$ and $y \in \mathcal{E}'$, $\langle y, x \rangle$ denotes the dual pairing between y and x . If $f : \mathcal{X} \rightarrow \mathbb{R}$ and $g : \mathcal{Y} \rightarrow \mathbb{R}$ are two real functions, then $f \oplus g : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ is the function defined as $f \oplus g(x, y) = f(x) + g(y)$. Analysis of optimality conditions and the derivation of dual problems rely on the transformation of functions into a dual space representation via the convex (or Fenchel) conjugate ϕ^* , defined in [19, Section 2.3]. For a function ϕ on a Banach space \mathcal{E} with dual \mathcal{E}' , it is defined for $y \in \mathcal{E}'$ as $\phi^*(y) = \sup_{x \in \mathcal{E}} (\langle y, x \rangle - \phi(x))$. Constraints are handled through the indicator function $\iota_{\mathcal{C}}$, which assigns a value of 0 for elements in the feasible set \mathcal{C} and $+\infty$ otherwise.

B. Assumptions on ϕ -divergences

The purpose of this section is to define the class of ϕ -divergences endowed with sufficient regularity to guarantee satisfactory theoretical properties to regularize OT problems. Let ϕ be a convex, proper, lower semi-continuous function from \mathbb{R} to $(-\infty, +\infty]$. Throughout this work, the real numbers a_ϕ and b_ϕ denote the following quantities

$$a_\phi \triangleq \inf \text{dom}(\phi); \text{ and} \quad (1)$$

$$b_\phi \triangleq \sup \text{dom}(\phi). \quad (2)$$

As stated in [20, Lemma 2.1], the following limits exist

$$\alpha_\phi \triangleq \lim_{x \rightarrow -\infty} \frac{\phi(x)}{x}; \text{ and} \quad (3)$$

$$\beta_\phi \triangleq \lim_{x \rightarrow +\infty} \frac{\phi(x)}{x}. \quad (4)$$

Note that if $\text{dom}(\phi) = (0, +\infty)$, then by convention $\alpha_\phi = -\infty$. It is shown in [20, Lemma 2.1] that the convex conjugate of ϕ , which is denoted by ϕ^* , verifies

$$\text{int}(\text{dom}(\phi^*)) = (\alpha_\phi, \beta_\phi). \quad (5)$$

The class of Legendre functions, characterized by strict convexity and essential smoothness, ensures a smooth one-to-one correspondence between primal and dual variables. The following definition characterizes functions of Legendre type.

Definition 1 ([20], Definition 2.5). *A function ϕ is said to be of Legendre type, if it is strictly convex on $\text{dom}(\phi)$; differentiable on $\text{int}(\text{dom}(\phi))$; and its derivative $\phi' : \text{int}(\text{dom}(\phi)) \rightarrow \mathbb{R}$ satisfies*

$$\lim_{x \rightarrow a_\phi} \phi'(x) = -\infty, \text{ if } a_\phi > -\infty; \text{ and} \quad (6)$$

$$\lim_{x \rightarrow b_\phi} \phi'(x) = +\infty, \text{ if } b_\phi < +\infty. \quad (7)$$

The specific class of Legendre functions possesses certain properties regarding conjugation, defined in [19, Section 2.3], as detailed in the following lemma.

Lemma 1 ([20], Lemma 2.6). *If $\phi : \mathbb{R} \rightarrow (-\infty, +\infty]$ is of Legendre type, then so is its convex conjugate ϕ^* . Moreover $\phi' : (a_\phi, b_\phi) \rightarrow (\alpha_\phi, \beta_\phi)$ and $\phi^{*'} : (\alpha_\phi, \beta_\phi) \rightarrow (a_\phi, b_\phi)$ are continuous, strictly increasing and mutually inverse maps.*

In the context of this work, the focus is on a specific subset of Legendre functions that are suitable for defining divergences between probability measures, which are defined as follows.

Definition 2 (Legendre-type generator). *A Legendre type function $\phi : \mathbb{R} \rightarrow (-\infty, +\infty]$ is said to be a Legendre-type generator when $\phi(1) = 0$; $a_\phi = 0$ and $b_\phi = +\infty$, with a_ϕ in (1) and b_ϕ in (2); and its derivative satisfies $\phi'(1) = 0$.*

Given a Legendre-type generator ϕ , it follows from (5) that the interior of the domain of ϕ^* is $(-\infty, \beta_\phi)$, and $\phi^{*'} : (-\infty, \beta_\phi) \rightarrow (0, +\infty)$ is a one-to-one strictly increasing map, the inverse of which is $\phi' : (0, +\infty) \rightarrow (-\infty, \beta_\phi)$. Note that a Legendre-type generator ϕ is always positive, as it is convex and verifies $\phi'(1) = 0$, thus its minimum is achieved at 1, and the function satisfies $\phi(1) = 0$. Moreover, note that the condition $\phi(1) = \phi'(1) = 0$ is not restrictive, as ϕ -divergences are invariant up to an affine transformation [21]. A comprehensive classification of the common ϕ -divergences generated by a Legendre-type generator is available in [22, Tables 1 and 2]. The following definition specifies the notion of ϕ -divergence adopted throughout this work.

Definition 3. *Given a Legendre-type generator ϕ and two probability measures P and Q defined on the same measurable space, the ϕ -divergence of P with respect to Q is*

$$D_\phi(P||Q) = \begin{cases} \int \phi\left(\frac{dP}{dQ}(\theta)\right) dQ(\theta), & \text{if } P \ll Q; \\ +\infty, & \text{otherwise,} \end{cases} \quad (8)$$

where $\frac{dP}{dQ}$ is the Radon-Nikodym derivative of P with respect to Q .

Definition 3 is not standard when the function ϕ is not assumed to be superlinear (i.e., $\lim_{x \rightarrow +\infty} \phi(x)/x = +\infty$). In the typical definition, certain probability measures may yield a finite ϕ -divergence value even when they are not absolutely continuous with respect to the reference measure; see [20], [23], [24]. However, when a ϕ -divergence is used as a regularization term in an optimization problem, allowing non-absolutely continuous measures may lead to situations in which the optimizer no longer admits a closed-form expression. This is due to the possible presence of a non-zero singular part, which typically cannot be expressed in closed form, as illustrated in [12, Theorem 3].

II. PROBLEM FORMULATION

This section establishes the mathematical formulation of optimal transport regularized by a ϕ -divergence. The primal problem is introduced, as well as its dual counterpart, and the necessary functional analytic tools. Throughout this section, the function ϕ is a Legendre-type generator (Definition 2). The measure $P_{\mathcal{X}\mathcal{Y}}$ in $\Delta(\mathcal{X} \times \mathcal{Y})$ is assumed to exhibit given marginal measures denoted by $P_X \in \Delta(\mathcal{X})$ and $P_Y \in \Delta(\mathcal{Y})$.

The constant λ is a strictly positive real number, and the function $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ is a bounded Borel measurable cost function. These elements constitute the parameters of the optimal transport problem with ϕ -divergence regularization, as presented in [11] and [12].

A. Primal Problem

The OT problem with ϕ -divergence regularization is formulated in [11] and [12] as

$$\inf_{P_{XY} \in \Pi(P_X, P_Y)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) dP_{XY}(x, y) + \lambda D_\phi(P_{XY} \| P_X P_Y). \quad (9)$$

From Definition 3, the divergence is infinite whenever P_{XY} is not absolutely continuous with respect to the product of its marginal measures, P_X and P_Y . Consequently, the analysis is restricted to absolutely continuous couplings. This allows the primal problem to be expressed directly in terms of the Radon-Nikodym derivative of P_{XY} with respect to the product measure $P_X P_Y$, which is a non-negative function in $L^1(\mathcal{X} \times \mathcal{Y})$. Hence,

$$\mathcal{P} \triangleq \inf_{r \geq 0} \int_{\mathcal{X} \times \mathcal{Y}} \left(c(x, y) r(x, y) + \lambda \phi(r(x, y)) \right) dP_X P_Y \quad (10a)$$

$$\text{s.t.} \quad \int_{\mathcal{Y}} r(\cdot, y) dP_Y(y) \stackrel{a.s.}{=} 1 \quad (10b)$$

$$\int_{\mathcal{X}} r(x, \cdot) dP_X(x) \stackrel{a.s.}{=} 1, \quad (10c)$$

where (10b) and (10c) enforce the marginal constraints on the Radon-Nikodym derivative of P_{XY} with respect to $P_X P_Y$; and r is a non-negative function in $L^1(\mathcal{X} \times \mathcal{Y})$.

B. Dual Problem

The dual problem is derived in [22, Section 2.2] using standard convex analysis tools, and is formulated over the product space $L^\infty(\mathcal{X}, P_X) \times L^\infty(\mathcal{Y}, P_Y)$ as follows

$$\mathcal{D} \triangleq \sup_{f, g} \int_{\mathcal{X}} f(x) dP_X(x) + \int_{\mathcal{Y}} g(y) dP_Y(y) - \lambda \int_{\mathcal{X} \times \mathcal{Y}} \phi^* \left(\frac{f(x) + g(y) - c(x, y)}{\lambda} \right) dP_X P_Y, \quad (11)$$

where the function ϕ^* is the convex conjugate of the function ϕ , as defined in [19, Section 2.3].

III. EXISTENCE AND OPTIMALITY PROPERTIES

A. Attainment of the dual problem

The existence of optimal potentials for the dual problem in (11) is established by extending the results of [11] and [12] to Polish spaces and bounded cost functions. This result follows from constructing maximizing sequences through alternating coordinate-wise updates [22, Section 2.4.1].

Theorem 1. *The supremum in problem (11) is attained for a unique pair of potentials (f_0, g_0) up to the transformation $(f, g) \mapsto (f + a, g - a)$, for $a \in \mathbb{R}$.*

Proof: The proof is presented in [22]. ■

The uniqueness is stated “up to a constant” because the dual functional depends on the potentials primarily through their sum $f(x) + g(y)$. Consequently, adding a scalar a to f and subtracting it from g leaves the objective value unchanged.

B. Characterization of dual optimizers

The characterization of optimal dual potentials relies on first-order optimality conditions obtained by differentiating the dual objective in (11). This requires differentiating the integral term involving the convex conjugate ϕ^* . Since the derivative $\phi^{* \prime}$ blows up as its argument approaches the boundary β_ϕ (Lemma 1), differentiation under the integral sign is only justified when the argument of ϕ^* remains uniformly bounded away from this singularity. To ensure this regularity, we restrict the analysis to a subset of admissible potentials on which the dual objective is differentiable. Specifically, define the set

$$\mathcal{A}_{\phi, c} \triangleq \{(f, g) \in L^\infty(\mathcal{X}, P_X) \times L^\infty(\mathcal{Y}, P_Y) : \exists \gamma < \beta_\phi, \text{Ran} \left(\frac{f \oplus g - c}{\lambda} \right) \stackrel{a.s.}{=}_{P_X P_Y} (-\infty, \gamma)\}. \quad (12)$$

Within this set, the first-order optimality conditions are well-defined and admit a simple characterization, as stated in the following lemma.

Lemma 2. *A pair of functions $(\hat{f}, \hat{g}) \in \mathcal{A}_{\phi, c}$, where $\mathcal{A}_{\phi, c}$ is defined in (12), are maximizing potentials for the dual problem in (11) if and only if:*

$$\int_{\mathcal{Y}} \phi'^{-1} \left(\frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda} \right) dP_Y(y) \stackrel{a.s.}{=}_{P_X} 1; \text{ and} \quad (13)$$

$$\int_{\mathcal{X}} \phi'^{-1} \left(\frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda} \right) dP_X(x) \stackrel{a.s.}{=}_{P_Y} 1. \quad (14)$$

Proof: The proof is presented in [22]. ■

C. Attainment of the primal problem

As shown in [12, Theorem 3], an optimal solution to (9) may not exist unless Definition 3 is extended to assign finite ϕ -divergence to measures admitting a non-zero singular part with respect to $P_X P_Y$. Under this extension, the optimal coupling itself may contain a singular component, which poses significant practical difficulties: this component has no closed-form representation and precludes the use of the Sinkhorn algorithm. Consequently, additional assumptions are required to guarantee the existence of a computable solution, as shown hereunder.

Theorem 2. *Suppose that there exists a pair of optimal potentials (\hat{f}, \hat{g}) in $\mathcal{A}_{\phi, c}$ for the dual problem in (11). Then, strong duality holds between the optimization problems (10) and (11), and the unique minimizer of the primal problem satisfies*

$$\frac{dP_{XY}^{(c, \phi)}}{dP_X P_Y}(x, y) \stackrel{a.s.}{=}_{P_X P_Y} \phi^{* \prime} \left(\frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda} \right). \quad (15)$$

Proof: The proof is presented in [22]. ■

D. Sufficient conditions for primal attainment

The validity of Theorem 2 relies on the assumption that there exists a pair of optimal potentials in $\mathcal{A}_{\phi,c}$. Intuitively, this condition requires that the optimal potentials stay uniformly bounded away from the singularity of the dual function ϕ^* . While verifying this property directly for general Polish spaces can be challenging, it holds automatically in several significant scenarios commonly encountered in applications. This section explores sufficient conditions ensuring this requirement is met.

The first sufficient condition concerns divergences with superlinear growth, such as the Kullback-Leibler divergence. In this setting, the domain of the conjugate function ϕ^* extends to the entire real line, effectively removing the boundary singularity.

Lemma 3. *If β_ϕ in (4) is equal to $+\infty$, any pair of optimal potentials is in $\mathcal{A}_{\phi,c}$.*

Proof: The proof is presented in [22]. ■

The second sufficient condition addresses the case of discrete OT, where the underlying spaces are finite sets. Here, the finiteness of the space prevents the values of the potentials from accumulating asymptotically against the boundary β_ϕ .

Lemma 4. *If the sets \mathcal{X} and \mathcal{Y} are both finite, then any pair of optimal potentials is in $\mathcal{A}_{\phi,c}$.*

Proof: The proof is presented in [22]. ■

IV. EQUIVALENCE OF THE PROBLEM TO REGULARIZATION ON THE FAMILY OF ϕ -DIVERGENCES

A. Main results

The following theorem constitutes the main result of this work. Intuitively, it states that the choice of divergence in regularized OT is not intrinsic to the problem: any effect of the regularization on the optimal coupling can equivalently be attributed to a modification of the transport cost. The result and its proof have been adapted from [16, Theorem 2], which develops in the context of empirical risk minimization with ϕ -divergence regularization. Although there are fundamental technical differences due to the presence of additional constraints in the case of optimal transport, the two results convey the same meaning in their corresponding scenarios.

Theorem 3. *Assume that the problem in (11) admits a pair of optimal potentials $(\hat{f}, \hat{g}) \in \mathcal{A}_{\phi,c}$, and let ψ be a Legendre-type generator (Definition 2). Then, there exists a bounded function $v : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ such that the problem in (9) is equivalent, in the sense that both problems admit the same minimizer, to the following optimization problem:*

$$\inf_{P_{XY} \in \Pi(P_X, P_Y)} \int_{\mathcal{X} \times \mathcal{Y}} v(x, y) dP_{XY}(x, y) + \lambda D_\psi(P_{XY} \| P_X P_Y). \quad (16)$$

Proof: Let $(\hat{f}, \hat{g}) \in \mathcal{A}_{\phi,c}$ be maximizing potentials for the dual problem in (11). Define the cost function v as

$$v(x, y) = -\lambda \psi' \left(\phi'^{-1} \left(\frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda} \right) \right) + 1. \quad (17)$$

Given that the functions c , \hat{f} and \hat{g} are bounded and that the range of $(x, y) \mapsto \frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda}$ is bounded and contained in some interval of the form $(-\infty, \gamma)$, with $\gamma < \beta_\phi$, the range of $(x, y) \mapsto \phi'^{-1} \left(\frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda} \right)$ is bounded and contained in an interval of the form $(\rho, +\infty)$ with $\rho > 0$. Since ψ' is C^1 on $(0, \infty)$, the function

$$(x, y) \mapsto \psi' \left(\phi'^{-1} \left(\frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda} \right) \right) \quad (18)$$

is also bounded. Thus, the function v is bounded, ensuring that problem (16) fits within the setting established in previous sections. Let $\check{f} \in L^\infty(\mathcal{X}, P_X)$ and $\check{g} \in L^\infty(\mathcal{Y}, P_Y)$ be defined by

$$\check{f}(x) = \frac{1}{2}, \quad \forall x \in \mathcal{X}; \quad \text{and} \quad \check{g}(y) = \frac{1}{2}, \quad \forall y \in \mathcal{Y}. \quad (19)$$

The proof proceeds by demonstrating that the functions \check{f}, \check{g} satisfy (13) and (14) for the problem in (16). Verification proceeds as follows: let $x \in \mathcal{X}$ be fixed, and compute the integral

$$\begin{aligned} & \int_{\mathcal{Y}} \psi'^{-1} \left(\frac{\check{f}(x) + \check{g}(y) - v(x, y)}{\lambda} \right) dP_Y(y) \\ &= \int_{\mathcal{Y}} \psi'^{-1} \left(\frac{1 + \lambda \psi'(\phi'^{-1}(\frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda})) - 1}{\lambda} \right) dP_Y(y) \quad (20) \end{aligned}$$

$$= \int_{\mathcal{Y}} \phi'^{-1} \left(\frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda} \right) dP_Y(y). \quad (21)$$

Since $(\hat{f}, \hat{g}) \in \mathcal{A}_{\phi,c}$ is the solution to (11), Lemma 2 implies

$$\int_{\mathcal{Y}} \phi'^{-1} \left(\frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda} \right) dP_Y(y) \stackrel{a.s.}{=} 1. \quad (22)$$

Combining (21) and (22) yields

$$\int_{\mathcal{Y}} \psi'^{-1} \left(\frac{\check{f}(x) + \check{g}(y) - v(x, y)}{\lambda} \right) dP_Y(y) \stackrel{a.s.}{=} 1. \quad (23)$$

Moreover, applying the same reasoning for a fixed $y \in \mathcal{Y}$ yields

$$\int_{\mathcal{X}} \psi'^{-1} \left(\frac{\check{f}(x) + \check{g}(y) - v(x, y)}{\lambda} \right) dP_X(x) \stackrel{a.s.}{=} 1. \quad (24)$$

From Lemma 2 (\hat{f}, \hat{g}) are optimal potentials for the dual of (16), which in turn requires verifying that $(\check{f}, \check{g}) \in \mathcal{A}_{\psi,v}$. This set is defined analogously to $\mathcal{A}_{\phi,c}$ in (12), by replacing the parameters ϕ and c with ψ and v respectively. This condition requires that the range of $l : (x, y) \mapsto \frac{\hat{f}(x) + \hat{g}(y) - v(x, y)}{\lambda}$ satisfies $\text{Ran}(l) \stackrel{a.s.}{\subset} (-\infty, \gamma)$ for some $\gamma < \beta_\psi$. Let $(x, y) \in \mathcal{X} \times \mathcal{Y}$ be fixed, and consider

$$\frac{\check{f}(x) + \check{g}(y) - v(x, y)}{\lambda} = \psi' \left(\phi'^{-1} \left(\frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda} \right) \right). \quad (25)$$

Since the pair (\hat{f}, \hat{g}) is in the set $\mathcal{A}_{\phi,c}$ defined in (12), and due to continuity and strict monotonicity of ϕ'^{-1} on

$(-\infty, \beta_\phi)$ (see Section I-B), the range of the function $(x, y) \mapsto \phi'^{-1}\left(\frac{\hat{f}(x)+\hat{g}(y)-c(x,y)}{\lambda}\right)$ is bounded and satisfies

$$\text{Ran}\left(\phi'^{-1}\left(\frac{\hat{f}(x)+\hat{g}(y)-c(x,y)}{\lambda}\right)\right) \stackrel{\text{a.s.}}{\subset}_{P_X P_Y} (\rho, +\infty), \quad (26)$$

for some $\rho > 0$. Given that ψ' is continuous and strictly increasing on $(0, +\infty)$, the range of $(x, y) \mapsto \psi'\left(\phi'^{-1}\left(\frac{\hat{f}(x)+\hat{g}(y)-c(x,y)}{\lambda}\right)\right)$ satisfies

$$\text{Ran}\left(\psi'\left(\phi'^{-1}\left(\frac{\hat{f}(x)+\hat{g}(y)-c(x,y)}{\lambda}\right)\right)\right) \stackrel{\text{a.s.}}{\subset}_{P_X P_Y} (-\infty, \gamma), \quad (27)$$

for some $\gamma < \beta_\psi$. Therefore, by (25), (\check{f}, \check{g}) belongs to $\mathcal{A}_{\psi, v}$ and satisfies (23) and (24). Lemma 2 implies that (\check{f}, \check{g}) are maximizing potentials for the dual problem of (16).

Finally, Theorem 2 gives the form of the solution of the problem in (16), denoted by $P_{XY}^{(v, \psi)}$ as

$$\frac{dP_{XY}^{(v, \psi)}}{dP_X P_Y}(x, y) = \psi'^{-1}\left(\frac{\check{f}(x) + \check{g}(y) - v(x, y)}{\lambda}\right) \quad (28)$$

$$= \psi'^{-1}\left(\psi'\left(\phi'^{-1}\left(\frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda}\right)\right)\right) \quad (29)$$

$$= \phi'^{-1}\left(\frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda}\right) = \frac{dP_{XY}^{(c, \phi)}}{dP_X P_Y}(x, y), \quad (30)$$

where $P_{XY}^{(c, \phi)}$ denotes the solution to the problem in (10). Thus, the two problems share the same minimizer, which completes the proof. ■

B. Examples

This section illustrates the structural equivalence established in Theorem 3 by considering fixed Polish spaces \mathcal{X} and \mathcal{Y} , a fixed bounded cost function c , and a fixed regularization parameter $\lambda > 0$. Let ϕ denote the generator of the Kullback–Leibler (KL) divergence, defined by $\phi(x) = x \log(x) - x + 1$. For this KL generator, the inverse of its derivative is given by

$$\phi'^{-1}(y) = \exp(y). \quad (31)$$

Moreover, the quantity β_ϕ in (4) is equal to $+\infty$. Therefore, it follows from Lemma 3 that the corresponding optimal potentials for the dual problem in (11), denoted by \hat{f} and \hat{g} , belong to the set $\mathcal{A}_{\phi, c}$ in (12). As a consequence, all the assumptions of Theorem 3 are satisfied for any target Legendre-type generator ψ satisfying Definition 2. In particular, for the transformed cost function defined in (17), the problems in (9) and (16) admit the same unique minimizer. The following examples illustrate the resulting transformed cost function v obtained via (17) for various choices of target regularization.

1) *From KL to reverse KL regularization:* Consider the target regularization defined by the reverse KL generator $\psi(x) = -\log(x) + x - 1$, whose derivative is $\psi'(x) = 1 - \frac{1}{x}$. From (17) and (31), the transformed cost function v ensuring that the initial problem (9) and the transformed problem (16) share the same unique minimizer is given by

$$v(x, y) = -\lambda \psi'\left(\exp\left(\frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda}\right)\right) + 1 \quad (32)$$

$$= -\lambda \left(1 - \exp\left(-\frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda}\right)\right) + 1 \quad (33)$$

$$= \lambda \exp\left(\frac{c(x, y) - \hat{f}(x) - \hat{g}(y)}{\lambda}\right) - \lambda + 1. \quad (34)$$

2) *From KL to Jensen-Shannon regularization:* In the case of a transition to Jensen-Shannon regularization, the target generator is $\psi(x) = x \log(x) - (x + 1) \log\left(\frac{x+1}{2}\right)$, with derivative $\psi'(x) = \log\left(\frac{2x}{x+1}\right)$. The transformed cost function v preserving the minimizer in that case is expressed as

$$v(x, y) = -\lambda \log\left(\frac{2 \exp\left(\frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda}\right)}{\exp\left(\frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda}\right) + 1}\right) + 1 \quad (35)$$

$$= \lambda \log\left(\frac{\exp\left(\frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda}\right) + 1}{2 \exp\left(\frac{\hat{f}(x) + \hat{g}(y) - c(x, y)}{\lambda}\right)}\right) + 1. \quad (36)$$

By simplifying the argument within the logarithm, the relation becomes

$$v(x, y) = \lambda \log\left(\frac{1 + \exp\left(\frac{c(x, y) - \hat{f}(x) - \hat{g}(y)}{\lambda}\right)}{2}\right) + 1. \quad (37)$$

These examples confirm that changing the regularizing divergence manifests as a non-linear adjustment of the cost function, ensuring that the unique minimizer of the regularized optimal transport problem remains unchanged.

V. CONCLUSIONS

This work investigates ϕ -divergence regularized optimal transport in Polish spaces with bounded cost functions. The existence of optimal dual potentials and unique primal couplings is established, providing closed-form expressions for the latter. The central contribution, Theorem 3, demonstrates a structural equivalence: a change in the regularization divergence can be compensated by a non-linear transformation of the cost function while preserving the same unique minimizer. This result remains primarily theoretical, since the transformed cost v in (17) explicitly depends on the dual potentials (\hat{f}, \hat{g}) of the original problem. Thus, it does not reduce the computational complexity of solving regularized OT problems. Nonetheless, it provides a unifying framework for regularized OT and offers new perspectives for the characterization and sampling of optimal couplings.

REFERENCES

- [1] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf, “Wasserstein auto-encoders,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, May 2018.
- [2] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN,” in *Proceedings of the International Conference on Machine Learning (ICML)*, Jul. 2017, pp. 214–223.
- [3] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, “Optimal transport for domain adaptation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1852–1865, Sep. 2017.
- [4] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. Poggio, “Learning with a Wasserstein loss,” in *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2*, Dec. 2015, pp. 2053–2061.
- [5] J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas, “Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains,” *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, pp. 1–11, Jul. 2015.
- [6] L. M. Pereira and M. H. Amini, “A survey on optimal transport for machine learning: Theory and applications,” *IEEE Access*, vol. 13, pp. 26 506–26 526, Jan. 2025.
- [7] C. Villani, *Optimal Transport: Old and New*, 1st ed. Berlin, Heidelberg: Springer-Verlag, 2009.
- [8] G. Peyré and M. Cuturi, *Computational Optimal Transport*, 1st ed. Hanover, MA, USA: Foundations and Trends in Machine Learning, 2020.
- [9] M. Blondel, V. Seguy, and A. Rolet, “Smooth and sparse optimal transport,” in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, Mar. 2018, pp. 314–323.
- [10] M. Cuturi, “Sinkhorn distances: Lightspeed computation of optimal transportation distances,” in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 2, Dec. 2013, pp. 2292–2300.
- [11] S. Di Marino and A. Gerolin, “Optimal transport losses and sinkhorn algorithm with general convex regularization,” *arXiv preprint arXiv:2007.00976*, Jul. 2020.
- [12] D. Terjék and D. González-Sánchez, “Optimal transport with f-divergence regularization and generalized sinkhorn algorithm,” in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 151, Mar. 2022, pp. 5135–5165.
- [13] O. Catoni, *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, 1st ed. Beachwood, OH, USA: Institute of Mathematical Statistics, 2007, vol. 56.
- [14] S. M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, “Empirical risk minimization with relative entropy regularization,” *IEEE Transactions on Information Theory*, vol. 70, no. 7, pp. 5122–5161, Jul. 2024.
- [15] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, “Empirical risk minimization with relative entropy regularization Type-II,” INRIA, Centre Inria d’Université Côte d’Azur, Sophia Antipolis, France, Tech. Rep. RR-9508, May 2023.
- [16] —, “Equivalence of empirical risk minimization to regularization on the family of f -divergences,” in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Jul. 2024, pp. 759–764.
- [17] —, “Empirical risk minimization with f-divergence regularization in statistical learning,” INRIA, Centre Inria d’Université Côte d’Azur, Sophia Antipolis, France, Tech. Rep. RR-9521, Oct. 2023.
- [18] —, “Asymmetry of the relative entropy in the regularization of empirical risk minimization,” *IEEE Transactions on Information Theory*, vol. 71, no. 8, pp. 6198–6226, Aug. 2025.
- [19] C. Zalinescu, *Convex analysis in general vector spaces*, 1st ed. Singapore: World Scientific, 2002.
- [20] J. M. Borwein and A. S. Lewis, “Partially-finite programming in l_1 and the existence of maximum entropy estimates,” *SIAM Journal on Optimization*, vol. 3, no. 2, pp. 248–267, May 1993.
- [21] E. M. Werner, “f-divergence for convex bodies,” in *Asymptotic Geometric Analysis: Proceedings of the Fall 2010 Fields Institute Thematic Program*. New York, NY: Springer, 2013, pp. 381–395.
- [22] M. Nicaise, Y. Bermudez, and S. M. Perlaza, “On optimal transport with f-divergence regularization,” INRIA, Centre Inria d’Université Côte d’Azur, Sophia Antipolis, France, Tech. Rep. RR-9607, January 2026.
- [23] I. Csiszár, “Information-type measures of difference of probability distributions and indirect observations,” *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 299–318, 1967.
- [24] R. Agrawal and T. Horel, “Optimal bounds between f-divergences and integral probability metrics,” *Journal of Machine Learning Research*, vol. 22, no. 1, pp. 5662–5720, Jan. 2021.