

A dimension-free method for robust statistics and machine learning via Schrödinger bridge

Davide La Vecchia^{1*} and Hang Liu²

¹*Geneva School of Economics and Management, University of Geneva,
Bld. du Pont d'Arve, Geneva, CH-1211, Switzerland.

²Department of Statistics and Finance, School of Management,
University of Science and Technology of China, Jinzhai Rd, Hefei,
230026, Anhui Province, China.

*Corresponding author(s). E-mail(s): davide.lavecchia@unige.ch;
Contributing authors: hliu01@ustc.edu.cn;

Abstract

We propose the Entropic-regularized Robust Optimal Transport (E-ROBOT) framework, a novel method that combines the robustness of ROBOT with the computational and statistical benefits of entropic regularization. We show that, rooted in the Schrödinger bridge problem theory, E-ROBOT defines the robust Sinkhorn divergence $\bar{W}_{\epsilon,\lambda}$, where the parameter λ controls robustness and ϵ governs the regularization strength. Letting $n \in \mathbb{N}$ denote the sample size, a central theoretical contribution is establishing that the sample complexity of $\bar{W}_{\epsilon,\lambda}$ is $\mathcal{O}(n^{-1/2})$, thereby avoiding the curse of dimensionality that plagues standard ROBOT. This dimension-free property unlocks the use of $\bar{W}_{\epsilon,\lambda}$ as a loss function in large-dimensional statistical and machine learning tasks. With this regard, we demonstrate its utility through four applications: goodness-of-fit testing; computation of barycenters for corrupted 2D and 3D shapes; definition of gradient flows; and image colour transfer. From the computation standpoint, a perk of our novel method is that it can be easily implemented by modifying existing (`Python`) routines. From the theoretical standpoint, our work opens the door to many research directions in statistics and machine learning: we discuss some of them.

Keywords: Optimal transport, Curse of Dimensionality, Goodness-of-Fit Test, Barycenters, Gradient flow, Outliers

1 Introduction

1.1 Related work

The notion of the Schrödinger bridge problem (SBP) originates in [Schrödinger \(1931\)](#), when Schrödinger investigated the most likely evolution of a cloud of independent Brownian particles subject to boundary conditions.

Intuitively, the SBP searches for the most likely joint distribution of initial and final cloud of particles that is consistent with the observed endpoint results (namely, the original and target measures). This entropy-minimization problem exhibits a profound unity across seemingly distinct fields, including statistical and quantum physics, probability, information theory, statistics, machine learning, and artificial intelligence. We refer the reader to [Bunne, Hsieh, Cuturi, and Krause \(2023\)](#); [Liu, Chen, So, and Theodorou \(2022\)](#); [Peyré and Cuturi \(2019\)](#); [Pooladian and Niles-Weed \(2025\)](#); [Rigollet and Weed \(2018\)](#); [Wang, Jiao, Xu, Wang, and Yang \(2021\)](#) and references therein.

One of the main attractive features of the SBP is its link to optimal transport (OT) problem; see [Léonard \(2013\)](#). This connection yields two significant advantages. The first advantage is theoretical: the sample complexity of the Sinkhorn divergence, which results from the combination of SBP and OT, scales at a rate better than that of typical unregularized OT distances, which suffer from the curse of dimensionality. Essentially, the problem with standard OT is that Wasserstein distance computed between two samples converges very slowly to its population counterpart; see [Genevay, Chizat, Bach, Cuturi, and Peyré \(2019\)](#). Moreover, [Feydy et al. \(2019\)](#) show that entropic-regularized OT (E-OT) interpolates between the Wasserstein distance and the Maximum Mean Discrepancy (MMD, see e.g. [Sriperumbudur, Fukumizu, and Lanckriet \(2011\)](#)). Specifically, it preserves the appealing geometric properties of OT losses, and, at the same time, it benefits from the low sample complexity of MMD norms. The second advantage is computational: the Sinkhorn algorithm accelerates the computation of an approximate transport plan, significantly expanding the range of OT applications, particularly in machine learning; see [Cuturi \(2013\)](#); [Peyré and Cuturi \(2019\)](#).

However, a well-known limitation of both OT and E-OT is their sensitivity to anomalous records and their requirement of finite moments—both issues stem from the use of an unbounded cost function in the transportation cost definition; see e.g. [Ma, Liu, La Vecchia, and Lerasle \(2025\)](#); [Mukherjee, Guha, Solomon, Sun, and Yurochkin \(2021\)](#); [Nietert, Goldfeld, and Cummings \(2022\)](#). To address these problems, [Ma et al. \(2025\)](#) introduce robust optimal transport (ROBOT) and study its statistical properties, proving its robustness to outliers. Despite these advantages and its good performance in many statistical and machine learning tasks, ROBOT still suffers from the curse of dimensionality and exhibits multi-scale behavior. Similarly to OT-based distances (see e.g. [Weed and Bach \(2019\)](#)), the sample complexity of ROBOT-based distances depends on the data dimension; see Theorem 10 in [Ma et al. \(2025\)](#). Given the widespread use of OT in statistics, time series analysis, and machine learning (see e.g. [Hallin \(2022\)](#); [Hallin, La Vecchia, and Liu \(2022, 2023\)](#); [Hallin and Liu \(2024\)](#);

[La Vecchia, Ronchetti, and Ilievski \(2022\)](#); [Peyré and Cuturi \(2019\)](#)), this limitation significantly restricts the applicability of ROBOT-based methods.

In the OT literature, solutions exist to mitigate the curse of dimensionality. For instance, the sliced-Wasserstein metric projects higher-dimensional distributions into one-dimensional representations and computes the distance as a functional of the Wasserstein distances between these projections; see e.g. [Kolouri, Park, Thorpe, Slepcev, and Rohde \(2017\)](#); [Peyré and Cuturi \(2019\)](#) for a review. Alternatively, Gaussian-smoothed OT applies isotropic Gaussian smoothing to the original measures, alleviating the curse of dimensionality while preserving structural properties of the Wasserstein distance; see [Goldfeld and Greenewald \(2020\)](#); [Nietert, Goldfeld, and Kato \(2021\)](#). Unfortunately, to the best of our knowledge, no robustness guarantees are available for these methods: although applicable in high dimensions, they still rely on unbounded cost functions and thus remain sensitive to outliers. This limits their applicability in the presence of anomalous records or when distributions lack finite moments.

1.2 Our contributions and a preview of some results

In this paper, we address both the dimensionality and robustness issues of OT: we propose a novel method that simultaneously handles both challenges. Our approach thus bridges a critical gap in the OT literature within machine learning, while also providing a significant contribution to multivariate and robust statistics. We consider both theoretical and computational aspects and we illustrate the ease-of-implementation of our methodology. By building on the existing OT literature, our theoretical developments yield the needed guarantees, ensuring that the method’s robustness and scalability are well-principled and not merely empirical.

To help the reader navigating through the paper, below we provide an overview of our contributions—we flag that Appendix A and Appendix B, available in the [Supplementary Material](#), contain all proofs and additional numerical results.

In §2 and §3, building on theory of SBP, we introduce the entropic-regularized robust optimal transport (E-ROBOT) framework, which combines the robustness of ROBOT with the computational and statistical benefits of entropic regularization. In §3, we derive key theoretical and methodological aspects of E-ROBOT. Specifically, we provide the functional form of its potentials and their properties (Propositions 1, 3, 4), its dual formulation (Proposition 2), and the convergence behavior of the transport plan as the regularization term vanishes (Proposition 5) and as the sample size increases (Proposition 6). We show how E-ROBOT defines a truncated robust Sinkhorn divergence ($\bar{W}_{\varepsilon,\lambda}$) that metrizes convergence in law (Proposition 7) and how the Schrödinger potentials define a Bregman-type divergence (Proposition 8). One of our main contributions is Theorem 9 and Corollary 10, where we derive the sample complexity of the robust Sinkhorn divergence and show that it achieves a dimension-free rate, similar to non-robust E-OT. Finally, we justify the use of E-ROBOT as a loss function for statistical inference and machine learning tasks (Proposition 11). While some results follow directly from existing SBP and OT theory, others require

careful adaptation to the E-ROBOT setting, revealing interesting theoretical implications, such as those related to new MMDs discussed after Proposition 7 and after Corollary 10 in Remark 1.

In §4, we illustrate the applicability and performance of E-ROBOT in various settings and tasks in statistics and machine learning. Readers primarily interested in computational aspects may proceed directly to that section. Our experiments demonstrate that implementing our method requires some modifications of existing `Python` libraries and `R` routines. Indeed, one needs to replace the unbounded cost matrix used in standard OT with the bounded cost matrix from (7). This yields ready-to-use tools that are reliable in the presence of outliers and in high dimensions. Furthermore, our experiments show that E-ROBOT-based procedures enable inference in large-dimensional settings where the underlying distribution lacks finite moments—even without outliers—such as multivariate t -distributions with small degrees of freedom (df). This is not possible with standard OT and E-OT based methods like those relying on the p -Wasserstein distance W_p , which all require finite moments of order $p \geq 1$.

In §5, we discuss how this work lays the foundation for several promising research avenues. These include new theoretical developments, such as the derivation of parametric inference based on our robust divergence, as well as methodological ones, such as the joint selection of the hyperparameters λ and ε . We regard these developments as natural and essential extensions of the framework presented here, though their detailed treatment constitutes a separate research program. The primary goal of this paper is to establish the theoretical and practical foundation of the E-ROBOT framework itself, thereby demonstrating its viability and advantages over existing methods.

To illustrate some of the advantages of E-ROBOT and the shortcomings of existing methods, we preview some of our results. Figure 3 shows outcomes for a (simple) hypothesis test of distribution equality; see [Hallin, Mordant, and Segers \(2021\)](#) for background. Testing distribution equality is relevant in statistics (e.g., goodness-of-fit) and machine learning (e.g., generative model training). We compare our $\bar{W}_{\varepsilon,\lambda}$ (continuous blue curve) with the W_1 -based test (dotted red curve) from [Hallin et al. \(2021\)](#); see §4.1 for details. To study the effect of moment existence and dimensionality, we use a large-dimensional setting: for a sample size $n = 50$, we generate samples from a d -variate, with $d = 50$, t -distribution and we consider different df . We fix the level at 5%: the results show that the W_1 -based test lacks power for $df = 1, 2$, gains some power at $df = 3$ (where first and second moments are finite), but is consistently outperformed by our test across all df values and alternatives. This demonstrates that standard procedures can struggle with moment requirements and dimensionality even in a simple setting (simple hypothesis and no outliers).

In the following pages, we detail the construction of the E-ROBOT framework and provide its theoretical underpinnings. We then illustrate its application not only in the same simple hypothesis testing problem of Figure 3, but also in more complex tasks, such as barycenters computation in 2D and 3D, gradient flows, and image color transfer.

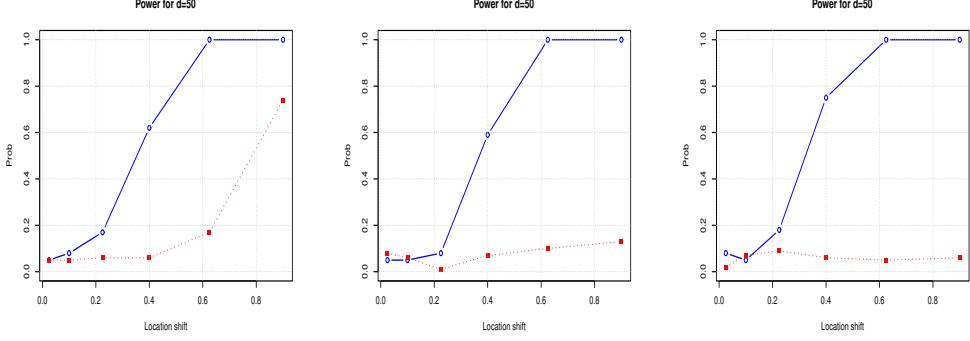


Fig. 1 Power curves for testing the null hypothesis of a d -variate t -distribution with different degrees-of-freedom (df), location zero, and identity scale matrix. The continuous (blue) curve with circles corresponds to our new test statistics based on $\bar{W}_{\varepsilon,\lambda}$, with $\varepsilon = 0.05$ and $\lambda = 3$; the dotted (red) curve with squares corresponds to the test statistics based on W_1 . Left plot is for $df = 3$, middle plot for $df = 2$, right plot for $df = 1$.

2 SBP, ROBOT, and E-ROBOT

In §2.1 and §2.2 we recall the key theoretical aspects of SBP and ROBOT that are needed for our developments; we refer to Léonard (2013) and to Ma et al. (2025) for more details. Then, in §2.3 we explain how SBP and ROBOT can be blended to obtain the E-ROBOT.

2.1 SBP

Let (X, μ) and (Y, ν) be separable probability spaces, and let $\mathcal{P}(X)$ denote the set of all probability measures on X . Let $R \in \mathcal{P}(X \times Y)$ be a reference measure. The goal is to find a coupling $\pi^* \in \Pi(\mu, \nu)$ minimizing the relative entropy with respect to R :

$$\pi^* = \arg \min_{\pi \in \Pi(\mu, \nu)} H(\pi \| R), \quad (1)$$

where $\Pi(\mu, \nu)$ is the set of couplings with fixed marginals μ and ν and $H(\pi \| R) = \mathbb{E}_\pi[\ln(\pi / R)]$ denotes the relative entropy (or Kullback-Leibler divergence) of π with respect to R . If $\Pi_{\text{fin}}(\mu, \nu) := \{\pi \in \Pi(\mu, \nu) : H(\pi \| R) < \infty\} \neq \emptyset$, then there exists a unique minimizer π^* . When R is absolutely continuous with respect to $\mu \otimes \nu$ (the product measure), the optimal π^* satisfies

$$\frac{d\pi^*}{dR} = e^{\varphi^*(x) + \psi^*(y)} \quad R\text{-a.s.}, \quad (2)$$

for some measurable functions $\varphi^* : X \rightarrow \mathbb{R}$ and $\psi^* : Y \rightarrow \mathbb{R}$, called Schrödinger potentials. These are unique up to an additive constant. Moreover, let $c : X \times Y \rightarrow$

$(-\infty, \infty]$ be a function and define

$$e^{-c(x,y)} = \frac{dR}{d(\mu \otimes \nu)}(x,y). \quad (3)$$

Then, the optimal potentials (φ^*, ψ^*) solve the Schrödinger system:

$$\varphi^*(x) = -\ln \int_Y e^{\psi^*(y)-c(x,y)} \nu(dy), \quad (4)$$

$$\psi^*(y) = -\ln \int_X e^{\varphi^*(x)-c(x,y)} \mu(dx). \quad (5)$$

We notice that the Schrödinger potentials ϕ^*, ψ^* satisfy the fixed-point equations in (4) and (5), which correspond to a SoftMax (log-sum-exp) operation. We will come back to this aspect in §2.3.

2.2 ROBOT

The ROBOT framework, as defined in Mukherjee et al. (2021), modifies the classical Kantorovich formulation of optimal transport (OT) by incorporating robustness to outliers via a total variation (TV) regularization. The primal formulation introduces an auxiliary perturbation s to the source measure μ , leading to the constrained problem:

$$\begin{aligned} & \min_{\pi, s} \int_X \int_Y c(x,y) \pi(x,y) dx dy + \lambda \|s\|_{\text{TV}} \\ \text{s.t. } & \int_Y \pi(x,y) dy = \mu(x) + s(x), \quad \int_X \pi(x,y) dx = \nu(y), \quad \int_X s(x) dx = 0, \end{aligned}$$

where $\lambda > 0$ controls the influence of outliers and $c(x,y) = d(x,y)$, with $d(x,y) = \|x - y\|$ being the Euclidean distance. This formulation effectively eliminates data points x such that $\mu(x) + s(x) = 0$, identifying them as outliers. A computationally efficient equivalent formulation is given by:

$$\inf \left\{ \int_{X \times Y} c_\lambda(x,y) d\pi(x,y) : \pi \in \Pi(\mu, \nu) \right\},$$

with $c_\lambda(x,y) := \tilde{d}_\lambda(x,y) := \min\{d(x,y), 2\lambda\}$, which replaces the original cost $c = d$ with a trimmed version c_λ that bounds the transport cost and introduces robustness. Ma et al. (2025) illustrate that the dual formulation of ROBOT, unlike classical OT, imposes a bounded range constraint on the Kantorovich potential ψ :

$$\sup \left\{ \int \psi d\mu - \int \psi d\nu : \psi \in \mathcal{C}_b(X), |\psi(x) - \psi(y)| \leq d(x,y), \text{range}(\psi) \leq 2\lambda \right\},$$

where for the sake of notation we omit the integration domain and \mathcal{C}_b denotes the set of bounded and continuous functions. Moreover, by noticing that $\tilde{d}_\lambda(x,y)$ is a metric,

one can construct the robust Wasserstein distance:

$$W^{(\lambda)}(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int c_\lambda(x, y) d\pi(x, y), \quad (6)$$

which is shown to be a proper metric on the space of probability measures $\mathcal{P}(X)$. Unlike standard Wasserstein distances, $W^{(\lambda)}$ remains well-defined for all probability measures, even without moment conditions, making it suitable for robust inference for heavy-tail data distributions. Additionally, $W^{(\lambda)}$ is continuous and monotonically non-decreasing with respect to $\lambda \in [0, \infty)$, and, if $W_1(\mu, \nu)$ exists, $\lim_{\lambda \rightarrow \infty} W^{(\lambda)}(\mu, \nu) = W_1(\mu, \nu)$. For a Polish space (X, d) , if we suppose that μ_k , (resp. ν_k) converges weakly to μ , (resp. ν) in $\mathcal{P}(X)$ as $k \rightarrow \infty$, then $W^{(\lambda)}(\mu_k, \nu_k) \rightarrow W^{(\lambda)}(\mu, \nu)$. Finally, Theorem 10 in Ma et al. (2025) proves that the sample complexity of $W^{(\lambda)}$ is:

$$\mathbb{E}[W^{(\lambda)}(\mu, \hat{\mu}_n)] = \begin{cases} \mathcal{O}(n^{-1/2}) & \text{when } d = 1, \\ \mathcal{O}(\frac{\ln n}{n^{1/2}}) & \text{when } d = 2, \\ \mathcal{O}(n^{-1/d}) & \text{when } d \geq 3, \end{cases}$$

which implies that ROBOT suffers from the curse of dimensionality: for $d > 2$ the dimension of the underlying space has an impact on the mean rate.

2.3 E-ROBOT

Let the product measure $P = \mu \otimes \nu \in \mathcal{P}(X \times Y)$ be the reference measure. Then the entropic optimal transport problem with regularization parameter $\varepsilon > 0$ is defined via the minimization over $\pi \in \Pi(\mu, \nu)$ of

$$C_\varepsilon(\mu, \nu, c, \pi) := \int c(x, y) d\pi(x, y) + \varepsilon H(\pi \| P).$$

To proceed further, let us consider the ROBOT cost function c_λ . For $\varepsilon > 0$, we define the rescaled and truncated cost function

$$\varepsilon^{-1} c_\lambda(x, y) = \varepsilon^{-1} \min(d(x, y), 2\lambda), \quad (7)$$

which allows to define the entropic regularized ROBOT (E-ROBOT) problem. Omitting the arguments for the sake of notation, we introduce the truncated Laplacian kernel

$$k_{\varepsilon, \lambda} = e^{-\varepsilon^{-1} c_\lambda} \quad (8)$$

(more generally, this is a Gibbs kernel) and define the reference (Gibbs) joint distribution

$$dR_\varepsilon := \frac{1}{\beta} k_{\varepsilon, \lambda} dP, \quad \text{where } \beta := \int k_{\varepsilon, \lambda} dP.$$

For any $\pi \in \mathcal{P}(X \times Y)$ with π absolutely continuous w.r.t. R_ε , namely $\pi \ll R_\varepsilon$, we have the equalities

$$\begin{aligned} H(\pi \| R_\varepsilon) &= \int \ln \left(\frac{d\pi}{dR_\varepsilon} \right) d\pi = \int \ln \left(\frac{d\pi}{dP} \frac{dP}{dR_\varepsilon} \right) d\pi \\ &= \int \ln \left(\frac{d\pi}{dP} \right) d\pi + \int \ln \left(\frac{dP}{dR_\varepsilon} \right) d\pi \\ &= H(\pi \| P) + \ln \beta + \int \varepsilon^{-1} c_\lambda d\pi, \end{aligned}$$

so

$$C_\varepsilon(\mu, \nu, c_\lambda, \pi) = \int c_\lambda d\pi + \varepsilon H(\pi \| P) = \varepsilon H(\pi \| R_\varepsilon) - \varepsilon \ln \beta. \quad (9)$$

The E-ROBOT problem is to minimize C_ε in (9), that is

$$\inf_{\pi \in \Pi(\mu, \nu)} C_\varepsilon(\mu, \nu, c_\lambda, \pi) = \inf_{\pi \in \Pi(\mu, \nu)} \varepsilon H(\pi \| R_\varepsilon) - \varepsilon \ln \beta, \quad (10)$$

which is a (static) Schrödinger bridge problem for R_ε , similarly to the problem stated in (1). When $\varepsilon > 0$, the E-ROBOT problem is strongly convex, so that the optimal plan is unique. Then, the next propositions can be proved along the same lines as in Thm. 4.2 and Thm. 4.7 in Nutz (2021), to which we refer for the mathematical detail.

Proposition 1 *Let c_λ be the ROBOT cost function and let (X, μ) and (Y, ν) be separable. Then:*

- (i) *there is a unique minimizer $\pi_\varepsilon^* \in \Pi(\mu, \nu)$ for the E-ROBOT problem in (10);*
- (ii) *There exist measurable functions $\varphi^* : X \rightarrow \mathbb{R}$, $\psi^* : Y \rightarrow \mathbb{R}$ (those are potentials which in fact depend on both ε and λ) such that*

$$\frac{d\pi_\varepsilon^*}{d(\mu \otimes \nu)} = e^{\varphi^* + \psi^* - \varepsilon^{-1} c_\lambda} \quad \mu \otimes \nu\text{-a.s.}$$

The potentials φ^, ψ^* are unique up to an additive constant, $\varphi^* \in L^1(\mu)$, $\psi^* \in L^1(\nu)$ and*

$$\varphi^*(x) = -\varepsilon \ln \int e^{\psi^*(y) - \varepsilon^{-1} c_\lambda(x, y)} \nu(dy) \quad \mu\text{-a.s.} \quad (11)$$

$$\psi^*(y) = -\varepsilon \ln \int e^{\varphi^*(x) - \varepsilon^{-1} c_\lambda(x, y)} \mu(dx) \quad \nu\text{-a.s.} \quad (12)$$

Conversely, if $\bar{\pi} \in \Pi(\mu, \nu)$ admits a density of the form

$$\frac{d\bar{\pi}}{d(\mu \otimes \nu)} = e^{\varphi + \psi - \varepsilon^{-1} c_\lambda} \quad \mu \otimes \nu\text{-a.s.}$$

for measurable functions φ, ψ satisfying (11) and (12), then $\bar{\pi} = \pi_\varepsilon^$.*

Eq. (11) and (12) make explicit the link between the E-ROBOT and the Schrödinger equations, as in (4) and (5). Moreover, since $c_\lambda \in L^1(\mu \otimes \nu)$, we state

Proposition 2 (E-ROBOT Dual problem) *The C_ϵ in (9) and the related E-ROBOT in (10) are such that:*

$$\inf_{\pi \in \Pi(\mu, \nu)} C_\epsilon = \sup_{\varphi \in L^1(\mu), \psi \in L^1(\nu)} \left\{ \int \varphi d\mu + \int \psi d\nu - \varepsilon \int e^{\varphi + \psi - \varepsilon^{-1} c_\lambda} d(\mu \otimes \nu) + \varepsilon \right\}, \quad (13)$$

and the supremum is attained by the E-ROBOT potentials $\varphi^*, \psi^* \in L^1(\mu) \times L^1(\nu)$, with

$$\inf_{\pi \in \Pi(\mu, \nu)} C_\epsilon = \varepsilon \left(\int \varphi^* d\mu + \int \psi^* d\nu \right). \quad (14)$$

The maximizers are a.s. unique up to an additive constant.

In the next section, equipped with the results of Propositions 1 and 2, we now study the main properties of the E-ROBOT problem and of some of its related notions.

3 Main results: key properties of E-ROBOT

3.1 Potentials: uniform boundedness and uniform convergence

The solution to the SBP described in Eq. (4) and (5) illustrates the pivotal role played by Schrödinger potentials. Propositions 1 and 2 further demonstrate that these potentials are fundamental to the solution of the E-ROBOT primal and dual problem. This leads to natural questions regarding the limiting behavior of the potentials in (11) and (12), as the regularization parameter approaches zero or as the sample size diverges. The next two propositions answer these questions. Specifically, Proposition 3 shows that the potentials are Lipschitz continuous functions, uniformly bounded.

Proposition 3 *Let $X, Y \subset \mathbb{R}^d$ and $c_\lambda : X \times Y \rightarrow \mathbb{R}$ be the continuous and bounded ROBOT cost function. Let $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$ be probability measures, and let φ^*, ψ^* be the optimal dual potentials for the entropic optimal transport problem with cost c_λ and regularization parameter $\varepsilon > 0$. Then φ^* and ψ^* : (i) are Lipschitz continuous functions on X and Y , respectively; (ii) $\varphi^* \in L^\infty(X)$ and $\psi^* \in L^\infty(Y)$, namely they are uniformly bounded.*

Besides being uniformly bounded, in the next proposition, we show also that the potentials converge uniformly as the sample size n diverges.

Proposition 4 *Let μ_n, ν_n be empirical measures based on n i.i.d. samples from compactly supported probability measures μ, ν on a compact sets $X, Y \subset \mathbb{R}^d$. Let ϕ_n^*, ψ_n^* be the Schrödinger potentials associated with the entropic optimal transport plan $\pi_n^* \in \Pi(\mu_n, \nu_n)$, and ϕ^*, ψ^* the potentials associated with the true plan $\pi^* \in \Pi(\mu, \nu)$, both for cost c_λ and regularization parameter $\varepsilon > 0$. Let c_λ be the ROBOT cost function. Then, as $n \rightarrow \infty$,*

$$\sup_{x \in X} |\phi_n^*(x) - \phi^*(x)| \rightarrow 0, \quad \sup_{y \in Y} |\psi_n^*(y) - \psi^*(y)| \rightarrow 0$$

i.e., the potentials converge uniformly.

3.2 Optimal regularized transport plan

It is easy to conjecture that the structure of the E-ROBOT potentials and their limiting behavior have implications on the limiting behavior of the related transportation plan. In this section, we study that aspect. More precisely, for fix λ , we discuss the limiting behavior as $\varepsilon \rightarrow 0$ of the E-ROBOT problem. The optimal transport plan π_0 of the corresponding unregularized OT problem is defined via:

$$C_0 = \inf_{\pi \in \Pi(\mu, \nu)} \int c_\lambda d\pi. \quad (15)$$

Since the value function $\inf_{\pi \in \Pi(\mu, \nu)} C_\varepsilon(\pi)$ of E-ROBOT decreases monotonically as $\varepsilon \downarrow 0$, we have

$$\lim_{\varepsilon \rightarrow 0} \inf_{\pi \in \Pi(\mu, \nu)} C_\varepsilon(\pi) \geq \inf_{\pi \in \Pi(\mu, \nu)} C_0(\pi).$$

We now show this inequality is an equality, meaning the E-ROBOT value converges to the ROBOT value as regularization vanishes (Proposition 5(i)). Moreover, the corresponding optimizer converges weakly: the optimal plan π_ε^* for E-ROBOT converges to an optimal plan π_0^* for the original ROBOT problem with cost c_λ (Proposition 5(ii)).

Proposition 5 *Let c_λ be the cost associated to the ROBOT problem for two measures μ and ν , and let π_0 be the optimal transport plan. If $H(\pi \| \mu \otimes \nu) < \infty$, then:*

(i) *we have that*

$$\lim_{\varepsilon \rightarrow 0} \inf_{\pi \in \Pi(\mu, \nu)} C_\varepsilon = C_0. \quad (16)$$

(ii) *if $\varepsilon_n \rightarrow 0$ and $\lim_{n \rightarrow \infty} \pi_{\varepsilon_n}^* = \pi_0$ weakly, we have that $\pi_0 \in \Pi(\mu, \nu)$ is the unique unconstrained ROBOT plan and it follows that $\lim_{\varepsilon \rightarrow 0} \pi_\varepsilon^* = \pi_0$ weakly.*

Finally, we study the large sample behaviour ($n \rightarrow \infty$) of the regularized optimal transport plan. To this end, we state:

Proposition 6 *Let $X, Y \subset \mathbb{R}^d$ be compact and $\mu_n \in \mathcal{P}(X), \nu_n \in \mathcal{P}(Y)$ be empirical measures based on i.i.d. samples from compactly supported probability measures μ, ν , such that $H(\pi \| \mu \otimes \nu) < \infty$. Let $\pi_n^* \in \Pi(\mu_n, \nu_n)$ be the entropic optimal transport plan with cost c_λ and regularization parameter $\varepsilon > 0$, and let $\pi^* \in \Pi(\mu, \nu)$ be the corresponding optimal plan for the true marginals. Then $\pi_n^* \rightarrow \pi^*$ weakly as $n \rightarrow \infty$.*

3.3 Robust Sinkhorn divergence and MMD

Given the cost function $c_\lambda(x, y)$ defined on $X \times Y$, with $Y = X$, and a regularization parameter $\varepsilon > 0$, the entropic regularized optimal transport cost between two probability measures μ and ν is denoted $W_{\varepsilon, \lambda}(\mu, \nu)$ and is defined as:

$$W_{\varepsilon, \lambda}(\mu, \nu) = \inf_{\pi} \int c_\lambda d\pi + \varepsilon H(\pi \| \mu \otimes \nu).$$

Due to the regularization, $W_{\varepsilon,\lambda}(\mu, \mu)$ is not granted to be zero and this entails the entropic bias. To correct for this bias and ensure that the loss vanishes when $\mu = \nu$, we propose the following modified unbiased version:

$$\bar{W}_{\varepsilon,\lambda}(\mu, \nu) := W_{\varepsilon,\lambda}(\mu, \nu) - \frac{1}{2} (W_{\varepsilon,\lambda}(\mu, \mu) + W_{\varepsilon,\lambda}(\nu, \nu)), \quad (17)$$

and we call it the robust Sinkhorn loss. When $\lambda \rightarrow \infty$, $c_\lambda \rightarrow c$ so we obtain the usual Sinkhorn loss, which is commonly referred to as the Sinkhorn divergence in the OT literature that has been studied in [Genevay et al. \(2019\)](#); [Genevay, Peyré, and Cuturi \(2018\)](#).

The following proposition states that the robust Sinkhorn divergence $\bar{W}_{\varepsilon,\lambda}$ defines a symmetric and positive definite loss function that is convex in each of its input variables. Moreover, it metrizes the convergence in law.

Proposition 7 *Let $X \subset \mathbb{R}^d$ and consider the cost function $c_\lambda(x, y)$. Then, for all probability measures μ and ν on X , $\bar{W}_{\varepsilon,\lambda}$ is such that:*

- (i) $0 = \bar{W}_{\varepsilon,\lambda}(\nu, \nu) \leq \bar{W}_{\varepsilon,\lambda}(\mu, \nu)$,
- (ii) $\mu = \nu \iff \bar{W}_{\varepsilon,\lambda}(\mu, \nu) = 0$,
- (iii) $\lim_{n \rightarrow \infty} \mu_n = \mu$ weakly $\iff \bar{W}_{\varepsilon,\lambda}(\mu_n, \mu) \rightarrow 0$.
- (iv) (Limiting behavior for $\varepsilon \rightarrow 0$) $\bar{W}_{\varepsilon,\lambda}(\mu, \nu) \rightarrow W_\lambda(\mu, \nu)$, as $\varepsilon \rightarrow 0$.

In addition to the limiting behavior for $\varepsilon \rightarrow 0$ discussed in Proposition 7, $\bar{W}_{\varepsilon,\lambda}$ has connections to MMD (see e.g. [Gretton, Borgwardt, Rasch, Schölkopf, and Smola \(2006\)](#)) and Bregman-type divergence (see e.g. [Pardo \(2018\)](#) for book-length introduction). [Gretton et al. \(2006\)](#) introduced MMD in machine learning and since then they have been applied for different tasks, like e.g. comparing distributions via distribution-free tests [Gretton, Borgwardt, Rasch, Schölkopf, and Smola \(2012\)](#), generative models [Li, Swersky, and Zemel \(2015\)](#), gradient flow and neural network optimization [Arbel, Korba, Salim, and Gretton \(2019\)](#). To illustrate the connection, let us consider

$$\|\mu - \nu\|_{-c_\lambda}^2 := \iint -c_\lambda(x, y) d(\mu - \nu)(x) d(\mu - \nu)(y), \quad (18)$$

which is the MMD with kernel $-c_\lambda$. We refer to Lemma 12 in Appendix A for the properties of this kernel, here we remark that expanding the last expression, we obtain

$$\|\mu - \nu\|_{-c_\lambda}^2 = \iint -c_\lambda d\mu(x) d\mu(y) + \iint -c_\lambda d\nu(x) d\nu(y) - 2 \iint -c_\lambda d\mu(x) d\nu(y). \quad (19)$$

From the definition of $W_{\lambda,\varepsilon}(\mu, \nu)$, when $\varepsilon \rightarrow +\infty$, the entropic term $\varepsilon H(\pi \| \mu \otimes \nu)$ dominates. The solution to the E-ROBOT problem is then close to the independent coupling and the transport cost term becomes $\iint c_\lambda(x, y) d\mu(x) d\nu(y)$. So, as $\varepsilon \rightarrow \infty$,

for each $W_{\lambda,\varepsilon}$ in (17) we have

$$\begin{aligned} W_{\lambda,\varepsilon}(\mu, \nu) &\rightarrow \iint c_\lambda(x, y) d\mu(x) d\nu(y), \\ W_{\lambda,\varepsilon}(\mu, \mu) &\rightarrow \iint c_\lambda(x, y) d\mu(x) d\mu(y), \\ W_{\lambda,\varepsilon}(\nu, \nu) &\rightarrow \iint c_\lambda(x, y) d\nu(x) d\nu(y). \end{aligned}$$

Thus, as $\varepsilon \rightarrow \infty$ we have

$$\overline{W}_{\varepsilon,\lambda}(\mu, \nu) \rightarrow \iint c_\lambda(x, y) d\mu(x) d\nu(y) - \frac{1}{2} \left(\iint c_\lambda(x, y) d\mu(x) d\mu(y) + \iint c_\lambda(x, y) d\nu(x) d\nu(y) \right), \quad (20)$$

and, comparing (20) to (19), we conclude that

$$\lim_{\varepsilon \rightarrow \infty} \overline{W}_{\varepsilon,\lambda}(\mu, \nu) = \frac{1}{2} \|\mu - \nu\|_{-c_\lambda}^2 \quad (21)$$

namely, in the large- ε limit, $\overline{W}_{\varepsilon,\lambda}(\mu, \nu)$ becomes half the squared MMD norm with kernel $-c_\lambda$. This results provide an interesting interpretation for the E-ROBOT: by changing ε , the E-ROBOT interpolates between the ROBOT and the MMD norm as obtained using the truncated cost as kernel.

In addition to this property, making use of the positive and c -universal kernel $k_{\varepsilon,\lambda}$ (see Lemma 12 in Appendix A), we define the MMD:

$$\|\mu\|_{k_{\varepsilon,\lambda}}^2 = \iint k_{\varepsilon,\lambda}(x, y) d\mu(x) d\mu(y) = \iint e^{-\frac{1}{\varepsilon} c_\lambda(x, y)} d\mu(x) d\mu(y), \quad (22)$$

and we define also the E-ROBOT negentropy $F_{\varepsilon,\lambda}(\mu)$ as a mapping from $\mathcal{P}(X)$ to \mathbb{R} :

$$F_{\varepsilon,\lambda} : \mu \mapsto -\frac{1}{2} W_{\varepsilon,\lambda}(\mu, \mu) = -\frac{1}{2} \inf_{\pi \in \Pi(\mu, \mu)} C_\varepsilon(\mu, \mu, c_\lambda, \pi) \quad (23)$$

Working along the same lines as Proposition 4 in Feydy et al. (2019) (see Appendix ??, Lemma ??), we can prove that

$$\frac{1}{\varepsilon} F_{\varepsilon,\lambda}(\mu) + \frac{1}{2} = \inf_{\xi \in \mathcal{P}(X)} \left\{ \int \ln \left(\frac{d\mu}{d\xi} \right) d\mu + \frac{1}{2} \|\mu\|_{k_{\varepsilon,\lambda}}^2 \right\}, \quad (24)$$

where $F_{\varepsilon,\lambda}$ is a strictly convex functional on $\mathcal{P}(X)$. Moreover, $F_{\varepsilon,\lambda}$ is differentiable in the following sense.

Let $\mathcal{C}(X)$ denote the set of continuous function on X , and define the operator $\langle \cdot, \cdot \rangle : \mathcal{P}(X) \times \mathcal{C}(X) \rightarrow \mathbb{R}$ as the mapping $(\mu, f) \mapsto \int f(x) d\mu(x)$. Recall that a function $F : \mathcal{P}(X) \rightarrow \mathbb{R}$ is differentiable at $\mu \in \mathcal{P}(X)$ if there exists a continuous function

(called gradient) $\nabla F(\mu) \in \mathcal{C}(X)$ such that for any $\xi = \nu_1 - \nu_2$ with $\nu_1, \nu_2 \in \mathcal{P}(X)$, we have

$$F(\mu + t\xi) = F(\mu) + t\langle \xi, \nabla F(\mu) \rangle + o(t).$$

Then, moving along the same lines as the proof in Appendix B2 of [Feydy et al. \(2019\)](#), we have the following proposition, which implies that the E-ROBOT negentropy $F_{\varepsilon, \lambda}$ is differentiable everywhere in $\mathcal{P}(X)$ with gradient $\nabla F_{\varepsilon, \lambda}(\mu)(x) = -\varphi^*(x)/2$ for $x \in X$.

Proposition 8 *Let $\mu, \nu \in \mathcal{P}(X)$ be probability measures on a compact set $X \subset \mathbb{R}^d$, and let $c_\lambda : X \times X \rightarrow \mathbb{R}$ be the ROBOT cost function. Then the entropic cost $W_{\varepsilon, \lambda}(\mu, \nu)$ is weak-* continuous and differentiable over $\mathcal{P}(X) \times \mathcal{P}(X)$, with the gradient given by the pair of Schrödinger potentials:*

$$\nabla W_{\varepsilon, \lambda}(\mu, \nu) = (\varphi^*, \psi^*).$$

In Section 4.3, we show that the formula of the gradient $\nabla F_{\varepsilon, \lambda}(\mu)$ can be applied to model the evolution of a distribution along the E-ROBOT gradient flow. Also, it allows us to define the E-ROBOT Hausdorff divergence

$$H_{\varepsilon, \lambda}(\mu, \nu) := \frac{1}{2} \langle \mu - \nu, \nabla F_{\varepsilon, \lambda}(\mu) - \nabla F_{\varepsilon, \lambda}(\nu) \rangle, \quad (25)$$

which is a Bregman-type divergence induced by the strictly convex functional $F_{\varepsilon, \lambda}$ and is therefore a positive definite quantity.

3.4 Sample complexity

The results of §3.3 illustrate the properties of $\bar{W}_{\varepsilon, \lambda}$ as a tool to measure the proximity between two (probability) measures. This is similar to W_p , W^λ , and W_ε . However, the results of §3.3 are stated at the population level. In applications, W_p and W^λ are estimated from samples. A well-known issue is that the error of these empirical estimates suffers from a serious dependence on dimension: the rate at which $W_p(\hat{\mu}_n, \mu)$ and $W^{(\lambda)}(\hat{\mu}_n, \mu)$ converge to 0 scales as $n^{-1/d}$ under mild moment conditions for $d \geq 3$. Thus, this rate (also called sample complexity) deteriorates poorly with dimension. As shown in [Genevay et al. \(2019\)](#), W_ε does not suffer from the same problem: its sample complexity scales with $n^{-1/2}$ in any dimension d . In Theorem 9 and Corollary 10, we show that $\bar{W}_{\varepsilon, \lambda}$ enjoys the same desirable property.

Before presenting the statements, we clarify an important theoretical point. Although our results on the sample complexity of $\bar{W}_{\varepsilon, \lambda}$ are consistent with those obtained for the Sinkhorn divergence in E-OT by [Genevay et al. \(2019\)](#) and for Gaussian-smoothed OT in [Nietert et al. \(2021\)](#), our proof requires a fundamentally different derivation and cannot directly build upon existing strategies. For instance, the proofs in [Genevay et al. \(2019\)](#) require that the cost function c is smooth, a condition not satisfied by our truncated cost c_λ . The proofs in [Nietert et al. \(2021\)](#) hinge on the specific idea of convolving measures with an isotropic Gaussian density, which is not a feature of our $\bar{W}_{\varepsilon, \lambda}$ framework. Therefore, to establish the sample complexity of $\bar{W}_{\varepsilon, \lambda}$, we must resort to different mathematical tools (empirical process theory), leveraging the special structure of the E-ROBOT potentials.

The novelty of our approach lies in this application of empirical process theory to the E-ROBOT potentials, whose regularity properties are guaranteed by the entropic regularization. To provide intuition for this strategy, recall that $\overline{W}_{\varepsilon,\lambda}(\mu, \nu)$ depends on $W_{\varepsilon,\lambda}(\mu, \nu)$, which in turn depends on the entropic regularization term $\varepsilon H(\pi \| \mu \otimes \nu)$. Now, recall that

$$H(\pi \| \mu \otimes \nu) = \int \ln \left(\frac{d\pi}{d(\mu \otimes \nu)} \right) d\pi,$$

where the optimal plan π^* is:

$$\frac{d\pi^*}{d(\mu \otimes \nu)}(x, y) = \exp \left(\varphi^*(x) + \psi^*(y) - \frac{1}{\varepsilon} c_\lambda(x, y) \right),$$

and similarly for π_n^* with potentials φ_n^*, ψ_n^* and empirical marginals μ_n, ν_n . Substituting into the entropy expression, we obtain:

$$H(\pi^* \| \mu \otimes \nu) = \int \left(\varphi^*(x) + \psi^*(y) - \frac{1}{\varepsilon} c_\lambda(x, y) \right) d\pi^*(x, y).$$

Now, since $\varphi_n^*, \psi_n^* \rightarrow \varphi^*, \psi^*$ uniformly (see Lemma 4), and c_λ is bounded and Lipschitz, the integrand $(x, y) \mapsto \varphi_n^*(x) + \psi_n^*(y) - c_\lambda(x, y)/\varepsilon$ is itself uniformly bounded and Lipschitz. Moreover, from Proposition 5, the plans $\pi_n^* \rightarrow \pi^*$ weakly, and the domain is compact. By standard empirical process theory (uniform convergence for Lipschitz function classes, see Sections 2.2 and 2.5.1 in Van Der Vaart and Wellner (1996)), we obtain $\mathbb{E}[|H(\pi_n^* \| \mu_n \otimes \nu_n) - H(\pi^* \| \mu \otimes \nu)|] = \mathcal{O}(n^{-1/2})$. This follows because the Schrödinger potentials φ_n^*, ψ_n^* are uniformly bounded and Lipschitz (see Proposition 3), and their uniform convergence (see Proposition 4) ensures that the integrand in the entropy expression is itself Lipschitz and bounded. Therefore, $H(\pi_n^* \| \mu_n \otimes \nu_n) - H(\pi^* \| \mu \otimes \nu)$ behaves like an empirical process indexed by a class of functions that is both Glivenko–Cantelli and Donsker, and its expectation converges at the rate $n^{-1/2}$.

The above derivation provides the basic intuition for how the SBP, the ROBOT, and the empirical process theory can be combined nicely to establish the sample complexity of $\overline{W}_{\varepsilon,\lambda}(\mu, \nu)$. With this regard, we state:

Theorem 9 *Let μ_n, ν_n be empirical measures based on n i.i.d. samples from compactly supported probability measures μ, ν on compact set $X \subset \mathbb{R}^d$. Let c_λ be a bounded Lipschitz cost function and $\varepsilon > 0$ a fixed regularization parameter. Then the expected deviation of the robust Sinkhorn loss satisfies:*

$$\mathbb{E}[|\overline{W}_{\varepsilon,\lambda}(\mu_n, \nu_n) - \overline{W}_{\varepsilon,\lambda}(\mu, \nu)|] = \mathcal{O}(n^{-1/2}).$$

Corollary 10 *Under the same assumptions as Theorem 9, we have $\mathbb{E}[\overline{W}_{\varepsilon,\lambda}(\mu_n, \mu)] = \mathcal{O}(n^{-1/2})$.*

Remark 1 The proofs of these results reveal an important aspect. A key distinction between E-ROBOT and ROBOT lies in the structure of their respective dual potentials. In E-ROBOT,

the Schrödinger potentials ϕ^*, ψ^* satisfy fixed-point equations in (4) and (5), which correspond to a SoftMax operation; see [Peyré and Cuturi \(2019\)](#), Ch. 4. This structure induces smoothness and regularity, ensuring that the potentials are uniformly bounded and Lipschitz continuous. As a result, the function class indexed by these potentials has finite entropy and is P-Donsker under compact support. In contrast, the ROBOT framework lacks entropic regularization, and its Kantorovich potentials arise from a linear program with Lipschitz and range constraints. While these potentials are bounded and Lipschitz, they do not enjoy the smoothing effects of the SoftMax structure. Consequently, uniform convergence and Donsker-type properties are not guaranteed in ROBOT when $d \geq 2$.

3.5 Truncated Laplace deconvolution

[Rigollet and Weed \(2018\)](#) give a statistical interpretation of E-OT for W_2 by showing that performing maximum-likelihood estimation for Gaussian deconvolution corresponds to calculating a projection with respect to the entropic optimal transport distance. The projection estimator as obtained using W_2 has been employed in the machine learning community as a smoothed version of a minimum Kantorovich distance estimator (MKE, [Bassetti, Bodini, and Regazzini \(2006\)](#); [Bassetti and Regazzini \(2006\)](#)) more suitable for automatic differentiation in generative models; see [Montavon, Müller, and Cuturi \(2016\)](#), [Genevay et al. \(2019, 2018\)](#). We now prove that this connection between EOT and maximum-likelihood deconvolution can be derived also for the E-ROBOT, considering the deconvolution with a truncated Laplace distribution.

To elaborate, we start by recalling that a class \mathcal{P} of probability measures is said to be closed under domination, if $\mu_1 \ll \mu_2$ for some $\mu_2 \in \mathcal{P}$ implies that $\mu_1 \in \mathcal{P}$. Moreover, let μ contain probability distributions over \mathbb{R}^d and let μ^* be an unknown distribution of an i.i.d. sample X_1, \dots, X_n . The deconvolution problems consists in estimating μ^* using the corrupted random observations (Q_1, \dots, Q_n) , where

$$Q_i = X_i + Z_i, \quad i = 1, \dots, n \tag{26}$$

and the errors Z_1, \dots, Z_n are independent of X_1, \dots, X_n . In what follows, the random variables $\{Z_i\}$ are assumed to be independent copies of a random variable Z with known truncated Laplace distribution: $Z \sim \mathcal{L}(0, \varepsilon, \lambda)$, where the location is zero, the scale is ε , and the truncation parameter is λ .

In this context, the distribution of Q_i has density $f_{(0, \varepsilon, \lambda)} \star d\mu^*$, where, for any $\mu \in \mathcal{P}$, we define

$$f_{(0, \varepsilon, \lambda)} \star d\mu(y) = \int f_{(0, \varepsilon, \lambda)}(y - x) d\mu(x)$$

and $f_{(0, \varepsilon, \lambda)}$ denotes the density of $Z \sim \mathcal{L}(0, \varepsilon, \lambda)$. Under these assumptions, we call (26) the truncated Laplacian deconvolution model. The maximum-likelihood estimator (MLE) $\hat{\mu}$ defined by

$$\hat{\mu} = \operatorname{argmax}_{\mu \in \mathcal{P}} \sum_{i=1}^n \ln f_{(0, \varepsilon, \lambda)} \star d\mu(Q_i)$$

is a natural candidate to estimate μ^* .

Equipped with these definitions, we state a proposition that makes the link between the E-ROBOT and the truncated Laplacian deconvolution model, showing that E-ROBOT is in fact implementing $\hat{\mu}$.

Proposition 11 *Let $\nu_n = \sum_{i=1}^n \delta_{q_i}/n$ be an empirical measure of the observations (q_1, q_2, \dots, q_n) . Let \mathcal{P} be a convex class of probability measures that is closed under domination. The maximum-likelihood estimator for the Laplace convolution model $Q = X + Z$, where the noise Z has truncated Laplace density $f_{(0,\varepsilon,\lambda)}(z) \propto \exp(-\varepsilon^{-1}c_\lambda(0,z))$, is given by*

$$\hat{\mu} = \arg \min_{\mu \in \mathcal{P}} W_{\lambda,\varepsilon}(\mu, \nu_n).$$

Proposition 11 implies that the maximum-likelihood estimator $\hat{\mu}$ is the projection of the empirical measure ν_n onto \mathcal{P} with respect to $W_{\lambda,\varepsilon}$. Differently from the W_2 case of Rigollet and Weed (2018), the E-ROBOT framework corresponds to a deconvolution problem with a specific, robust noise model: the noise variable Z is assumed to follow a truncated Laplacian distribution. This distribution has the following properties: (i) for $\|z\| \leq 2\lambda$, $f_{0,\varepsilon,\lambda}(z) \propto \exp(-\|z\|/\varepsilon)$, identical to a standard Laplace distribution; (ii) for $\|z\| > 2\lambda$, $f_{0,\varepsilon,\lambda}(z) \propto \exp(-2\lambda/\varepsilon)$, a constant value. This implies a uniform distribution on the tails beyond the radius 2λ ; (iii) the log-probability of any large value is bounded below by $-2\lambda/\varepsilon$, making the model robust to outlying observations. The parameter λ controls the robustness (truncation point), while ε controls the scale (dispersion) of the core Laplace distribution component. Proposition 11 justifies the use of E-ROBOT for inference and prediction tasks, like those described in Ma et al. (2025)—e.g. minimum Kantorovich distance estimation, generative models, domain adaptation, and outliers detection.

4 Numerical illustrations

We illustrate the benefits and ease-of-use of E-ROBOT with four examples, covering key statistical inference issues and typical machine learning problems. All calculations were performed on a standard laptop with a 2.4 GHz 8-Core Intel Core i9 processor, with each exercise requiring only a few minutes. Code to replicate our results is available on GitHub at <https://github.com/dvdlvc/E-ROBOT> and can be combined with the ROBOT code available at <https://github.com/dvdlvc/Robust-optimal-transportation>.

A methodological note on implementation. The application of E-ROBOT requires selecting the hyper-parameters λ (from ROBOT) and ε (from E-OT). A theoretically grounded, general-purpose procedure for this joint selection remains a fundamental open challenge—rather than a limitation specific to this work. The literature, in fact, offers scant guidance even on selecting these parameters in isolation, with some exceptions like the ROBOT of Ma et al. (2025). We regard the derivation of such a joint selection criterion as an *essential but separate* line of theoretical research—see §5. In practice, we emphasize that E-ROBOT is highly operable. We demonstrate that small values of ε (e.g., on the order of $1E^{-2}$), consistent with standard E-OT procedures (see e.g. Peyré and Cuturi (2019)), yield excellent results across all our experiments (e.g.,

the power curves in §4.1 and barycenter calculations in §4.2.2) even for large dimensions. Furthermore, inheriting the stability of its ROBOT predecessor, the method performs effectively across a wide range of λ values. For implementation, we advise analysts to inspect the truncated cost matrix: the distribution of its entries provides immediate, empirical insight into the scale of outlier-induced costs and a suitable range for λ .

4.1 Nonparametric tests for equality of multivariate distributions

The $\overline{W}_{\lambda,\epsilon}$ and/or the related MMD or divergences described in §3.3 can be applied to devise testing procedures aimed to determine whether given two distribution are the same. More generally, they can be used as a loss functions in various statistics machine learning tasks such as density estimation, domain adaptation, and generative models. Among these different uses, we consider the Goodness-of-Fit (GoF) test problem discussed in §2 of [Hallin et al. \(2021\)](#) uses the empirical Wasserstein distance between the empirical distribution $\hat{\mu}_n$ and a fully specified null distribution μ_0 . Their test statistic is based on $W_p^p(\hat{\mu}_n, \mu_0)$, with critical values determined via Monte Carlo simulation under the null. While this approach is fully nonparametric, it is sensitive to outliers and suffers from the curse of dimensionality. To address these limitations, we propose replacing the classical Wasserstein distance with the robust Sinkhorn distance $\overline{W}_{\varepsilon,\lambda}$ from the E-ROBOT framework. Specifically, we define the test statistic as:

$$\overline{T}_n := \overline{W}_{\varepsilon,\lambda}(\hat{\mu}_n, \mu_0). \quad (27)$$

This modification inherits the robustness properties of ROBOT and the statistical regularity of entropic optimal transport, including dimension-independent convergence rates and uniform convergence of the associated Schrödinger potentials.

Under the conditions of Proposition 7, one can prove a consistency result analogous to Proposition 1 in [Hallin et al. \(2021\)](#): for any fixed alternative $\mu \neq \mu_0$, the test based on \overline{T}_n rejects the null hypothesis with probability tending to one as $n \rightarrow \infty$: $\lim_{n \rightarrow \infty} \mathbb{P}(\overline{T}_n > c_n(\alpha)) = 1$, where $c_n(\alpha)$ is the Monte Carlo critical value at level α . This follows from the convergence of $\overline{W}_{\varepsilon,\lambda}(\hat{\mu}_n, \mu)$ to $\overline{W}_{\varepsilon,\lambda}(\mu, \mu_0)$ and the fact that $\overline{W}_{\varepsilon,\lambda}(\mu, \mu_0) > 0$ under the alternative.

To implement the test, we compute \overline{T}_n as in (27) using an adaptation of the Sinkhorn algorithm (see e.g. [Peyré and Cuturi \(2019\)](#)) for the ROBOT setting, as in the pseudocode available in Algorithm 1. Then, we apply it to the mentioned GoF problem. Specifically, we test two null hypotheses: first, $\mathcal{H}_0 : \mu_0 = \mathcal{N}(0_d, I_d)$ (a d -variate standard normal); second, $\mathcal{H}_0 : \mu_0 = t(0_d, I_d, 1)$, namely a d -variate t -distribution with $df = 1$, location 0_d (the d -dimensional vector of zeros), and scale $d \times d$ -matrix I_d . To examine the role of dimensionality d , we study power curves for $d \in \{2, 10, 15\}$, sample size $n = 50$, significance level 5%, and a sequence of local alternatives obtained by shifting the location parameter equally in each dimension. For comparison, we also include the test statistic based on the W_1 distance from [Hallin et al. \(2021\)](#). Note that in this setting, $\overline{W}_{\lambda,\epsilon}$ is well-defined for all considered distributions, whereas W_1

is well-defined for the Gaussian cases but not for the t -distributions (where the first moment does not exist).

Algorithm 1 E-ROBOT Sinkhorn Algorithm

```

1: Input:
2:    $1_n = (1, \dots, 1)^\top$ 
3:   Source marginal:  $\mu \in \mathbb{R}_+^n$ , where  $\mu^\top 1_n = 1$ 
4:   Target marginal:  $\nu \in \mathbb{R}_+^n$ , where  $\nu^\top 1_n = 1$ 
5:   Regularization parameter:  $\varepsilon > 0$ 
6:   Robustness parameter:  $\lambda > 0$ 
7: Output:
8:   Approximate optimal transport plan:  $\pi^{(t)}$ 
9:   Final scaling vectors:  $u^{(t)}, v^{(t)}$ 
10:  procedure EROBOT_SINKHORN( $\mu, \nu, \varepsilon, \lambda$ )
11:    // Precompute the Gibbs kernel matrix  $K$ 
12:    for  $i = 1$  to  $n$  do
13:      for  $j = 1$  to  $n$  do
14:         $c_\lambda(i, j) \leftarrow \min(\|x_i - y_j\|, 2\lambda)$            ▷ Eq. (7): ROBOT cost
15:         $K(i, j) \leftarrow \exp(-c_\lambda(i, j)/\varepsilon)$            ▷ Eq. (8): E-ROBOT kernel
16:      end for
17:    end for
18:    // Initialize scaling vectors
19:     $u^{(0)} \leftarrow 1_n$ 
20:    // Main Sinkhorn iteration loop
21:    for  $t = 0, 1, 2, \dots$  do                                ▷ Iterate until convergence
22:       $v^{(t)} \leftarrow \nu / (K^\top u^{(t)})$            ▷ Element-wise division
23:       $u^{(t+1)} \leftarrow \mu / (K v^{(t)})$            ▷ Element-wise division
24:    end for
25:    // Form the approximate optimal transport plan
26:     $\pi^{(t)} \leftarrow \text{diag}(u^{(t)}) K \text{diag}(v^{(t)})$ 
27:    return  $\pi^{(t)}, u^{(t)}, v^{(t)}$ 
28: end procedure

```

Figure 2 displays the results for the multivariate normal case. The left plot shows that our test and the W_1 -based test have similar power for $d = 2$. However, as d increases, the power of \bar{T}_n exceeds that of the W_1 -based test. For the t -distribution (Figure 3), the W_1 -based test performs near its level, while \bar{T}_n maintains good power across all dimensions. These results complement the preview of our findings that we provided in §1.2.

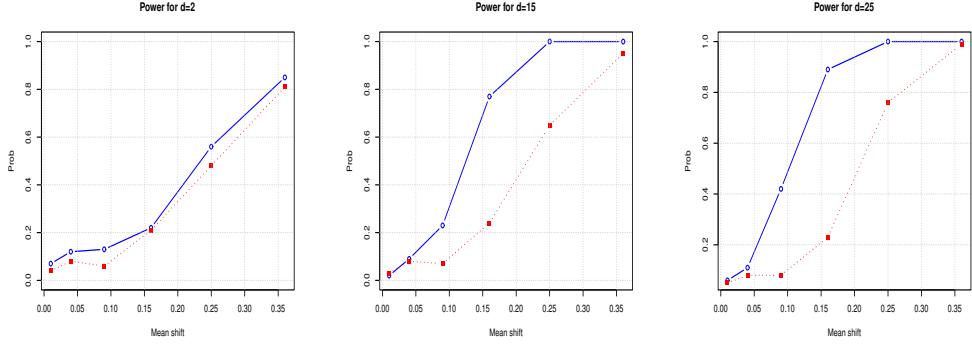


Fig. 2 GoF power curves for testing the null of a d -variate standard normal. The continuous (blue) curve with circles corresponds to the \bar{T}_n statistic using the entropic regularized Sinkhorn distance with $\varepsilon = 5$ and $\lambda = 10$; the dotted (red) curve with squares corresponds to the test based on W_1 .

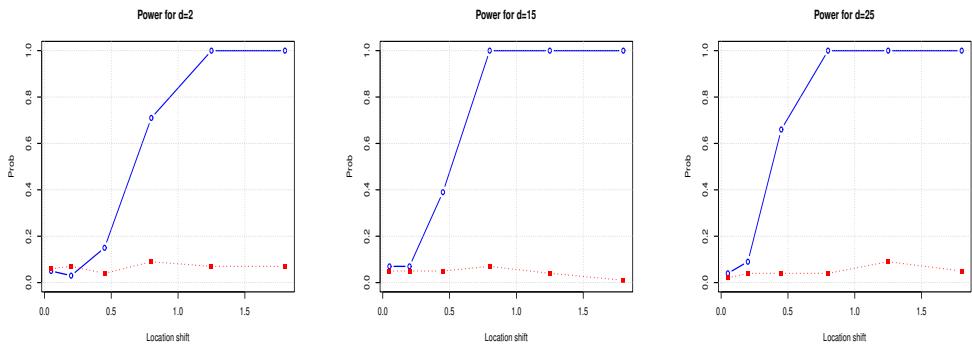


Fig. 3 GoF power curves for testing the null of a d -variate t -distribution with one degree of freedom, location zero, and identity scale matrix. The continuous (blue) curve with circles corresponds to the \bar{T}_n statistic using the entropic regularized Sinkhorn distance with $\varepsilon = 0.05$ and $\lambda = 3$; the dotted (red) curve with squares corresponds to the test based on W_1 .

4.2 E-ROBOT barycenters

4.2.1 IBP for E-ROBOT

Computing Wasserstein barycenters is a fundamental task in OT, with machine learning applications ranging from shape analysis and image synthesis to generative modeling. It provides a principled way to define a central representative of a collection of probability measures, generalizing the notion of the Fréchet mean to the Wasserstein space.

Let $\Sigma_n := \{\mu \in \mathbb{R}_+^n : \mu^\top 1_n = 1\}$ denote the simplex in \mathbb{R}^n . In the classical entropic barycenter formulation, the barycenter $\mu \in \Sigma_n$ of input measures $\mu_m \in \Sigma_n, m = 1, \dots, M$, is defined as the solution to $\min_{\mu \in \Sigma_n} \sum_{m=1}^M \alpha_m W_\varepsilon(\mu_m, \mu)$. Given

$(\mu, \nu) \in \Sigma_n \times \Sigma_n$, the polytope of couplings between μ and ν is defined as

$$\Pi(\mu, \nu) := \left\{ \gamma \in \mathbb{R}_+^{n \times n} : \gamma \mathbf{1}_n = \mu, \gamma^\top \mathbf{1}_n = \nu \right\}.$$

As expressed in (9), the E-OT problem can be re-cast in the form $\min_{\gamma \in \mathcal{S}} \text{KL}(\gamma | \xi)$ where ξ is a given point in $\mathbb{R}_+^{n \times n}$, and \mathcal{S} is an intersection of two closed convex sets, namely $\mathcal{S} = \mathcal{S}_1 \cap \mathcal{S}_2$, such that \mathcal{S} has nonempty intersection with $\mathbb{R}_+^{n \times n}$, where $\mathcal{S}_1 = \{\gamma = (\gamma_m)_{m=1}^M \in (\Sigma_n)^M : \gamma_m^\top \mathbf{1}_n = \mu_m \quad \forall m\}$ and $\mathcal{S}_2 = \{\gamma = (\gamma_m)_{m=1}^M \in (\Sigma_n)^M : \exists \mu \in \mathbb{R}_+^n, \gamma_m \mathbf{1}_n = \mu \quad \forall m\}$. We focus on the case where the convex sets \mathcal{S}_ℓ , for $\ell = 1, 2$ are affine subspaces. In this case, it is possible to solve the KL-minimization problem by simply using iterative KL-projections. For the sake of completeness, we recall that, given a convex set D and a reference measure ξ , the KL-projection of ξ onto D is defined as

$$P_D^{\text{KL}}(\xi) := \arg \min_{\gamma \in D} \text{KL}(\gamma \| \xi), \quad \text{KL}(\gamma \| \xi) = \sum_{i,j} \gamma(i,j) \left(\ln \left(\frac{\gamma(i,j)}{\xi(i,j)} \right) - 1 \right).$$

To compute the KL-projection, we start from $\gamma^{(0)} = \xi$, and compute $\forall t > 0$, $\gamma^{(t)} := P_{\mathcal{S}_{\text{mod2}}}^{\text{KL}}(\gamma^{(t-1)})$. One can show that $\gamma^{(t)}$ converges towards the unique solution $\gamma^{(t)} \rightarrow P_{\mathcal{S}}^{\text{KL}}(\xi)$ as $t \rightarrow \infty$; see [Benamou, Carlier, Cuturi, Nenna, and Peyré \(2015\)](#).

The computation of the barycenter in the E-ROBOT setting, requires the computation of a Wasserstein barycenter of input probability measures $\mu_1, \dots, \mu_M \in \Sigma_n$ with weights $\alpha_1, \dots, \alpha_M$ (such that $\sum_m \alpha_m = 1$), using $W_{\lambda, \varepsilon}$. To achieve this goal, we resort on KL-projections using the truncated cost matrix $C_\lambda \in \mathbb{R}^{N \times N}$ whose (i, j) -entry is $C_\lambda(i, j) = \min(C(i, j), 2\lambda)$, where $C(i, j) = d(x_i, y_j)$, and the associated Laplace kernel $\xi_k = \xi = k_{\lambda, \varepsilon}$ as in (8), for every $m = 1, 2, \dots, M$. Indeed, as in the E-OT case, the E-ROBOT barycenter problem can be re-formulated as:

$$\min_{\gamma} \sum_{m=1}^M \alpha_m \text{KL}(\gamma_m \| \xi_m), \quad \text{s.t.} \quad \gamma \in \mathcal{S}_1 \cap \mathcal{S}_2. \quad (28)$$

Then, the Iterative Bregman Projection (IBP) scheme (see [Benamou et al. \(2015\)](#)) can be applied. More specifically, we first compute the projection onto the constraint set \mathcal{S}_1 for each k , then we compute the projection onto the constraint set \mathcal{S}_2 , which enforces a shared left marginal μ across all couplings. This procedures leads to iterates $\gamma^{(t)} = (\gamma_m^{(t)})_m$ which satisfy, for each m , $\gamma_m^{(t)} = \text{diag}(u_m^{(t)}) \xi \text{diag}(v_m^{(t)})$ for two vectors $(u_m^{(t)}, v_m^{(t)}) \in \mathbb{R}^n \times \mathbb{R}^n$, initialized as $v_m^{(0)} = \mathbf{1}_n$ for all m , and computed with the iterations:

$$u_m^{(t)} = \frac{\mu^{(t)}}{\xi v_m^{(t)}}, \quad v_m^{(t+1)} = \frac{\mu_m}{\xi^\top u_m^{(t)}},$$

where $\mu^{(t)}$ is the current estimate of the barycenter obtained as: $\mu^{(t)} = \prod_{m=1}^N \left(u_m^{(t)} \odot (\xi v_m^{(t)}) \right)^{\alpha_m}$. Operations are to be interpreted element-wise, namely, for

vectors $(a, b) \in \mathbb{R}^n \times \mathbb{R}^n$, we denote entry-wise multiplication and division as

$$a \odot b := (a_i b_i)_i \in \mathbb{R}^n \quad \text{and} \quad \frac{a}{b} := (a_i / b_i)_i \in \mathbb{R}^n.$$

These alternating projections are repeated until convergence: the resulting shared left marginal μ is the robust entropic barycenter. To summarize the described methodology, in Algorithm 2, we provide the pseudo code for E-ROBOT barycenters via IBP calculation.

In the next two subsections, we illustrate how the E-ROBOT barycenters perform on 2D and 3D shapes in the presence of anomalous records. To implement our method, we resort on the Python library `ot`, which contains the routine `bregman` that computes Bregman projections for E-OT. That routine requires as an input a user-specified cost matrix. Therefore, to implement our E-ROBOT method we need to input in `ot.bregman` the matrix resulting from the application of the truncated cost function c_λ to the 2D and 3D data, computing the distance via the routine `cdist` and trimming the entries of the resulting matrix via 2λ .

4.2.2 Barycenters for corrupted 2D and 3D shapes

In the next numerical example, we illustrate the use of the IBP and E-ROBOT in the computation of barycenters for 2D images. We consider two shapes. *Shape 1 (Source)* is a red circle with a radius of 4.5 pixels which we contaminate with 10 outliers in the top-right corner. *Shape 2 (Target)* is a blue square with a side length of 9 pixels which we contaminate with 10 outliers in the bottom-right corner. The resulting images are normalized to represent probability distributions.

We compute entropic barycenters, which can be interpreted as interpolated shapes between the source and target distributions. For the sake of visualization, we consider weights $t = 0.25, 0.5, 0.75$, with corresponding the weights for the barycenter calculation being $\{1 - t, t\}$. We consider the E-ROBOT and the EOT. In the EOT case, the cost matrix is based on the Euclidean distance between pixel coordinates as implemented in the Python routine `entropic_barycenter` in the `ot` library. To implement the E-ROBOT barycenters, we modify this routine introducing a truncation parameter λ , which trims the entries of this matrix. This simple modification is central to the implementation. Indeed, it allows for the comparison between two distinct methods: the standard EOT and our novel E-ROBOT. In Figure 4, we display two sets of plots, one for each method.

The five top plots, corresponding to the E-ROBOT with a small $\lambda = 4$, show a clean barycenter computation process: the red circle smoothly transforms into the blue square, while the outliers appear to simply fade in and out at their fixed locations without being transported. This demonstrates that E-ROBOT successfully isolates the barycenter computation of the main shapes from the influence of the outliers. Differently, the five bottom plots, which is for the E-ROBOT with a large λ , illustrate that the outliers significantly impact the barycenter calculation: as the circle becomes the square, the red outliers from the top-right corner are transported to the location of the blue outliers in the bottom-right corner. This results in particularly noticeable at $t = 0.5$, where a blurred mass appears to be streaking between the outlier locations.

Algorithm 2 E-ROBOT Barycenter via Iterative Bregman Projections (IBP)

1: **Input:**

2: Input measures: $\mu_1, \mu_2, \dots, \mu_M \in \Sigma_n$
3: Weights: $\alpha_1, \alpha_2, \dots, \alpha_M$ where $\sum_{m=1}^M \alpha_m = 1$
4: Truncated cost matrix: $C_\lambda \in \mathbb{R}^{n \times n}$, where $C_\lambda(i, j) = \min(C(i, j), 2\lambda)$
5: Regularization parameter: $\varepsilon > 0$
6: Convergence threshold: $\delta > 0$

7: **Output:**

8: Robust entropic barycenter: $\mu^* \in \Sigma_n$

9: **procedure** IBP_EROBOT_BARYCENTER($\{\mu_m\}, \{\alpha_m\}, C_\lambda, \varepsilon, \delta$)

10: Precompute the Gibbs kernel:
11: $K \leftarrow \exp(-C_\lambda/\varepsilon)$ ▷ Element-wise exponentiation
12: Initialize scaling vectors:
13: **for** $m = 1$ to M **do**
14: $v_m \leftarrow 1_n$ ▷ Initialize to vector of ones
15: **end for**
16: $\mu \leftarrow 1_n/n$ ▷ Initialize barycenter uniformly
17: $\Delta \leftarrow \infty$ ▷ Initialize Δ to a very large number
18: $t \leftarrow 0$
19: **while** $\Delta > \delta$ **do**
20: $t \leftarrow t + 1$
21: First projection: enforce marginal constraints
22: **for** $m = 1$ to M **do**
23: $u_m \leftarrow \mu_m / (K v_m)$ ▷ Element-wise division
24: **end for**
25: Second projection: enforce shared barycenter constraint
26: **for** $m = 1$ to M **do**
27: $\tilde{v}_m \leftarrow \mu / (K^\top u_m)$ ▷ Temporary update (element-wise division)
28: **end for**
29: Update barycenter estimate:
30: $\mu_{\text{new}} \leftarrow 1_n$ ▷ Initialize to ones
31: **for** $m = 1$ to M **do**
32: $\mu_{\text{new}} \leftarrow \mu_{\text{new}} \odot (u_m \odot (K \tilde{v}_m))^{\alpha_m}$ ▷ Element-wise operations
33: **end for**
34: $\mu_{\text{new}} \leftarrow \mu_{\text{new}} / \sum \mu_{\text{new}}$ ▷ Normalize
35: Update scaling vectors:
36: **for** $m = 1$ to M **do**
37: $v_m \leftarrow \mu_m / (K^\top u_m)$ ▷ Direct update from constraint (element-wise division)
38: **end for**
39: $\Delta \leftarrow \|\mu_{\text{new}} - \mu\|_1$ ▷ Compute L_1 norm
40: $\mu \leftarrow \mu_{\text{new}}$
41: **end while**
42: $\mu^* \leftarrow \mu$ ▷ Assign final estimate to output variable
43: **return** μ^*
44: **end procedure**

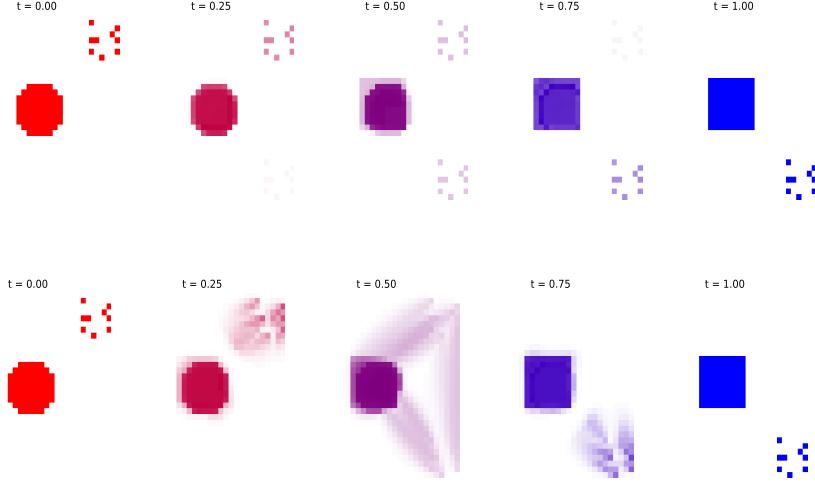


Fig. 4 E-ROBOT Barycenters via IBP for 2D shapes: top panels for $\lambda = 4$ and bottom for $\lambda = 4 \times 10^7$. Shapes are on a 32×32 grid. The entropic regularization parameter is $\varepsilon = 0.15$.

This is due to the fact that the E-ROBOT with a large λ behaves like the standard EOT approach: it attempts to transport all mass, including the anomalous records, as part of the overall transformation.

Similar considerations hold for the case of 3D shapes, where we compute the barycenter for values of $t = 0, 0.5, 1$, between a cube and torus. The original shapes are contaminated by outliers, mimicking the logic as in the 2D case. In Figure 5, we display the results for the E-ROBOT with small λ (three top plots) and the E-ROBOT with large λ (three bottom panels). Also in the 3D case, the key point is that in the E-ROBOT with large λ , the IBP algorithm considers the transport of both the shapes and the outliers, treating them as part of a single, coherent distribution. In contrast, the E-ROBOT with small λ limits the maximum cost of transporting mass, particularly between the far-apart main shapes and the outliers. This is evident comparing the top and bottom middle plots.

4.3 Gradient flows for corrupted 2D shapes

Let $\mu \in \mathcal{P}(X)$ be a probability measure on a compact set $X \subset \mathbb{R}^d$. Gradient flows are the continuous-time analogue of gradient descent: they describe how a probability distribution μ_t evolves with time t . In optimal transport, this evolution takes place in the Wasserstein space of probability measures. Such flows appear naturally in machine learning, for example in generative modeling (evolving a model distribution toward data), domain adaptation (aligning source and target), and clustering/density estimation (regularizing or smoothing empirical measures).

To compare the de-biased robust Sinkhorn divergence $\overline{W}_{\varepsilon,\lambda}$ with the standard Sinkhorn-type loss for W_1 , a natural experiment is to let a model distribution μ_t evolve along the E-ROBOT gradient flow of a data-fit loss that drives it toward a

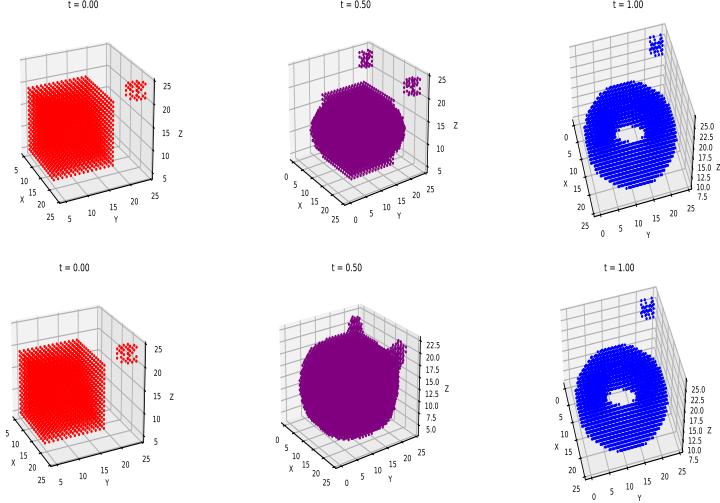


Fig. 5 E-ROBOT barycenters via IBP for 3D shapes, for weights $t = 0, 0.5, 1$: top 3 panels are for $\lambda = 0.1$ and bottom 3 panels are $\lambda = 10$). Shapes are on a $22 \times 22 \times 22$ grid. The entropic regularization parameter is $\varepsilon = 0.05$.

target distribution (Santambrogio (2015)). This non-parametric fitting setup makes explicit the role of the robustness parameter λ in the presence of outliers. Classical EOT flows suffer from well-documented artefacts: an entropic shrinkage bias (working on the entropic regularized OT alone pulls μ toward an over-concentrated measure), and – on the other end of the spectrum – kernel MMDs can show vanishing gradients near the support’s extremes; see Feydy et al. (2019). These issues motivate the use of a robust, regularized: $\overline{W}_{\varepsilon, \lambda}$.

More in detail, recall the E-ROBOT negentropy functional $F_{\varepsilon, \lambda}(\mu)$ in (23). This functional is strictly convex and differentiable; when $\mu = \nu$, the Schrödinger potentials coincide, $\phi^* = \psi^*$. Consequently, the (first-variation) gradient of the E-ROBOT negentropy reads $\nabla F_{\varepsilon, \lambda}(\mu) = -\phi^*/2$, and the corresponding Wasserstein gradient flow is $d\mu_t/dt = -\nabla F_{\varepsilon, \lambda}(\mu_t)$. To simulate this continuous-time PDE, one may resort on the explicit (forward) Euler scheme with step size $\tau > 0$, where the current distribution is pushed forward by a small displacement along the negative gradient field. If $\mu_k = n^{-1} \sum_{i=1}^n \delta_{x_i^k}$ is an empirical measure at the k -th step, this reduces to particle updates of the familiar gradient descent form $x_i^{k+1} = x_i^k - \tau \nabla_{x_i} F_{\varepsilon, \lambda}(\mu_k)$, for $i = 1, \dots, n$. In Python code (as in the `GeomLoss` package), one evaluates the loss with parameters (ε, λ) , back-propagates to obtain $\nabla_{x_i} F_{\lambda, \varepsilon}$, and applies the Euler step above to all particles. The step size τ controls stability and small values of τ imply slower but more faithful to the continuous flow. This explicit discretization exactly mirrors the procedure used in Section 4 of Feydy et al. (2019), which contrasts kernel MMD, biased entropic OT, and the de-biased Sinkhorn divergence on toy registration and gradient flow tasks.

For the sake of visualisation and following the existing examples available in the `GeomLoss`, we focus on 2D shapes. In Figure 6 we display the gradient flows for the

E-ROBOT with a small value of λ (top 6 panels) and a with large λ (bottom panels). The plots showcase how the cost truncation parameter λ controls the E-ROBOT method's robustness to outliers. For a large value of λ (bottom panels), the E-ROBOT approach behaves similarly to standard EOT because the cost truncation is effectively removed. In this case, the gradient flow illustrates that the outliers (black stars) are transported and embedded into the central blob of mass as the shapes evolve. This is visually evident as the black stars, initially separate, are drawn into the evolving mass for $t \geq 0.5$ (they are getting closer and closer to the blobs). This aspect highlights the sensitivity of standard EOT to anomalous data points. This is evident also in the bottom panels, which are for the EOT with W_2 , as obtained using the command `gradient_flow` in `GeomLoss`—which remains the faster approach, as indicated by the time for each iteration. In contrast, for a small λ (top panels), the truncated cost function limits the maximum transport cost between far-apart points. The gradient flow for small λ demonstrates the E-ROBOT method's robustness: the flow of the main blob proceeds without being influenced by the distant outliers. The gradient flow effectively excludes these outliers from the optimal transport plan, allowing the algorithm to focus on the smooth transformation of the core blobs. This behaviour confirms that the E-ROBOT method successfully mitigates the influence of outliers by preventing them from being part of the transport. However, an import remark is in order. A closer inspection of the gradient flow in the top panels of Figure 6 reveals a subtle but important phenomenon: the outliers are not perfectly stationary from the very beginning. For early timesteps ($t = 0.25$ or $t = 0.5$), they exhibit some minor movement before settling into a fixed configuration. This initial movement is a natural consequence of the gradient descent optimization process. The algorithm, starting from the initial configuration at $t = 0$, is in the process of finding the optimal transport plan. At this early stage, there is a weak but non-zero influence from the distant points (the main blob) on the outliers. The gradient flow is essentially taking small, exploratory steps in the direction of steepest descent for the overall system. However, the core mechanism of the E-ROBOT method – the truncated cost function – quickly takes precedence. The algorithm realises that the cost of moving the outliers is prohibitively high due to the small λ value: this leads to outliers motion becoming negligible and the gradient flow for the outliers effectively vanishes as the algorithm converges on the optimal, outlier-excluding solution.

Referring to the plots in Appendix B, the key conclusion from the experiment is that the use of $\overline{W}_{\lambda,\varepsilon}$ is preferable to the MMDs for gradient flow. Indeed, $\overline{W}_{\lambda,\varepsilon}$ provides a more reliable, namely a more geometry aware, flow and more robust gradient signal for moving probability distributions toward each other, especially when they are initially far apart and contain outliers.

4.4 Image color transfer

Color transfer is a fundamental and highly applicable task in computer vision and computer graphics. Its importance stems from its role as a core problem in image manipulation and its wide range of practical applications. For instance, it can be used for aesthetic enhancement of photographs and for accessibility purposes, such as

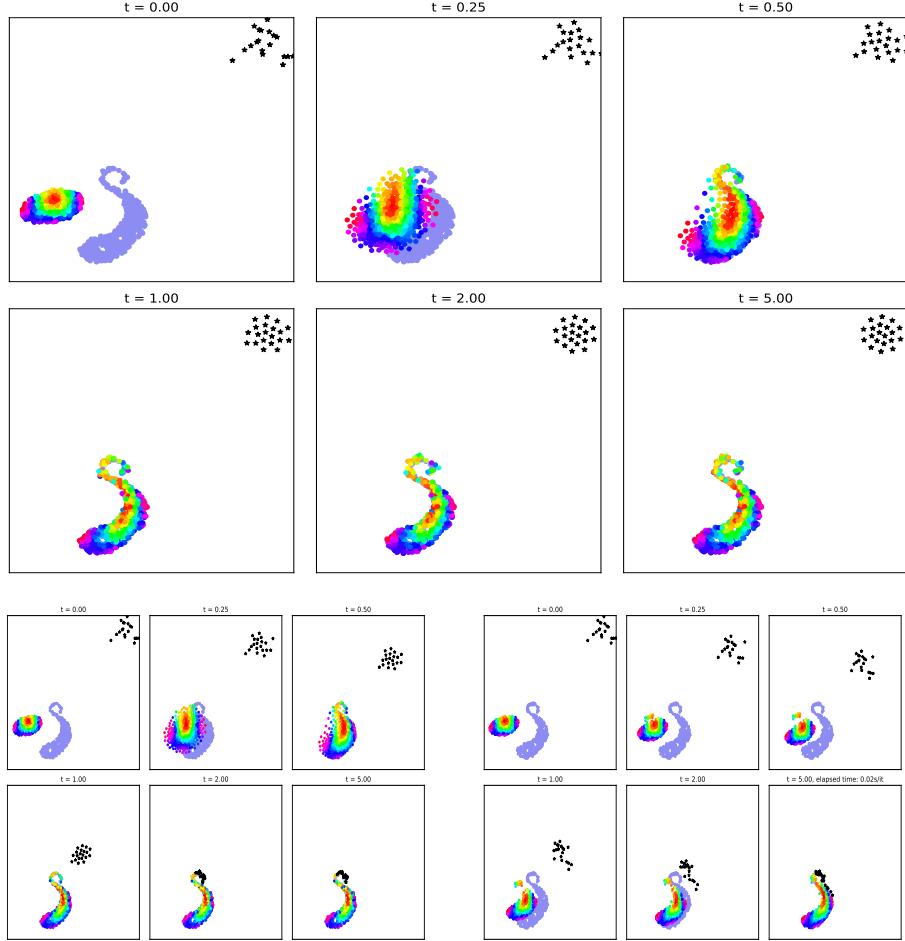


Fig. 6 Gradient flows for 2D shapes via entropic regularized OT. Top panels: E-ROBOT with $\lambda = 0.6$, bottom left 6 panels E-ROBOT with $\lambda = 6$, and bottom right 6 panels EOT with W_2 . The entropic regularization parameter is $\varepsilon = 0.05$. The learning rate, i.e. time step τ , is set equal to 0.05, similarly to the default value in `GeomLoss`.

modifying color schemes for color-blind viewers or restoring the faded color palette of historical images.

In machine learning, from a computational perspective, color transfer serves as a canonical example of distribution alignment. Indeed, the problem of matching the color distribution of a source image to a target is analogous to aligning the feature distributions of two datasets. Methods that solve color transfer robustly and in large-dimensions, like E-ROBOT, provide valuable insights and tools for this machine learning task. With this regard, it is well-known that the regularization of the transport plan helps to remove colorization artifacts due to noise amplification. More precisely, both OT and ROBOT map between complicated densities can be non smooth. As a result, using directly their transport plans to perform color transfer amplifies the noise

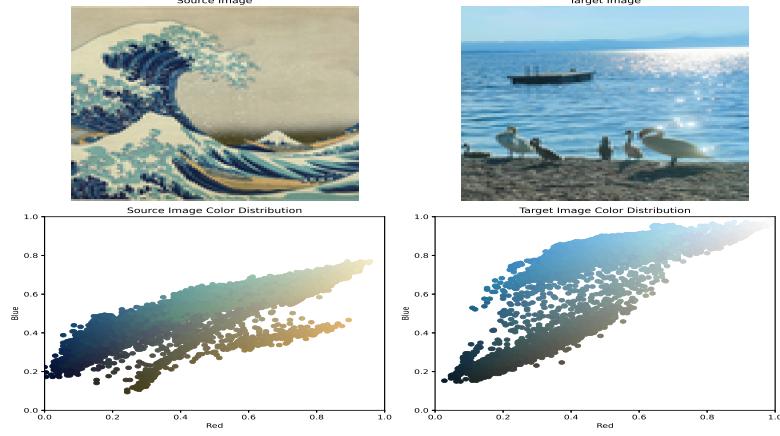


Fig. 7 Source (left) and target (right) distributions. Source: *The Great Wave off Kanagawa* by Hokusai, picture credit: Wikipedia. Target: A scene on Lac Léman, Switzerland, picture credit: Nadia La Vecchia.

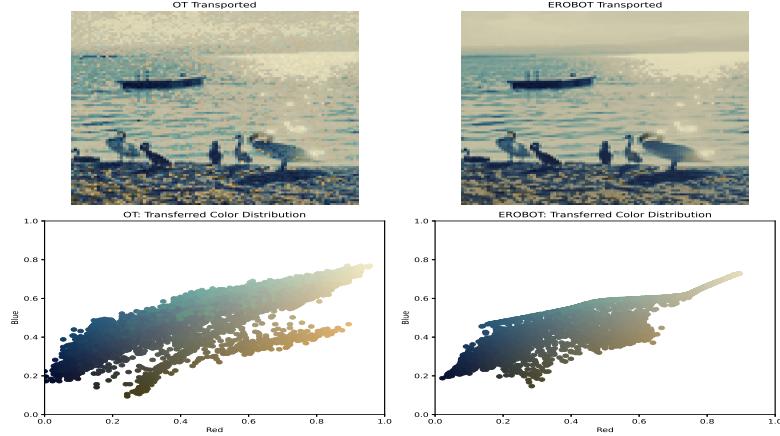


Fig. 8 Results of color transfer. Top panel: Output images for OT (left) and E-ROBOT (right), with $\lambda = 20$ and $\varepsilon = 0.01$. Bottom: Their corresponding color distributions.

in flat areas of the image (namely, it creates colour artifacts). We refer to [Ferradans, Papadakis, Peyré, and Aujol \(2014\)](#) for a discussion.

We demonstrate the practical efficacy of E-ROBOT on a color transfer task, adapting the palette of Hokusai's *The Great Wave off Kanagawa* (source image) to that of a scene on Lake Léman (target image), in the Geneva (Switzerland) area; see Figure 7 for the pictures (top panels), with the corresponding color distributions (bottom panels), as obtained using a Red-Blue representation. The ultimate goal is to impose on the target image the colours of Hokusai's paint, while preserving the smoothness of both images and avoiding the creation of artifacts.

The bottom plots of Figure 7 highlight that the source image contains some colours that pose significant challenges for traditional OT. For instance, there are extreme pigment variations (the vibrant Prussian blue used in the woodblock print creates saturated blue values that fall outside the natural color distribution of the target landscape photograph) or compositional elements (the distinctive white foam splashes and the yellow boat in Hokusai’s work create isolated color clusters that don’t correspond to any elements in the target image).

The results in Figure 8 reveal striking differences in how OT and E-ROBOT handle these challenges. The OT result exhibits notable artifacts, including a speckled noise pattern indicative of unstable pixel assignments, evidenced by a persistent secondary peak in the red region of its distribution. This secondary peak represents OT’s attempt to overfit to red/yellow colours, forcibly matching anomalous source colours to inappropriate target regions. The resulting distribution shows issues in the red channel, where the artificial peak creates unnatural tones and distorted color balances.

In contrast, the E-ROBOT transformation (right panels) demonstrates remarkable robustness. The resulting color distribution illustrates that the robust Sinkhorn divergence effectively downweights the influence of extreme colours. The algorithm recognizes that certain source colours (like the intense Prussian blue or extreme tones of yellow) entail large transport cost, and thus prevents them from distorting the overall color mapping. This robustness is visually apparent in the transformed image. For instance, the sky tones transition naturally from Hokusai’s distinctive palette to the target’s atmospheric blues without the artificial color casts that plague the OT result via a number of grey untied dots. Moreover, thanks to the regularization, the E-ROBOT features smoothness in the final picture, confirming its statistical superiority for tasks involving real-world images with complex color distributions and inherent outliers.

5 Conclusion and possible developments

We introduce the E-ROBOT framework, which combines the outlier robustness of ROBOT with the computational efficiency of entropic regularization via the Schrödinger bridge. The resulting robust Sinkhorn divergence $\bar{W}_{\varepsilon,\lambda}$ achieves a dimension-free sample complexity of $\mathcal{O}(n^{-1/2})$, overcoming the curse of dimensionality. We demonstrate its applicability in high-dimensional, heavy-tailed settings for tasks like goodness-of-fit testing, barycenter computation, and gradient flows. E-ROBOT can be implemented with simple modifications to existing optimal transport algorithms.

Besides these theoretical and methodological results, other future developments can be envisioned. Hereunder we mention some of them.

(i) *Parametric inference.* The robust Sinkhorn divergence $\bar{W}_{\varepsilon,\lambda}$ can be applied as loss function to conduct parametric inference in statistics similarly to MKE (see [Bassetti et al. \(2006\)](#)) or in generative modeling (see [Genevay et al. \(2018\)](#)). Specifically, as in [Ma et al. \(2025\)](#), one may define a parametric model as $\{\mu_\theta, \theta \in \Theta \subset \mathbb{R}^q, q \geq 1\}$.

Then, the minimum robust Sinkhorn estimator is defined as

$$\hat{\theta}_n^{\lambda, \varepsilon} = \operatorname{argmin}_{\theta \in \Theta} \overline{W}_{\varepsilon, \lambda}(\hat{\mu}_n, \mu_\theta).$$

When there is no explicit expression for the probability measure characterizing the parametric model (e.g. in complex generative models), the computation of $\hat{\theta}_n^{\lambda, \varepsilon}$ can be difficult. To cope with this issue, one may think of using the minimum expected robust Sinkhorn estimator defined as

$$\hat{\theta}_{n,m}^{\lambda, \varepsilon} = \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}_m [\overline{W}_{\varepsilon, \lambda}(\hat{\mu}_n, \hat{\mu}_{\theta, m})],$$

where the expectation \mathbb{E}_m is taken over the distribution $\mu_\theta^{(m)}$. To implement $\hat{\theta}_{n,m}^{\lambda}$ one can rely on Monte Carlo methods and approximate numerically $\mathbb{E}_m[\overline{W}_{\varepsilon, \lambda}(\hat{\mu}_n, \hat{\mu}_{\theta, m})]$. The existence, measurability, and consistency of the resulting estimator should be proved as in [Bernton, Jacob, Gerber, and Robert \(2019\)](#). Moreover, since the sample complexity of $\overline{W}_{\varepsilon, \lambda}$ scales with $n^{-1/2}$, we conjecture that it is possible to prove root- n consistency and asymptotic normality of $\hat{\theta}_{n,m}^{\lambda}$, similarly to the results for W_2 in [del Barrio and Loubes \(2019\)](#)—a conjecture that does not make sense for OT- and ROBOT-based estimators in multivariate setting. For the resulting estimators, one may think also of deriving small-sample approximation to their distributions via saddlepoint techniques, exploring connections in the setting of dependent data (as in [Jiang, La Vecchia, Ronchetti, and Scaillet \(2023\); La Vecchia and Ronchetti \(2019\)](#)) or of independent data (as in [La Vecchia et al. \(2022\)](#)).

(ii) *GoF via Bregman-type divergence and MMD.* Following the same logic as the GoF test described in §4.1, one can also define test statistics using either the Hausdorff divergence in (25) or the MMD using $k_{\lambda, \varepsilon}$. In the case of simple hypothesis, the distribution of these statistics can be obtained via Monte Carlo methods, as in §4.1. Moreover, one may think of considering also composite hypotheses and derive the distribution of the test statistics via bootstrap methods, whose statistical guarantees may be proved building on the results in Theorem 9 and on [Klatt, Tameling, and Munk \(2020\)](#). The resulting approach based on $\overline{W}_{\varepsilon, \lambda}$ may offer an alternative to the recent developments in [Hu and Lin \(2025\)](#), where the max-sliced Wasserstein distance is applied.

(iii) *Relaxing the assumptions.* Some of our results of this paper assume compact support: this is common in the OT literature for technical convenience. With this regard, we notice that some of the results in the paper (like Proposition 1 and 2) already hold for noncompact spaces. Thus, one may think of relaxing the compactness assumption. In addition, we highlight that many of our theoretical and methodological results can be extended to the case where we use a generic cost function is $\varepsilon^{-1}c_\lambda$, with $c_\lambda = \tilde{d}^p$ and $\tilde{d}_\lambda = \min(d, 2\lambda)$. The use of this cost function could lead to an interesting research topic: extending $W_{\varepsilon, \lambda}$ to $W_{\varepsilon, \lambda, p}$, analogous to the order- p Wasserstein distance W_p in classical optimal transport. Both these extensions require modified assumptions and suitable modifications of our proofs.

(iv) *Selection of λ and ε .* A key challenge for practitioners is the joint selection of the robustness parameter λ and the entropic regularization parameter ε . We note that this problem is not only open but also fundamentally new, as the literature on OT offers limited guidance even for choosing these parameters individually. The selection of ε in E-OT is often heuristic (see e.g. [Goldfeld and Greenewald \(2020\)](#)), and robust OT methods frequently lack a general data-driven procedure for λ (see e.g. [Nietert et al. \(2022\)](#)). Therefore, the joint calibration required for E-ROBOT operates on new ground and highlights a clear gap in the current literature that we are planning to fill. While beyond the scope of this paper, which introduces and validates the E-ROBOT framework itself, we conjecture that a solution could be based on a concentration inequality derived from the principles in Theorem 9. We posit this as a critical avenue for future work.

References

- Arbel, M., Korba, A., Salim, A., Gretton, A. (2019). Maximum mean discrepancy gradient flow. *Advances in Neural Information Processing Systems*, 32, 1-11
- Bassetti, F., Bodini, A., Regazzini, E. (2006). On minimum Kantorovich distance estimators. *Statistics & probability letters*, 76(12), 1298–1302
- Bassetti, F., & Regazzini, E. (2006). Asymptotic properties and robustness of minimum dissimilarity estimators of location-scale parameters. *Theory of Probability & Its Applications*, 50(2), 171–186
- Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., Peyré, G. (2015). Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2), A1111–A1138
- Bernton, E., Jacob, P.E., Gerber, M., Robert, C.P. (2019). On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4), 657–676
- Bunne, C., Hsieh, Y.-P., Cuturi, M., Krause, A. (2023). The Schrödinger bridge between gaussian measures has a closed form. *International conference on artificial intelligence and statistics* (pp. 5802–5833).
- Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 26). Curran Associates, Inc.

del Barrio, E., & Loubes, J.-M. (2019). Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47(2), 926–951

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

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouvé, A., Peyré, G. (2019). Interpolating between optimal transport and MMD using Sinkhorn divergences. *The 22nd international conference on artificial intelligence and statistics* (pp. 2681–2690).

Genevay, A., Chizat, L., Bach, F., Cuturi, M., Peyré, G. (2019). Sample complexity of Sinkhorn divergences. *The 22nd international conference on artificial intelligence and statistics* (pp. 1574–1583).

Genevay, A., Peyré, G., Cuturi, M. (2018). Learning generative models with Sinkhorn divergences. *International conference on artificial intelligence and statistics* (pp. 1608–1617).

Goldfeld, Z., & Greenewald, K. (2020). Gaussian-smoothed optimal transport: Metric structure and statistical efficiency. *International conference on artificial intelligence and statistics* (pp. 3327–3337).

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A. (2006). A kernel method for the two-sample-problem. B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems* (Vol. 19). MIT Press.

Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A. (2012). A kernel two-sample test. *The journal of machine learning research*, 13(1), 723–773

Hallin, M. (2022). Measure transportation and statistical decision theory. *Annual Review of Statistics and Its Application*, 9, 401–424

Hallin, M., La Vecchia, D., Liu, H. (2022). Center-outward R-estimation for semiparametric VARMA models. *Journal of the American Statistical Association*, 117(538), 925–938

Hallin, M., La Vecchia, D., Liu, H. (2023). Rank-based testing for semiparametric VAR models: a measure transportation approach. *Bernoulli*, 29(1), 229–273

- Hallin, M., & Liu, H. (2024). Quantiles and quantile regression on riemannian manifolds: a measure-transportation-based approach. *arXiv preprint arXiv:2410.15711*, 1–42
- Hallin, M., Mordant, G., Segers, J. (2021). Multivariate goodness-of-fit tests based on Wasserstein distance. *Electronic Journal of Statistics*, 15, 1328–1371,
- Hu, X., & Lin, Z. (2025). Two-sample distribution tests in high dimensions via max-sliced Wasserstein distance and bootstrapping. *Biometrika*, 112(2), asaf001,
- Jiang, C., La Vecchia, D., Ronchetti, E., Scaillet, O. (2023). Saddlepoint approximations for spatial panel data models. *Journal of the American Statistical Association*, 118(542), 1164–1175
- Klatt, M., Tameling, C., Munk, A. (2020). Empirical regularized optimal transport: Statistical theory and applications. *SIAM Journal on Mathematics of Data Science*, 2(2), 419–443
- Kolouri, S., Park, S.R., Thorpe, M., Slepcev, D., Rohde, G.K. (2017). Optimal mass transport: Signal processing and machine learning applications. *IEEE signal processing magazine*, 34(4), 43–59
- La Vecchia, D., & Ronchetti, E. (2019). Saddlepoint approximations for short and long memory time series: A frequency domain approach. *Journal of Econometrics*, 213(2), 578–592
- La Vecchia, D., Ronchetti, E., Ilievski, A. (2022). On some connections between Esscher's tilting, saddlepoint approximations, and optimal transportation: A statistical perspective. *Statistical Science*, 1(1), 1–22
- Léonard, C. (2013). A survey of the Schrödinger problem and some of its connections with optimal transport. *arXiv:1308.0215*, 1–40
- Li, Y., Swersky, K., Zemel, R. (2015). Generative moment matching networks. *International conference on machine learning* (pp. 1718–1727).
- Liu, G.-H., Chen, T., So, O., Theodorou, E. (2022). Deep generalized Schrödinger bridge. *Advances in Neural Information Processing Systems*, 35, 9374–9388

Ma, Y., Liu, H., La Vecchia, D., Lerasle, M. (2025). Inference via robust optimal transportation: theory and methods. *International Statistical Review*(available on line), 1-39

Montavon, G., Müller, K.-R., Cuturi, M. (2016). Wasserstein training of restricted Boltzmann machines. *Advances in Neural Information Processing Systems*, 29, 1-9

Mukherjee, D., Guha, A., Solomon, J.M., Sun, Y., Yurochkin, M. (2021). Outlier-robust optimal transport. *Proceedings of the 38th international conference on machine learning* (Vol. 139, pp. 7850–7860).

Nietert, S., Goldfeld, Z., Cummings, R. (2022). Outlier-robust optimal transport: Duality, structure, and statistical analysis. *International conference on artificial intelligence and statistics* (pp. 11691–11719).

Nietert, S., Goldfeld, Z., Kato, K. (2021). Smooth p -Wasserstein distance: structure, empirical approximation, and statistical applications. *International conference on machine learning* (pp. 8172–8183).

Nutz, M. (2021). *Introduction to entropic optimal transport*. Lecture notes, Columbia University.

Pardo, L. (2018). *Statistical inference based on divergence measures*. Chapman and Hall/CRC.

Peyré, G., & Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6), 355–607

Pooladian, A.-A., & Niles-Weed, J. (2025). Plug-in estimation of Schrödinger bridges. *SIAM Journal on Mathematics of Data Science*, 7(3), 1315–1336

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

Santambrogio, F. (2015). *Optimal transport for applied mathematicians* (Vol. 55; N. Birkhäuser, Ed.) (No. 58-63). Springer.

Schrödinger, E. (1931). Über die umkehrung der naturgesetze. *Sitzungsberichte der Preussischen Akademie der Wissenschaften, Physikalisch-Mathematische Klasse*, 144–153

Sriperumbudur, B.K., Fukumizu, K., Lanckriet, G.R. (2011). Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(7), 2389–2410

Van Der Vaart, A.W., & Wellner, J.A. (1996). *Weak convergence and empirical processes*. Springer.

Wang, G., Jiao, Y., Xu, Q., Wang, Y., Yang, C. (2021). Deep generative learning via Schrödinger bridge. *International conference on machine learning* (pp. 10794–10804).

Weed, J., & Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25, 2620–2648