QUARTERLY OF APPLIED MATHEMATICS
VOLUME LXXXIV, NUMBER 4
DECEMBER 2026, PAGES 743–786
https://doi.org/10.1090/qam/1749
Article electronically published on August 19, 2026

PROXIMAL OPTIMAL TRANSPORT DIVERGENCES

By

RICARDO BAPTISTA (*Department of Statistical Sciences, California Institute of Technology, Pasadena California 91125*),

PANAGIOTA BIRMPA (*Actuarial Mathematics & Statistics, School of Mathematical & Computer Sciences, Heriot-Watt University and Maxwell Institute, Edinburgh EH14 4AS, United Kingdom*),

MARKOS A. KATSOULAKIS (*Department of Mathematics & Statistics, University of Massachusetts Amherst, Amherst, Massachusetts 01003*),

LUC REY-BELLET (*Department of Mathematics & Statistics, University of Massachusetts Amherst, Amherst, Massachusetts 01003*),

AND

BENJAMIN J. ZHANG (*Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912*)

We dedicate this paper to Paul Dupuis for his tremendous contributions to his community within mathematics and across science and engineering. His seminal work in mathematical control, applied probability, and Monte Carlo methods, together with his generous mentorship, has profoundly shaped the thinking of his collaborators and mentees. Many of the ideas in this paper trace directly to his visionary work by Dupuis and Mao [ESAIM Control Optim. Calc. Var. 28 (2022), p. 38].

Abstract. We introduce the proximal optimal transport divergence, a novel discrepancy measure that interpolates between information divergences and optimal transport distances via an infimal convolution formulation. This divergence provides a principled foundation for optimal transport proximals and proximal optimization methods frequently used in generative modeling. We explore its mathematical properties, including smoothness, boundedness, and computational tractability, and establish connections

Received August 4, 2025, and, in revised form, April 16, 2026.
2020 *Mathematics Subject Classification.* Primary 94A17, 49Q22, 90C25; Secondary 35Q89, 68T07.
The first author was supported by the von Kármán instructorship at Caltech and a Department of Defense Vannevar Bush Faculty Fellowship (award N00014-22-1-2790) held by Andrew M. Stuart. The third, fourth, and fifth authors were partially funded by AFOSR grant FA9550-21-1-0354. The third and fourth authors were partially funded by NSF DMS-2307115.
Email address: rsb@caltech.edu
Email address: P.Birmpa@hw.ac.uk
Email address: markos@umass.edu
Email address: luc@math.umass.edu
Current address: School of Data Sciences and Society, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27514.
Email address: bjz@unc.edu

©2026 Brown University

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

to primal-dual formulations and adversarial learning. The proximal operator associated with the proximal optimal transport divergence can be interpreted as a transport map that pushes a reference distribution toward the optimal generative distribution, which approximates the target distribution that is only accessible through data samples. Building on the Benamou-Brenier dynamic formulation of classical optimal transport, we also establish a dynamic formulation for proximal OT divergences. The resulting dynamic formulation is a first order mean-field game whose optimality conditions are governed by a pair of nonlinear partial differential equations: a backward Hamilton-Jacobi equation and a forward continuity equation. Our framework generalizes existing approaches while offering new insights and computational tools for generative modeling, distributionally robust optimization, and gradient-based learning in probability spaces.

1. Introduction. Divergences and metrics are tools for measuring the dissimilarity between probability distributions, enabling the comparison of models with data. They are fundamental in data science and machine learning, where tasks such as sampling from unnormalized distributions or generative modeling depend on selecting a suitable discrepancy measure for learning models. The two primary classes of measures for comparing probability distributions are information divergences and optimal transport distances. Information divergences, such as the Kullback-Leibler (KL) and f-divergences, are based on expectations of likelihood-related quantities between distributions. Their ease of computation and optimization through only samples of a target distribution makes them widely used in machine learning. However, these divergences are meaningful only when the distributions share the same support. In contrast, optimal transport distances do not require overlapping supports, making them more flexible in certain applications. However, this advantage comes at the cost of significantly higher computational complexity.

In this paper, we introduce the *proximal optimal transport (OT) divergence* between two probability measures P and Q on some space X and Y, denoted by $\mathfrak{D}_\varepsilon^c$ and defined as an infimal convolution

$$\mathfrak{D}_\varepsilon^c(P\|Q) = \inf_R \left\{T_c(P,R) + \varepsilon\mathfrak{D}(R\|Q)\right\}, \tag{1.1}$$

where the infimum is over all probability distributions R on Y, T_c is the optimal transport cost corresponding to some cost function $c : X \times Y \to [0,\infty)$ and $\mathfrak{D}$ is an information divergence. The positive parameter ε controls the trade-off between the information divergence and the optimal transport cost. Under mild assumptions, this is an information divergence. Infimal convolutions have been previously utilized to define novel forms of divergences based on the dual representation of the 1-Wasserstein distance. For instance, the recently introduced (f,Γ)-divergence, [7,20], combines the dual form of the 1-Wasserstein distance with f-divergences, demonstrating the computational ease and flexibility of f-divergences with the advantages of the 1-Wasserstein distance. This paper extends that concept to encompass general optimal transport distances. Importantly, this work reveals that the proximal OT divergence is the foundational mathematical principle behind optimal transport proximals and proximal optimization methods that are frequently employed in generative modeling. By framing proximal OT divergences in the context of generative modeling, we gain deeper insights that allow for the expansion of

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

definitions, properties, and algorithms, encompassing both primal and dual formulations and extending beyond them. Despite its fundamental role, the proximal OT divergence has not been explicitly introduced as an independent concept until now.

1.1. *Why proximal optimal transport divergences?* Proximal OT divergences possess several appealing mathematical properties that make them particularly useful for comparing data to models or for comparing two datasets, even when their supports lie on different manifolds. These properties enable robust and stable learning of target distributions, especially in settings where the data is supported on lower-dimensional spaces. We outline the key properties and inherited computational benefits of the divergences below.

- **Proximal OT divergences are finite.** The divergences are bounded above by the optimal transport distance and the information divergences, i.e.,

$$\mathfrak{D}^c_\varepsilon(P\|Q) \leq \min\left\{T_c(P,Q), \varepsilon\mathfrak{D}(P\|Q)\right\}. \tag{1.2}$$

  This bound ensures finiteness even in the absence of absolute continuity, which enhances the suitability of the divergences for practical applications. For instance, in generative modeling where P represents the generative distribution and Q is the unknown target distribution accessible only through samples, the proximal OT divergence remains finite. In contrast, any information divergence relying on the likelihood ratio dP/dQ may be infinite in such settings as P and Q can be mutually singular.
- **Transport and re-weighting property.** The interpolation via the infimal convolution given in (1.1) can be interpreted as a two-step procedure. The first step involves optimally transporting the probability distribution P into an intermediate probability measure R that is absolutely continuous with respect to Q with an associated cost given by $T_c(P,R)$. In the second step, R is re-distributed to Q using the information divergence, incurring a cost $\mathfrak{D}(R\|Q)$. The optimal intermediate measure denoted by R^* minimizes the total cost and determines the optimal proximal OT divergence in (1.1). This property makes them powerful tools for learning manifolds [32], as evidenced by related proximal concepts [7, 8, 43, 60].
- **Smoothness.** Another key property is that for uniformly continuous cost functions c and f-divergences in (1.1), the variational derivative of $\mathfrak{D}^c_{f,\varepsilon}(P\|Q)$ with respect to P remains well-defined without requiring absolutely continuous perturbations relative to Q (e.g. singular ones), which is primarily treated as the target distribution in this paper. This contrasts with classical KL or f-divergences, which impose such restrictions. This aspect is critical, as variational derivatives play a central role in machine learning tasks, such as the construction of loss functions, generative adversarial networks (GANs), optimal transport methods, and Wasserstein gradient flows, ensuring stability under general conditions.
- **Transport optimal operator.** Under mild assumptions, we establish uniqueness of the optimal intermediate measure R^* solving (1.1). We formulate the intermediate measure as the output of the following *transport proximal operator*

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

acting on the measure P, i.e.,

$$R^* = \mathbf{prox}^c_{\varepsilon\mathfrak{D}(\cdot\|Q)}(P) \coloneqq \operatorname*{argmin}_{R\in\mathcal{P}(X)} \left\{T_c(P,R) + \varepsilon\mathfrak{D}(R\|Q)\right\}. \tag{1.3}$$

This formulation aligns with the broader concept of proximal operators [41], reinforcing its theoretical significance and practical utility in generative modeling [60]. The uniqueness of R^* and the transport proximal operator are particularly relevant for generative modeling, where the objective is to approximate an unknown data distribution Q that is prescribed using only samples. A common approach is to start with a reference distribution P that is easy to sample from, and then to optimally transport it to obtain R^*; see second bullet above. When ε is large, the divergence will dominate in (1.3) and $R^* \approx Q$. Thus, the distribution R^* can serve as a generative model to approximate the data distribution. This perspective was used in [22,26,40] to learn a regularized neural ODEs for the generative model by using the proximal operator in (1.3) with the Kullback-Leibler divergence for $\mathfrak{D}$.

- **Computational aspects.** The proximal OT divergences have primal and dual formulations that yield computational tools for evaluating the divergence. In particular, the dual formulation offers a variational representation involving test functions. For example, when the information divergence is the KL divergence, the dual representation is given by

$$\mathfrak{D}^c_{\mathrm{KL},\varepsilon}(P\|Q) = \sup_{\substack{(\phi,\psi)\in C_b(X\times X),\\ \phi(x)+\psi(y)\le c(x,y)}} \left\{\mathbb{E}_P[\phi] - \varepsilon\ln\mathbb{E}_Q[e^{-\psi/\varepsilon}]\right\}, \tag{1.4}$$

where $C_b(X\times Y)$ is the set of continuous and bounded functions on $X\times Y$. This representation introduces discriminators, drawing an analogy to adversarial learning methods [3,24,38]. From a computational perspective, the dual formulation enables efficient implementations using convex neural networks [6]. Moreover, the computation of the variational derivative of proximal OT divergences involves the discriminator, defined as the optimal ϕ^* in (1.4), which allows us to build gradient flows of the proximal OT divergence, and the associated gradient descent algorithm in probability space [16,23,25].
- **Dynamic formulation of proximal OT divergences.** The primal formulation along with the Benamou-Brenier formulation [5] can be interpreted through the lens of mean-field games (MFG) [29], in which the MFG framework provides an analogue Benamou-Brenier formula for proximal OT divergences. While deeper connections to Hamilton-Jacobi-Bellman (HJB) theory and MFG-based theoretical analysis could provide further insights, we do not fully explore them here. The MFG formulation enables the use of generative modeling tools [59], to compute the proximal OT divergence and optimal intermediate measure in the primal formulation (1.1), such as generative flows like neural ODEs [14] and OT-flows [40].

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

1.2. *Related work.* Special cases of the proximal OT divergences have already been extensively analyzed in both theoretical and computational contexts. In particular, when T_c is the 1-Wasserstein distance (i.e., $c(x, y) = d(x, y)$) and $\mathfrak{D}$ is a f or Rényi divergence, the proximal OT divergence recovers the (f, Γ)-divergence where Γ is the set of all 1-Lipschitz functions or the 1-Wasserstein regularized Rényi divergence, that were introduced in [7, 20] and [10], respectively. In general, these divergences interpolate between integral probability metrics (IPMs) over function spaces Γ, and f and Rényi divergences. An interesting case is the maximum mean discrepancy (MMD) regularized f or Rényi divergence where Γ is the unit ball in a reproducing kernel Hilbert space (RKHS), which was studied in [7] and [10] and relates to early foundational work on the dual representation of MMDs developed in [36]. Another closely related class of divergences to proximal OT divergences is the Moreau-Yoshida f-divergences developed in [53].

All of the aforementioned divergences, both in their general form and specific cases, have been extensively used as objective functionals in generative adversarial networks (GANs) and gradient flows, showing their ability for stable training across a broad range of target distributions [9, 10, 23, 25], including those with heavy tails [16] and those supported on a lower-dimensional manifold. In particular, [26] applies the combination of 1-Wasserstein and 2-Wasserstein proximals to stably learn distributions supported on manifolds. Optimal transport proximals have also been instrumental in numerical methods for nonlinear partial differential equations and gradient flows, notably in the celebrated Jordan-Kinderlehrer-Otto (JKO) scheme [28], which employs the Wasserstein proximal operator of the KL divergence. For recent advancements in numerical JKO-based schemes, see [13]. More recently, optimal transport proximals have attracted significant interest in generative modeling, particularly in the use of 2-Wasserstein regularized KL divergences for training continuous normalizing flows (CNFs) [22, 40]. Approximations of 2-Wasserstein proximals for both linear and nonlinear energy functionals have also been explored in [30], while score-based generative models have been shown to correspond to the 2-Wasserstein proximal of cross-entropy [59, 60]. As proximal OT divergences can be naturally interpreted as a form of entropic regularization, they share strong connections with Schrödinger bridge problems [39] and various other frameworks based on entropic regularization of optimal transport problems [15, 18, 42].

Distributionally robust optimization (DRO) [45] is another related field, which relies on constructing appropriate ambiguity sets to ensure robustness against distributional uncertainty. Common approaches for modeling misspecification include likelihood-based methods, which use information divergences to define the ambiguity set, and distance-based methods, which use Wasserstein distances. Recent advancements have sought to bridge these two perspectives. A novel method in [11] extends OT-DRO by incorporating conditional moment constraints, while [8] introduces optimal-transport-regularized divergences. The divergences have been employed to enhance the adversarial robustness of deep learning models for applications in privacy and robustness against adversarial attacks. These developments align with special cases of our framework, particularly the finite-dimensional f-divergence and KL reformulations. Particular forms of proximal OT divergences have also been applied in Bayes-PAC analysis [27], resulting in the tractable bounds presented in [55].

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

1.3. *Main results.* Our main contributions can be summarized as follows. For each $\varepsilon > 0$:

- **Dynamic/mean-field-game formulation.** Theorem 2.9 shows that when the transport cost is a p-Wasserstein distance, the proximal OT divergence admits a dynamic formulation similar to the Benamou-Brenier formula. The resulting system is a first order mean-field game, in which the information divergence $\mathfrak{D}(\rho_T \| Q)$ plays the role of a relaxed terminal constraint:

$$\mathfrak{D}^p_\varepsilon(P\|Q) = \min_{v,\rho} \left\{ \varepsilon \mathfrak{D}(\rho_1\|Q) + \int_0^1 \mathbb{E}_{\rho_t}\left[|v_t|^p\right] dt : \begin{array}{c} \partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0 \\ \rho_0 = P \end{array} \right\}.$$

  The optimality conditions form a coupled backward Hamilton-Jacobi and forward continuity system of PDEs, connecting proximal OT divergences to HJB theory and to recent generative modeling frameworks.
- **Duality and discriminator representation.** Theorems 3.1 and 3.3 show that proximal OT divergences admits a dual variational representation (1.4) over pairs of test functions (ϕ, ψ) constrained by $\phi \oplus \psi \le c$. The optimal ψ^* determines the proximal minimizer R^* through an explicit Gibbs formula, which describes a transport and re-weighting interpretation for proximal OT divergences and enables computation via input-convex neural networks.
- **Variational derivative as the discriminator.** Theorem 4.1 shows that the first variation of proximal OT divergence $\delta \mathfrak{D}^c_{\mathrm{KL},\varepsilon}/\delta P$ is given by an extension of the optimal potential arising from the divergence's dual formulation. This variational derivative is well-defined along singular perturbations of p and is the analytical foundation for stable gradient-based learning and gradients flows in probability space developed in Section 6.

Organization. The remainder of this paper is organized as follows. In Section 2 we present the definition and properties of the proximal OT divergence, showing how it interpolates between transport distances and divergences. Section 3 uses duality to provide a variational representation of the divergence, and Section 4 leverages the variational representation to compute the first variation of the divergence with respect to perturbations of one input measure. Section 5 provides the formulations of algorithmic approaches to compute the proximal OT divergence, and Section 6 shows connections to gradient flows. The proofs for all results are relegated to Appendices A–E. Explicit calculations of the proximal OT divergences for univariate and multivariate Gaussian distributions are in Appendix F.

Notation. Let X be a Polish space (i.e., a complete, separable metric space), equipped with Borel σ-algebra $\mathcal{B}_X$ and metric $d : X \times X \to [0, \infty)$. We denote $\mathcal{M}(X)$ to be the set of all finite (signed) Borel measures on X and $\mathcal{P}(X)$ to be the set of all probability measures on X. If X and Y are two Polish spaces, and $P \in \mathcal{P}(X)$, $Q \in \mathcal{P}(Y)$, then $\Pi(P, Q)$ is the set of all couplings between P and Q, that is, the set of probability measures $\pi \in \mathcal{P}(X \times Y)$ such that $\pi(A \times Y) = P(A)$ and $\pi(X \times B) = Q(B)$ for all Borel sets $A \in \mathcal{B}_X$, $B \in \mathcal{B}_Y$. We denote $C(X)$ to be the set of continuous functions and

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

$C_b(X)$ to be the set of bounded continuous functions on X, respectively. For functions $\phi \in C(X)$, $\psi \in C(Y)$, denote $\phi \oplus \psi \coloneqq \phi(x) + \psi(y) \in C(X \times Y)$.

2. Proximal optimal transport divergences. We now formally introduce the proximal optimal transport (OT) divergences in Section 2.1 and establish some key elementary properties that make them useful in computational applications in Sections 2.2 and 2.3. Building on the Benamou-Brenier dynamic formulation of optimal transport, we also establish a dynamic formulation in Subsection 2.4 for proximal OT divergences. The resulting dynamic formulation is a first order *mean-field game* [29] whose optimality conditions are governed by a pair of nonlinear partial differential equations, a backward Hamilton-Jacobi (HJ) equation and a forward continuity equation. In this section, we consider general information divergences. In subsequent sections, our results focus specifically on the Kullback-Leibler divergence, though the results extend analogously to any f-divergence.

2.1. *Definition of the proximal OT divergence.* A proximal OT divergence interpolates between an optimal transport distance and an information divergence via an infimal convolution formulation, preserving the advantageous properties of both. An information divergence $\mathfrak{D}(P\|Q)$ provides a notion of "distance" between probability distributions, satisfying $\mathfrak{D}(P\|Q) \geq 0$ and $\mathfrak{D}(P\|Q) = 0$ if and only if $P = Q$. However, they are generally not considered probability metrics, as they violate symmetry and the triangle inequality. Examples of information divergences include Kullback-Leibler (KL) divergence, and more generally, f-divergences.

Given a cost function $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$, an optimal transport cost (see e.g. [47, 56, 57]) between distributions P and Q is defined as

$$T_c(P, Q) = \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_\pi[c], \tag{2.1}$$

where the infimum is taken over all couplings $\pi \in \Pi(P, Q)$, defined as the set of distributions with marginals P and Q. Therefore, this formulation seeks the optimal coupling π that transports P to Q with minimal average cost with respect to c. The definition allows for X and Y to be different spaces, though they are often the same. For simplicity, we will assume that $X = Y$ going forward. Let (X, d) be a metric space, then a special class of optimal transport distances is the p-Wasserstein distances denoted by $W_p(P, Q)$ for $p \geq 1$, in which $c(x, y) = d(x, y)^p$ (note that when $X = \mathbb{R}^d$, $c(x, y) = |x - y|_p^p$). That is,

$$W_p(P, Q) = \inf_{\pi \in \Pi(P,Q)} \left(\int_{\Omega \times \Omega} d(x, y)^p \, d\pi(x, y) \right)^{\frac{1}{p}}, \tag{2.2}$$

where $P, Q \in \mathcal{P}_p(X) = \left\{ P \in \mathcal{P} : \int d(x, x_0)^p dP < \infty \right\}$, and x_0 is an arbitrary point in X. The metric space $\mathcal{P}_p(X)$ endowed with $W_p(P, Q)$ is called the p-Wasserstein space. The proximal OT divergence is defined as the *infimal convolution* of the optimal transport distance with the information divergence as given below.

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

DEFINITION 2.1. Let $c : X \times Y \to [0, \infty)$ be a cost function, $\mathfrak{D}(\cdot\|\cdot)$ be an information divergence and $\varepsilon > 0$. The *proximal optimal transport divergence* of $P \in \mathcal{P}(X)$ with respect to $Q \in \mathcal{P}(Y)$ is defined as the infimum convolution

$$\mathfrak{D}^c_\varepsilon(P\|Q) \coloneqq \inf_{R\in\mathcal{P}(Y)} \{T_c(P, R) + \varepsilon\mathfrak{D}(R\|Q)\}. \tag{2.3}$$

For $P, Q \in \mathcal{P}_p(X)$ and $T_c(P, Q) = W^p_p(P, Q)$, we use the notation

$$\mathfrak{D}^p_\varepsilon(P\|Q) \coloneqq \inf_{R\in\mathcal{P}(X)} \left\{W^p_p(P, R) + \varepsilon\mathfrak{D}(R\|Q)\right\}. \tag{2.4}$$

EXAMPLE 2.2 (Examples of proximal OT divergences).

(1) When T_c is the Wasserstein-1 distance, i.e., $c(x, y) = d(x, y)$ and $\mathfrak{D}$ is a f-divergence, the formulation recovers the Lipschitz-regularized f-divergences. These divergences have been extensively studied for their significance in the design and theoretical analysis of generative models, particularly in the context of stability and robustness in [7, 16, 20, 25].

(2) When T_c is the p-Wasserstein distance with $p \geq 2$, and $\mathfrak{D}$ is the KL divergence, the formulation leads to the p-Wasserstein regularized KL divergence. In Section 2.4, we derive the dynamic formulation of this specific proximal OT divergence through its Benamou-Brenier representation. The dynamic formulation yields optimal trajectories that follow straight-line paths with constant velocity [22]. This property aligns with the geodesic structure of the p-Wasserstein space and has been effectively utilized in generative modeling, particularly in [40] and is discussed in Section 5.3.

(3) Recently, the novel approach in [26] combines 1-Wasserstein and 2-Wasserstein proximals, leading to a proximal OT divergence where T_c corresponds to the 2-Wasserstein distance and $\mathfrak{D}$ is a Lipschitz-regularized f-divergence. This formulation ensures the well-posedness of generative flows and enables the robust learning of distributions supported on manifolds.

2.2. *Properties of proximal OT divergences.* We now derive some fundamental properties of the proximal OT divergence (2.3), showing its desirable and useful characteristics. To proceed, we impose certain assumptions on the optimal transport cost and the information divergence.

Assumption A-T. The transport cost $T_c(P, Q)$ is a divergence. Moreover, the map

$$(P, Q) \mapsto T_c(P, Q) \tag{2.5}$$

is convex and lower semicontinuous in the weak topology (this holds if c is bounded below and lower semicontinuous [47]).

Assumption A-D. The map

$$(P, Q) \mapsto \mathfrak{D}(P\|Q) \tag{2.6}$$

is jointly convex and lower semicontinuous in the weak topology. Moreover, we assume that the level sets

$$\{R : \mathfrak{D}(R\|Q) \leq \alpha\}, \tag{2.7}$$

for any $\alpha \geq 0$, are pre-compact in the weak topology and strictly convex.

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

These assumptions are satisfied by the KL divergence. Moreover, under these assumptions, the proximal OT divergence is indeed a divergence, inheriting the basic convexity and lower semicontinuity from the original information divergence. In particular, compactness and strict convexity guarantee the existence of a unique minimizer R^* in the infimal convolution of the proximal OT divergence. These results are stated in the next two theorems, which are proved in Appendix A.

THEOREM 2.3 (Divergence properties). If assumptions **A-T** and **A-D** are satisfied, then for every $\varepsilon > 0$, $\mathfrak{D}_\varepsilon^c$ has the following properties.

(1) **Divergence.** $\mathfrak{D}_\varepsilon^c(P\|Q)$ is a divergence, i.e.,

$$\mathfrak{D}_\varepsilon^c(P\|Q) \geq 0 \quad \text{and} \quad \mathfrak{D}_\varepsilon^c(P\|Q) = 0 \iff P = Q. \tag{2.8}$$

(2) **Convexity and lower semicontinuity.** The map

$$(P, Q) \mapsto \mathfrak{D}_\varepsilon^c(P\|Q) \tag{2.9}$$

is convex and lower semi-continuous in the weak topology.

THEOREM 2.4 (Proximal properties). If assumptions **A-T** and **A-D** are satisfied, then for every $\varepsilon > 0$, $\mathfrak{D}_\varepsilon^c$ has the following properties.

(1) **Upper bound.** The proximal OT divergence is bounded above by the optimal transport distance and the information divergences, i.e.,

$$\mathfrak{D}_\varepsilon^c(P\|Q) \leq \min\left\{T_c(P, Q), \varepsilon\mathfrak{D}(P\|Q)\right\}. \tag{2.10}$$

(2) **Proximal minimizer.** If $\mathfrak{D}_\varepsilon^c(P\|Q) < \infty$, then there is a unique minimizer $R^* \in \mathcal{P}(Y)$ of (2.3) such that

$$\mathfrak{D}_\varepsilon^c(P\|Q) = T_c(P, R^*) + \varepsilon\mathfrak{D}(R^*\|Q). \tag{2.11}$$

The upper bound (2.10) follows directly by simply setting $R = P$ and $R = Q$ in (2.3). As an important consequence, the proximal divergence is finite whenever either the divergence $\mathfrak{D}(P\|Q)$ or the optimal transport is finite. Notably, this holds without requiring absolute continuity, a crucial distinction from the classical KL divergence, where finiteness typically depends on this condition. For example, when considering the p-Wasserstein metrics with the KL divergence, we observe $\mathfrak{D}^p_{\mathrm{KL},\varepsilon}(P\|Q)$ is finite whenever P and Q have p-th finite moments without any assumption on absolute continuity of P with respect to Q. This property makes proximal OT divergences well-suited for comparing data to models or between datasets, including when data distributions are concentrated on low-dimensional structures, as suggested by the "manifold hypothesis" in machine learning.

Furthermore, the proximal OT divergences yield a coarser topology that admits a larger set of target distributions beyond absolute continuity. Both the ball based on the KL divergence $B_{\mathrm{KL}}(Q) = \{\mu : \mathfrak{D}_{\mathrm{KL}}(\mu\|Q) \leq 1\}$ and the ball based on the optimal transport cost $B_{T_c}(Q) = \{\mu : T_c(\mu\|Q) \leq 1\}$ are subsets of the proximal OT divergence ball $B_{\mathrm{ProxOT}}(Q) = \{\mu : \mathfrak{D}_\varepsilon^c(\mu\|Q) \leq 1\}$. While $B_{T_c}(Q)$ includes probability distributions that may not be absolutely continuous with respect to Q, B_{KL} excludes them. Meanwhile, $B_{T_c}(Q)$ may fail to capture heavy-tailed distributions (see Example 1 in [16]), while B_{KL} includes them. $B_{\mathrm{ProxOT}}(Q)$ includes both of these types of distributions. This highlights

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

the advantage of proximal OT divergences in enhancing robustness, both in generative modeling tasks and in the construction of ambiguity sets for DRO.

The existence of a unique proximal minimizer follows naturally from the compactness of the level sets of the map $R \mapsto \mathfrak{D}(R\|Q)$ in the weak topology. The minimizer R^* is crucial for proving many of the results in this paper, and also plays a significant role in applications and algorithms. For example, in generative modeling applications, where P is a reference probability distribution which is easy to sample from and Q is an unknown data distribution accessible only through samples, the probability distribution R^* often serves as the approximate target distribution that is produced by the generative algorithm [22, 26, 40]. This insight motivates the definition of a proximal operator that directly maps the reference measure P to the minimizer R^*. This transport proximal operator is a generalization of the proximal operator in $\mathbb{R}^d$ [41]. When the transport metric is the Wasserstein distance, it is known as the Wasserstein proximal operator and has been studied in [30].

DEFINITION 2.5 (Transport proximal operator). Let Q be a fixed probability distribution, $\varepsilon > 0$, and let assumptions **A-T** and **A-D** be satisfied. We define the *transport proximal operator*

$$\mathbf{prox}^c_{\varepsilon\mathfrak{D}(\cdot\|Q)} : \mathcal{P}(X) \to \mathcal{P}(Y) \tag{2.12}$$

as the mapping that assigns each measure $P \in \mathcal{P}(X)$ to the minimizer

$$R^* = \mathbf{prox}^c_{\varepsilon\mathfrak{D}(\cdot\|Q)}(P) \coloneqq \operatorname*{argmin}_{R\in\mathcal{P}(Y)} \{T_c(P, R) + \varepsilon\mathfrak{D}(R\|Q)\}. \tag{2.13}$$

When T_c is the p-Wasserstein distance, we denote this operator by $\mathbf{prox}^p_{\varepsilon\mathfrak{D}(\cdot\|Q)}$ and refer to it as *Wasserstein proximal operator*. When Q is a fixed probability distribution, we abbreviate the notation for the operator to $\mathbf{prox}^c_{\varepsilon\mathfrak{D}}$.

REMARK 2.6 (Transport and re-weighting). The proximal minimizer elucidates how the divergence compares two distributions: the measure P is first (optimally) transported to the intermediate measure R^*, which is absolutely continuous with respect to Q, even if P and Q are not mutually absolutely continuous. Moreover, the support of R^* can differ significantly from P. The cost incurred by this movement is the optimal transport cost and essentially depends on the distance over which the mass must be moved. Subsequently, R^* is redistributed to Q using the information divergence. In this stage, the support of R^* remains contained within the support of Q; any attempt to expand or shift it beyond the support of Q would result in an infinite information divergence. This structural property implies that proximal OT divergences can effectively capture data supported on manifolds [26]. Notably, related proximal divergence frameworks have shown strong manifold learning capabilities, as seen in [7, 8, 43, 60].

REMARK 2.7 (Connection to the JKO scheme). The Jordan-Kinderlehrer-Otto (JKO) scheme is a time-discretization approach for solving the gradient flow associated with an energy functional $\mathcal{E}(\rho)$ in the 2-Wasserstein space of probability measures $\mathcal{P}_2(X)$. It is closely linked to the transport proximal operator, providing a variational framework that iteratively minimizes a combination of the energy functional and a regularizer given by the 2-Wasserstein distance. The continuous-time gradient flow is defined as the solution

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

of the PDE

$$\frac{\partial \rho(x,t)}{\partial t} = -\nabla \cdot \left(\rho \nabla \frac{\delta \mathcal{E}(\rho)}{\delta \rho} \right), \qquad \rho(\cdot, 0) = \rho_0(\cdot), \tag{2.14}$$

where $\rho(x,t)$ is the probability density evolving over time starting from an initial probability density ρ_0, and $\frac{\delta \mathcal{E}}{\delta \rho}$ is the variational derivative of the energy functional. The right-hand side of (2.14) represents the Wasserstein gradient flow. Instead of directly solving the continuous-time gradient flow in (2.14), the JKO scheme discretizes time into small intervals and finds an approximate solution at each time step. Specifically, the scheme with a discrete time step of size τ is formulated as an iterative minimization

$$\rho^*(\cdot, t_{n+1}) = \operatorname*{argmin}_{\rho} \left\{ \mathcal{E}(\rho) + \frac{1}{2\tau} W_2^2(\rho^*(\cdot, t_n), \rho) \right\}. \tag{2.15}$$

The continuous-time limit of the discrete scheme as $\tau \to 0$ recovers the solution of the gradient flow PDE (2.14); see [28] for the original work and [2] for a rigorous overview. Popular examples are the heat equation and the Fokker-Planck equation, which can be seen as the gradient flow of the entropy functional and the entropy plus a potential energy, respectively. In the latter example, by a simple calculation:

$$\mathcal{E}(\rho) = \int f(\rho(x))\rho(x)dx + \int V(x)\rho(x)dx = \mathfrak{D}_{\mathrm{KL}}(\rho \| q) = \int \rho(x) \log \frac{\rho(x)}{q(x)} dx,$$

with $f(t) = t \log t$ and $q(x) \propto e^{-V(x)}$.

The connection between the JKO scheme and the proximal OT divergences follows by interpreting (2.15) using operator (2.13), i.e.,

$$\rho^*(\cdot, t_{n+1}) = \frac{1}{2\tau} \mathbf{prox}^2_{2\tau \mathfrak{D}_{\mathrm{KL}}(\cdot \| q)}(\rho^*(\cdot, t_n))$$

starting from some initial probability density $\rho(\cdot, 0) = \rho_0$.

2.3. *Proximal OT divergences interpolate between optimal transport distances and information divergences.* The parameter ε plays a crucial role in proximal OT divergences. It controls the trade-off between the information divergence term and the optimal transport cost, thereby determining the relative importance of re-weighting versus transport. More precisely, ε scales the information divergence term, affecting how much it contributes to the total cost. For example, in the divergence-dominant regime, i.e., when ε is large, the information divergence term becomes more significant and even a small value of divergence $\mathfrak{D}(R\|Q)$ contributes significantly to the overall objective as it is magnified by ε. This effect is formalized in Theorem 2.8, demonstrating that the proximal OT divergence is not only a lower bound for both the divergence and the optimal transport (see (2.10)), but also serves as a continuous regularization of both terms.

THEOREM 2.8 (Interpolation properties). Under assumptions **A-T** and **A-D**, the following holds:

(1) **Divergence regularization by OT**: As $\varepsilon \searrow 0$, $\frac{1}{\varepsilon}\mathfrak{D}^c_\varepsilon$ is increasing and

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \mathfrak{D}^c_\varepsilon(P\|Q) = \mathfrak{D}(P\|Q). \tag{2.16}$$

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

(2) **OT regularization by divergence**: As $\varepsilon \nearrow \infty$, $\mathfrak{D}^c_\varepsilon$ is increasing and

$$\lim_{\varepsilon\to\infty} \mathfrak{D}^c_\varepsilon(P\|Q) = T_c(P,Q). \tag{2.17}$$

The proof of this theorem is given in Appendix B and is essentially the same as the one given in [7, 20] for the family of (f,Γ)-divergences introduced therein.

2.4. *Dynamic formulation of proximal OT divergences and mean-field games.* Here, we demonstrate that the proximal OT divergences have a continuous-time, flow-based formulation that augments the classical Benamou-Brenier formulation of optimal transport [5] with a mean-field game (MFG) structure. Specifically, let $X = Y = \mathbb{R}^d$ and consider the transport cost $c(x,y) = |x-y|^p$ for $p > 1$. According to the Benamou-Brenier formula, see e.g. [47], optimal transport can be interpreted as the problem of finding an optimal vector field $v_t : \mathbb{R}^d \to \mathbb{R}^d$ for all $t \in [0,1]$ that continuously transports the probability measure P into the measure R, i.e.,

$$W_p^p(P,R) = \min_{\rho,v} \left\{ \int_0^1 \mathbb{E}_{\rho_t}\left[|v_t|^p\right] dt : \begin{array}{c} \partial_t\rho_t + \nabla\cdot(\rho_t v_t) = 0 \\ \rho_0 = P,\ \rho_1 = R \end{array} \right\}. \tag{2.18}$$

This leads immediately to Theorem 2.9 that provides an equivalent formulation of the divergence based on dynamic optimal transport.

THEOREM 2.9 (Dynamic formulation of proximal OT divergences). For $p > 1$ and $\varepsilon > 0$, the proximal OT divergence $\mathfrak{D}^p_\varepsilon(P\|Q)$ has the following representation:

$$\begin{aligned} \mathfrak{D}^p_\epsilon(P\|Q) &= \inf_{R\in\mathcal{P}(\mathbb{R}^d)} \left\{\varepsilon\mathfrak{D}(R\|Q) + W_p^p(P,R)\right\} \\ &= \min_{v,\rho} \left\{ \varepsilon\mathfrak{D}(\rho_1\|Q) + \int_0^1 \mathbb{E}_{\rho_t}\left[|v_t|^p\right] dt : \begin{array}{c} \partial_t\rho_t + \nabla\cdot(\rho_t v_t) = 0 \\ \rho_0 = P \end{array} \right\}. \end{aligned} \tag{2.19}$$

The choice of the terminal time T, chosen to be $T = 1$ in Theorem 2.9, is entirely arbitrary. A more natural choice is $T = \varepsilon$, and by rescaling time for the flow accordingly, we obtain the divergence

$$\begin{aligned} \mathfrak{D}^p_T(P\|Q) &= T \inf_{R\in\mathcal{P}(\mathbb{R}^d)} \left\{ \mathfrak{D}(R\|Q) + \frac{1}{T} W_p^p(P,R) \right\} \\ &= T \min_{v,\rho} \left\{ \mathfrak{D}(\rho_T\|Q) + \frac{1}{T^2}\int_0^T \mathbb{E}_{\rho_t}\left[|v_t|^p\right] dt : \begin{array}{c} \partial_t\rho_t + \nabla\cdot(\rho_t v_t) = 0 \\ \rho_0 = P \end{array} \right\}, \end{aligned} \tag{2.20}$$

where the extra factor $1/T$ arises from the time-substitution $t \to Tt$. (2.20) results in a more convenient form for time-discretization and learning algorithms. We can interpret the parameter T as the time-horizon allotted for the Wasserstein component of the proximal OT divergence to carry out the transport.

REMARK 2.10 (Proximal in generative flows). The dynamic formulation of proximal OT divergence is particularly useful in generative modeling algorithms which learn the optimal vector field v_t^* in (2.20) and thus the flow driving the source P toward the proximal distribution $\rho_T^* = R^* = \mathbf{prox}^c_{\varepsilon\mathfrak{D}}(P)$. The interpolation properties established in Theorem 2.8 indicate that ρ_T^* is a good approximation of Q when $T > 1$ in (2.20). Specifically, Theorem 2.8 states that $\lim_{T\to\infty} \mathfrak{D}^c_T(P\|Q) = W_p^p(P,Q)$. Thus, by the

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

approximation

$$\mathfrak{D}_T^c(P\|Q) = T\mathfrak{D}(R^*\|Q) + W_p^p(P, R^*) \approx W_p^p(P, Q), \qquad T > 1. \tag{2.21}$$

Hence, R^* is a good approximation of Q, and ρ_T^* serves as a good surrogate of Q.

REMARK 2.11 (The mean-field game for proximal OT divergences). The dynamical formulation of the proximal OT divergence closely resembles the Benamou-Brenier formulation of the Wasserstein transport cost [5]. However, the divergence term and the optimization over the intermediate measure $\rho_T = R$ in (2.19) serves as a relaxation of the terminal condition in the Benamou-Brenier formulation. Mathematically, this turns the optimization problem (2.19) into a mean-field game (MFG) [29], with $\mathfrak{D}(\rho_T\|Q)$ and $|v|^p$ playing the role of the terminal cost and action cost in the MFG formulation, respectively. The optimality conditions are a coupled system of nonlinear PDEs: a backward Hamilton-Jacobi-Bellman (HJB) equation that governs how an individual agent optimally controls its movement while considering the evolving population distribution $\rho(\cdot, t)$, with a forward continuity equation that describes the dynamics of the density of the agents under a particular velocity field [59]. For example, if the Wasserstein-2 distance is chosen to be the optimal transport cost and $\varepsilon = 2/T$, the mean-field game yields the following system of coupled PDEs

$$\begin{cases} -\dfrac{\partial U}{\partial t} + \dfrac{1}{2}|\nabla U|^2 = 0, \\ \dfrac{\partial \rho}{\partial t} - \nabla \cdot (\rho \nabla U) = 0, \\ U(x, T) = 1 + \log \dfrac{\rho(x, T)}{Q(x)},\ \rho(x, 0) = P(x). \end{cases} \tag{2.22}$$

Moreover, the optimal velocity is a time-dependent gradient field $v^*(x, t) = -\nabla U(x, t)$.

3. Variational representations for proximal OT divergences. In this section, we show that the dual formulation of the proximal OT divergences provides a variational representation of the proximal OT divergences based on a suitable classes of test functions. Variational representations of divergences, integral probability metrics, and optimal transport costs play an important theoretical and practical role. For example, these representations are an essential ingredient in generative adversarial networks (GANs) which are constructed by solving a minmax game that minimize over parameters of the generator and maximize over the function space in the variational representation of the divergence to identify the discriminator [24]. Here in particular, the dual formulation allows us to compute proximal OT divergences using convex neural networks as discriminators; see Section 5.2 for a complete discussion. In Section 3.1 we review variational representations of information divergences and optimal transport distances. We then derive variational representations for the proximal OT divergence in Section 3.2, which we use to derive a data processing inequality in Section 3.3.

3.1. *Duality provides alternative representations of optimal transport distances and information divergences.* In this section, we choose the KL divergence as an information divergence to present our results. One could also consider more general f-divergences. We begin by revisiting some well-known facts about optimal transport and the KL divergence.

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

Duality for optimal transport. We summarize some well-known results in optimal transport; see [47, 56, 57] for proofs. Given a cost function $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$, we consider the function space

$$\Phi_c = \{(\phi, \psi) \in C_b(X \times Y), \phi \oplus \psi \leq c\} \tag{3.1}$$

(recall $\phi \oplus \psi$ denotes the function $\phi(x)+\psi(y)$). A fundamental result in optimal transport is the following duality result: suppose c is bounded below and lower semicontinuous, then we have

$$T_c(P, Q) = \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_\pi[c] = \sup_{(\phi,\psi)\in\Phi_c} \{\mathbb{E}_P[\phi] + \mathbb{E}_Q[\psi]\}. \tag{3.2}$$

Regarding the existence of optimizers, an optimal coupling π^* always exists given that X and Y are Polish. If c is uniformly continuous and bounded, then there exist optimizers $(\phi^*, \psi^*) \in \Phi_c$. For general cost functions c, we cannot expect the optimizers to be bounded but there exists optimizers $\phi \in L^1(P)$, $\psi \in L^1(Q)$ satisfying $\phi \oplus \psi \leq c$ provided we assume that

$$c(x, y) \leq a(x) + b(y) \quad \text{ with } a \in L^1(P), b \in L^1(Q). \tag{3.3}$$

For example if $c(x, y) = d(x, y)^p$, this condition holds for all probability measures in the p-Wasserstein space defined in (2.2).

Duality for the KL divergence. The KL divergence between two probability measures $R \in \mathcal{P}(Y)$ and $Q \in \mathcal{P}(Y)$ is given by

$$\mathfrak{D}_{\mathrm{KL}}(R\|Q) = \begin{cases} \mathbb{E}_R\left[\log \frac{dR}{dQ}\right] & \text{if } R \ll Q \\ +\infty & \text{otherwise} \end{cases}. \tag{3.4}$$

The Donsker-Varadhan formula provides a variational representation of the KL divergence (see [19] for a proof), which is based on the convex duality between the divergence and the cumulant generating function $\Lambda(\psi) = \log \mathbb{E}_Q[e^\psi]$. That is,

$$\sup_{\psi \in C_b(Y)} \left\{ \int_Y \psi \, dR - \log \mathbb{E}_Q[e^\psi] \right\} = \begin{cases} \mathfrak{D}_{\mathrm{KL}}(R\|Q), & R \in \mathcal{P}(Y) \\ +\infty, & R \in \mathcal{M}(Y) \setminus \mathcal{P}(Y). \end{cases} \tag{3.5}$$

The reader should note that this variational representation implies that we define $\mathfrak{D}_{\mathrm{KL}}(R\|Q)$ as $+\infty$ if R is not a probability measure. While one could extend $\mathfrak{D}_{\mathrm{KL}}(R\|Q)$ to $\mathcal{M}(Y)$ using the definition in the first line (3.4) to obtain a finite value, this leads to another convex duality representation (see e.g. [7]). However, the representation in (3.5) is the one that will be essential for our purposes.

We will also use the (convex) dual formula to (3.5) referred to as the Gibbs variational principle: if $\psi \in C_b(Y)$ then

$$\log \mathbb{E}_Q[e^\psi] = \sup_{R \in \mathcal{P}(Y)} \{\mathbb{E}_R[\psi] - \mathfrak{D}_{\mathrm{KL}}(R\|Q)\}, \tag{3.6}$$

and the supremum is attained when R is such that

$$\frac{dR}{dQ} = \frac{e^\psi}{\mathbb{E}_Q[e^\psi]}. \tag{3.7}$$

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

3.2. *Duality for proximal OT divergences via variational representations.* In this section, we establish a dual formulation for the proximal OT divergence when KL divergence is combined with an optimal transport cost. This is achieved by leveraging the variational representations given in (3.6) and (3.2) respectively. In particular, the infimum convolution formula yields the following duality result.

THEOREM 3.1 (Duality for proximal OT divergence). Suppose X and Y are Polish spaces and that c is lower-semicontinuous and bounded below. Given $P \in \mathcal{P}(X)$ and $Q \in \mathcal{P}(Y)$ suppose that $c(x,y) \le a(x) + b(y)$ with $\mathbb{E}_P[a] < \infty$ and $\mathbb{E}_Q[b] < \infty$. Then, for every $\varepsilon > 0$, we have the following variational representations

$$\begin{aligned} \mathfrak{D}^c_{\mathrm{KL},\varepsilon}(P\|Q) &= \inf_{R\in\mathcal{P}(Y)} T_c(P,R) + \varepsilon \mathfrak{D}_{\mathrm{KL}}(R\|Q) \\ &= \sup_{(\phi,\psi)\in\Phi_c} \left\{ \mathbb{E}_P[\phi] - \varepsilon \log \mathbb{E}_Q[e^{-\psi/\varepsilon}] \right\}. \end{aligned} \tag{3.8}$$

The proof of Theorem 3.1, at this level of generality, is rather long and is given in Appendix C. The general idea is to use a convex analysis argument in the case where X and Y are compact and c is continuous and then prove the general case using several approximation arguments and lower semicontinuity. In this respect, the strategy of the proof follows quite closely the proof of duality for optimal transport given in [56] and uses many of the technical estimates and ideas therein.

For now, we are content with a formal proof based on the dual formulation of the transport cost $T_c(P,R)$ and the Gibbs variational principle for the KL divergence. Without rigorous justification, observe that, by exchanging infimum and supremum—which could likely be justified for compact spaces using an appropriate minimax theorem—we obtain

$$\begin{aligned} \mathfrak{D}^c_{\mathrm{KL},\varepsilon}(P\|Q) &= \inf_{R\in\mathcal{P}(Y)} \{T_c(P,R) + \varepsilon \mathfrak{D}_{\mathrm{KL}}(R\|Q)\} \\ &= \inf_{R\in\mathcal{P}(Y)} \sup_{(\phi,\psi)\in\Phi_c} \{\mathbb{E}_P[\phi] + \mathbb{E}_R[\psi] + \varepsilon \mathfrak{D}_{\mathrm{KL}}(R\|Q)\} \quad (\text{ by (3.2)}) \\ &= \sup_{(\phi,\psi)\in\Phi_c} \inf_{R\in\mathcal{P}(Y)} \{\mathbb{E}_P[\phi] + \mathbb{E}_R[\psi] + \varepsilon \mathfrak{D}_{\mathrm{KL}}(R\|Q)\} \quad (\inf \leftrightarrow \sup) \\ &= \sup_{(\phi,\psi)\in\Phi_c} \left\{ \mathbb{E}_P[\phi] - \varepsilon \sup_{R\in\mathcal{P}(Y)} \{\mathbb{E}_R[-\psi/\varepsilon] - \mathfrak{D}_{\mathrm{KL}}(R\|Q)\} \right\} \\ &= \sup_{(\phi,\psi)\in\Phi_c} \left\{ \mathbb{E}_P[\phi] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi/\varepsilon}] \right\} \quad (\text{by (3.6)}). \end{aligned}$$

EXAMPLE 3.2 (Lipschitz-regularized f-divergences). In Example 2.2, the Lipschitz-regularized f-divergences [25] arise as a special case of proximal OT divergences, where the information divergence is given by an f-divergence and the 1-Wasserstein is the optimal transport cost. It can be readily shown that a c-concave function, defined in Appendix C, corresponds exactly to a 1-Lipschitz function, and vice versa. Moreover, in this case, we have $\psi = \phi^c = -\phi$ (see Lemma C.1) and thus

$$\mathfrak{D}_{\mathrm{KL}}^{\mathrm{Lip}\frac{1}{\varepsilon}(X)}(P\|Q) \coloneqq \sup_{\phi\in\mathrm{Lip}\frac{1}{\varepsilon}(X)} \left\{ \mathbb{E}_P[\phi] - \ln \mathbb{E}_Q[e^{\phi}] \right\} = \frac{1}{\varepsilon} \mathfrak{D}^1_{\mathrm{KL},\varepsilon}(P\|Q),$$

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

where $\mathrm{Lip}_{\frac{1}{\varepsilon}}(X)$ is the set of Lipschitz functions on X with Lipschtz constant $1/\varepsilon$. This can be extended to the dual form of the Lipschitz-regularized f-divergences as $\mathfrak{D}_f^{\mathrm{Lip}_{\frac{1}{\varepsilon}}(X)}(P\|Q) = \frac{1}{\varepsilon}\mathfrak{D}^1_{f,\varepsilon}(P\|Q)$.

Next, we derive the existence and characterization of the solution R^* to the infimal convolution problem (2.3) as well as the optimizers (ϕ^*, ψ^*) in the dual representation (3.8). Furthermore, the three optimizers R^*, ϕ^*, ψ^* are related in such a way that they reinforce the transport and re-weighting interpretation of proximal OT divergences as discussed in Remark 2.6 and stated in Theorem 3.3 (d) and further explained thereafter.

THEOREM 3.3 (Optimizers for the proximal OT divergence). Assume that X and Y are Polish spaces and that c is lower-semicontinuous and bounded below. Let $P \in \mathcal{P}(X)$ and $Q \in \mathcal{P}(Y)$ and assume that $c(x,y) \le a(x) + b(y)$ with $\mathbb{E}_P[a] < \infty$ and $\mathbb{E}_Q[b] < \infty$. Then the following statements hold for every $\varepsilon > 0$.

(a) There exists a unique minimizer R^* in (2.3).

(b) If c is bounded and uniformly continuous then there exists a maximizer pair $(\phi^*, \psi^*) \in \Phi_c$ given in (3.8). Up to a constant shift $(\phi^*, \psi^*) \to (\phi^* - a, \psi^* + a)$, ϕ^* is unique P almost surely and ψ^* is unique Q almost surely.

(c) In general, the supremum is attained for a pair $(\phi^*, \psi^*) \in L^1(P) \times L^1(Q)$ with the same uniqueness property as for c continuous.

(d) The optimizer R^*, ϕ^*, ψ^* are related in the following manner:

(1) The pair (ϕ^*, ψ^*) is the optimizer in the dual representation of the $T_c(P, R^*)$, i.e.

$$T_c(P, R^*) = \mathbb{E}_P[\phi^*] + \mathbb{E}_{R^*}[\psi^*]\,. \tag{3.9}$$

(2) The Radon-Nikodym derivative of R^* with respect to Q is given by

$$\frac{dR^*}{dQ} = \frac{e^{-\psi^*/\varepsilon}}{\mathbb{E}_Q[e^{-\psi^*/\varepsilon}]}. \tag{3.10}$$

Note that all optimizers R^*, ϕ^*, ψ^* depend on ε; however, for notational simplicity, we omit this dependence throughout. The proof of this theorem is given in Appendix C. The technical part consists in proving the existence of optimizers. Here, we focus on explaining the relationship between the optimizers. Assuming we have established both the duality and the existence of maximizers (ϕ^*, ψ^*) (3.8), then we have

$$\mathbb{E}_{\pi^*}[c] + \varepsilon\mathfrak{D}(R^*\|Q) = \mathbb{E}_P[\phi^*] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi^*/\varepsilon}], \tag{3.11}$$

where π^* is the optimal transport plan between P and R^*. Equivalently

$$\underbrace{\mathbb{E}_{\pi^*}[c - \phi^* - \psi^*]}_{\ge 0} + \varepsilon \underbrace{\left\{\mathfrak{D}(R^*\|Q) - \mathbb{E}_R[-\psi^*/\varepsilon] + \log \mathbb{E}_Q[e^{-\psi^*/\varepsilon}]\right\}}_{\ge 0} = 0 \tag{3.12}$$

and, since both terms must vanish, we get exactly the characterizations of the optimizers in (3.9) and (3.10) in Theorem 3.3.

Part (d) in Theorem 3.3 can be interpreted as a transport and re-weighting process (see also Remark 2.6) where (3.9) is the transport of P to R^* and the emerging Gibbs distribution in (3.10) describes the re-weighting of R^* to Q with the weighting determined

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

by ψ^*/ε. In generative modeling, this process enables the generation of new data by transporting and re-weighting samples.

Lastly, we show that the optimizer ϕ^* given in Part (d) of Theorem 3.3 corresponds to an optimal transport map from P to the intermediate measure R^*. To do so, we recall the Monge formulation of the optimal transport problem between measures $P, R^* \in \mathcal{P}(\mathbb{R}^d)$ that seeks a measurable map $T : \mathbb{R}^d \to \mathbb{R}^d$ pushing P onto R^* while minimizing the optimal transport cost:

$$T_c(P, R^*) = \inf_T \int_{\mathbb{R}^d} c(x, T(x)) dP(x). \tag{3.13}$$

For the special case $c(x, y) = \frac{1}{2}|x - y|^2$, under suitable conditions for the existence of a map (see Theorem 4.2 in [48]), the optimal transport map is given by $T^*(x) = x - \nabla\phi^*(x)$ where ϕ^* is given in Part (d) of Theorem 3.3.

REMARK 3.4 (Alternative expression for the solution of JKO scheme at each step). The iterative minimization in the JKO scheme can be expressed using the transport proximal operator:

$$\rho^*(\cdot, t_{n+1}) = \frac{1}{2\tau} \mathbf{prox}^2_{2\tau\mathfrak{D}_{\mathrm{KL}}(\cdot\|q)}(\rho^*(\cdot, t_n)),$$

see Remark 2.7. From Theorem 3.3(d)(2), the corresponding proximal minimizer satisfies

$$\frac{\rho^*(\cdot, t_n)}{q} = \frac{e^{-\psi_n^*/2\tau}}{2\tau \mathbb{E}_Q[e^{-\psi_n^*/2\tau}]} \text{ which implies } \rho^*(\cdot, t_n) \propto e^{-\psi_n^*/2\tau - V}. \tag{3.14}$$

This expression aligns with Proposition 8.7 in [47].

3.3. *Applications of duality.* In this section, we derive further properties of the proximal OT divergence through its variational formulation. All proofs can be found in Appendix D. We first show that the proximal p-Wasserstein divergence is invariant under metric isomorphism.

THEOREM 3.5 (Invariance under isomorphisms). Suppose T is a metric isomorphism of X (i.e. $T : X \to X$ is invertible and preserves distance, i.e., $d(Tx, Ty) = d(x, y)$ for all $x, y \in X$). Then, for every $\varepsilon > 0$,

$$\mathfrak{D}^p_\varepsilon(T_\# P \| T_\# Q) = \mathfrak{D}^p_\varepsilon(P\|Q). \tag{3.15}$$

Next, we establish a version of the data processing inequality [54]. Suppose that $K(x, dx')$ is a kernel from X to X' such that for each x, $K(x, dx') \in \mathcal{P}(X')$ and $K(x, A)$ is measurable for every x and every Borel set $A \subset X'$. The kernel K induces a map that takes functions on X' to functions of X by $K\phi(x) = \int_{X'} K(x, dx')\phi(x')$. Moreover, K maps probability measures on X to probability measures on X' by $K_\# P(A) = \int_X K(x, A) dP(x)$.

THEOREM 3.6 (Data processing inequality). Suppose $X = Y$ and K is a kernel from X to X' which maps $C_b(X')$ into $C_b(X)$. Then, for every $\varepsilon > 0$, we have

$$\mathfrak{D}^c_\varepsilon(K_\# P \| K_\# Q) \le \mathfrak{D}^{Kc}_\varepsilon(P\|Q), \tag{3.16}$$

where Kc is the cost function on C defined by

$$Kc(x, y) = \int_X \int_Y c(x, y) K(x, dx) K(y, dy). \tag{3.17}$$

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

Finally, we prove that the proximal OT divergence with p-Wasserstein $\mathfrak{D}_p^\varepsilon$ is additive for product measures.

THEOREM 3.7 (Additivity for product measures). Assume $X = Y = \mathbb{R}^n$ and $c(x-y) = \|x-y\|^p$. If $P = P_1 \times P_2$ and $Q = Q_1 \times Q_2$ are product measures with $P_1, Q_1 \in \mathcal{P}(\mathbb{R}^{n_1})$ and $P_2, Q_2 \in \mathcal{P}(\mathbb{R}^{n_2})$, then for every $\varepsilon > 0$, the proximal OT is additive

$$\mathfrak{D}^p_{\mathrm{KL},\varepsilon}(P_1 \times P_2 \| Q_1 \times Q_2) = \mathfrak{D}^p_{\mathrm{KL},\varepsilon}(P_1 \| Q_1) + \mathfrak{D}^p_{\mathrm{KL},\varepsilon}(P_2 \| Q_2). \tag{3.18}$$

The proof relies on both the primal and dual representations of the divergence. By restricting the infimum to product intermediate measures, we obtain one side of the inequality in (3.18), while restricting the supremum to functions of the form $\phi(x) = \phi_1(x_1) + \phi_2(x_2)$ (and similarly for ψ) yields the reverse inequality. The full details are provided in Appendix D.

4. The first variation of the proximal OT divergence. In this section we show that the variational derivative of the proximal OT divergence around a fixed measure Q exists under very broad conditions and is practically computable. The existence and the evaluation of variational derivatives of proximal OT divergences involves the discriminator (i.e., the optimal ϕ^* in (3.8)) in the dual variational representation of proximal OT divergences. Here we show that the variational derivative is always well-defined for any probability measure perturbations for which the proximal OT divergence remains finite. We note that such perturbations can be singular and do not necessarily need to be absolutely continuous with respect to Q where the variational derivative is calculated. Therefore, variational derivatives of proximal OT divergences exist in "all directions," not just along absolutely continuous perturbations of the baseline Q directions, which is the case, for instance, in KL and other f-divergences. A key practical outcome of this observation is that machine learning algorithms that involve the optimization of f-divergences of any type will be *provably* stabilized when a proximal OT divergence is deployed instead of an f-divergence. For instance, in gradient-based algorithms used to solve the optimization problem $\min_\theta \mathfrak{D}^c_{\mathrm{KL},\varepsilon}(P^\theta \| Q)$, it is necessary to compute the gradient

$$\nabla_\theta \mathfrak{D}^c_{\mathrm{KL},\varepsilon}(P^\theta \| Q) = \int \frac{\delta \mathfrak{D}^c_{\mathrm{KL},\varepsilon}(P\|Q)}{\delta P}(P^\theta(x)) \nabla_\theta P^\theta(x) dx,$$

which involves the variational derivative $\frac{\delta \mathfrak{D}^c_{\mathrm{KL},\varepsilon}(P\|Q)}{\delta P}$ which must be well-defined for the gradient to be properly evaluated. Moreover, the construction of gradient flows in the space of probability measures also relies critically on this variational derivative as discussed in detail in Section 6.

4.1. *Proximal OT divergences have well-defined variational derivative.* From Theorem 3.3, the optimizers (ϕ^*, ψ^*) are uniquely determined in $\mathrm{supp}(P)$ and $\mathrm{supp}(Q)$ respectively, up to additive constant, i.e., that is we can replace (ϕ^*, ψ^*) by $(\phi^* + a, \psi^* - a)$ for any $a \in \mathbb{R}$. To compute the directional derivative of the optimal transport divergence, we extend ϕ^* to all of X. This extension follows a standard approach repeatedly used in

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

optimal transport theory, i.e., we define

$$\widehat{\phi}^*(x) = \inf_{y \in \mathrm{supp}(Q)} \{c(x,y) - \psi^*(y)\} \,. \tag{4.1}$$

Clearly $\widehat{\phi}^*(x) = \phi^*(x)$ for all $x \in \mathrm{supp}(P)$ and by definition $\widehat{\phi}^*(x)$ is the largest function $h(x)$ such that $h(x) = \phi^*(x)$ on $\mathrm{supp}(P)$ and $h(x) + \psi^*(y) \le c(x,y)$ for $x \in X$ and $y \in \mathrm{supp}(Q)$. Our next theorem states that the variational derivative is equal to $\widehat{\phi}^*(x)$ and extends the results proved in [20, 25] for the proximal 1-Wasserstein divergence. For simplicity, we focus on the case where c is bounded and uniformly continuous, though the result likely extends to more general cost functions as well.

THEOREM 4.1 (Variational derivative of the proximal OT divergence). Suppose c is uniformly continuous. Let $\rho \in \mathcal{M}(X)$ be a finite signed measure with total mass 0, $\rho(X) = 0$, such that $\rho = \rho_+ - \rho_-$ and $\rho_\pm \in \mathcal{P}(X)$ are mutually singular. Assume that for some $\alpha_0 > 0$, $P + \alpha\rho \in \mathcal{P}(X)$ if $\alpha \le \alpha_0$. Then, for every $\varepsilon > 0$, we have

$$\lim_{\alpha \to 0} \frac{\mathfrak{D}^c_{\mathrm{KL},\varepsilon}(P + \alpha\rho \| Q) - \mathfrak{D}^c_{\mathrm{KL},\varepsilon}(P\|Q)}{\alpha} = \int \widehat{\phi}^* d\rho, \tag{4.2}$$

where $\widehat{\phi}^*$ is given by (4.1) and (ϕ^*, ψ^*) are the optimizer in (3.8).

The proof of the theorem is given in Appendix E. In the sequel, we will use the following symbolic notation involving the variational derivative,

$$\frac{\delta \mathfrak{D}^c_{\mathrm{KL},\varepsilon}(P\|Q)}{\delta P} = \widehat{\phi}^*. \tag{4.3}$$

This allows to express Theorem 4.1 in a concise form.

REMARK 4.2. From Theorem 4.1 (applied to ρ and $-\rho$), we conclude that if $P + \alpha\rho \in \mathcal{P}(X)$ for α in a neighborhood of 0 then $\mathfrak{D}^c_{\mathrm{KL},\varepsilon}(P + \alpha\rho\|Q)$ is actually differentiable at 0. However, for differentiability to hold, ρ must be absolutely continuous with respect to P. In the general case, we only obtain a one-sided derivative, but this is precisely what is required for constructing a gradient descent algorithm in probability space.

5. Algorithmic aspects of proximal OT divergences. The computational aspects of proximal divergences can be broadly categorized into two interrelated challenges: (1) computing or estimating the proximal OT divergence itself using its primal and dual formulation, and (2) computing the transport proximal operator $\mathbf{prox}^c_{\varepsilon\mathfrak{D}}$ given in Definition 2.5. Notably, there are computable formulations of the proximal OT divergence and its associated operators, for which various numerical methods have been proposed in the literature (see, e.g., [14, 40]).

5.1. *Proximal OT divergences for discrete distributions.* Proximal OT divergences are computationally tractable due to their dual and dynamic formulations, which offer two alternative algorithmic, as we discuss in the following sections. We first focus on computing the proximal OT divergence in the discrete setting, particularly when P and Q are empirical measures. In this case, the dual variational formulation of the proximal OT divergence directly translates into a convex optimization problem, allowing for efficient

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

numerical computation. Specifically, let $P = \sum_{i=1}^N a_i \delta_{x^i}, Q = \sum_{j=1}^M b_j \delta_{y^i}$ be two empirical measures where $(a_i)_{i=1}^N, (b_j)_{j=1}^M$ are weights in the probability simplex. We write the intermediate measure as $R = \sum_{i=1}^N w_i \delta_{y_j}$ ensuring that R remains absolutely continuous with respect to Q. The primal formulation of the proximal OT divergence with the KL divergence is given by

$$\mathfrak{D}_\varepsilon^c(P\|Q) = \min_{(w_i)} \left\{ \min_\pi \langle C, \pi \rangle + \epsilon \sum_{i=1}^M w_i \log \frac{w_i}{b_i} \right\}, \tag{5.1}$$

where $C \in \mathbb{R}^{N\times M}$ is the matrix with entries $C_{i,j} = c(x^i, y^j)$ and $\pi \in \mathbb{R}^{N\times M}$ is a (discrete) transport plan from P to Q, i.e., $\pi \in \Pi(a, c)$ where

$$\Pi(a, c) = \left\{ \pi \in \mathbb{R}^{N\times M},\ \pi \geq 0,\ \pi 1_N = a,\ \pi^T 1_M = c \right\}. \tag{5.2}$$

Alternatively, the proximal OT divergence can be written using the dual formulation in terms of potential $\phi \in \mathbb{R}^N, \psi \in \mathbb{R}^M$ as

$$\mathfrak{D}_\varepsilon^c(P\|Q) = \max_{\phi\oplus\psi\leq C} \left\{ \sum_{i=1}^N \phi_i a_i - \epsilon \log \sum_{j=1}^M b_j e^{-\psi_j/\epsilon} \right\}, \tag{5.3}$$

where the constraint means that $\phi_i + \psi_j \leq C_{ij}$ for $i = 1, \ldots, N$ and $j = 1, \ldots, M$. Moreover, the weights w_i of R are given by

$$w_j = \frac{e^{-\psi_j/\varepsilon}}{\sum_l e^{-\psi_l/\varepsilon}} \tag{5.4}$$

which comes from (3.10). The problem in (5.3) has a concave objective (by noting that the log-sum-exp function is convex) with linear inequality constraints. The problem can be solved using convex programming tools.

5.2. *Neural network approximations to proximal OT divergences.* Let us consider the proximal OT divergence with the KL divergence and 2-Wasserstein distance, i.e., the objective for the quadratic cost $c(x, y) = \frac{1}{2}|x - y|^2$. The dual representation (3.8) for proximal OT divergences requires optimizing over the set Φ_c which poses significant computational challenges. To address this, we use the convexification trick introduced in [56]. Building on the approach in [33], we reformulate the problem by enforcing convexity constraints on the function space. This reformulation allows us to express proximal OT divergences as a minimax optimization over convex functions, enabling the use of input convex neural networks (ICNNs) for efficient computation.

Specifically, we perform the change of variables $\widetilde{\varphi}(x) := \frac{1}{2}\|x\|^2 - \varphi(x)$ and $\widetilde{\psi}(y) := \frac{1}{2}\|y\|^2 - \psi(y)$ and then, the constraint in the dual problem can be written as $\widetilde{\varphi}(x) + \widetilde{\psi}(y) \geq xy$. Moreover, it is not too difficult to show that $\widetilde{\varphi}$ is convex and $\frac{1}{2}\|y\|^2 - \psi(y) = \frac{1}{2}\|y\|^2 - \varphi^c(y) = \widetilde{\varphi}^*(y)$ is the Legendre transform of $\widetilde{\varphi}$ at optimality, see Proposition 1.21 in [47]. Then the optimization problem (3.8) can be expressed as

$$\begin{aligned} &\sup_{\widetilde{\varphi}} \left\{ \mathbb{E}_P[-\widetilde{\varphi}] - \epsilon \log \mathbb{E}_Q[e^{\widetilde{\varphi}^*/\epsilon} e^{-\|y\|^2/2\epsilon}] \right\} + C \\ &\qquad = -\inf_{\widetilde{\varphi}} \left\{ \mathbb{E}_P[\widetilde{\varphi}] + \epsilon \log \mathbb{E}_Q[e^{\widetilde{\varphi}^*/\epsilon} e^{-\|y\|^2/2\epsilon}] \right\} + C, \end{aligned} \tag{5.5}$$

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

where $C = \frac{1}{2}\mathbb{E}_P\|x\|^2$ and $\widetilde{\varphi}$ is a convex and continuous function on X. This reformulation can also be interpreted as expressing the objective in terms of a modified cost function $c(x, y) = -xy$ and a perturbed measure Q obtained by multiplying Q by a standard Gaussian density with variance ϵ followed by re-normalization.

In practice, we solve this by parameterizing $\widetilde{\varphi}$ with an input-convex neural network in the class `ICNN`. From the proof of Theorem 3.3 in [33] we have that

$$\langle y, \nabla g(y)\rangle - \widetilde{\varphi}(\nabla g(y)) \le \langle y, \nabla\widetilde{\varphi}^*(y)\rangle - \widetilde{\varphi}(\nabla\widetilde{\varphi}^*(y)) = \widetilde{\varphi}^*(y),$$

for all differentiable $g, \widetilde{\varphi}$. Moreover, the supremum of the left-hand side over convex functions $g \in \texttt{CVX}$ achieves the Legendre transform $g = \widetilde{\varphi}^*$. Given that $x \mapsto \epsilon \log \mathbb{E}_Q[e^{x/\epsilon}e^{-\|y\|^2/2\epsilon}]$ is an increasing function, we have that

$$\sup_{g\in\texttt{CVX}} \epsilon \log \mathbb{E}_Q[e^{(\langle y,\nabla g(y)\rangle - \widetilde{\varphi}(\nabla g(y)))/\epsilon}e^{-\|y\|^2/2\epsilon}] \le \epsilon \log \mathbb{E}_Q[e^{-\widetilde{\varphi}^*(y)}e^{-\|y\|^2/2\epsilon}].$$

Using this result, we can compute the OT divergence in (5.5) by solving the variational problem

$$\inf_{\widetilde{\varphi}\in\texttt{ICNN}} \sup_{g\in\texttt{ICNN}} \left\{\mathbb{E}_P[\widetilde{\varphi}] + \epsilon \log \mathbb{E}_Q[e^{(\langle y,\nabla g(y)\rangle - \widetilde{\varphi}(\nabla g(y)))/\epsilon}e^{-\|y\|^2/2\epsilon}]\right\}. \tag{5.6}$$

5.3. *Generative flow-based algorithms.* Here, we demonstrate that a specific class of normalizing flows, known as the optimal-transport flow (OT-flow) introduced in [40], is derived from a proximal OT divergence. The OT-flow is a type of generative model that was formulated as a stable version of the normalizing flows [46, 51, 52]. Given only samples from the distribution Q, a normalizing flow is used to learn a transformation that maps these samples to a simpler, typically known distribution, such as a standard Gaussian. The learned transformation is called the flow. The inverse of this flow serves as the generative model, which transforms samples from the simpler distribution P into new samples that approximate the complex distribution Q. In [40], the flow is found via maximum likelihood estimation, which is regularized by the Wasserstein distance and the connection with the proximal OT divergences is immediate:

$$\begin{aligned}\min_{v,\rho}\Bigg\{\mathbb{E}_{\rho(x,T)}\left[\log\frac{\rho(x,T)}{P(x)}\right] + \int_0^T\int_{\mathbb{R}^d}\frac{1}{2}|v(x,t)|^2\rho(x,t)dx\,dt \\ : \frac{\partial\rho}{\partial t} + \nabla\cdot(v\rho) = 0,\ \rho(x,0) = Q(x)\Bigg\} \\ = 2T\inf_R\left[\mathfrak{D}_{\mathrm{KL}}(R\|P) + \frac{1}{2T}W_2^2(Q,R)\right] = \mathfrak{D}^2_{\mathrm{KL},2T}(Q\|P).\end{aligned} \tag{5.7}$$

This framework allows us to compute our divergences and the proximal minimizer when KL and 2-Wasserstein efficiently by neural ODEs [14] and OT-flow implemented by the tools outlined in Section 3 of [40]. Training an OT-flow requires both forward and backward simulations of invertible flows, and can be unstable in high-dimensional examples where data often lies on a lower-dimensional manifold. In fact, the choice of KL in (5.7) does not guarantee the existence of an optimizer due to the lack of absolute continuity, and there is no invertible mapping between the reference distribution being often a d-dimensional Gaussian and the target distribution supported on a lower-dimensional

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

manifold. $W_1 \bigoplus W_2$ proximal generative flows in [26] address these challenges, as they do not require the inversion of the flow or absolute continuity between the measures.

5.4. *Application to conditional sampling.* Finally, we demonstrate how the divergence can be used to seek generative models that sample target conditional distributions. Let $x \in X$ and $z \in Z$ be two variables that follow a joint distribution $Q(x, z)$. For example, x may be parameter in an inverse problem, endowed with a prior distribution $Q(x)$, and z may be observation, which is related to x by a stochastic forward model, i.e., $z = G(x) + \eta$, where η is an (e.g., additive Gaussian) observational noise variable.

Here we seek an approximation to Q that factorizes as $Q(x, z) = Q(x|z)Q(z)$ in order to extract the conditional distribution for the parameter given the observation, which is known as the *posterior distribution* in Bayesian inverse problems. The approximate distribution may be parameterized by a triangular transport map $T(x, z) = [T_x(x, z), T_z(z)]$, where $T_x : X \times Z \to X$ and $T_z : Z \to Z$ such that T pushes forward any simple product distribution $P(x)P(z)$ to $Q(x, z)$. In particular, setting $P(z) = Q(z)$ and $T_z(z) = Id(z)$, then $T_x(\cdot, z)$ pushes forward $P(x)$ to $Q(x|z)$ for all z and can be used for conditional sampling; see [4] for more details. Similarly to the unconditional setting in Section 5.3, we can seek the map T_x as a flow via maximum likelihood estimation regularized by the Wasserstein-2 distance as in (5.7). The variational problem for the flow can be written as

$$\min_{v,\rho} \left\{ \mathbb{E}_{\rho(x,z,T)} \left[\log \frac{\rho(x, z, T)}{P(x)P(z)} \right] + \int_0^T \int \int_{\mathbb{R}^d} \frac{1}{2} |v(x, z, t)|^2 \rho(x, z, t) dx dz\, dt \right.$$
$$\left. : \frac{\partial \rho}{\partial t} + \nabla \cdot (v\rho) = 0,\ \rho(x, z, 0) = Q(x, z) \right\} = \mathfrak{D}^2_{\mathrm{KL},2T}(Q\|P)\,, \quad (5.8)$$

where the velocity field $v : X \times Z \times [0, T] \to X$ is evolving the x parameter for each observation z. We emphasize that the constraint on the flow to act only on x in (5.8) results in a sequence of marginal distributions of the form $\rho(x, z, t) = \rho(x|z, t)Q(z)$ for all z; see [58] for an instantiation and application of the regularized divergence in (5.8) for solving Bayesian inverse problems.

6. Gradient flows for proximal OT divergences and connections to probability flows. In this section, we consider Q as a fixed target measure—typically arising from observed data—and explore how the proximal OT divergence $\mathfrak{D}^c_\varepsilon(P\|Q)$ can be leveraged to transport a source measure P toward Q by building gradient flows for $\mathfrak{D}^c_\varepsilon(P\|Q)$. This has been extensively studied in the recent work [25] for the special case of the 1-Wasserstein proximal divergence under the assumption that Q has a finite first moment, and [16] where the authors demonstrate the robustness of the 1-Wasserstein α-divergences in generative modeling without requiring any assumptions on Q. Finally, we establish connections with probability flow ODEs. These ODEs provide a deterministic framework for evolving distributions over time and have recently emerged as powerful tools in generative modeling.

A central step in constructing the gradient flow is the computation of the variational derivative of the divergence, denoted informally as $\frac{\delta \mathfrak{D}^c_\varepsilon(P\|Q)}{\delta P}$ given in Theorem 4.1, not

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

only facilitates the definition of the flow but also sheds light on the regularizing influence of the entropic term. In particular, the entropic regularization ensures the existence and uniqueness of a well-defined variational derivative—something that is generally lacking in the classical optimal transport setting (see, e.g., Section 7.2 of [47]). Conversely, adding the optimal transport term to the KL divergence alters its properties in a meaningful way, given that the KL divergence itself already has a well-defined variational derivative, i.e., $\frac{\delta \mathfrak{D}_{\mathrm{KL}}(P\|Q)}{\delta P} = 1 + \log \frac{dP}{dQ}$ provided that P is absolutely continuous to Q. Regularizing with the optimal transport, the variational derivative can be extended to apply to more general, non-absolutely continuous measures P. Let P be a source probability distribution and Q an arbitrary target distribution. At the formal level, a gradient flow on the space of probability measures is obtained as the solution to the following partial differential equation

$$\partial_t P_t = \operatorname{div}\left(P_t \nabla \frac{\delta \mathfrak{D}^c_\varepsilon(P_t\|Q)}{\delta P_t}\right), \quad P_0 = P. \tag{6.1}$$

As noted in Remark 2.7, the gradient flow (2.14) in the JKO scheme shares the same structure as (6.1) and finds the approximate solution at each time step by solving (2.15). Here, we aim to leverage the formulation in (6.1) to construct generative particle algorithms as discussed next. Recalling the variational derivative of the proximal OT divergence with the KL divergence given in Theorem 4.1, i.e.,

$$\frac{\delta \mathfrak{D}^c_{\mathrm{KL},\varepsilon}(P\|Q)}{\delta P} = \widehat{\phi}^* \quad \text{with} \quad \widehat{\phi}^*(x) = \inf_{y\in\operatorname{supp}(Q)} \{c(x,y) - \psi^*(y)\},$$

and

$$(\phi^*, \psi^*) = \operatorname{argmax}_{(\phi,\psi)\in\Phi_c} \left\{\mathbb{E}_P[\phi] - \varepsilon \log \mathbb{E}_Q[e^{-\psi/\varepsilon}]\right\},$$

we construct the gradient flows for proximal OT divergences as follows

$$\begin{aligned} \partial_t P_t &= \operatorname{div}\left(P_t \nabla \widehat{\phi}^*_t\right), \quad P_0 = P, \\ \widehat{\phi}^*_t(x) &= \inf_{y\in\operatorname{supp}(Q)} \{c(x,y) - \psi^*_t(y)\} \\ (\phi^*_t, \psi^*_t) &= \operatorname{argmax}_{(\phi,\psi)\in\Phi_c} \left\{\mathbb{E}_{P_t}[\phi] - \varepsilon \log \mathbb{E}_Q[e^{-\psi/\varepsilon}]\right\}. \end{aligned} \tag{6.2}$$

In the context of generative models, both the target distribution Q and the generative model P_t in (6.1) are accessible only through samples and their corresponding empirical distributions. It has been mentioned that $\mathfrak{D}^c_{\mathrm{KL},\varepsilon}(P\|Q)$ can directly compare such singular distributions as ensured by the upper bound in (2.10). We now turn to the Lagrangian formulation of the PDE (6.1), which corresponds to its underlying ODE/variational representation.

$$\begin{aligned} \frac{dY_t}{dt} &= -\nabla \widehat{\phi}^*_t(Y_t), \quad Y_0 \sim P_0 = P \\ \widehat{\phi}^*_t(x) &= \inf_{y\in\operatorname{supp}(Q)} \{c(x,y) - \psi^*_t(y)\} \\ (\phi^*_t, \psi^*_t) &= \operatorname{argmax}_{(\phi,\psi)\in\Phi_c} \left\{\mathbb{E}_{P_t}[\phi] - \varepsilon \log \mathbb{E}_Q[e^{-\psi/\varepsilon}]\right\}. \end{aligned} \tag{6.3}$$

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

To construct a particle algorithm based on (6.3), we begin with the input data $(X^{(i)})_{i=1}^N$ from the unknown target distribution Q and $(Y^{(j)})_{j=1}^M$ drawn from a source distribution P. The corresponding empirical measures are $\hat{Q}^N$ and $\hat{P}_0^M$. The Euler time discretization of (6.3) takes the form:

$$\begin{aligned} Y_{n+1}^{(j)} &= Y_n^{(j)} - \Delta t \nabla \widehat{\phi}_n^*(Y_n^{(j)}), \quad Y_0^{(j)} = Y^{(j)} \sim P_0, \qquad j = 1, \dots, M \\ \phi_n^*(Y_n^{(j)}) &= \min_{i=1,\dots,N} \left\{ c(Y_n^{(j)}, X^{(i)}) - \psi_n^*(X^{(i)}) \right\}, \qquad j = 1, \dots, M \\ (\phi_n^*, \psi_n^*) &= \operatorname{argmax}_{(\phi,\psi)\in\Phi_c} \left\{ \frac{\sum_{j=1}^M \phi(Y_n^{(i)})}{N} - \varepsilon \log \frac{\sum_{i=1}^N e^{-\psi(X^{(i)})/\varepsilon}}{N} \right\}, \end{aligned} \tag{6.4}$$

where we remind that $\Phi_c = \{(\phi, \psi) \in C_b(\mathbb{R}^d \times \mathbb{R}^d), \phi \oplus \psi \le c\}$. The discrete scheme (6.4) can be interpreted as a generative algorithm that transports the initial source samples $(Y_0^{(j)})_{j=1}^M$ from P_0 (typically a distribution that is easy to sample) over a time horizon $T = n_T \Delta t$, where n_T is the total number of steps, to a new set of generated data $(Y_{n_T}^{(j)})_{j=1}^M$ that approximate samples from Q.

In [25], the particular case of $\mathfrak{D}^1_{\mathrm{KL},\varepsilon}$—the proximal OT divergence with 1-Wasserstein distance—is studied. Consequently, the optimization problem in (6.4) is formulated over the space of Lipschitz functions. To make this tractable in practice, this function space $\mathrm{Lip}_{\frac{1}{\varepsilon}}(\mathbb{R}^{\mathrm{d}})$ is approximated using neural networks $\Gamma^{NN}_{\frac{1}{\varepsilon}}$, i.e.,

$$\phi_n^* = \operatorname{argmax}_{\phi\in\Gamma^{NN}_{\frac{1}{\varepsilon}}} \left\{ \frac{\sum_{j=1}^M \phi(Y_n^{(i)})}{N} - \varepsilon \log \frac{\sum_{i=1}^N e^{\phi(X^{(i)})/\varepsilon}}{N} \right\},$$

where the Lipschitz constraint is enforced through spectral normalization techniques [35]. Moreover, the optimization problem in the discrete scheme (6.4) for the gradient flow of $\mathfrak{D}^2_{\mathrm{KL},\varepsilon}$ can be implemented by solving the variational problem (5.6) using input convex neural networks. The gradient particle scheme in (6.4) can also be augmented with a force-matching step to recover a continuous spatiotemporal vector field that approximates $-\nabla\widehat{\phi}_n^*$ throughout space and time, not just along particle trajectories. This extension enables interpolation, finer temporal resolution, and makes the method fully generative, allowing simulation from arbitrary initial data; see [17]. Force matching, conceptually related to score matching (6.6), was originally introduced in molecular dynamics to reconstruct interatomic potentials from *ab initio* force data [21], and has since found widespread application in coarse-graining of molecular simulations, see for example the recent review [37].

Connections to probability flows and score matching. Next, we explore the connection between gradient flows and probability flow ODEs (i.e., deterministic sampling) [34]. Given an initial condition p_0, the Fokker-Planck equation

$$\partial_t p(t,y) = -\operatorname{div}(f(t,y)p(t,y)) + \frac{1}{2} g^2(t) \Delta p(t,y)$$

describes the evolution of the marginal density p_t for the random variable Y_t following the forward SDE

$$dY_t = f(t, Y_t)dt + g(t)dW_t,$$

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

where $f(t,\cdot):\mathbb{R}^d\to\mathbb{R}^d$ and $g(t)\in\mathbb{R}$ are the drift and diffusion coefficient respectively, and W_t is a standard Brownian motion. The Laplacian of p can be written as a transport term driven by the score function $\nabla\log p$, i.e.,

$$\Delta p=\operatorname{div}(p\nabla\log p)$$

and the Fokker-Planck equation is written as

$$\frac{\partial p_t(y)}{\partial t}=-\operatorname{div}\left(p_t(y)v(t,y)\right),\qquad v(t,y)=f(t,y)-\frac{g(t)^2}{2}\underbrace{\nabla\log p_t(y)}_{\text{score function of }p_t}.$$

The corresponding probability flow ODE is given by

$$\frac{dY_t}{dt}=v(t,Y_t),\quad Y_0\sim p_0. \tag{6.5}$$

This perspective also forms the core idea in [50], which used as a deterministic alternative to the reverse-time stochastic differential equation (SDE), offering a new way to generate samples and compute exact likelihoods. In fact, the authors first use the forward SDE to gradually add noise to the data, then train a time-dependent score-based model $s_\theta(y)$ by minimizing the Fisher divergence between the model and the data distributions defined as

$$\mathbb{E}_{p(y)}\left[\|s_\theta(y)-\nabla\log p_t(y)\|_2^2\right] \tag{6.6}$$

using score-matching techniques (e.g., [49]), and finally generate samples by solving the corresponding probability flow ODE. Flow matching [31] and stochastic interpolants [1] are related frameworks that also seek a velocity field of the form in (6.5) to match a prescribed path of marginal distributions across time.

In our gradient flow perspective, starting from the simple case of the classical KL divergence and computing its variational derivative reveals a key connection to score-based modeling. Specifically, the gradient of the variational derivative of the KL divergence—equal to the gradient of the discriminator—can be written in terms of the score function: $\nabla\frac{\delta\mathfrak{D}_{\mathrm{KL}}(p_t\|q)}{\delta p_t}=\nabla\phi_t^*=\nabla\log\frac{p_t}{q}=\nabla\log p_t-\nabla\log q$ and thus

$$\text{score function of } p_t\ =\nabla\log p_t=\nabla\phi_t^*+\nabla\log q. \tag{6.7}$$

This formulation illustrates how score-based modeling naturally arises from the KL divergence, thereby linking our gradient flow framework to the probability flow ODE. In fact, for specific choices of the functions $g(t)$ and $f(t,y)$, this yields the gradient flow of the KL divergence, which is further generalized by the gradient flow associated with proximal OT divergences in (6.1). From a computational standpoint, rather than training a score-based model directly, we solve the variational problem defined by (6.4) using convex optimization techniques, as detailed in Section 5.1 and neural networks and discriminators in analogy to GANs [25].

In fact, our methods are a type of probability flow—in the sense that we solve deterministic dynamics for the particles—but the velocities fields are informed by proximal OT divergences rather than the score [50] or the KL divergence [34]. In general, the probability flow corresponding to the gradient flow of any proximal OT divergence can be computed through approximating discriminators ϕ_t^* without explicitly knowing the

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

distribution q. The resulting probability flow does recover the Wasserstein gradient flow corresponding to the usual KL divergence, i.e., (6.7) reduces to the usual score function.

Appendix A. Proofs of the divergence theorems in Section 2.2. We start with Theorem 2.4 which is proved by using the lower semicontinuity of the KL divergence and transport cost together with the compactness of the level sets of the KL divergence.

Proof of Theorem 2.4. The upper bound (2.10) is easily obtained by picking $R = P$ and $R = Q$ and using the divergence property for $T_c(P, Q)$ and $P\|Q)$.

To show the existence of the minimizer, let R_n be a minimizing sequence such that

$$T_c(P, R_n) + \epsilon\mathfrak{D}(R_n\|Q) \leq \mathfrak{D}^c_\varepsilon(P\|Q) + \frac{1}{n}. \tag{A.1}$$

Since the level sets $\cdot\|Q)$ are precompact in the weak topology there exists a convergent subsequence (which we denote again by R_n) such that $R_n \to R^*$ weakly. Since both $T_c(P, \cdot)$ and $\mathfrak{D}(\cdot\|Q)$ are lower semicontinuous in the weak topology we have

$$\begin{aligned} T_c(P, R^*) + \varepsilon\mathfrak{D}(R^*\|Q) &\leq \liminf_{n\to\infty} \{T_c(P, R^n) + \varepsilon\mathfrak{D}(R^n\|Q)\} \\ &\leq \mathfrak{D}^c_\varepsilon(P\|Q) \end{aligned} \tag{A.2}$$

which shows that R^* is a minimizer.

The uniqueness follows from the strict convexity of $\cdot\|Q)$ and the convexity of $T_c(P, \cdot)$. If R^*_1 and R^*_2 are distinct minimizers then set $R^*_3 = \frac{1}{2}(R^*_1 + R^*_2)$. We have then

$$\begin{aligned} &T_c(P, R^*_3) + \varepsilon\mathfrak{D}(R^*_3\|Q) \\ &< \frac{1}{2}\left(T_c(P, R^*_1) + \epsilon\mathfrak{D}(R^*_1\|Q) + T_c(P, R^*_2) + \epsilon\mathfrak{D}(R^*_2\|Q)\right) \\ &= \mathfrak{D}^c_\varepsilon(P\|Q)\,. \end{aligned} \tag{A.3}$$

This contradicts the fact that R^*_1 and R^*_2 are minimizers. □

Proof of Theorem 2.3. The divergence property, convexity and lower semiconitnuity are proved as follows:

Divergence property. For the divergence property the nonnegativity in (2.8) follows from the nonnegativity of $T_c(P, R)$ and $\mathfrak{D}(R\|Q)$. If the proximal OT divergence $\mathfrak{D}^c_\varepsilon(P\|Q) = 0$ then using the minimizer R^* provided by Theorem 2.4 we have

$$0 = T_c(P, R^*) + \mathfrak{D}(R^*\|Q) \tag{A.4}$$

and thus $R^* = Q = P$ and thus $P = Q$.

Convexity. Consider two pairs or probability measures P_1, Q_1 and P_2, Q_2 and the corresponding minimizers R^*_1 and R^*_2. Note also that the function $F(P, Q, R) = T_C(P, R) + \varepsilon\mathfrak{D}(R\|Q)$ is jointly convex. We have then

$$\begin{aligned} &\mathfrak{D}^c_\varepsilon(\alpha P_1 + (1-\alpha)P_2\|\alpha Q_1 + (1-\alpha Q_2) \\ &\leq F(\alpha(P_1, Q_1, R^*_1) + (1-\alpha)(P_2, Q_2, R^*_2)) \\ &\leq \alpha F(P_1, Q_1, R^*_1) + (1-\alpha)F(P_2, Q_2, R^*_2) \\ &= \alpha\mathfrak{D}^c_\varepsilon(P_1\|Q_1) + (1-\alpha)\mathfrak{D}^c_\varepsilon(P_2\|Q_2) \end{aligned} \tag{A.5}$$

which proves convexity.

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

Lower semicontinuity. For the lower semicontinuity property let $(P_n, Q_n) \in \mathcal{P}(X) \times \mathcal{P}(Y)$ a weakly convergent sequence and suppose n_k is such that

$$\liminf_n \mathfrak{D}_\varepsilon^c(P_n\|Q_n) = \lim_k \mathfrak{D}_\varepsilon^c(P_{n_k}\|Q_{n_k})\,. \tag{A.6}$$

If the limit is infnite there is nothing to prove so we can assume the limit is finite. Denoting by $R^*_{n_k}$ the minimizer, the sequence

$$\mathfrak{D}_\varepsilon^c(P_{n_k}\|Q_{n_k}) = T_c(P_{n_k}, R^*_{n_k}) + \varepsilon\mathfrak{D}(R^*_{n_k}\|Q_{n_k}) \tag{A.7}$$

is bounded and therefore the sequence $R^*_{n_k}$ has a convergent subsequence (also denoted by $R^*_{n_k}$) with limit R. Therefore, using the lower semicontinuity of T_c and $\mathfrak{D}$ in both argument we have

$$\begin{aligned}\liminf_n \mathfrak{D}_\varepsilon^c(P_n\|Q_n) &= \lim_k \mathfrak{D}_\varepsilon^c(P_{n_k}\|Q_{n_k})\\ &= \lim_k T_c(P_{n_k}, R^*_{n_k}) + \varepsilon\mathfrak{D}(R^*_{n_k}\|Q_{n_k})\\ &\geq T_c(P,R) + \varepsilon\mathfrak{D}(R\|Q)\\ &\geq \mathfrak{D}_\varepsilon^c(P\|Q)\end{aligned} \tag{A.8}$$

which proves lower semicontinuity and concludes the proof. □

Appendix B. Proof of the interpolation theorem in Section 2.3.

Proof of Theorem 2.8. The proof of the limits $\varepsilon \to 0$ and $\varepsilon \to \infty$ go as follows:

The limit $\varepsilon \to 0$. Assume first that $P\|Q) < \infty$. Then we have

$$\frac{1}{\varepsilon}\mathfrak{D}_\varepsilon^c(P\|Q) = \inf_{R\in\mathcal{P}(Y)} \left\{\mathfrak{D}(R\|Q) + \frac{1}{\varepsilon}T_c(P,R)\right\} \leq \mathfrak{D}(P\|Q) < \infty\,. \tag{B.1}$$

This implies that

$$\limsup_{\varepsilon\to 0} \frac{1}{\varepsilon}\mathfrak{D}_\varepsilon^c(P\|Q) \leq \mathfrak{D}(P\|Q)\,. \tag{B.2}$$

Suppose $\epsilon_n \to 0$ and let R^*_n be the corresponding minimizers in (B.1). The bound on the KL divergence and (B.1) implies that the sequence $\{R^*_n\}$ is weakly compact and so passing to a subsequence we can assume that $R^*_n \to R$. By the lower semicontinuity of T_c and (B.1) we have

$$T_c(P,R) \leq \liminf_n \varepsilon_n \frac{1}{\varepsilon_n} T_c(P, R^*_n) \leq \lim_n \varepsilon_n \mathfrak{D}(P\|Q) = 0 \tag{B.3}$$

and therefore $R^* = P$. Finally, by the lower semicontinuity of $\mathfrak{D}(P\|Q)$

$$\begin{aligned}\liminf_n \frac{1}{\varepsilon_n}\mathfrak{D}_{\varepsilon_n}^c(P\|Q) &= \liminf_n \left\{\mathfrak{D}(R^*_n\|Q) + \frac{1}{\varepsilon_n}T_c(P,R^*_n)\right\}\\ &\geq \liminf_n \mathfrak{D}(R^*_n\|Q) \geq \mathfrak{D}(P\|Q)\,.\end{aligned} \tag{B.4}$$

Since the the sequence ϵ_n is arbitrary this concludes the proof in the case where $\mathfrak{D}(P\|Q) < \infty$.

Next let us consider the case where $\mathfrak{D}(P\|Q) = \infty$ and so we need to show that $\liminf_{\varepsilon\to 0} \frac{1}{\varepsilon}\mathfrak{D}_\varepsilon^c(P\|Q) = \infty$. By contradiction assume it does not and thus there exists a sequence $\varepsilon_n \to 0$ such that $\lim_n \frac{1}{\varepsilon_n}\mathfrak{D}_{\varepsilon_n}^c(P\|Q) = D < \infty$. For this sequence we can repeat the argument used in the case of finite $\mathfrak{D}(P\|Q) = \infty$ (see equations (B.3) and

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

(B.4)) to show that, passing to a subsequence if necessary we have $\lim_n \frac{1}{\varepsilon_n}\mathfrak{D}_c^{\varepsilon_n}(P\|Q) \geq \mathfrak{D}(P\|Q) = +\infty$ which is a contradiction.

The limit $\epsilon \to \infty$. Assume first that $T_c(P,Q) < \infty$. The argument is very similar to the limit $\varepsilon \to 0$. We have

$$\mathfrak{D}_\varepsilon^c(P\|Q) \leq \inf_{R\in\mathcal{P}(Y)} \{T_c(P,R) + \varepsilon\mathfrak{D}(R\|Q)\} \leq T_c(P,Q) < \infty\,. \tag{B.5}$$

This implies immediately that $\limsup_{\varepsilon\to\infty} \mathfrak{D}_\varepsilon^c(P\|Q) \leq T_c(P,Q)$. Arguing as above given any sequence $\varepsilon_n \to \infty$ the minimizers R_n^* in (B.5) have a convergent subsequence $R_n^* \to R$. The lower semicontinuity of $\mathfrak{D}(\cdot\|Q)$ implies that

$$\mathfrak{D}(R\|Q) \leq \liminf_n \frac{1}{\varepsilon_n}\varepsilon_n\mathfrak{D}(R_n^*\|Q)) \leq \lim_n \frac{1}{\varepsilon_n}T_c(P,Q) = 0, \tag{B.6}$$

and thus $R = Q$. This implies that

$$\begin{aligned}\liminf_n \mathfrak{D}_{\varepsilon_n}^c(P\|Q) &= \liminf_n \{T_c(P,R_n^*) + \varepsilon_n\mathfrak{D}(R_n^*\|Q)\}\\ &\geq \liminf_n T_c(P,R_n^*) \geq T_c(P,Q).\end{aligned} \tag{B.7}$$

Since the the sequence ϵ_n is arbitrary this concludes the proof. The case $T_c(P,R) = \infty$ is argued similarly as above this concludes the proof. □

Appendix C. Proofs of the duality results in Section 3. In this section we prove Theorem 3.1. We start by proving this theorem in the special case where X and Y are compact and c is continuous, and we prove it by using similar convex optimization tools as the ones in [47] (see Theorem 1.46).

We start by showing that problem has a maximizer. The proof is a slight generalization of the corresponding result in optimal transport and use the notion of c-transform which we now recall. For $\phi : X \to \bar{\mathbb{R}}$, we define

$$\phi^c(y) = \inf_x c(x,y) - \phi(x) \tag{C.1}$$

and for $\psi : Y \to \bar{\mathbb{R}}$, we define

$$\psi^{\overline{c}}(x) = \inf_y c(x,y) - \psi(y). \tag{C.2}$$

We call a function $\psi(y)$ c-concave if $\psi = \phi^c$ for some function $\phi(x)$ (similarly for $\overline{c}$-concavity). The key property of c-concave functions on compact sets is that they form an equicontinuous family (see e.g [47], p.12 and the argument below).

Lemma C.1. If X and Y are compact and c is continuous then there exists a solution to the problem

$$\sup_{(\phi,\psi)\in\Phi_c} \left\{\mathbb{E}_P[\phi] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi/\varepsilon}]\right\} \tag{C.3}$$

and φ is c-concave, ψ is $\overline{c}$-concave and $\psi = \varphi^c$.

Proof. Consider a pair (ϕ,ψ) with $\phi \oplus \psi \leq c$, that is $\psi(y) \leq c(x,y) - \phi(x)$. Since the map

$$\psi \mapsto -\varepsilon \ln \mathbb{E}_Q[e^{-\psi/\varepsilon}]$$

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

is increasing, the optimized functional in (C.3) always increases if ψ increases and so we can always replace in the optimization a pair (φ, ψ) by (φ, φ^c). So we can always assume that ψ is c-concave. Similarly we can always replace a pair (φ, ψ) by $(\psi^{\overline{c}}, \psi)$ and assume that ϕ is $\overline{c}$-concave.

Given a maximizing sequence (ϕ_n, ψ_n) in (C.3) and applying c- and $\overline{c}$-transforms we obtain another maximizing sequence (for which we use the same symbols). Since X and Y are compact the function c is uniformly continuous we denote by ω be its modulus on continuity (that is $|c(x,y) - c(x',y')| \leq \omega(d(x,x') + d(y,y'))$ for all $x, x' y, y'$).

By definition of the c-transform any c-concave function ϕ^c has the same modulus of continuity as c. This implies that the family (ϕ_n, ψ_n) is equicontinuous.

To get equiboundedness we note that replacing (ϕ_n, ψ_n) by $(\phi_n + a_n, \psi_n - a_n)$ for an arbitrary constant a_n does not change the value of the functional in (C.3). By choosing the constants a_n appropriately we can assume that $\phi_n(x) \geq 0$ and thus, by equicontinuity, we can assume that $0 \leq \phi_n \leq \omega(\mathrm{diam}(X))$ and since $\psi_n = \phi_n^c$ we get $\psi_n \in [\min(c) - \omega(\mathrm{diam}(X)), \max(c)]$. This shows that the sequence (ϕ_n, ψ_n) can be chosen to be equibounded. By Arzela-Ascoli theorem there is a convergent subsequence converging uniformly to the limit (ϕ, ψ) which still satisfies $\phi \oplus \psi \leq c$. By uniform continuity we have

$$\mathbb{E}_P[\phi_n] - \varepsilon \log \mathbb{E}_Q[e^{-\psi_n/\varepsilon}] \to \mathbb{E}_P[\phi] - \varepsilon \log \mathbb{E}_Q[e^{-\psi/\varepsilon}] \tag{C.4}$$

and thus the supremum is attained in (C.3). □

Let us consider the following map $F : C(X \times Y) \to \mathbb{R}$ given by

$$F(p) = - \max_{(\phi,\psi) \in \Phi_{c-p}} \left\{ \mathbb{E}_P[\varphi] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi/\varepsilon}] \right\} \tag{C.5}$$

denotes a certain class of functions. By Lemma C.1, the map F is well-defined, and in particular, $F(0)$ corresponds to the negative of the right-hand side of (3.8).

To establish duality, we will employ convex duality, specifically showing that $F = F^{**}$ (the double Legendre transform). To proceed, we first need Lemma C.2.

LEMMA C.2. Suppose X and Y are compact. The map $F : C(X \times Y) \to \mathbb{R}$ is convex and lower semi-continuous and thus by Legendre-Fenchel theorem we have $F = F^{**}$

Proof. For the convexity let $p_0, p_1 \in C(X \times y)$ let (φ_0, ψ_0) and (φ_1, ψ_1) the corresponding maximizers (given by Lemma C.1. Let $p_t = (1-t)p_0 + tp_1$ and set $\varphi_t = (1-t)\varphi_0 + t\varphi_1$ and $\psi_t = (1-t)\psi_0 + t\psi_t$.

The map $(\phi, \psi) \to \mathbb{E}_P[\varphi] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi/\varepsilon}]$ is concave and thus

$$\mathbb{E}_P[\varphi_t] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi_t/\varepsilon}] \geq -(1-t)F(p_0) - tF(p_1) \tag{C.6}$$

and thus F is convex.

For the lower semicontinuity suppose $p_k \to p$. Passing to a subsequence if necessary (which we denote by p_k again) we can assume that $\liminf F(p_k) = \lim_k F(p_k)$. By the converse statement in Arzela-Ascoli theorem the sequence $\{p_k\}$ is equibounded and equicontinuous. We now use the Lemma C.1 with cost $c - p_k$ and obtain maximizers (φ_k, ψ_k) which are $c - p_k$ and $\overline{c - p_k}$ concave. This implies that $(\varphi_{n_k}, \psi_{n_k})$are also equicontinuous and equibounded (using the same arguments as in the proof of Lemma C.1. By

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

Arzela-Ascoli again, passing to another subsequence, we can assume that $\varphi_k \to \varphi$ and $\psi_k \to \psi$ uniformly. Since $\varphi_k \oplus \psi_k \leq c - p_k$ we have then $\varphi \oplus \psi \leq c - p$. So

$$\begin{aligned} \liminf F(p_n) &= \lim_k F(p_k) = -\lim_k \left(\mathbb{E}_P[\varphi_k] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi_k/\varepsilon}] \right) \\ &= -\left(\mathbb{E}_P[\varphi] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi/\varepsilon}] \right) \geq F(p) \end{aligned} \tag{C.7}$$

and thus F is lower semicontinuous. □

With these two technical lemmas in place, the core argument establishing duality is contained in Proposition C.3.

PROPOSITION C.3. For the map $F(p)$ given in (C.5) we have

$$F^{**}(0) = -\inf_{R \in \mathcal{P}(Y)} \{W(P, R) + R, Q)\} \tag{C.8}$$

Proof. Let us first compute the Legendre-Fenchel transform of $F^*(\pi)$ of F. For $\pi \in \mathcal{M}(X \times Y)$ we have

$$\begin{aligned} F^*(\pi) &= \sup_{p \in C(X \times Y)} \left\{ \int p d\pi - F(p) \right\} \\ &= \sup_{p \in C(X \times Y)} \left\{ \int p d\pi + \sup_{\varphi \oplus \psi \leq c - p} \{ \mathbb{E}_P[\varphi] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi/\varepsilon}] \} \right\} \end{aligned} \tag{C.9}$$

which can be written as a unique sup over p, ϕ, ψ. If $\pi \in \mathcal{M}(X \times Y)$ is not a positive Borel measure then these exist $q \leq 0$ such that $\int q d\pi > 0$. If we take then $\varphi = 0$, $\psi = 0$ and $p = c + nq$ then we find $F^*(\pi) \geq \int p d\pi + n \int q d\pi$ and thus $F^*(\pi) = +\infty$.

If $\pi \in \mathcal{M}(X \times Y)$ is a positive Borel measure and (ϕ, ψ) is fixed to take the supremum over p we should take p as large as possible, that is $\phi \oplus \psi = c - p$ and thus

$$\begin{aligned} F^*(\pi) &= \sup_{\varphi, \psi} \left\{ \int (c - \varphi \oplus \psi) d\pi + \mathbb{E}_P[\varphi] - \varepsilon \log \mathbb{E}_Q[e^{-\psi/\varepsilon}] \right\} \\ &= \int_{X \times Y} c d\pi + \sup_{\varphi} \left\{ \mathbb{E}_P[\varphi] - \int_{X \times Y} \varphi d\pi \right\} + \varepsilon \sup_{\psi} \left\{ \int (-\psi/\varepsilon) d\pi - \ln \mathbb{E}_Q[E^{-\psi/\varepsilon}] \right\}. \end{aligned} \tag{C.10}$$

Denoting by π_X and π_Y the marginals of π, we have for the supremum over φ

$$\sup_{\varphi} \left\{ \mathbb{E}_P[\varphi] - \int_{X \times Y} \varphi d\pi \right\} = \begin{cases} 0 & \text{if } \pi_X = P, \\ \infty & \text{otherwise.} \end{cases} \tag{C.11}$$

For the supremum over ψ, the Donsker-Varadhan representation of the KL divergence gives

$$\sup_{\psi} \left\{ \int (-\psi/\varepsilon) d\pi - \ln \mathbb{E}_Q[E^{-\psi/\varepsilon}] \right\} = \begin{cases} \pi_Y \| Q) & \text{if } \pi_Y \in \mathcal{P}(Y) \\ +\infty & \text{if } \pi_Y \in \mathcal{M}(Y) \setminus \mathcal{P}(Y) \end{cases}. \tag{C.12}$$

Therefore we find

$$F^*(\pi) = \begin{cases} \mathbb{E}_\pi[c] + \varepsilon \pi_Y \| Q) & \text{if } \pi_X = P \text{ and } \pi_Y \in \mathcal{P}(Y), \\ +\infty & \text{otherwise.} \end{cases} \tag{C.13}$$

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

Therefore

$$\begin{aligned} F^{**}(0) &= \sup_{\pi} \int 0 d\pi - F^*(\pi) \\ &= - \inf_{R \in \mathcal{P}(Y)} \inf_{\pi \in \Pi(P,R)} \{\mathbb{E}_\pi[c] + \varepsilon \pi_Y \| Q)\} \\ &= - \inf_{R \in \mathcal{P}(Y)} \{W(P,R) + \varepsilon \mathfrak{D}(R,Q)\} \end{aligned} \tag{C.14}$$

and this concludes the proof. □

Proof of Theorem 3.3. Let R^* be the (unique) minimizer in (2.3), π^* an optimal transport plan between P and R^* and (ϕ^*, ψ^*) a maximizer for the (C.3). We have then

$$\mathbb{E}_{\pi^*}[c] + \varepsilon \mathfrak{D}(R^* \| Q) = \mathbb{E}_P[\phi^*] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi^*/\varepsilon}] \tag{C.15}$$

which we can rewrite as

$$\underbrace{\mathbb{E}_{\pi^*}[c - \phi^* - \psi^*]}_{\geq 0} + \varepsilon \underbrace{\left\{ \mathfrak{D}(R^* \| Q) - \mathbb{E}_R[-\psi^*/\varepsilon] + \log \mathbb{E}_Q[e^{-\psi^*/\varepsilon}] \right\}}_{\geq 0} = 0. \tag{C.16}$$

The first term is non-negative by definition of the transport cost and the second by Donsker-Varadhan's formula. For the second term to be zero both equations (3.9) and (3.10) to hold and this concludes the proof. We note further that ψ^* is uniquely determined Q almost surely (up to a constant) and thus, by continuity, uniquely determined in $\mathrm{supp}(Q)$. From the first term in (3.12) we see that $\phi \oplus \psi = c$ π^* almost surely and thus ϕ^* is uniquely determined P almost surely (up to a constant) and also in $\mathrm{supp}(P)$. This concludes the proof of Theorem 3.3. □

To extend the duality formula from the compact case to the general case we proceed in two steps. We first relax the compactness assumption on X and Y but assume that the cost c is bounded and continuous.

THEOREM C.4. Assume X and Y are Polish space and c is bounded continuous then we have

$$\mathfrak{D}^c_\varepsilon(P \| Q) = \sup_{(\phi,\psi) \in \Phi_c} \left\{ \mathbb{E}_P[\phi] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi/\varepsilon}] \right\}. \tag{C.17}$$

Proof. Recall that by duality for optimal transport and the KL divergence we have for any $R \in \mathcal{P}(Y)$

$$\begin{aligned} \mathbb{E}_P[\phi] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi/\varepsilon}] &= \mathbb{E}_P[\phi] + \mathbb{E}_R[\psi] + \varepsilon \left(\mathbb{E}_R[-\psi/\varepsilon] - \ln \mathbb{E}_Q[e^{-\psi/\varepsilon}] \right) \\ &\leq T_c(P,R) + \mathfrak{D}(R \| Q) \end{aligned} \tag{C.18}$$

and thus

$$\mathfrak{D}^c_\varepsilon(P \| Q) \geq \sup_{(\phi,\psi) \in \Phi_c} \left\{ \mathbb{E}_P[\phi] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi/\varepsilon}] \right\}. \tag{C.19}$$

We show the reverse inequality. Since X and Y are Polish spaces, we pick a sequence $\delta_n \to 0$ and compact sets X_n, Y_n such that $P(X \setminus X_n) \leq \delta_n$ and $Q(Y \setminus Y_n) \leq \delta_n$. We denote by P_n and Q_n the restriction of P to X_n (i.e., $P_n(A) = P(A \cap X_n)/P(X_n)$). It is easy to verify that P_n and Q_n converge weakly to P and Q respectively. We also denote c_n the restriction of c to $X_n \times Y_n$ and we have then

$$\mathfrak{D}^c_\varepsilon(P_n \| Q_n) = \inf_{R \in \mathcal{P}(Y_n)} \{T_{c_n}(P_n, R) + \varepsilon \mathfrak{D}(R \| Q_n)\} = \mathfrak{D}^{c_n}_\varepsilon(P_n \| Q_n) \tag{C.20}$$

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

and we can use the results obtained in the compact case.

Using the lower semicontinuity of $\mathfrak{D}_\varepsilon^c(P\|Q)$ given an arbitrary $\alpha > 0$ we can pick N such that

$$\begin{aligned}\mathfrak{D}_\varepsilon^c(P\|Q) &\leq \mathfrak{D}_\varepsilon^{c_N}(P_N\|Q_N) + \alpha \\ &= \sup_{(\phi,\psi)\leq\Phi_{c_N}} \left\{\mathbb{E}_{P_N}[\phi] - \varepsilon \ln \mathbb{E}_{Q_N}[e^{-\psi/\varepsilon}]\right\} + \alpha \\ &= \mathbb{E}_{P_N}[\phi_N^*] - \varepsilon \ln \mathbb{E}_{Q_N}[e^{-\psi_N^*/\varepsilon}] + \alpha,\end{aligned} \tag{C.21}$$

where we have used the duality for the compact case and the existence of a maximizers (Lemma C.1). Recall that $\phi_N^*(x) + \psi_N^*(y) = c(x,y)$ is $x \in \operatorname{supp}(P_N)$ and $y \in \operatorname{supp}(Q_N)$.

Next (using the standard c-transform trick in optimal transport) we extend the pair $(\phi_N^*(x), \psi_N^*(y))$ to X and Y by setting

$$\begin{aligned}\widetilde{\phi}(x) &= \inf_{y\in X_N} \{c(x,y) - \psi_N^*(y)\} \\ \widetilde{\psi}(y) &= \inf_{y\in Y} \left\{c(x,y) - \widetilde{\phi}(y)\right\}.\end{aligned} \tag{C.22}$$

This definition ensures that $\phi(x) = \phi_N(x)$ on $\operatorname{supp}(P_N)$ and $\psi(x) = \psi_N(y)$ on $\operatorname{supp}(Q_N)$ and moreover $\phi(x) + \psi(y) \leq c(x,y)$ for all x, y.

The fact that c is bounded provides a bound on ϕ and ψ. Without loss of generality we can assume that $c \geq 0$ and we pick x_0, y_0 such that $\phi_N(x_0) + \psi_N(x_0) = c(x_0, y_0) \geq 0$. The pair ϕ_N, ψ_N is determined up to a constant which we can choose so that $\phi_N(x_0) \geq 0$ and $\psi_N(x_0) \geq 0$. From (C.22) we get first

$$\phi(x) \leq c(x, y_0), \quad \psi(y) \leq c(x_0, y) \tag{C.23}$$

and then

$$\phi(x) \geq \inf_y \{c(x,y) - c(x_0,y)\}, \quad \psi(x) \geq \inf_x c(x,y) - c(x,y_0). \tag{C.24}$$

In particular we have $\|\phi\|_\infty \leq \|c\|_\infty$ and $\|\psi\|_\infty \leq \|c\|_\infty$. Since P_N and Q_N converge weakly to P and Q by choosing N large enough we have then form (C.21)

$$\begin{aligned}\mathfrak{D}_\varepsilon^c(P\|Q) &\leq \mathbb{E}_P[\widetilde{\phi}] - \varepsilon \ln \mathbb{E}_Q[e^{-\widetilde{\psi}/\varepsilon}] + 2\alpha \\ &\leq \sup_{(\phi,\psi)\in\Phi_c} \left\{\mathbb{E}_P[\widetilde{\phi}] - \varepsilon \ln \mathbb{E}_Q[e^{-\widetilde{\psi}/\varepsilon}]\right\} + 2\alpha.\end{aligned} \tag{C.25}$$

Since α is arbitrary this concludes the proof of the theorem. □

Finally, we apply an additional approximation argument to handle general lower semicontinuous cost functions.

THEOREM C.5. Assume that X and Y are Polish spaces and that c is bounded below and lower semicontinuous and that $c \leq a \oplus b$ with $a \in L^1(P)$ and $b \in L^1(Q)$. Then duality holds

$$\mathfrak{D}_\varepsilon^c(P\|Q) = \sup_{(\phi,\psi)\in\Phi_c} \left\{\mathbb{E}_P[\phi] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi/\varepsilon}]\right\}. \tag{C.26}$$

Proof. The condition $c \leq a \oplus b$ with $a \in L^1(P)$ implies that $\mathfrak{D}_\varepsilon^c(P\|Q)$ is finite since

$$\mathfrak{D}_\varepsilon^c(P\|Q) \leq T_c(P,Q) \leq \inf_{\pi\in\Pi(P,Q)} \int c d\pi \leq \mathbb{E}_P[a] + \mathbb{E}_Q[b] < \infty.$$

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

We can write the lower semicontinuous c as an increasing limit of nonnegative continuous function c_n and by replacing c_n by $\max\{c_n, n\}$ we can assume that the functions c_n are bounded and continuous.

We already know that

$$\mathfrak{D}_\varepsilon^c(P\|Q) \geq \sup_{(\phi,\psi)\in\Phi_c} \left\{\mathbb{E}_P[\phi] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi/\varepsilon}]\right\}. \tag{C.27}$$

By Theorem C.4 and because c_n is increasing to c we have

$$\begin{aligned}\mathfrak{D}_\varepsilon^{c_n}(P\|Q) &= \sup_{(\phi,\psi)\in\Phi_{c_n}} \left\{\mathbb{E}_P[\phi] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi/\varepsilon}]\right\} \\ &\leq \sup_{(\phi,\psi)\in\Phi_c} \left\{\mathbb{E}_P[\phi] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi/\varepsilon}]\right\},\end{aligned} \tag{C.28}$$

and therefore it is enough to show that

$$\lim_{n\to\infty} \mathfrak{D}_\varepsilon^{c_n}(P\|Q) \geq \mathfrak{D}_\varepsilon^c(P\|Q)\,. \tag{C.29}$$

Since c_n is increasing then the sequence $\mathfrak{D}_\varepsilon^{c_n}(P\|Q)$ is increasing in n and so the limit exists. We have, using the optimal measure R_n and an optimal transport plan $\pi_n \in \Pi(P, R_n)$

$$\mathfrak{D}_\varepsilon^{c_n}(P\|Q) = \int c_n d\pi_n + \varepsilon \mathfrak{D}_{\mathrm{KL}}(R_n\|Q)\,. \tag{C.30}$$

We have

$$\varepsilon \mathfrak{D}_{\mathrm{KL}}(R_n\|Q) \leq \mathfrak{D}_\varepsilon^{c_n}(P\|Q) \leq \mathfrak{D}_\varepsilon^c(P\|Q)\,, \tag{C.31}$$

and therefore the sequence $\{R_n\}$ is tight as well. This implies that the sequence of optimal transport plans $\{\pi_n\}$ is tight and therefore there exist accumulation points π^*, R^* with $\pi^* \in \Pi(P, R^*)$. Since c_n is increasing we have for $n \geq m$

$$\mathfrak{D}_\varepsilon^{c_n}(P\|Q) \geq \int c_m d\pi_n + \varepsilon \mathfrak{D}_{\mathrm{KL}}(R_n\|Q) \tag{C.32}$$

and thus for any m

$$\begin{aligned}\lim_{n\to\infty} \mathfrak{D}_\varepsilon^{c_n}(P\|Q) &\geq \limsup_{n\to\infty} \int c_m d\pi_n + \varepsilon \mathfrak{D}_{\mathrm{KL}}(R_n\|Q) \\ &\geq \int c_m d\pi^* + \varepsilon \mathfrak{D}_{\mathrm{KL}}(R^*\|Q).\end{aligned} \tag{C.33}$$

By monotone convergence $\lim_{m\to\infty} \int c_m d\pi^* = \int c d\pi^*$ and thus

$$\lim_{n\to\infty} \mathfrak{D}_\varepsilon^{c_n}(P\|Q) \geq \int c d\pi^* + \varepsilon \mathfrak{D}_{\mathrm{KL}}(R^*\|Q) \geq \mathfrak{D}_c^\varepsilon(P\|Q)\,. \tag{C.34}$$

This concludes the proof. □

Finally we prove the existence of optimizers in the general case. Since c is only lower semi-continuous we will now take $\phi \in L^1(P)$ and $\psi \in L^1(Q)$ instead of bounded and continuous. We will use a weak convergence argument in L^1, due to [44] (see also [56]) and which we modify accordingly.

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

THEOREM C.6. Assume that X and Y are Polish spaces and that c is bounded below and lower semicontinuous and that $c \leq a \oplus b$ with $a \in L^1(P)$ and $b \in L^1(Q)$. Then the dual problem

$$\mathfrak{D}_\varepsilon^c(P\|Q) = \sup_{\substack{\phi\in L^1(P),\psi\in L^1(Q)\\ \phi\oplus\psi\leq c}} \left\{\mathbb{E}_P[\phi] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi/\varepsilon}]\right\} \tag{C.35}$$

possesses a maximizer (ϕ^*, ψ^*).

Proof. Our assumption on c insures that $\mathfrak{D}_\varepsilon^c(P\|Q)$ is finite and let us consider a optimizing sequence (ϕ_n, ψ_n) such that

$$\lim_{n\to\infty} \left(\mathbb{E}_P[\phi_n] - \varepsilon \log \mathbb{E}_Q[e^{-\varepsilon\psi_n/\varepsilon}]\right) = \mathfrak{D}_\varepsilon^c(P\|Q). \tag{C.36}$$

As usual we can assume that $\phi_n = \psi_n^c$ and therefore we get the upper bounds

$$\begin{aligned} \phi_n(x) &\leq c(x, y_0) - \psi_n(y_0) \leq a(x) + b(y_0) - \psi_n(y_0) \equiv \tilde{a}(x) \in L^1(P)\\ \psi_n(x) &\leq c(x_0, y) - \phi_n(x_0) \leq b(Y) + a(x_0) - \phi_n(x_0) \equiv \tilde{b}(y) \in L^1(Q) \end{aligned} \tag{C.37}$$

and so the sequences ϕ_n and ψ_n are bounded above by L^1 function.

Since (ϕ_n, ψ_n) is an optimizing sequence and using Jensen inequality

$$\mathfrak{D}_\varepsilon^c(P\|Q) - \alpha < \mathbb{E}_P[\phi_n] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi_n/\varepsilon}] \leq \mathbb{E}_P[\phi_n] + E[\psi_n] \leq T_c(P, Q) < \infty. \tag{C.38}$$

Together with the bound on ϕ_n and ψ_n this implies that $\phi_n \in L^1(P)$ and $\psi_n \in L^1(Q)$. Moreover since we can always replace (ϕ_n, ψ_n) by $(\phi_n + a_n, \psi_n - a_n)$ we can assume that the sequences $\{\phi_n\}$ and $\{\psi_n\}$ are bounded in $L^1(P)$ and $L^1(Q)$ respectively.

We show next that ϕ_n and ψ_n have suitable limits. To do this let us set

$$f_n = \tilde{a} - \phi_n \geq 0 \quad g_n = \tilde{b} - \psi_n \geq 0 \tag{C.39}$$

and $f_n \in L^1(P)$ and $g_n \in L^1(Q)$. We truncate f_n and g_n by setting

$$f_n^{(l)} = \max\{f_n, l\} \quad g_n^{(l)} = \max\{g_n, l\}. \tag{C.40}$$

Note that

$$\begin{aligned} 0 \leq f_n^{(l)} \leq l \quad \text{and} \quad f_n^{(l)} \leq f_n^{(l+1)} \leq \cdots\\ 0 \leq g_n^{(l)} \leq l \quad \text{and} \quad g_n^{(l)} \leq g_n^{(l+1)} \leq \cdots . \end{aligned} \tag{C.41}$$

If we now set $\phi_n^{(l)} = \tilde{a} - f_n^{(l)}$ and $\psi_n^{(l)} = \tilde{b} - f_n^{(l)}$ we have

$$\begin{aligned} \tilde{a} \geq \cdots \geq \phi_n^{(l)} \geq \phi_n^{(l+1)} \geq \cdots \geq \phi_n\\ \tilde{b} \geq \cdots \geq \psi_n^{(l)} \geq \psi_n^{(l+1)} \geq \cdots \geq \psi_n \end{aligned} \tag{C.42}$$

and in particular

$$\mathbb{E}_P[\phi_n^{(l)}] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi_n^{(l)}/\varepsilon}] \geq \mathbb{E}_P[\phi_n] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi_n/\varepsilon}]. \tag{C.43}$$

Moreover since

$$\phi_n^{(l)} = \tilde{a} + O(l) \tag{C.44}$$

the sequence $\{\phi_n^{(l)}\}_{n\geq 1}$ is uniformly integrable in $L^1(P)$, hence weakly compact. Up to extraction of a subsequence we can thus assume that $\phi_n^{(l)}$ converges to weakly in L^1 to some $\phi^{(l)} \in L^1(P)$. By using a diagonal sequence extraction we can then assume that

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

$\phi_n^{(l)}$ converges to $\phi^{(l)}$ for every l. Since weak convergence preserve ordering we have $\tilde{a} \geq \phi^{(l)} \geq \phi^{(l+1)} \geq \cdots$. The same results holds of course for $\psi_n^{(l)}$.

Recall the weak lower semicontinuity of convex functions [12] (that is if u_n converges weakly in L^1 to u and F is convex then $\int F(u)dQ \leq \liminf_n \int F(u_n)dQ$). From this and from (C.43) we obtain

$$\begin{aligned}\mathbb{E}_P[\phi^{(l)}] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi^{(l)}/\varepsilon}] &\geq \limsup_n \left(\mathbb{E}_P[\phi_n^{(l)}] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi_{n_k}^{(l)}/\varepsilon}]\right) \\ &\geq \limsup_n \left(\mathbb{E}_P[\phi_n] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi_n/\varepsilon}]\right) \\ &= \mathfrak{D}_\varepsilon^c(P\|Q).\end{aligned} \tag{C.45}$$

Finally we note that the sequences $\phi^{(l)}$ and $\psi^{(l)}$ are bounded in L^1, are decreasing in l, and are bounded above by integrable functions. Therefore we can apply the monotone convergence and deduce the existence of L^1 limits ϕ and ψ, defined almost everywhere

$$\phi = \lim_{l\to\infty} \phi^{(l)}, \quad \psi = \lim_{l\to\infty} \psi^{(l)}. \tag{C.46}$$

By the monotone convergence theorem we obtain that

$$\mathbb{E}_P[\phi] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi/\varepsilon}] = \lim_{l\to\infty} \mathbb{E}_P[\phi^{(l)}] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi^{(l)}/\varepsilon}] \geq \mathfrak{D}_\varepsilon^c(P\|Q). \tag{C.47}$$

□

Appendix D. Proof of DPI and additivity in Section 3.3. We first give the proof of the invariance of the proximal p-Wasserstein divergence under isomorphism.

Proof of Theorem 3.5. We use the dual variational representation. Since $c(Tx, Ty) = c(x, y)$, $\phi \oplus \psi \leq c$ implies that $\phi \circ T \oplus \psi \circ T \leq c$. Therefore

$$\begin{aligned}\mathfrak{D}^p_{\mathrm{KL},\varepsilon}(T_\# P\|T_\# Q) &= \sup_{\phi\oplus\psi\leq c} \mathbb{E}_P[\phi \circ T] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi\circ T/\varepsilon}] \\ &\leq \sup_{\phi'\oplus\psi'\leq c} \mathbb{E}_P[\phi'] - \varepsilon \ln \mathbb{E}_Q[e^{-\psi'/\varepsilon}] \\ &= \mathfrak{D}^p_{\mathrm{KL},\varepsilon}(P\|Q).\end{aligned} \tag{D.1}$$

We have inequality since in the first supremum we only consider ϕ' and ψ' which have the form $\phi' = \phi \circ T$ and $\psi' = \psi \circ T$. Since T is assumed to be invertible however, we obtain equality. □

The proof of the data processing inequality in Theorem 3.6 follows by a similar argument and is therefore omitted. Finally we prove the additivity property of the proximal OT divergence.

Proof of Theorem 3.7. Since $c(x, y) = \|x-y\|_p^p = \|x_1-y_1\|_p^p+\|x_2-y_2\|_p^p = c_1(x_1, y_1)+c_2(x_2, y_2)$ let us use the primal representation for optimal cost and for the proximal OT divergence and restrict both infima to product measures and product coupling. By using

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

the additivity of relative entropy we obtain then the lower bound

$$\begin{aligned}
&\mathfrak{D}^p_{\mathrm{KL},\varepsilon}(P_1 \times P_2 \| Q_1 \times Q_2) \\
&\leq \inf_{R=R_1\times R_2} \left\{ W_p^p(P_1 \times P_2, R_1 \times R_2) + \varepsilon \mathfrak{D}(R_1 \times R_2 \| Q_1 \times Q_2) \right\} \\
&\leq \inf_{R_1,R_2} \inf_{\substack{\pi_1 \in \Pi(P_1,R_1) \\ \pi_2 \in \Pi(P_2,R_2)}} \left\{ \int_{X_1} \|x_1 - y_1\|_p^p d\pi_1 + \int_{X_2} \|x_2 - y_2\|_p^p d\pi_2 \right. \\
&\qquad\qquad \left. + \varepsilon \mathfrak{D}(R_1 \| Q_1) + \varepsilon \mathfrak{D}(R_2 \| Q_2) \right\} \\
&= \mathfrak{D}^p_{\mathrm{KL},\varepsilon}(P_1 \| Q_1) + \mathfrak{D}^p_{\mathrm{KL},\varepsilon}(P_2 \| Q_2).
\end{aligned} \tag{D.2}$$

On the other hand using the dual variational representation and considering only ϕ and ψ of the form $\phi(x) = \phi_1(x_1) + \phi_2(x_2)$ and $\psi(y) = \psi_1(y_1) + \psi_2(x_2)$ with $\phi_1(x_1) + \psi_1(x_1) \leq \|x_1 - y_1\|_p^p$ and $\phi_2(x_2) + \psi_2(x_2) \leq \|x_2 - y_2\|_p^p$. We have then the upper bound

$$\begin{aligned}
&\mathfrak{D}^p_{\mathrm{KL},\varepsilon}(P_1 \times P_2 \| Q_1 \times Q_2) \\
&\geq \sup_{\substack{\phi_1 \oplus \psi_1 \leq c_1 \\ \phi_2 \oplus \psi_2 \leq c_2}} \left\{ \mathbb{E}_{P_1 \times P_2}[\phi_1 + \phi_2] - \varepsilon \log \mathbb{E}_{Q_1 \times Q_2}[e^{-(\psi_1+\psi_2)/\varepsilon}] \right\} \\
&= \sup_{\phi_1 \oplus \psi_1 \leq c_1} \left\{ \mathbb{E}_{P_1}[\phi_1] - \varepsilon \log \mathbb{E}_{Q_1}[e^{-(\psi_1)\varepsilon}] \right\} \\
&\qquad + \sup_{\phi_2 \oplus \psi_2 \leq c_2} \left\{ \mathbb{E}_{P_2}[\phi_2] - \varepsilon \log \mathbb{E}_{Q_2}[e^{-\psi_2/\varepsilon}] \right\} \\
&= \mathfrak{D}^p_{\mathrm{KL},\varepsilon}(P_1 \| Q_1) + \mathfrak{D}^p_{\mathrm{KL},\varepsilon}(P_2 \| Q_2).
\end{aligned} \tag{D.3}$$

This concludes the proof. □

Appendix E. Proof of Theorem 4.1. We use the Jordan decomposition $\rho = \rho_+ - \rho_-$ where $\rho_\pm \in \mathcal{P}(X)$ are mutually singular so there exists two disjoint sets $X_\pm$ such that $\rho_\pm(A) = \rho_\pm(A \cap X_\pm)$ for all measurable sets A. The measure $P + \alpha(\rho_+ - \rho_-)$ has total mass 1 but to be a probability measure we need that $\alpha\rho_-(A) \leq (P + \alpha\rho_+)(A)$ holds for all A. This implies that ρ_- is absolutely continuous with respect to P. Indeed if $P(A) = 0$ then

$$\alpha\rho_-(A) = \alpha\rho_-(A \cap X_-) \leq P(A \cap X_-) + \alpha\rho_+(A \cap X_-) \leq P(A) = 0. \tag{E.1}$$

By Theorem 2.3 the map $F(\alpha) = D^\varepsilon_{c,\mathrm{KL}}(P + \alpha\rho \| Q)$ is lower semicontinuous and convex. By assumption it is also finite (and thus also continuous) on the interval $[0, \alpha_0)$. As a finite convex function on an interval the map $F(\alpha)$ has left derivative $F'_-(\alpha)$ and right derivative $F'_+(\alpha)$ at every point $\alpha \in (0, \alpha_0)$ and the right derivative $F'_+(0)$ exists. Moreover the derivative $F'(\alpha)$ exists and $\lim_{\alpha \searrow 0} F'_-(\alpha) = F'_+(0)$.

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

For $0 < \alpha < \alpha_0$ let us denote $(\phi^*_\alpha, \psi^*_\alpha)$ the optimizers for the dual representation of $D^\varepsilon_{c,\mathrm{KL}}(P + \alpha\rho\|Q)$ which are unique up to constant in $\operatorname{supp} P + \rho_+$ and $\operatorname{supp}(Q)$ respectively. For h sufficiently small we have

$$\begin{aligned} F(\alpha \pm h) &= \sup_{(\phi,\psi)\in\Phi_c} \left\{ \mathbb{E}_{P+(\alpha\pm h)\rho}[\phi] - \varepsilon \log \mathbb{E}_Q[e^{-\psi/\varepsilon}] \right\} \\ &\geq \mathbb{E}_{P+(\alpha\pm h)\rho}[\phi^*_\alpha] - \varepsilon \log \mathbb{E}_Q[e^{-\psi^*_\alpha/\varepsilon}] \\ &= F(\alpha) \pm h \int \phi^*_\alpha d\rho \end{aligned} \tag{E.2}$$

from which we conclude that

$$F'_-(\alpha) \leq \int \widehat{\phi}^*_\alpha d\rho \leq F'_+(\alpha)\,. \tag{E.3}$$

For $\alpha = 0$ a similar computation using the optimizer $\widehat{\phi}^*$ defined in (4.1) we find that

$$F'_+(0) \geq \int \widehat{\phi}^* d\rho\,. \tag{E.4}$$

Let us now pick a sequence $\alpha_n \searrow 0$ such that F is differentiable at all α_n. By (E.3) we have

$$F'_+(0) = \lim_{n\to\infty} F'(\alpha_n) = \lim_{n\to\infty} \int \widehat{\phi}^*_{\alpha_n} d\rho. \tag{E.5}$$

Since c is bounded, arguing as in Theorem C.4 (see equations (C.23) and (C.24)) the sequences ($\widehat{\phi}^*_{\alpha_n}$ and $\psi^*_{\alpha_n}$) are equibounded (by $\|c\|_\infty$).

Since c is uniformly continuous it has a modulus of continuity, i.e. $|c(x,y) = c(x',y')| \leq \omega(\mathfrak{D}(x,x') + d(y,y'))$ for some function $\omega : \mathbb{R}^+ \to \mathbb{R}^+$ which satisfies $\omega(0) = 0$ and is continous at 0. Since we can assume that $\widehat{\phi}^*_{\alpha_n}$ and $\psi^*_{\alpha_n}$ are c-transform of each other, they inherit the modulus of continuity from c. By an Arzela-Ascoli type argument this implies that the sequences $\widehat{\phi}^*_{\alpha_n}$ and $\psi^*_{\alpha_n}$ converge, uniformly on compact sets, to some functions $\widetilde{\phi}$ and $\widetilde{\psi}$. We have then, by lower semicontinuity,

$$\begin{aligned} \mathfrak{D}^c_{\mathrm{KL},\varepsilon}(P\|Q) &= \lim_{n\to\infty} \mathfrak{D}^c_{\mathrm{KL},\varepsilon}(P + \alpha_n\rho\|Q) \\ &= \lim_{n\to\infty} \left\{ \mathbb{E}_{P+\alpha_n\rho}[\widehat{\phi}^*_{\alpha_n}] - \varepsilon \log \mathbb{E}_Q[e^{-\psi^*_{\alpha_n}/\varepsilon}] \right\} \\ &= \mathbb{E}_P[\widetilde{\phi}] - \varepsilon \log \mathbb{E}_Q[e^{-\widetilde{\psi}/\varepsilon}] \\ &\leq \mathfrak{D}^c_{\mathrm{KL},\varepsilon}(P\|Q) \end{aligned} \tag{E.6}$$

and thus $(\widetilde{\phi}, \widetilde{\psi})$ must be an optimizer. By definition $\widetilde{\phi}(x) = \widehat{\phi}^*(x)$ for $x \in \operatorname{supp}(P)$ and $\widetilde{\phi}(x) \leq \widehat{\phi}^*(x)$ for all $x \in X$. Therefore, using that ρ_- is absolutely continuous with respect to P

$$F'_+(0) = \int \widetilde{\phi}(x) d\rho = \int \widetilde{\phi}(x) d\rho_+ - \int \widetilde{\phi}(x) d\rho_- \leq \int \widehat{\phi}^* d\rho. \tag{E.7}$$

Combine with (E.4) this shows that $F'_+(0) = \int \widehat{\phi}^* d\rho$.

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

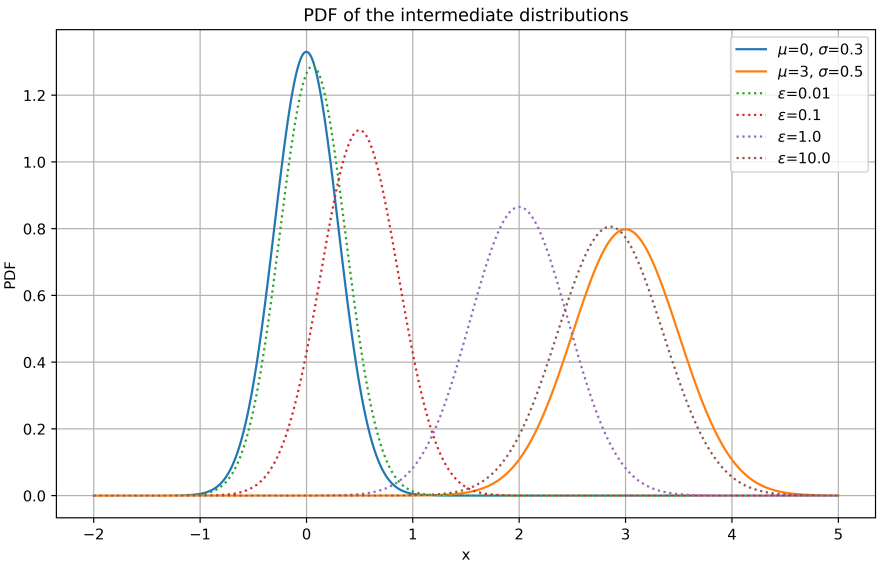


FIG. 1. Intermediate measure for several values of ε for 1-d Gaussian

Appendix F. Example: Proximal OT divergences for Gaussians. In this section, we derive an explicit expression for the proximal 2-Wasserstein divergence between two Gaussian distributions. We begin with the one-dimensional case and then extend the analysis to the multivariate setting with commuting covariance matrices.

THEOREM F.1. Let $P = \mathcal{N}(m_1, \sigma_1^2)$ and $Q = \mathcal{N}(m_2, \sigma_2^2)$. Then the OT divergence made by 2-Wasserstein and the KL divergence is given by choosing the intermediary distribution to be the Gaussian distribution $R = \mathcal{N}(m_R, \sigma_R^2)$ with

$$m_R = \frac{m_1 + \frac{\varepsilon}{2\sigma_2^2} m_2}{1 + \frac{\varepsilon}{2\sigma_2^2}} \tag{F.1}$$

$$\sigma_R = \frac{\sigma_1 + \sqrt{\sigma_1^2 + 2\varepsilon(1 + \frac{\varepsilon}{2\sigma_2^2})}}{2(1 + \frac{\varepsilon}{2\sigma_2^2})} \tag{F.2}$$

and the proximal OT divergence takes the form

$$\mathfrak{D}^2_{\mathrm{KL},\varepsilon}(P\|Q) = \frac{\frac{\epsilon}{2\sigma_2^2}}{1 + \frac{\epsilon}{2\sigma_2^2}}(m_1 - m_2)^2 + \sigma_1(\sigma_1 - \sigma_R) + \varepsilon \ln \frac{\sigma_2}{\sigma_R}. \tag{F.3}$$

We illustrate Theorem F.1 in Figure 1 by displaying the intermediate measure $R = R_\varepsilon$ for several values of ε.

Proof. The idea is that the intermediate measure R should be Gaussian itself. If this is the case, using the explicit formulas for the 2-Wasserstein distance and the KL divergence

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

between Gaussian distributions we find

$$\begin{aligned} &W_2^2(P,R)+\varepsilon\mathfrak{D}_{\mathrm{KL}}(R\|Q)\\ &=(m_1-m_R)^2+(\sigma_1-\sigma_R)^2+\varepsilon\left(\ln\frac{\sigma_2}{\sigma_R}+\frac{1}{2}\left(\frac{\sigma_R^2+(m_R-m_2)^2}{\sigma_2^2}-1\right)\right). \end{aligned} \tag{F.4}$$

Optimizing over m_R and σ_R along with straightforward computations yields the expressions in (F.1)–(F.3). To establish that the infimum is attained by an intermediate Gaussian distribution, we rely on the dual variational representation (3.8). To get a lower bound on the OT divergence we take the infimum over gaussian $R=\mathcal{N}(m_R,\sigma_R)$. We denote by $\tilde{P}=\mathcal{N}(0,\sigma_1)$ the shifted version of P and similarly for $\tilde{Q}$ and $\tilde{R}$.

We have

$$\begin{aligned} &W_2^2(P,R)+\varepsilon\mathfrak{D}_{\mathrm{KL}}R\|Q)\\ &=(m_1-m_R)^2+(\sigma_1-\sigma_R)^2+\varepsilon\left(\ln\frac{\sigma_2}{\sigma_R}+\frac{1}{2}\left(\frac{\sigma_R^2+(m_R-m_2)^2}{\sigma_2^2}-1\right)\right)\\ &=(m_1-m_R)^2+\varepsilon\frac{(m_R-m_2)^2}{\sigma_2^2}+W_2^2(\tilde{P},\tilde{R})+\varepsilon\tilde{R}\|\tilde{Q}) \end{aligned} \tag{F.5}$$

as a sum of two nonnegative terms and so we can optimize over the mean m_R and the variance σ_R separately. Optimizing over m_R gives (F.1) and

$$(m_1-m_R)^2+\varepsilon\frac{(m_R-m_2)^2}{\sigma_2^2}=\frac{\frac{\epsilon}{2\sigma_2^2}}{1+\frac{\epsilon}{2\sigma_2^2}}(m_1-m_2)^2.$$

Minimzing over σ_R gives the stationarity condition

$$2(\sigma_R-\sigma_1)-\varepsilon\frac{1}{\sigma_R}+\varepsilon\frac{\sigma_R}{\sigma_2^2}=0. \tag{F.6}$$

This gives a quadratic equation, and taking the positive root we find (F.2). Note also that the stationarity condition can be rewritten as

$$\frac{\varepsilon}{2}\left(\frac{\sigma_R^2}{\sigma_2^2}-1\right)=\sigma_1\sigma_R-\sigma_R^2 \tag{F.7}$$

and so we find that for the optimal σ_R we have

$$W_2^2(\tilde{P},\tilde{R})+\varepsilon\tilde{R}\|\tilde{Q})=\sigma_1(\sigma_1-\sigma_R)+\epsilon\ln\frac{\sigma_2}{\sigma_R}. \tag{F.8}$$

To find an upper bound on the OT divergence we use the dual representation. Since we are dealing with Gaussian it is natural to take $\phi(x)$ and $\psi(y)$ to be quadratic polynomial and since we can assume that the mean is 0 we can simply take quadratic. In view of Theorem 3.1 we choose to write ψ as

$$\psi(y)=\varepsilon\left(\frac{y^2}{2\sigma_R^2}-\frac{y^2}{2\sigma_2^2}\right). \tag{F.9}$$

Next, we show that if σ_R satisfies the stationary condition (F.6) then

$$E_{\tilde{P}}[\psi^c(x)]-\varepsilon\ln\mathbb{E}_{\tilde{Q}}[e^{-\psi(y)/\varepsilon}]=W_2^2(\tilde{P},\tilde{R})+\varepsilon\tilde{R}\|\tilde{Q})$$

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

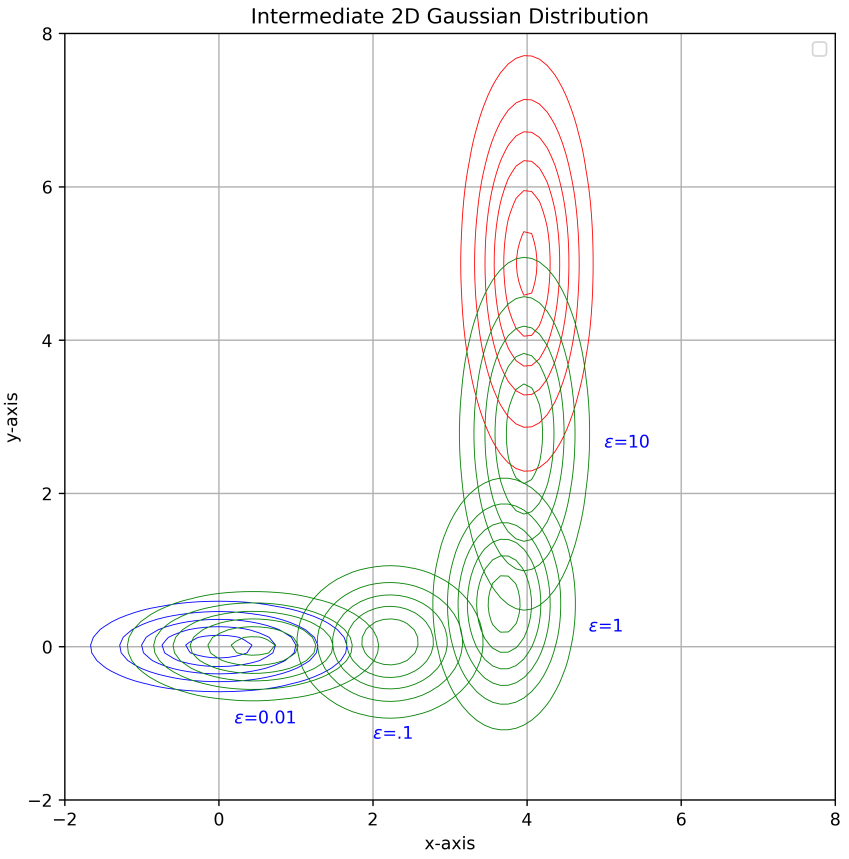


FIG. 2. Intermediate measures for 2D Gaussian for several values of ε

holds which completes the proof. First note that

$$E_{\tilde{Q}}[e^{-\psi/\varepsilon}] = \frac{1}{\sqrt{2\pi\sigma_2^2}} \int e^{-y^2/2\sigma_R^2} dy = \varepsilon \ln \frac{\sigma_R}{\sigma_2}$$

so that $-\varepsilon \log \mathbb{E}_{\tilde{Q}}[e^{-\psi/\varepsilon}] = \varepsilon \log \frac{\sigma_2}{\sigma_R}$ which is the second term in (F.8). Computing the c transform of $\psi(y) = ay^2$ we find

$$\psi^c(x) = \inf_y \{(x-y)^2 - ay^2\} = x^2 - \frac{1}{1-a} x^2 \tag{F.10}$$

for our choice of a, if we use the stationarity condition in the form of (F.7)

$$a = \varepsilon \left(\frac{1}{2\sigma_R^2} - \frac{1}{2\sigma_2^2} \right) = \frac{1}{\sigma_R^2} \frac{\epsilon}{2} \left(1 - \frac{\sigma_R^2}{\sigma_2^2} \right) = \frac{\sigma_R - \sigma_1}{\sigma_R}$$

and thus $1 - a = \frac{\sigma_1}{\sigma_R}$. Using this we obtain

$$E_{\tilde{P}}[\psi^c(x)] = \sigma_1^2 - \frac{1}{1-a} \sigma_1^2 = \sigma_1^2 - \sigma_R \sigma_1$$

which is the first term in (F.8). This concludes the proof. □

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

For multivariate Gaussians with commuting covariance matrices, the sequence of calculations from the one-dimensional case extends directly. The general case of non-commuting covariances is left for future work.

THEOREM F.2. Let $P = \mathcal{N}(m_1, \Sigma_1)$ and $Q = \mathcal{N}(m_2, \Sigma_2)$, where Σ_1 and Σ_2 commute. Then the proximal OT divergence $\mathfrak{D}^2_{\mathrm{KL},\varepsilon}$ is given by choosing the intermediary distribution to be the Gaussian $R = \mathcal{N}(m_R, \Sigma_R)$ with

$$\begin{aligned} m_R &= \left(I + \frac{\varepsilon}{2}\Sigma_2^{-1}\right)^{-1}\left(m_1 + \frac{\varepsilon}{2}\Sigma_2^{-1}\mu_2 +\right) \\ \Sigma_R^{1/2} &= \left(I + \frac{\varepsilon}{2}\Sigma_2^{-1}\right)^{-1}\frac{1}{2}\left(\Sigma_1^{1/2} + \left(\Sigma_1 + 2\varepsilon\left(I + \frac{\varepsilon}{2}\Sigma_2^{-1}\right)^{-1}\right)^{1/2}\right). \end{aligned} \tag{F.11}$$

Proof. Since Σ_1 and Σ_2 commute, we find an orthogonal transformation which diagonalizes Σ_1 and Σ_2 simultaneously. There exists an orthogonal transformation that simultaneously diagonalizes both matrices. By Theorem 3.5, we may therefore assume without loss of generality that both Σ_1 and Σ_2 are diagonal, implying that P and Q are product measures. The result then follows from the additivity property (Theorem 3.7) combined with the one-dimensional case established in Theorem F.1. In Figure 2, we also illustrate the intermediate measures for several values of ϵ. □

REFERENCES

[1] M. S. Albergo and E. Vanden-Eijnden, *Building normalizing flows with stochastic interpolants*, In The Eleventh International Conference on Learning Representations, 2023.

[2] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows in metric spaces and in the space of probability measures*, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 2005. MR2129498

[3] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein generative adversarial networks*, In Doina Precup and Yee Whye Teh, eds, Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, PMLR, Cambridge, MA, 06–11 Aug 2017, pp. 214–223.

[4] R. Baptista, B. Hosseini, N. B. Kovachki, and Y. M. Marzouk, *Conditional sampling with monotone GANs: from generative models to likelihood-free inference*, SIAM/ASA J. Uncertain. Quantif. **12** (2024), no. 3, 868–900, DOI 10.1137/23M1581546. MR4783615

[5] J.-D. Benamou and Y. Brenier, *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem*, Numer. Math. **84** (2000), no. 3, 375–393, DOI 10.1007/s002110050002. MR1738163

[6] Y. Bengio, N. Roux, P. Vincent, O. Delalleau, and P. Marcotte, *Convex neural networks*, Advances in neural information processing systems **18** (2006), 123. MIT; 1998.

[7] J. Birrell, P. Dupuis, M. A. Katsoulakis, Y. Pantazis, and L. Rey-Bellet, (f, Γ)*-divergences: interpolating between* f*-divergences and integral probability metrics*, J. Mach. Learn. Res. **23** (2022), Paper No. [39], 70. MR4420764

[8] Jeremiah Birrell and Mohammadreza Ebrahimi, *Adversarially robust deep learning with optimal-transport-regularized divergences*, 2023.

[9] Jeremiah Birrell, Markos A. Katsoulakis, Luc Rey-Bellet, and Wei Zhu, *Structure-preserving GANs*, In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, eds, International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, vol. 162 of Proceedings of Machine Learning Research, 2022, pp. 1982–2020.

[10] Jeremiah Birrell, Yannis Pantazis, Paul Dupuis, Luc Rey-Bellet, and Markos Katsoulakis, *Function-space regularized Rényi divergences*, In International Conference on Learning Representations (ICLR), 2023, `https://openreview.net/forum?id=89GT-S49mGd`.

[11] J. Blanchet, D. Kuhn, J. Li, and B. Taskesen, *Unifying distributionally robust optimization via optimal transport theory*, `arXiv:2308.05414`, 2023.

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

[12] H. Brezis, *Analyse fonctionnelle* (French), Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree], Masson, Paris, 1983. Théorie et applications. [Theory and applications]. MR697382

[13] J. A. Carrillo, K. Craig, L. Wang, and C. Wei, *Primal dual methods for Wasserstein gradient flows*, Found. Comput. Math. **22** (2022), no. 2, 389–443, DOI 10.1007/s10208-021-09503-1. MR4407747

[14] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, *Neural ordinary differential equations*, In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds, Advances in Neural Information Processing Systems, vol. 31, Curran Associates, Inc., Red Hook, NY, 2018.

[15] Y. Chen, T. T. Georgiou, and M. Pavon, *On the relation between optimal transport and Schrödinger bridges: a stochastic control viewpoint*, J. Optim. Theory Appl. **169** (2016), no. 2, 671–691, DOI 10.1007/s10957-015-0803-z. MR3489825

[16] Z. Chen, H. Gu, M. A. Katsoulakis, L. Rey-Bellet, and W. Zhu, *Robust generative learning with Lipschitz-regularized α-divergences allows minimal assumptions on target distributions*, Inf. Inference **14** (2025), no. 4, Paper No. iaaf028, 43, DOI 10.1093/imaiai/iaaf028. MR4968758

[17] Y.-C. Cheng, H. Gu, T. O. McDonald, W. Wu, S. Tripathi, C. Guarducci, D. Russo, D. L. Abravanel, M. Bailey, Y. Wang, Y. Zhang, Y. Pantazis, H. Levine, R. Jeselsohn, M. A. Katsoulakis, and F. Michor, *Profet predicts continuous gene expression dynamics from scrna-seq data to elucidate heterogeneity of cancer treatment responses*, Preprint, bioRxiv, 2025, DOI 10.1101/2025.06.27.662030.

[18] S. Chewi, J. Niles-Weed, and P. Rigollet, *Statistical optimal transport*, Lecture Notes in Mathematics, vol. 2364, Springer, Cham, [2025] ©2025. École d'Été de Probabilités de Saint-Flour XLIX—2019; École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School], DOI 10.1007/978-3-031-85160-5. MR4901218

[19] P. Dupuis and R. S. Ellis, *A weak convergence approach to the theory of large deviations*, Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons, Inc., New York, 1997. A Wiley-Interscience Publication, DOI 10.1002/9781118165904. MR1431744

[20] P. Dupuis and Y. Mao, *Formulation and properties of a divergence used to compare probability measures without absolute continuity*, ESAIM Control Optim. Calc. Var. **28** (2022), Paper No. 10, 38, DOI 10.1051/cocv/2022002. MR4371079

[21] F. Ercolessi and J. B. Adams, *Interatomic potentials from first-principles calculations: the force-matching method*, Europhys. Lett. **26** (1994), no. 8, 583–588.

[22] C. Finlay, J.-H. Jacobsen, L. Nurbekyan, and A. M. Oberman, *How to train your neural ODE: the world of Jacobian and kinetic regularization*, In International Conference on Machine Learning, PMLR, 2020, pp. 3154–3164.

[23] P. Glaser, M. Arbel, and A. Gretton, *Kale flow: a relaxed KL gradient flow for probabilities with disjoint support*, In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, eds, Advances in Neural Information Processing Systems, vol. 34, Curran Associates, Inc., Red Hook, NY, 2021, pp. 8018–8031.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, Communications of the ACM **63** (2020), no. 11, 139–144. ACM New York, NY.

[25] H. Gu, P. Birmpa, Y. Pantazis, L. Rey-Bellet, and M. A. Katsoulakis, *Lipschitz-regularized gradient flows and generative particle algorithms for high-dimensional scarce data*, SIAM J. Math. Data Sci. **6** (2024), no. 4, 1205–1235, DOI 10.1137/23M1587841. MR4836657

[26] H. Gu, M. A Katsoulakis, L. Rey-Bellet, and B. J Zhang, *Combining Wasserstein-1 and Wasserstein-2 proximals: robust manifold learning via well-posed generative flows*, `arXiv:2407.11901`, 2024.

[27] B. Guedj, *A primer on PAC-Bayesian learning*, SMF 2018: Congrès de la Société Mathématique de France, Sémin. Congr., vol. 33, Soc. Math. France, Paris, [2019] ©2019, pp. 391–413. MR4508923

[28] R. Jordan, D. Kinderlehrer, and F. Otto, *The variational formulation of the Fokker-Planck equation*, SIAM J. Math. Anal. **29** (1998), no. 1, 1–17, DOI 10.1137/S0036141096303359. MR1617171

[29] J.-M. Lasry and P.-L. Lions, *Mean field games*, Jpn. J. Math. **2** (2007), no. 1, 229–260, DOI 10.1007/s11537-007-0657-8. MR2295621

[30] W. Li, S. Liu, and S. Osher, *A kernel formula for regularized Wasserstein proximal operators*, Res. Math. Sci. **10** (2023), no. 4, Paper No. 43, 16, DOI 10.1007/s40687-023-00407-w. MR4650053

[31] Y. Lipman, R. TQ Chen, H. Ben-Hamu, M. Nickel, and M. Le, *Flow matching for generative modeling*, In The Eleventh International Conference on Learning Representations, 2023.

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

[32] G. Loaiza-Ganem, B. Leigh Ross, R. Hosseinzadeh, A. L. Caterini, and J. C. Cresswell, *Deep generative models through the lens of the manifold hypothesis: a survey and new connections*, Preprint, arXiv:2404.02954, 2024.
[33] A. Makkuva, A. Taghvaei, S. Oh, and J. Lee, *Optimal transport mapping via input convex neural networks*, In International Conference on Machine Learning, PMLR, Cambridge, MA, 2020, pp. 6672–6681.
[34] D. Maoutsa, S. Reich, and M. Opper, *Interacting particle solutions of Fokker-Planck equations through gradient-log-density estimation*, Entropy **22** (2020), no. 8, Paper No. 802, 35, DOI 10.3390/e22080802. MR4220298
[35] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, *Spectral normalization for generative adversarial networks*, Preprint, arXiv:1802.05957, 2018.
[36] X. Nguyen, M. J. Wainwright, and M. I. Jordan, *Estimating divergence functionals and the likelihood ratio by convex risk minimization*, IEEE Trans. Inform. Theory **56** (2010), no. 11, 5847–5861, DOI 10.1109/TIT.2010.2068870. MR2808937
[37] W. G. Noid, R. J. Szukalo, K. M. Kidder, and M. C. Lesniewski, *Rigorous progress in coarse-graining*, Annu. Rev. Phys. Chem. **75** (June 2024), no. 1, 21–45.
[38] S. Nowozin, B. Cseke, and R. Tomioka, *F-GAN: training generative neural samplers using variational divergence minimization*, In Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, Curran Associates Inc., Red Hook, NY, USA, 2016, pp. 271–279.
[39] M. Nutz and J. Wiesel, *On the martingale Schrödinger bridge between two distributions*, arXiv:2401.05209, 2024.
[40] D. Onken, S. Wu Fung, X. Li, and L. Ruthotto, *OT-flow: fast and accurate continuous normalizing flows via optimal transport*, In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 9223–9232.
[41] N. Parikh and S. Boyd, *Proximal algorithms*, Found. Trends Optim. **1** (2014), no. 3, 127–239.
[42] G. Peyré and M. Cuturi, *Computational optimal transport: with applications to data science*, Found. Trends Mach. Learn. **11** (2019), no. 5-6, 355–607.
[43] J. Pidstrigach, *Score-based generative models detect manifolds*, Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 35852–35865.
[44] S. T. Rachev and L. Rüschendorf, *Mass transportation problems. Vol. II*, Probability and its Applications (New York), Springer-Verlag, New York, 1998. Applications. MR1619171
[45] H. Rahimian and S. Mehrotra, *Frameworks and results in distributionally robust optimization*, Open J. Math. Optim. **3** (2022), Art. No. 4, 85, DOI 10.5802/ojmo.15. MR4459982
[46] D. Rezende and S. Mohamed, *Variational inference with normalizing flows*, In International Conference on Machine Learning, PMLR, 2015, pp. 1530–1538.
[47] F. Santambrogio, *Optimal transport for applied mathematicians*, Progress in Nonlinear Differential Equations and their Applications, vol. 87, Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling, DOI 10.1007/978-3-319-20828-2. MR3409718
[48] F. Santambrogio, {*Euclidean, metric, and Wasserstein*} *gradient flows: an overview*, Bull. Math. Sci. **7** (2017), no. 1, 87–154, DOI 10.1007/s13373-017-0101-1. MR3625852
[49] Y. Song, S. Garg, J. Shi, and S. Ermon, *Sliced score matching: a scalable approach to density and score estimation*, In Uncertainty in Artificial Intelligence, PMLR, 2020, pp. 574–584.
[50] Y. Song, J. N. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, *Score-based generative modeling through stochastic differential equations*, arXiv:2011.13456, 2021.
[51] E. G. Tabak and C. V. Turner, *A family of nonparametric density estimation algorithms*, Comm. Pure Appl. Math. **66** (2013), no. 2, 145–164, DOI 10.1002/cpa.21423. MR2999294
[52] E. G. Tabak and E. Vanden-Eijnden, *Density estimation by dual ascent of the log-likelihood*, Commun. Math. Sci. **8** (2010), no. 1, 217–233, DOI 10.4310/cms.2010.v8.n1.a11. MR2655907
[53] D. Terjék, *Moreau-Yosida f-divergences*, In International Conference on Machine Learning, PMLR, 2021, pp. 10214–10224.
[54] T. van Erven and P. Harremoës, *Rényi divergence and Kullback-Leibler divergence*, IEEE Trans. Inform. Theory **60** (2014), no. 7, 3797–3820, DOI 10.1109/TIT.2014.2320500. MR3225930
[55] P. Viallard, M. Haddouche, U. Şimşekli, and B. Guedj, *Tighter generalisation bounds via interpolation*, Preprint, arXiv:2402.05101, 2024.
[56] C. Villani, *Topics in optimal transportation*, Graduate Studies in Mathematics, vol. 58, American Mathematical Society, Providence, RI, 2003, DOI 10.1090/gsm/058. MR1964483

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.

[57] C. Villani, *Optimal transport*, Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 338, Springer-Verlag, Berlin, 2009. Old and new, DOI 10.1007/978-3-540-71050-9. MR2459454

[58] Z. O. Wang, R. Baptista, Y. Marzouk, L. Ruthotto, and D. Verma, *Efficient neural network approaches for conditional optimal transport with applications in Bayesian inference*, SIAM J. Sci. Comput. **47** (2025), no. 4, C979–C1005, DOI 10.1137/24M1678659. MR4945564

[59] B. J. Zhang and M. A. Katsoulakis, *A mean-field games laboratory for generative modeling*, `arXiv:2304.13534`, 2023.

[60] B. J. Zhang, S. Liu, W. Li, M. A. Katsoulakis, and S. J. Osher, *Wasserstein proximal operators describe score-based generative models and resolve memorization*, `arXiv:2402.06162`, 2024.

This is a free offprint provided to the author by the publisher. Copyright restrictions may apply.