Unbalanced Optimal Transport meets Sliced-Wasserstein

Anonymous Authors[1]

Abstract

Optimal transport (OT) has emerged as a powerful framework to compare probability measures, a fundamental task in many statistical and machine learning problems. Substantial advances have been made over the last decade in designing OT variants which are either computationally and statistically more efficient, or more robust to the measures/datasets to compare. Among them, sliced OT distances have been extensively used to mitigate optimal transport's cubic algorithmic complexity and curse of dimensionality. In parallel, unbalanced OT was designed to allow comparisons of more general positive measures, while being more robust to outliers. In this paper, we propose to combine these two concepts, namely slicing and unbalanced OT, to develop a general framework for efficiently comparing positive measures. We propose two new loss functions based on the idea of slicing unbalanced OT, and study their induced topology and statistical properties. We then develop a fast Frank-Wolfe-type algorithm to compute these losses, and show that our methodology is modular as it encompasses and extends prior related work. We finally conduct an empirical analysis of our loss functions and methodology on both synthetic and real datasets, to illustrate their relevance and applicability.

1. Introduction

Positive measures are ubiquitous in various fields, including data sciences and machine learning (ML) where they commonly serve as data representations. A common example is the density fitting task, which arises in generative modeling (Arjovsky et al., 2017; De Bortoli et al., 2021): the observed samples can be represented as a discrete positive measure α and the goal is to find a parametric measure β_η which fits the best α. This can be achieved by training a model that minimizes a loss function over η, usually defined as a distance between α and β_η. Therefore, it is important to choose a meaningful discrepancy with desirable statistical, robustness and computational properties. In particular, some settings require comparing arbitrary positive measures, *i.e.* measures whose total mass can have an arbitrary value, as opposed to probability distributions, whose total mass is equal to 1. In cell biology (Schiebinger et al., 2019), for example, measures are used to represent and compare gene expressions of cell populations, and the total mass represents the population size.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the Workshop on New Frontiers in Learning, Control, and Dynamical Systems at the International Conference on Machine Learning (ICML). Do not distribute.

(Unbalanced) Optimal Transport. Optimal transport has been chosen as a loss function in various ML applications. OT defines a distance between two positive measures of same mass α and β (*i.e.* $m(\alpha) = m(\beta)$) by moving the mass of α toward the mass of β with least possible effort. The mass equality can nevertheless be hindering by imposing a normalization of α and β to enforce $m(\alpha) = m(\beta)$, which is potentially spurious and makes the problem less interpretable. In recent years, OT has then been extended to settings where measures have different masses, leading to the *unbalanced OT* (UOT) framework (Liero et al., 2018; Kondratyev et al., 2016; Chizat et al., 2018b). An appealing outcome of this new OT variant is its robustness to outliers which is achieved by discarding them before transporting α to β. UOT has been useful for many theoretical and practical applications, e.g. theory of deep learning (Chizat & Bach, 2018; Rotskoff et al., 2019), biology (Schiebinger et al., 2019; Demetci et al., 2022) and domain adaptation (Fatras et al., 2021). We refer to (Séjourné et al., 2022a) for an extensive survey of UOT. Computing OT requires to solve a linear program whose complexity is in $\mathcal{O}(n^3 \log n)$. Besides, accurately estimating OT distances through empirical disributions is challenging as OT suffers from the curse of dimension (Dudley, 1969). A common workaround is to rely on OT variants with lower complexities and better statistical properties. Among the most popular, we can list entropic OT (Cuturi, 2013), minibatch OT (Fatras et al., 2020) and sliced OT (Radon, 2005; Bonneel et al., 2015). In this paper, we will focus on the latter.

Slicing (U)OT and related work. Sliced OT leverages the OT 1D closed-form solution to define a new cost. It averages the OT cost between projections of (α, β) on 1D subspaces of $\mathbb{R}^d$. For 1D data, the OT solution can be computed through a sort algorithm, leading to an appealing $\mathcal{O}(n\log(n))$ complexity (Peyré et al., 2019). Furthermore, it has been shown to lift useful topological and statistical properties of OT from 1-dimensional to multi-dimensional settings (Bayraktar & Guo, 2021; Nadjahi et al., 2020; Goldfeld & Greenewald, 2021). It therefore helps to mitigate the curse of dimensionality making SOT-based algorithms theoretically-grounded, statistically efficient and efficiently solvable even on large-scale settings. These appealing properties motivated the development of several variants and generalizations, e.g. to different types or distributions of projections (Kolouri et al., 2019; Deshpande et al., 2019; Nguyen et al., 2020; Ohana et al., 2023) and non-Euclidean data (Bonet et al., 2023a; 2022a; 2023b). The slicing operation has also been applied to partial OT (Bonneel & Coeurjolly, 2019; Bai et al., 2022; Sato et al., 2020), a particular case of UOT, in order to speed up comparisons of unnormalized measures at large scale. However, while (sliced) partial OT allows to compare measures with different masses, it assumes that each input measure is discrete and supported on points that all share the same mass (typically 1). In contrast, the Gaussian-Hellinger-Kantorovich (GHK) distance (Liero et al., 2018), another popular formulation of UOT, allows to compare measures with different masses *and* supported on points with varying masses, and has not been studied jointly with slicing.

Contributions. This paper presents the first general framework combining UOT and slicing. Our main contribution is the introduction of two novel sliced variants of UOT, respectively called *Sliced UOT (*SUOT*)* and *Unbalanced Sliced OT (*USOT*)*. SUOT and USOT both leverage one-dimensional projections and the newly-proposed implementation of UOT in 1D (Séjourné et al., 2022b), but differ in the penalization used to relax the constraint on the equality of masses: USOT essentially performs a global reweighting of the inputs measures (α,β), while SUOT reweights each projection of (α,β). Our work builds upon the Frank-Wolfe-type method (Frank & Wolfe, 1956) recently proposed in (Séjourné et al., 2022b) to efficiently compute GHK between univariate measures, an instance of UOT which has not yet been combined with slicing. We derive the associated theoretical properties, along with the corresponding fast and GPU-friendly algorithms. We demonstrate its versatility and efficiency on challenging experiments, where slicing is considered on a non-Euclidean hyperbolic manifold, as a similarity measure for document classification, or for computing barycenters of geoclimatic data.

Outline. In Section 2, we provide background knowledge on UOT and sliced OT (SOT). In Section 3, we define our two new loss functions (SUOT and USOT) and prove their metric, topological, statistical and duality properties in wide generality. We then detail in Section 4 the numerical implementation of SUOT and USOT based on the Frank-Wolfe algorithm. We investigate their empirical performance on hyperbolic and geophysical data as well as document classification in Section 5.

2. Background

Unbalanced Optimal Transport. We denote by $\mathcal{M}_+(\mathbb{R}^d)$ the set of all positive Radon measures on $\mathbb{R}^d$. For any $\alpha\in\mathcal{M}_+(\mathbb{R}^d)$, $\mathrm{supp}(\alpha)$ is the support of α and $m(\alpha)=\int_{\mathbb{R}^d}\mathrm{d}\alpha(x)$ the mass of α. We recall the standard formulation of unbalanced OT (Liero et al., 2018), which uses *φ-divergences* for regularization.

Definition 2.1. (Unbalanced OT) Let $\alpha,\beta\in\mathcal{M}_+(\mathbb{R}^d)$. Let $\varphi:\mathbb{R}\to\mathbb{R}\cup\{+\infty\}$ be an *entropy function*, *i.e.* φ is convex, lower semicontinuous, $\mathrm{dom}(\varphi)\triangleq\{x\in\mathbb{R},\,\varphi(x)<+\infty\}\subset[0,+\infty)$ and $\varphi(1)=0$. Denote $\varphi'_\infty\triangleq\lim_{x\to+\infty}\varphi(x)/x$. The *$\varphi$-divergence* between α and β is defined as,

$$\mathrm{D}_\varphi(\alpha|\beta)\triangleq\int_{\mathbb{R}^d}\varphi\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}(x)\right)\mathrm{d}\beta(x)+\varphi'_\infty\int_{\mathbb{R}^d}\mathrm{d}\alpha^\perp(x)\,, \tag{1}$$

where $\alpha^\perp$ is defined as $\alpha=(\mathrm{d}\alpha/\mathrm{d}\beta)\beta+\alpha^\perp$. Given two entropy functions (φ_1,φ_2) and a cost $\mathrm{C}_d:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$, the unbalanced OT problem between α and β reads

$$\mathrm{UOT}(\alpha,\beta)\triangleq\inf_{\pi\in\mathcal{M}_+(\mathbb{R}^d\times\mathbb{R}^d)}\int\mathrm{C}_d(x,y)\mathrm{d}\pi(x,y)+\mathrm{D}_{\varphi_1}(\pi_1|\alpha)+\mathrm{D}_{\varphi_2}(\pi_2|\beta)\,, \tag{2}$$

where (π_1,π_2) denote the marginal distributions of π.

When $\varphi_1=\varphi_2$ and $\varphi_1(x)=0$ for $x=1$, $\varphi_1(x)=+\infty$ otherwise, (2) boils down to the Kantorovich formulation of OT (or *balanced OT*), which we denote by $\mathrm{OT}(\alpha,\beta)$. Indeed, in that case, $\mathrm{D}_{\varphi_1}(\pi_1|\alpha)=\mathrm{D}_{\varphi_2}(\pi_2|\beta)=0$ if $\pi_1=\alpha$ and $\pi_2=\beta$, $\mathrm{D}_{\varphi_1}(\pi_1|\alpha)=\mathrm{D}_{\varphi_2}(\pi_2|\beta)=+\infty$ otherwise.

Under suitable choices of entropy functions (φ_1,φ_2), $\mathrm{UOT}(\alpha,\beta)$ allows to compare α and β even when $m(\alpha)\neq m(\beta)$ and can discard outliers, which makes it more robust than $\mathrm{OT}(\alpha,\beta)$. Two common choices are $\varphi(x)=\rho\,|x-1|$ and $\varphi(x)=\rho(x\log(x)-x+1)$, where $\rho>0$ is a *characteristic radius* w.r.t. C_d. They respectively correspond to $\mathrm{D}_\varphi=\rho\mathrm{TV}$ (total variation distance (Chizat et al., 2018a)) and $\mathrm{D}_\varphi=\rho\mathrm{KL}$ (*Kullback-Leibler divergence*).

The UOT problem has been shown to admit an equivalent formulation obtained by deriving the dual of (2) and proving strong duality. Based on Proposition 2.2, computing $\mathrm{UOT}(\alpha,\beta)$ consists in optimizing a pair of continuous functions (f,g).

Proposition 2.2. *(Liero et al., 2018, Corollary 4.12) The* UOT *problem* (2) *can equivalently be written as*

$$\mathrm{UOT}(\alpha,\beta) = \sup_{f\oplus g\leq \mathrm{C}_d} \int \varphi_1^\circ(f(x))\mathrm{d}\alpha(x) + \int \varphi_2^\circ(g(y))\mathrm{d}\beta(y), \tag{3}$$

where for $i \in \{1,2\}$, $\varphi_i^\circ(x) \triangleq -\varphi_i^*(-x)$ *with* $\varphi_i^*(x) \triangleq \sup_{y\geq 0} xy - \varphi_i(y)$ *the* Legendre transform *of* φ_i*, and* $f \oplus g \leq \mathrm{C}_d$ *means that for* $(x,y) \sim \alpha\otimes\beta$, $f(x) + g(y) \leq \mathrm{C}_d(x,y)$.

In this paper, we mainly focus on the *GHK setting*, both theoretically and computationally. It corresponds to (2) with $\mathrm{C}_d(x,y) = ||x-y||^2$, $\mathrm{D}_{\varphi_i} = \rho_i\mathrm{KL}$, leading to $\varphi_i^\circ(x) = \rho_i(1-e^{-x/\rho_i})$. $\mathrm{UOT}(\alpha,\beta)$ is known to be computationally intensive (Pham et al., 2020), thus motivating the development of methods that can scale to dimensions and sample sizes encountered in ML applications.

Sliced Optimal Transport. Among the many workarounds that have been proposed to overcome the OT computational bottleneck (Peyré et al., 2019), Sliced OT (Rabin et al., 2012) has attracted a lot of attention due to its computational benefits and theoretical guarantees. We define it below.

Definition 2.3 (Sliced OT)**.** Let $\mathbb{S}^{d-1} \triangleq \{\theta\in\mathbb{R}^d : \|\theta\| = 1\}$ be the unit sphere in $\mathbb{R}^d$. For $\theta \in \mathbb{S}^{d-1}$, denote by $\theta^\star : \mathbb{R}^d \to \mathbb{R}$ the linear map such that for $x\in\mathbb{R}^d$, $\theta^\star(x) \triangleq \langle\theta, x\rangle$. Let $\boldsymbol{\sigma}$ be the uniform probability over $\mathbb{S}^{d-1}$. For $\alpha,\beta\in\mathcal{M}_+(\mathbb{R}^d)$, the *Sliced OT* problem reads

$$\mathrm{SOT}(\alpha,\beta) \triangleq \int_{\mathbb{S}^{d-1}} \mathrm{OT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta)\mathrm{d}\boldsymbol{\sigma}(\theta)\,, \tag{4}$$

where for any measurable function f and $\xi\in\mathcal{M}_+(\mathbb{R}^d)$, $f_\sharp\xi$ is the *push-forward measure* of ξ by f, *i.e.* for any measurable set $\mathrm{A}\subset\mathbb{R}$, $f_\sharp\xi(\mathrm{A}) \triangleq \xi(f^{-1}(\mathrm{A}))$, $f^{-1}(\mathrm{A}) \triangleq \{x\in\mathbb{R}^d : f(x)\in\mathrm{A}\}$.

Note that $\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta$ are two measures supported on $\mathbb{R}$, therefore $\mathrm{OT}(\theta^\star_\sharp\mu, \theta^\star_\sharp\nu)$ is defined in terms of a cost function $\mathrm{C}_1 : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$. Since OT between univariate measures can be efficiently computed, $\mathrm{SOT}(\alpha,\beta)$ can provide significant computational advantages over $\mathrm{OT}(\alpha,\beta)$ in large-scale settings. In practice, if α and β are discrete measures supported on $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ respectively, the standard procedure for approximating $\mathrm{SOT}(\alpha,\beta)$ consists in (i) sampling m i.i.d. samples $\{\theta_j\}_{j=1}^m$ from $\boldsymbol{\sigma}$, (ii) computing $\mathrm{OT}((\theta_j^\star)_\sharp\alpha, (\theta_j^\star)_\sharp\beta)$, $j = 1,\dots,m$. Computing OT between univariate discrete measures amounts to sorting (Peyré et al., 2019, Section 2.6), thus step (ii) involves $\mathcal{O}(n\log n)$ operations for each θ_j.

$\mathrm{SOT}(\alpha,\beta)$ is defined in terms of the Kantorovich formulation of OT, hence inherits the following drawbacks: $\mathrm{SOT}(\alpha,\beta) < +\infty$ only when $m(\alpha) = m(\beta)$, and may not provide meaningful comparisons in presence of outliers. To overcome such limitations, prior work have proposed sliced versions of partial OT (Bonneel & Coeurjolly, 2019; Bai et al., 2022), a particular instance of UOT. However, their contributions only apply to measures whose samples have constant mass. We generalize their line of work in the next section.

3. Sliced Unbalanced OT and Unbalanced Sliced OT: Theoretical Analysis

We propose two strategies to make unbalanced OT scalable, by leveraging sliced OT. We formulate two loss functions (Definition 3.1), then study their theoretical properties and discuss their implications.

Definition 3.1. Let $\alpha,\beta\in\mathcal{M}_+(\mathbb{R}^d)$. The **Sliced Unbalanced OT** loss (SUOT) and the **Unbalanced Sliced OT** loss (USOT) between α and β are defined as,

$$\mathrm{SUOT}(\alpha,\beta) \triangleq \int_{\mathbb{S}^{d-1}} \mathrm{UOT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta)\mathrm{d}\boldsymbol{\sigma}(\theta)\,, \tag{5}$$

$$\mathrm{USOT}(\alpha,\beta) \triangleq \inf_{(\pi_1,\pi_2)\in\mathcal{M}_+(\mathbb{R}^d)\times\mathcal{M}_+(\mathbb{R}^d)} \mathrm{SOT}(\pi_1,\pi_2) + \mathrm{D}_{\varphi_1}(\pi_1|\alpha) + \mathrm{D}_{\varphi_2}(\pi_2|\beta)\,. \tag{6}$$

$\mathrm{SUOT}(\alpha,\beta)$ compares α and β by solving the UOT problem between $\theta^\star_\sharp\alpha$ and $\theta^\star_\sharp\beta$ for $\theta\sim\boldsymbol{\sigma}$. Note that SUOT extends the sliced partial OT problem (Bonneel & Coeurjolly, 2019; Bai et al., 2022) (where $\mathrm{D}_{\varphi_i} = \rho_i\mathrm{TV}$) by allowing the use of arbitrary φ-divergences. On the other hand, USOT is a completely novel approach and stems from the following property on UOT (Liero et al., 2018, Equations (4.21)): $\mathrm{UOT}(\alpha,\beta) = \inf_{(\pi_1,\pi_2)\in\mathcal{M}_+(\mathbb{R}^d)^2} \mathrm{OT}(\pi_1,\pi_2) + \mathrm{D}_{\varphi_1}(\pi_1|\alpha) + \mathrm{D}_{\varphi_2}(\pi_2|\beta)$.

SUOT vs. USOT. As outlined in Definition 3.1, SUOT and USOT differ in how the transportation problem is penalized: $\mathrm{SUOT}(\alpha,\beta)$ regularizes the marginals of π_θ for $\theta\sim\boldsymbol{\sigma}$ where π_θ denotes the solution of $\mathrm{UOT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta)$, while $\mathrm{USOT}(\alpha,\beta)$ operates a geometric normalization directly on (α,β). We illustrate this difference on the following practical setting: we consider $(\alpha,\beta)\in\mathcal{M}_+(\mathbb{R}^2)$ where α is polluted with some outliers, and we compute $\mathrm{SUOT}(\alpha,\beta)$ and $\mathrm{USOT}(\alpha,\beta)$. We plot the input measures and the sampled projections $\{\theta_k\}_k$ (Figure 1, left), the marginals of π_{θ_k} for SUOT and the marginals of $(\theta_k)^\star_\sharp\pi$ for USOT (Figure 1, right). As expected, SUOT marginals change for each θ_k. We also observe that the source outliers have successfully been removed for any θ when using USOT, while they may still appear with SUOT (e.g. for $\theta = 120°$): this is a direct consequence of the penalization terms D_{φ_i} in USOT, which operate on (α,β) rather than on their projections.

Theoretical analysis. In the rest of this section, we prove a set of theoretical properties of SUOT and USOT. All proofs

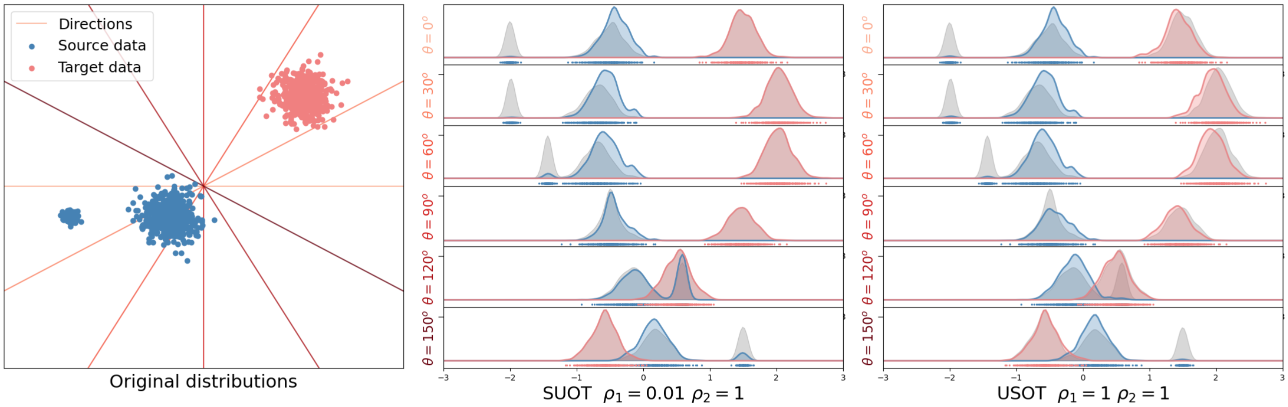


Figure 1: **Toy illustration** on the behaviors of SUOT and USOT. *(left)* Original 2D samples and slices used for illustration. KDE density estimations of the projected samples: grey, original distributions, colored, distributions reweighed by SUOT *(center)*, and reweighed by USOT *(right)*.

are provided in Appendix A. We first identify the conditions on the cost C_1 and entropies φ_1, φ_2 under which the infimum is attained in $\mathrm{UOT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta)$ for $\theta \in \mathbb{S}^{d-1}$ and in $\mathrm{USOT}(\alpha, \beta)$: the formal statement is given in Appendix A. We also show that these optimization problems are convex, both SUOT and USOT are jointly convex w.r.t. their input measures, and that strong duality holds (Theorem 3.7).

Next, we prove that both SUOT and USOT preserve some topological properties of UOT, starting with the metric axioms as stated in the next proposition.

Proposition 3.2. ***(Metric properties)*** (i) *Suppose* UOT *is non-negative, symmetric and/or definite on* $\mathcal{M}_+(\mathbb{R}) \times \mathcal{M}_+(\mathbb{R})$*. Then,* SUOT *is respectively non-negative, symmetric and/or definite on* $\mathcal{M}_+(\mathbb{R}^d) \times \mathcal{M}_+(\mathbb{R}^d)$*. If there exists* $p \in [1, +\infty)$ *s.t. for any* $(\alpha, \beta, \gamma) \in \mathcal{M}_+(\mathbb{R})$*,* $\mathrm{UOT}^{1/p}(\alpha, \beta) \leq \mathrm{UOT}^{1/p}(\alpha, \gamma) + \mathrm{UOT}^{1/p}(\gamma, \beta)$*, then* $\mathrm{SUOT}^{1/p}(\alpha, \beta) \leq \mathrm{SUOT}^{1/p}(\alpha, \gamma) + \mathrm{SUOT}^{1/p}(\gamma, \beta)$.

(ii) *For* $\alpha, \beta \in \mathcal{M}_+(\mathbb{R}^d)$, $\mathrm{USOT}(\alpha, \beta) \geq 0$*. If* $\varphi_1 = \varphi_2$*,* USOT *is symmetric. If* $\mathrm{D}_{\varphi_1}, \mathrm{D}_{\varphi_2}$ *are definite, so is* USOT.

By Proposition 3.2(i), establishing the metric axioms of UOT between *univariate* measures (e.g., as detailed in (Séjourné et al., 2022a, Section 3.3.1)) suffices to prove the metric axioms of SUOT between *multivariate* measures. Since e.g. GHK (Liero et al., 2018, Theorem 7.25) is a metric for $p = 2$, then so is the associated SUOT.

In our next theorem, we show that SUOT, USOT and UOT are equivalent, under certain assumptions on the entropies (φ_1, φ_2), cost functions, and input measures (α, β).

Theorem 3.3. ***(Equivalence of*** **SUOT, USOT, UOT*)*** *Let* $\mathsf{X} \subset \mathbb{R}^d$ *be a compact set with radius* R*. Let* $p \in [1, +\infty)$*. Assume* $\mathrm{C}_1(x, y) = |x - y|^p$, $\mathrm{C}_d(x, y) = \|x - y\|^p$, $\mathrm{D}_{\varphi_1} = \mathrm{D}_{\varphi_2} = \rho\mathrm{KL}$*. Then, for* $\alpha, \beta \in \mathcal{M}_+(\mathsf{X})$*,*

$$\mathrm{SUOT}(\alpha, \beta) \leq \mathrm{USOT}(\alpha, \beta) \leq \mathrm{UOT}(\alpha, \beta)\,, \quad \textit{and} \tag{7}$$

$$\mathrm{UOT}(\alpha, \beta) \leq c(m(\alpha), m(\beta), \rho, R)\mathrm{SUOT}(\alpha, \beta)^{1/(d+1)}\,, \tag{8}$$

where $c(m(\alpha), m(\beta), \rho, R)$ *is constant depending on* $m(\alpha), m(\beta), \rho, R$*, which is non-decreasing in* $m(\alpha)$ *and* $m(\beta)$*. Additionally, assume there exists* $M > 0$ *s.t.* $m(\alpha) \leq M, m(\beta) \leq M$*. Then,* $c(m(\alpha), m(\beta), \rho, R)$ *no longer depends on* $m(\alpha), m(\beta)$*, which proves the equivalence of* SUOT, USOT *and* UOT.

Theorem 3.3 is an application of a more general result, which we derive in the appendix. In particular, we show that the first two inequalities in (7) hold under milder assumptions on φ_1, φ_2 and $\mathrm{C}_1, \mathrm{C}_d$. The equivalence of SUOT, USOT and UOT is useful to prove that SUOT and USOT *metrize the weak* convergence* when UOT does, e.g. in the GHK setting (Liero et al., 2018, Theorem 7.25). Before formally stating this result, we recall that a sequence of positive measures $(\alpha_n)_{n \in \mathbb{N}^*}$ converges weakly to $\alpha \in \mathcal{M}_+(\mathbb{R}^d)$ (denoted $\alpha_n \rightharpoonup \alpha$) if for any continuous $f : \mathbb{R}^d \to \mathbb{R}$, $\lim_{n \to +\infty} \int f \mathrm{d}\alpha_n = \int f \mathrm{d}\alpha$.

Theorem 3.4. ***(Weak* metrization)*** *Assume* $\mathrm{D}_{\varphi_1} = \mathrm{D}_{\varphi_2} = \rho\mathrm{KL}$*. Let* $p \in [1, +\infty)$ *and consider* $\mathrm{C}_1(x, y) = |x - y|^p$*,* $\mathrm{C}_d(x, y) = \|x - y\|^p$*. Let* (α_n) *be a sequence of measures in* $\mathcal{M}_+(\mathsf{X})$ *and* $\alpha \in \mathcal{M}_+(\mathsf{X})$*, where* $\mathsf{X} \subset \mathbb{R}^d$ *is compact with radius* $R > 0$*. Then,* $\alpha_n \rightharpoonup \alpha \Leftrightarrow \lim_{n \to +\infty} \mathrm{SUOT}(\alpha_n, \alpha) = 0 \Leftrightarrow \lim_{n \to +\infty} \mathrm{USOT}(\alpha_n, \alpha) = 0$.

The metrization of weak* convergence is an important property when comparing measures. For instance, it can be leveraged to justify the well-posedness of approximating an unbalanced Wasserstein gradient flow (Ambrosio et al., 2005) using SUOT, as done in (Bonet et al., 2022b; Candau-Tilh,

2020) for SOT. Unbalanced Wasserstein gradient flows have been a key tool in deep learning theory, e.g. to prove global convergence of 1-hidden layer neural networks (Chizat & Bach, 2018; Rotskoff et al., 2019).

We now specialize some metric and topological properties to sliced partial OT, a particular case of SUOT. Theorem 3.5 shows that our framework encompasses existing approaches and more importantly, helps complement their analysis (Bonneel & Coeurjolly, 2019; Bai et al., 2022).

Theorem 3.5. ***(Properties of Sliced Partial OT)*** *Assume* $\mathrm{C}_1(x,y) = |x-y|$ *and* $D_{\varphi_1} = D_{\varphi_2} = \rho\mathrm{TV}$. *Then,* USOT *satisfies the triangle inequality. Additionally, for any* $(\alpha,\beta) \in \mathcal{M}_+(\mathsf{X})$ *where* $\mathsf{X} \subset \mathbb{R}^d$ *is compact with radius* R, $\mathrm{UOT}(\alpha,\beta) \leq c(\rho,R)\,\mathrm{SUOT}(\alpha,\beta)^{1/(d+1)}$, *and* USOT *and* SUOT *both metrize the weak* convergence.*

We move on to the statistical properties and prove that SUOT offers important statistical benefits, as it lifts the *sample complexity* of UOT from one-dimensional setting to multi-dimensional ones. In what follows, for any $\alpha \in \mathcal{M}_+(\mathbb{R}^d)$, we use $\hat{\alpha}_n$ to denote the empirical approximation of α over $n \geq 1$ i.i.d. samples, *i.e.* $\hat{\alpha}_n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$, $Z_i \sim \alpha$.

Theorem 3.6. ***(Sample complexity)*** *If for* $\mu,\nu \in \mathcal{M}_+(\mathbb{R})$, $\mathbb{E}|\mathrm{UOT}(\mu,\nu) - \mathrm{UOT}(\hat{\mu}_n,\hat{\nu}_n)| \leq \kappa(n)$, *then for* $\alpha,\beta \in \mathcal{M}_+(\mathbb{R}^d)$, $\mathbb{E}|\mathrm{SUOT}(\alpha,\beta) - \mathrm{SUOT}(\hat{\alpha}_n,\hat{\beta}_n)| \leq \kappa(n)$.

If for $\mu,\nu \in \mathcal{M}_+(\mathbb{R})$, $\mathbb{E}|\mathrm{UOT}(\mu,\hat{\mu}_n)| \leq \xi(n)$, *then for* $\alpha,\beta \in \mathcal{M}_+(\mathbb{R}^d)$, $\mathbb{E}|\mathrm{SUOT}(\alpha,\hat{\alpha}_n)| \leq \xi(n)$.

Theorem 3.6 means that SUOT enjoys a *dimension-free* sample complexity, even when comparing multivariate measures: this advantage is recurrent of sliced divergences (Nadjahi et al., 2020) and further motivates their use on high-dimensional settings. The sample complexity rates $\kappa(n)$ or $\xi(n)$ can be deduced from the literature on UOT for univariate measures, for example we refer to (Vacher & Vialard, 2022) for the GHK setting. Establishing the statistical properties of USOT may require extending (Nietert et al., 2022): we leave this question for future work.

We conclude this section by deriving the dual formulations of SUOT, USOT and proving that strong duality holds. We will consider that $\boldsymbol{\sigma}$ is approximated with $\hat{\boldsymbol{\sigma}}_K = \frac{1}{K}\sum_{k=1}^K \delta_{\theta_k}$, $\theta_k \sim \boldsymbol{\sigma}$. This corresponds to the routine case in practice, as practitioners usually resort to a Monte Carlo approximation to estimate the expectation w.r.t. $\boldsymbol{\sigma}$ defining sliced OT.

Theorem 3.7. ***(Strong duality)*** *For* $i \in \{1,2\}$, *let* φ_i *be an entropy function s.t.* $\mathrm{dom}(\varphi_i^*) \cap \mathbb{R}_-$ *is non-empty, and either* $0 \in \mathrm{dom}(\varphi_i)$ *or* $m(\alpha), m(\beta) \in \mathrm{dom}(\varphi_i)$. *Define* $\mathcal{E} \triangleq \{\forall\theta \in \mathrm{supp}(\sigma_K),\, f_\theta \oplus g_\theta \leq \mathrm{C}_1\}$. *Let* $f_{avg} \triangleq \int_{\mathbb{S}^{d-1}} f_\theta \mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta)$, $g_{avg} \triangleq \int_{\mathbb{S}^{d-1}} g_\theta \mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta)$.

Then, SUOT (5) *and* USOT (6) *can be equivalently written for* $\alpha,\beta \in \mathcal{M}_+(\mathbb{R}^d)$ *as,*

$$\begin{aligned}&\mathrm{SUOT}(\alpha,\beta)\\ &= \sup_{(f_\theta),(g_\theta)\in\mathcal{E}} \int_{\mathbb{S}^{d-1}} \Big(\int \varphi_1^\circ\big(f_\theta \circ \theta^\star(x)\big)\mathrm{d}\alpha(x) \\ &\qquad + \int \varphi_2^\circ\big(g_\theta \circ \theta^\star(y)\big)\mathrm{d}\beta(y)\Big)\mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta)\end{aligned} \tag{9}$$

$$\begin{aligned}&\mathrm{USOT}(\alpha,\beta)\\ &= \sup_{(f_\theta),(g_\theta)\in\mathcal{E}} \int \varphi_1^\circ\big(f_{avg} \circ \theta^\star(x)\big)\mathrm{d}\alpha(x) \\ &\qquad + \int \varphi_2^\circ\big(g_{avg} \circ \theta^\star(y)\big)\mathrm{d}\beta(y)\end{aligned} \tag{10}$$

We conjecture that strong duality also holds for $\boldsymbol{\sigma}$ Lebesgue over $\mathbb{S}^{d-1}$, and discuss this aspect in Appendix A. Theorem 3.7 has important pratical implications, since it justifies the Frank-Wolfe-type algorithms that we develop in Section 4 to compute SUOT and USOT in practice.

4. Computing SUOT and USOT with Frank-Wolfe algorithms

We propose two algorithms by leveraging our strong duality result (Theorem 3.7) along with a Frank-Wolfe algorithm (FW, Frank & Wolfe (1956)) introduced in (Séjourné et al., 2022b) to optimize UOT dual (3). Our methods, summarized in Algorithms 1 and 2, can be applied for smooth $\mathrm{D}_{\varphi_1}, \mathrm{D}_{\varphi_2}$: this is satisfied for GHK (where $\mathrm{D}_{\varphi_i} = \rho_i\mathrm{KL}$), but not for sliced partial OT (where $\mathrm{D}_{\varphi_i} = \rho_i\mathrm{TV}$, Bai et al. (2022)). We refer to Appendix B for more technical details on our methodology and its theoretical justification.

FW is an iterative procedure which aims at maximizes a functional $\mathcal{H}$ over a compact convex set $\mathcal{E}$, by maximizing a linear approximation $\nabla\mathcal{H}$: given iterate x^t, FW solves the linear oracle $r^{t+1} \in \arg\max_{r\in\mathcal{E}} \langle\nabla\mathcal{H}(x^t), r\rangle$ and performs a convex update $x^{t+1} = (1-\gamma_{t+1})x^t + \gamma_{t+1}r^{t+1}$, with γ_{t+1} typically chosen as $\gamma_{t+1} = 2/(2+t+1)$. We call this step `FWStep` in our pseudo-code. When applied in (Séjourné et al., 2022b) to compute $\mathrm{UOT}(\alpha,\beta)$ dual (3), `FWStep` updates (f_t, g_t) s.t. $f_t \oplus g_t \leq \mathrm{C}_d$, and the linear oracle is the balanced dual of $\mathrm{OT}(\alpha_t,\beta_t)$ where (α_t,β_t) are normalized versions of (α,β). Updating (α_t,β_t) involves (f_t,g_t) and $\boldsymbol{\rho} = (\rho_1,\rho_2)$: we refer to this routine as $\mathtt{Norm}(\alpha,\beta,f_t,g_t,\boldsymbol{\rho})$ and report the closed-form updates in Appendix B. In other words, computing UOT amounts to solve a sequence of OT problems, which can efficiently be done for univariate measures (Séjourné et al., 2022b).

Analogously to UOT, and by Theorem 3.7, we propose to compute $\mathrm{SUOT}(\alpha,\beta)$ and $\mathrm{USOT}(\alpha,\beta)$ based on their dual forms. FW iterates consists in solving a sequence of sliced OT problems. We derive the updates for the `FWStep`

Algorithm 1 – SUOT

Input: $\alpha, \beta, F, (\theta_k)_{k=1}^K, \boldsymbol{\rho} = (\rho_1, \rho_2)$
Output: $\mathrm{SUOT}(\alpha, \beta), (f_\theta, g_\theta)$

$(f_\theta, g_\theta) \leftarrow (0, 0)$
for $t = 0, 1, \ldots, F-1$, **for** $\theta \in (\theta_k)_{k=1}^K$ **do**
  $(\alpha_\theta, \beta_\theta) \leftarrow \mathtt{Norm}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta, f_\theta, g_\theta, \boldsymbol{\rho})$
  $(r_\theta, s_\theta) \leftarrow \mathtt{SlicedDual}(\alpha_\theta, \beta_\theta)$
  $(f_\theta, g_\theta) \leftarrow \mathtt{FWStep}(f_\theta, g_\theta, r_\theta, s_\theta, \gamma_t)$
end for
Return $\mathrm{SUOT}(\alpha, \beta), (f_\theta, g_\theta)$ as in (9)

Algorithm 2 – USOT

Input: $\alpha, \beta, F, (\theta_k)_{k=1}^K, \boldsymbol{\rho} = (\rho_1, \rho_2)$
Output: $\mathrm{USOT}(\alpha, \beta), (f_{avg}, g_{avg})$

$(f_\theta, g_\theta, f_{avg}, g_{avg}) \leftarrow (0, 0, 0, 0)$
for $t = 0, 1, \ldots, F-1$, **for** $\theta \in (\theta_k)_{k=1}^K$ **do**
  $(\pi_1, \pi_2) \leftarrow \mathtt{Norm}(\alpha, \beta, f_{avg}, g_{avg}, \boldsymbol{\rho})$
  $(r_\theta, s_\theta) \leftarrow \mathtt{SlicedDual}(\theta^\star_\sharp\pi_1, \theta^\star_\sharp\pi_2)$
  $r_{avg}, s_{avg} \leftarrow \mathtt{AvgPot}(r_\theta), \mathtt{AvgPot}(s_\theta)$
  $(f_{avg}, g_{avg}) \leftarrow \mathtt{FWStep}(f_{avg}, g_{avg}, r_{avg}, s_{avg}, \gamma_t)$
end for
Return $\mathrm{USOT}(\alpha, \beta), (f_{avg}, g_{avg})$ as in (10)

tailored for SUOT and USOT in Appendix B, and re-use the aforementioned `Norm` routine. For USOT, we implement an additional routine called **`AvgPot`**$\big((f_\theta)\big)$ to compute $\int f_\theta \mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta)$ given the sliced potentials (f_θ).

A crucial difference is the need of SOT dual potentials (r_θ, s_θ) to call `Norm`. However, past implementations only return the loss $\mathrm{SOT}(\alpha, \beta)$ for e.g. training models (Deshpande et al., 2019; Nguyen et al., 2020). Thus we designed two novel (GPU) implementations in PyTorch (Paszke et al., 2019) which return them. The first one leverages that the gradient of $\mathrm{OT}(\alpha, \beta)$ w.r.t. (α, β) are optimal (f, g), which allows to backpropagate $\mathrm{OT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta)$ w.r.t. (α, β) to obtain (r_θ, s_θ). The second implementation computes them in parallel on GPUs using their closed form, which to the best of our knowledge is a new sliced algorithm. We call $\mathtt{SlicedDual}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta)$ the step returning optimal (r_θ, s_θ) solving $\mathrm{OT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta)$ for all θ. Both routines preserve the $O(N \log N)$ per slice time complexity and can be adapted to any SOT variant. Thus, our FW approach is modular in that one can reuse the SOT literature. We illustrate this by computing USOT between distributions in the hyperbolic Poincaré disk. (Figure 2).

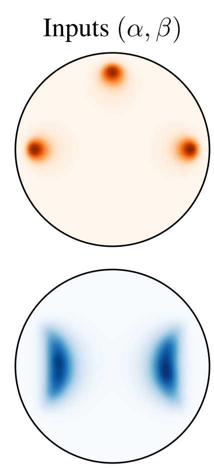
Algorithmic complexity. FW algorithms and its variants have been widely studied theoretically. Computing `SlicedDual` has a complexity $O(KN \log N)$, where N is the number of samples, and K the number of projections of $\hat{\boldsymbol{\sigma}}_K$. The overall complexity of SUOT and USOT is thus $O(FKN \log N)$, where F is the number of FW iterations needed to reach convergence. Our setting falls under the assumptions of (Lacoste-Julien & Jaggi, 2015, Theorem 8), thus ensuring fast convergence of our methods. We plot in Appendix B empirical evidence that a few iterations of FW ($F \leq 20$) suffice to reach numerical precision.

Outputing marginals of SUOT and USOT. The optimal primal marginals of UOT (therefore, SUOT and USOT) are geometric normalizations of inputs (α, β) with discarded outliers. Their computation involves the `Norm` routine, using optimal dual potentials. This is how we compute marginals in Figures 1, 2 and 4: see Appendix B.

Stochastic USOT. In practice, $\hat{\boldsymbol{\sigma}}_K = \frac{1}{K}\sum_i^K \delta_{\theta_i}$ is fixed, and (f_{avg}, g_{avg}) are computed w.r.t. $\hat{\boldsymbol{\sigma}}_K$. However, $\mathbb{E}_{\theta_k \sim \boldsymbol{\sigma}}[\hat{\boldsymbol{\sigma}}_K] = \boldsymbol{\sigma}$. Thus, assuming Theorem 3.7 holds for $\boldsymbol{\sigma}$, we have $\mathbb{E}_{\theta_k \sim \boldsymbol{\sigma}}[f_{avg}(x)] = \int f_\theta(\theta^\star(x))\mathrm{d}\boldsymbol{\sigma}(\theta)$ if we sample a new $\hat{\boldsymbol{\sigma}}_K$ at each FW step. This approach, which we refer to as, *Stochastic* USOT, should output a more accurate estimate of the USOT w.r.t. $\boldsymbol{\sigma}$, but is more expensive: we need to sort projected data w.r.t new projections at each iteration. More importantly, for balanced OT ($\varphi^\circ(x) = x$), USOT $=$ SOT and this idea remains valid for sliced OT. See Section 5 for applications.

5. Experiments

Comparing hyperbolic datasets. We display in Figure 2 the impact of the parameter $\rho = \rho_1 = \rho_2$ on the optimal marginals of USOT. To illustrate the modularity of our FW algorithm, our inputs are synthetic mixtures of Wrapped Normal Distribution on the 2-hyperbolic manifold $\mathbb{H}$ (Nagano et al., 2019), so that the FW oracle is hyperbolic sliced OT (Bonet et al., 2022a). The parameter θ characterizes on $\mathbb{H}$ any geodesic curve passing through the origin, and each sample is projected by taking the shortest path to such geodesics. Once projected on a geodesic curve, we sort data and compute SOT w.r.t. hyperbolic metric $d_{\mathbb{H}}$.



We display the 2-hyperbolic manifold on the Poincaré disc. The measure α (in red) is a mixture of 3 isotropic normal distributions, with a mode at the top of the disc playing the role of an outlier. The measure β is a mixture of two anisotropic normal distributions, whose means are close to two modes of α, but are slightly shifted at the disc's center. We illustrate several take-home messages, stated in Section 3. First, the optimal

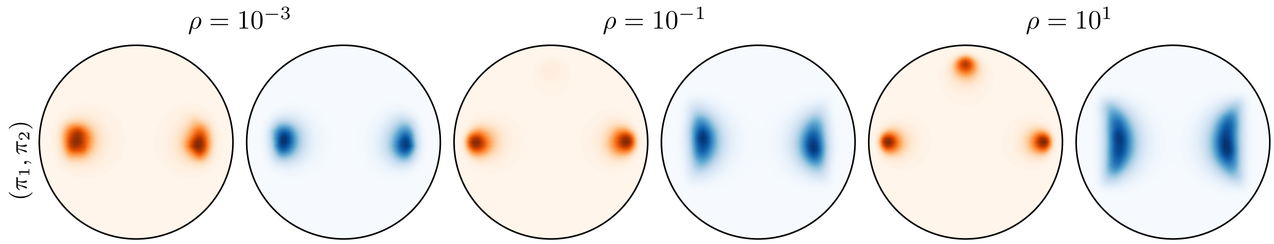


Figure 2: KDE estimation (kernel $e^{-d_{\mathbb{H}}^2/\sigma}$) of optimal (π_1, π_2) of $\text{USOT}(\alpha, \beta)$ when $\text{D}_{\varphi_i} = \rho\text{KL}$.

marginals (π_1, π_2) are renormalisation of (α, β) accounting for their geometry, which are able to remove outliers for properly tuned ρ. When ρ is large, $(\pi_1, \pi_2) \simeq (\alpha, \beta)$ and we retrieve SOT. When ρ is too small, outliers are removed, but we see a shift of the modes, so that modes of (π_1, π_2) are closer to each other, but do not exactly correspond to those of (α, β). Second, note that such plot cannot be made with SUOT, since the optimal marginals depend on the projection θ (see Figure 1). Third, we are indeed able to reuse any variant of SOT existing in the literature.

Document classification. To show the benefits of our proposed losses over SOT, we consider a document classification problem (Kusner et al., 2015). Documents are represented as distributions of words embedded with *word2vec* (Mikolov et al., 2013) in dimension $d = 300$. Let D_k be the k-th document and $x_1^k, \dots, x_{n_k}^k \in \mathbb{R}^d$ be the set of words in D_k. Then, $D_k = \sum_{i=1}^{n_k} w_i^k \delta_{x_i^k}$ where w_i^k is the frequency of x_i^k in D_k normalized s.t. $\sum_{i=1}^{n_k} w_i^k = 1$. Given a loss function L, the document classification task is solved by computing the matrix $\big(\text{L}(D_k, D_\ell)\big)_{k,\ell}$, then using a k-nearest neighbor classifier. Since a word typically appears several times in a document, the measures are not uniform and sliced partial OT (Bonneel & Coeurjolly, 2019; Bai et al., 2022) cannot be used in this setting. The aim of this experiment is to show that by discarding possible outliers using a well chosen parameter ρ, USOT is able to outperform SOT and SUOT on this task. We consider BBCSport dataset (Kusner et al., 2015), Movies reviews (Pang et al., 2002) and the Goodreads dataset (Maharjan et al., 2017) on two tasks (genre and likability). We report in Table 1 the accuracy of SUOT, USOT and the stochastic USOT (SUSOT) compared with SOT, OT and UOT computed with the majorization minimization algorithm (Chapel et al., 2021) or approximated with the Sinkhorn algorithm (Pham et al., 2020). All the benchmark methods are computed using the POT library (Flamary et al., 2021). For sliced methods (SOT, SUOT, USOT and SUSOT), we average over 3 computations of the loss matrix and report the standard deviation in Table 1. The number of neighbors was selected via cross validation. The results in Table 1 are reported for ρ yielding the best accuracy, and we display an ablation of this parameter on the BBCSport dataset in Figure 3. We observe that when ρ is tuned, USOT outperforms SOT, just as UOT outperforms OT. Note that OT and UOT cannot be used in large scale settings (typically large documents) as their complexity scale cubically. We report in Appendix C runtimes on the Goodreads dataset. In particular, computing the OT matrix took 3 times longer than computing the USOT matrix on GPU. Morever, we were unable to run UOT using POT on the Movies and Goodreads datasets in a reasonable amount of time, due to their computational complexity.

Barycenter on geophysical data. OT barycenters have been an important topic of interest (Bonet et al., 2022b; Le et al., 2021). To compute barycenters under the USOT geometry on a fixed grid, we employ a mirror-descent strategy similar to (Cuturi & Doucet, 2014a, Algorithm (1)): see Appendix C. We showcase unbalanced sliced OT barycenter using climate model data. Ensembles of multiple models are commonly employed to reduce biases and evaluate uncertainties in climate projections (*e.g.* (Sanderson et al., 2015; Thao et al., 2022)). The commonly used Multi-Model Mean approach assumes models are centered around true values and averages the ensemble with equal or varying weights. However, spatial averaging may fail in capturing specific characteristics of the physical system at stake. We propose to use USOT barycenter here instead. We consider the ClimateNet dataset (Prabhat et al., 2021), and more specifically the TMQ (precipitable water) indicator. The ClimateNet dataset is a human-expert-labeled curated dataset that captures notably tropical cyclones (TCs). In order to simulate the output of several climate models, we take a specific instant (first date of 2011) and deform the data with the elastic deformation from TorchVision (Paszke et al., 2019), in an area located close to the eastern part of the U.S. We obtain 4 different TCs (Figure 4, first row). As expected, the classical L2 spatial mean (Figure 4, second row) reveals 4 different TCs centers/modes, which is undesirable. Since the total TMQ mass in the considered zone varies between the different models, a direct application of SOT is impossible, or requires a normalization of the mass that has undesired effect as can be seen on the second picture of the second row. Finally, we show the result of the USOT barycenter with $\rho_1 = 1e1$ (related to the data) and $\rho_2 = 1e4$ (related to the barycenter). As a result, the corresponding barycenter has

Table 1: Accuracy on document classification

	BBCSport	Movies	Goodreads genre	Goodreads like
OT	91.64	68.88	52.75	70.60
UOT	96.27	-	-	-
Sinkhorn UOT	93.64	63.8	42.55	66.06
SOT	$89.39_{\pm 0.76}$	$66.95_{\pm 0.45}$	$50.09_{\pm 0.51}$	$65.60_{\pm 0.20}$
SUOT	$90.12_{\pm 0.15}$	$67.84_{\pm 0.37}$	$50.15_{\pm 0.04}$	$66.72_{\pm 0.38}$
USOT	$92.36_{\pm 0.07}$	$69.21_{\pm 0.37}$	$51.87_{\pm 0.56}$	$67.41_{\pm 1.06}$
SUSOT	$92.45_{\pm 0.39}$	$69.53_{\pm 0.53}$	$51.93_{\pm 0.53}$	$67.33_{\pm 0.26}$

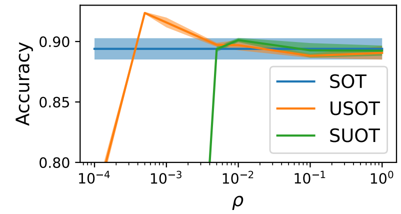


Figure 3: Ablation on BBCSport of the parameter ρ.

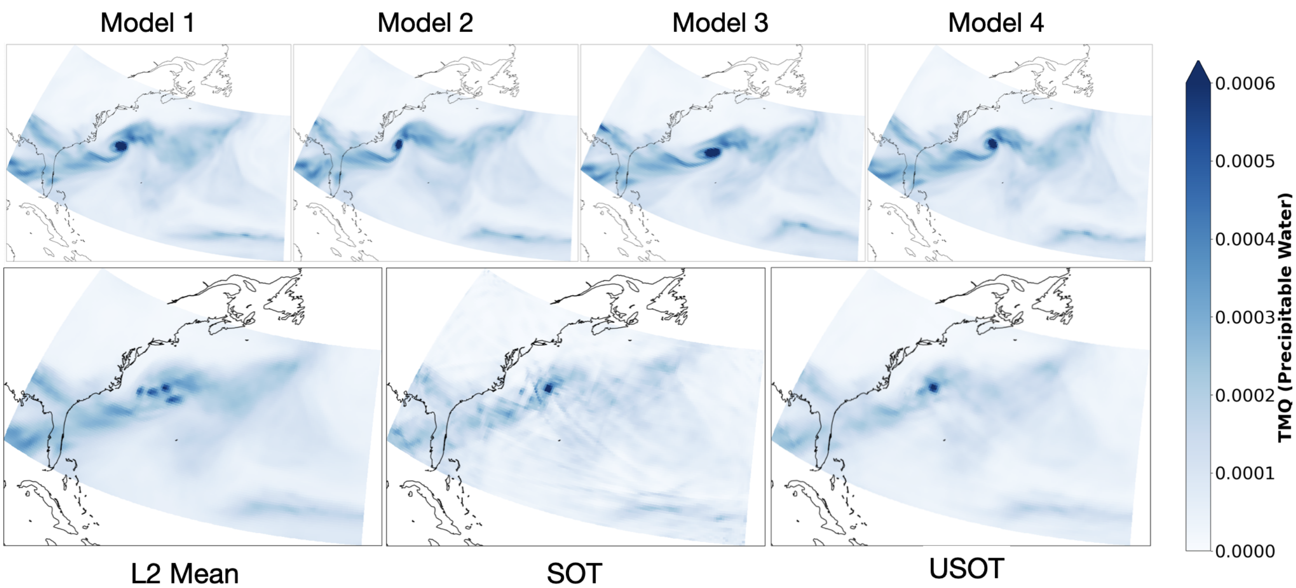


Figure 4: **Barycenter of geophysical data**. (*First row*) Simulated output of 4 different climate models depicting different scenarios for the evolution of a tropical cyclone (*Second row*) Results of different averaging/aggregation strategies.

only one apparent mode which is the expected behaviour. The considered measures have a size of 100×200, and we run the barycenter algorithm for 500 iterations (with $K = 64$ projections), which takes 3 minutes on a commodity GPU. UOT barycenters for this size of problems are untractable, and to the best of our knowledge, this is the first time such large scale unbalanced OT barycenters can be computed. This experiment encourages an in-depth analysis of the relevance of this aggregation strategy for climate modeling and related problems.

6. Conclusion and Discussion

We proposed two losses merging unbalanced and sliced OT, with theoretical guarantees and an efficient Frank-Wolfe algorithm which allows to reuse any sliced OT variant. We highlighted experimentally the performance improvement over SOT, and described novel applications of unbalanced OT barycenters of positive measures, with a new case study on geophysical data. These novel results and algorithms pave the way to numerous new applications of sliced variants of OT: we believe our contributions will motivate practitioners to further explore their use in ML applications, without having to pre-process probability measures.

An immediate drawback is the induced additional computational cost w.r.t. SOT. While our empirical results show that SUOT and USOT significantly outperform SOT, and though the complexity is sub-quadratic in the number of samples, our FW approach uses SOT as a subroutine, rendering it necessarily more expensive. Another practical burden comes from the introduction of hyperparameters (ρ_1, ρ_2), which requires cross-validation when possible. A future direction would be to derive efficient strategies to tune (ρ_1, ρ_2), maybe w.r.t. the applicative context, and complement possible interpretations of ρ as a "threshold" for the geometric information encoded by C_1, C_d. On the other hand, while OT between univariate measures defines a reproducing kernel and sliced OT takes advantage of this property (Kolouri et al., 2016; Carriere et al., 2017), some of our experiments suggest this no longer holds for UOT (therefore, for SUOT, USOT). This leaves as an open direction the design of OT-based kernel methods between arbitrary positive measures.

References

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Bai, Y., Schmitzer, B., Thorpe, M., and Kolouri, S. Sliced optimal partial transport. *arXiv preprint arXiv:2212.08049*, 2022.

Bayraktar, E. and Guo, G. Strong equivalence between metrics of Wasserstein type. *Electronic Communications in Probability*, 26(none):1 – 13, 2021. doi: 10.1214/21-ECP383.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Bogachev, V. I. and Ruas, M. A. S. *Measure theory*, volume 1. Springer, 2007.

Bonet, C., Chapel, L., Drumetz, L., and Courty, N. Hyperbolic sliced-wasserstein via geodesic and horospherical projections. *arXiv preprint arXiv:2211.10066*, 2022a.

Bonet, C., Courty, N., Septier, F., and Drumetz, L. Efficient gradient flows in sliced-wasserstein space. *Transactions on Machine Learning Research*, 2022b.

Bonet, C., Berg, P., Courty, N., Septier, F., Drumetz, L., and Pham, M.-T. Spherical sliced-wasserstein. In *The Eleventh International Conference on Learning Representations*, 2023a.

Bonet, C., Malézieux, B., Rakotomamonjy, A., Drumetz, L., Moreau, T., Kowalski, M., and Courty, N. Sliced-wasserstein on symmetric positive definite matrices for m/eeg signals. In *Proceedings of the 40th International Conference on Machine Learning*, 2023b.

Bonneel, N. and Coeurjolly, D. Spot: sliced partial optimal transport. *ACM Transactions on Graphics (TOG)*, 38(4):1–13, 2019.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

Bonnotte, N. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Université Paris Sud-Paris XI; Scuola normale superiore (Pise, Italie), 2013.

Candau-Tilh, J. Wasserstein and sliced-wasserstein distances. Master's thesis, Université Pierre et Marie Curie, 2020.

Carriere, M., Cuturi, M., and Oudot, S. Sliced wasserstein kernel for persistence diagrams. In *International conference on machine learning*, pp. 664–673. PMLR, 2017.

Chapel, L., Flamary, R., Wu, H., Févotte, C., and Gasso, G. Unbalanced optimal transport through non-negative penalized linear regression. *Advances in Neural Information Processing Systems*, 34:23270–23282, 2021.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. An interpolating distance between optimal transport and fisher–rao metrics. *Foundations of Computational Mathematics*, 18:1–44, 2018a.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018b.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Cuturi, M. and Doucet, A. Fast computation of wasserstein barycenters. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 685–693, Bejing, China, 22–24 Jun 2014a. PMLR.

Cuturi, M. and Doucet, A. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pp. 685–693. PMLR, 2014b.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

Demetci, P., Santorella, R., Chakravarthy, M., Sandstede, B., and Singh, R. Scotv2: Single-cell multiomic alignment with disproportionate cell-type representation. *Journal of Computational Biology*, 29(11):1213–1228, 2022.

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

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

Fatras, K., Zine, Y., Flamary, R., Gribonval, R., and Courty, N. Learning with minibatch wasserstein : asymptotic and gradient properties. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2131–2141, Online, 26–28 Aug 2020. PMLR. URL http://proceedings.mlr.press/v108/fatras20a.html.

Fatras, K., Sejourne, T., Flamary, R., and Courty, N. Unbalanced minibatch optimal transport; applications to domain adaptation. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3186–3197. PMLR, 18–24 Jul 2021. URL http://proceedings.mlr.press/v139/fatras21a.html.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., et al. Pot: Python optimal transport. *The Journal of Machine Learning Research*, 22(1):3571–3578, 2021.

Frank, M. and Wolfe, P. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2): 95–110, 1956.

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

Goldfeld, Z. and Greenewald, K. Sliced mutual information: A scalable measure of statistical dependence. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 17567–17578. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/92c4661685bf6681f6a33b78ef729658-Paper.pdf.

Kolouri, S., Zou, Y., and Rohde, G. K. Sliced wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5258–5267, 2016.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. Generalized sliced wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

Kondratyev, S., Monsaingeon, L., and Vorotnikov, D. A fitness-driven cross-diffusion system from population dynamics as a gradient flow. *Journal of Differential Equations*, 261(5):2784–2808, 2016.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. From word embeddings to document distances. In *International conference on machine learning*, pp. 957–966. PMLR, 2015.

Lacoste-Julien, S. and Jaggi, M. On the global linear convergence of frank-wolfe optimization variants. *Advances in neural information processing systems*, 28, 2015.

Le, K., Nguyen, H., Nguyen, Q. M., Pham, T., Bui, H., and Ho, N. On robust optimal transport: Computational complexity and barycenter computation. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 21947–21959, 2021.

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

Maharjan, S., Arevalo, J., Montes, M., González, F. A., and Solorio, T. A multi-task approach to predict likability of books. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 1217–1227, 2017.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

Nadjahi, K., Durmus, A., Chizat, L., Kolouri, S., Shahrampour, S., and Simsekli, U. Statistical and topological properties of sliced probability divergences. *Advances in Neural Information Processing Systems*, 33:20802–20812, 2020.

Nagano, Y., Yamaguchi, S., Fujita, Y., and Koyama, M. A wrapped normal distribution on hyperbolic space for gradient-based learning. In *International Conference on Machine Learning*, pp. 4693–4702. PMLR, 2019.

Nguyen, K., Ho, N., Pham, T., and Bui, H. Distributional sliced-wasserstein and applications to generative modeling. *arXiv preprint arXiv:2002.07367*, 2020.

Nietert, S., Goldfeld, Z., and Cummings, R. Outlier-robust optimal transport: Duality, structure, and statistical analysis. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. PMLR, 2022.

Ohana, R., Nadjahi, K., Rakotomamonjy, A., and Ralaivola, L. Shedding a pac-bayesian light on adaptive sliced-wasserstein distances. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Pang, B., Lee, L., and Vaithyanathan, S. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pp. 79–86, 2002.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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

Pham, K., Le, K., Ho, N., Pham, T., and Bui, H. On unbalanced optimal transport: An analysis of Sinkhorn algorithm. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7673–7682. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/pham20a.html.

Piccoli, B. and Rossi, F. Generalized wasserstein distance and its application to transport equations with source. *Archive for Rational Mechanics and Analysis*, 211:335–358, 2014.

Prabhat, Kashinath, K., Mudigonda, M., Kim, S., Kapp-Schwoerer, L., Graubner, A., Karaismailoglu, E., von Kleist, L., Kurth, T., Greiner, A., Mahesh, A., Yang, K., Lewis, C., Chen, J., Lou, A., Chandran, S., Toms, B., Chapman, W., Dagon, K., Shields, C. A., O'Brien, T., Wehner, M., and Collins, W. Climatenet: an expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. *Geoscientific Model Development*, 14(1):107–124, 2021. doi: 10.5194/gmd-14-107-2021. URL https://gmd.copernicus.org/articles/14/107/2021/.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pp. 435–446. Springer, 2012.

Radon, J. 1.1 über die bestimmung von funktionen durch ihre integralwerte längs gewisser mannigfaltigkeiten. *Classic papers in modern diagnostic radiology*, 5(21):124, 2005.

Rotskoff, G., Jelassi, S., Bruna, J., and Vanden-Eijnden, E. Global convergence of neuron birth-death dynamics. *arXiv preprint arXiv:1902.01843*, 2019.

Sanderson, B. M., Knutti, R., and Caldwell, P. A representative democracy to reduce interdependency in a multimodel ensemble. *Journal of Climate*, 28(13):5171–5194, 2015.

Santambrogio, F. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Sato, R., Yamada, M., and Kashima, H. Fast unbalanced optimal transport on a tree. *Advances in neural information processing systems*, 33:19039–19051, 2020.

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

Séjourné, T., Feydy, J., Vialard, F.-X., Trouvé, A., and Peyré, G. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958*, 2019.

Séjourné, T., Peyré, G., and Vialard, F.-X. Unbalanced optimal transport, from theory to numerics. *arXiv preprint arXiv:2211.08775*, 2022a.

Séjourné, T., Vialard, F.-X., and Peyré, G. Faster unbalanced optimal transport: Translation invariant sinkhorn and 1-d frank-wolfe. In *International Conference on Artificial Intelligence and Statistics*, pp. 4995–5021. PMLR, 2022b.

Simons, S. *Minimax and monotonicity*. Springer, 2006.

Thao, S., Garvik, M., Mariethoz, G., and Vrac, M. Combining global climate models using graph cuts. *Climate Dynamics*, 59:2345–2361, 2022. URL https://hal.science/hal-03620538.

Vacher, A. and Vialard, F.-X. Stability of semi-dual unbalanced optimal transport: fast statistical rates and convergent algorithm. 2022.

Xi, J. and Niles-Weed, J. Distributional convergence of the sliced wasserstein process. *arXiv preprint arXiv:2206.00156*, 2022.

A. Postponed proofs for Section 3

A.1. Existence of minimizers

We provide the formal statement and detailed proof on the existence of a solution for both SUOT and USOT, as mentioned in Section 3.

Proposition A.1. ***(Existence of minimizers)*** *Assume that* C_1 *is lower-semicontinuous and that either (i)* $\varphi'_{1,\infty} = \varphi'_{2,\infty} = +\infty$*, or (ii)* C_1 *has compact sublevels on* $\mathbb{R}\times\mathbb{R}$ *and* $\varphi'_{1,\infty} + \varphi'_{2,\infty} + \inf \mathrm{C}_1 > 0$*. Then the solution of* $\mathrm{SUOT}(\alpha,\beta)$ *and* $\mathrm{USOT}(\alpha,\beta)$ *exist, i.e. the infimum in* (5) *and* (6) *is attained. More precisely, there exists* (π_1,π_2) *which attains the infimum for* $\mathrm{USOT}(\alpha,\beta)$ *(see Equation* (6)*). Concerning* $\mathrm{SUOT}(\alpha,\beta)$*, there exists for any* $\theta\in\mathrm{supp}(\boldsymbol{\sigma})$ *a plan* π_θ *attaining the infimum in* $\mathrm{UOT}(\theta^\star_\sharp\alpha,\theta^\star_\sharp\beta)$ *(see Equation* (2)*).*

Proof. We leverage (Liero et al., 2018, Theorem 3.3) to prove this proposition. In the setting of SUOT, if such assumptions (i) or (ii) are satisfied for (α,β), then they also hold for $(\theta^\star_\sharp\alpha,\theta^\star_\sharp\beta)$ for any $\theta\in\mathbb{S}^{d-1}$. Hence, $\mathrm{UOT}(\theta^\star_\sharp\alpha,\theta^\star_\sharp\beta)$ admits a solution π^θ.

Concerning USOT, note that one necessarily has $m(\pi_1)=m(\pi_2)$, otherwise $\mathrm{SOT}(\pi_1,\pi_2)=+\infty$. From (Liero et al., 2018, Equation (3.10)), that for any admissible (π_1,π_2,π), one has

$$\mathrm{USOT}(\alpha,\beta) \geq m(\pi)\inf\mathrm{C}_1 + m(\alpha)\varphi_1(\tfrac{m(\pi)}{m(\alpha)}) + m(\beta)\varphi_2(\tfrac{m(\pi)}{m(\beta)}).$$

In both settings the above bounds implies coercivity of the functional of USOT w.r.t. the masses of the measures (π_1,π_2,π). Thus there exists $M>0$ such that $m(\pi_1)=m(\pi_2)=m(\pi)<M$, otherwise $\mathrm{USOT}(\alpha,\beta)=+\infty$. By the Banach-Alaoglu theorem, the set of bounded measures (π_1,π_2) is compact, and the set of plans π with such marginals is also compact because $\mathbb{R}^d$ is Polish and C_1 is lower-semicontinuous (Santambrogio, 2015, Theorem 1.7). Because the functional of USOT is lower-semicontinuous in (π_1,π_2,π) and we can restrict optimization over a compact set, we have existence of minimizers for USOT by standard proofs of calculus of variations. □

A.2. Metric properties: Proof of Proposition 3.2

Proof of Proposition 3.2. **Metric properties of SUOT.** Symmetry and non-negativity are immediate. Assume $\mathrm{SUOT}(\alpha,\beta)=0$. Since $\boldsymbol{\sigma}$ is the uniform distribution on $\mathbb{S}^{d-1}$, then for any $\theta\in\mathbb{S}^{d-1}$, $\mathrm{UOT}(\theta^\star_\sharp\alpha,\theta^\star_\sharp\beta)=0$, and since UOT is assumed to be definite, then $\theta^\star_\sharp\alpha=\theta^\star_\sharp\beta$. By (Bogachev & Ruas, 2007, Proposition 3.8.6), this implies that α and β have the same Fourier transform. By injectivity of the Fourier transform, we conclude that $\alpha=\beta$, hence SUOT is definite. The triangle inequality results from applying the Minkowski inequality then the triangle inequality for $\mathrm{UOT}^{1/p}$ for $p\in[1,+\infty)$: for any $\alpha,\beta,\gamma\in\mathcal{M}_+(\mathbb{R}^d)$,

$$\begin{aligned}
&\mathrm{SUOT}^{1/p}(\alpha,\beta)\\
&=\left(\int_{\mathbb{S}^{d-1}}\mathrm{UOT}(\theta^\star_\sharp\alpha,\theta^\star_\sharp\beta)\mathrm{d}\boldsymbol{\sigma}(\theta)\right)^{1/p}\\
&\leq\left(\int_{\mathbb{S}^{d-1}}\left[\mathrm{UOT}^{1/p}(\theta^\star_\sharp\alpha,\theta^\star_\sharp\gamma)+\mathrm{UOT}^{1/p}(\theta^\star_\sharp\gamma,\theta^\star_\sharp\beta)\right]^p\mathrm{d}\boldsymbol{\sigma}(\theta)\right)^{1/p}\\
&\leq\left(\int_{\mathbb{S}^{d-1}}\left[\mathrm{UOT}^{1/p}(\theta^\star_\sharp\alpha,\theta^\star_\sharp\gamma)\right]^p\mathrm{d}\boldsymbol{\sigma}(\theta)\right)^{1/p}+\left(\int_{\mathbb{S}^{d-1}}\left[\mathrm{UOT}^{1/p}(\theta^\star_\sharp\gamma,\theta^\star_\sharp\beta)\right]^p\mathrm{d}\boldsymbol{\sigma}(\theta)\right)^{1/p}\\
&=\mathrm{SUOT}^{1/p}(\alpha,\gamma)+\mathrm{SUOT}^{1/p}(\gamma,\beta).
\end{aligned}$$

Metric properties of USOT. Let $(\alpha,\beta)\in\mathcal{M}_+(\mathbb{R}^d)$. Non-negativity is immediate, as USOT is defined as a program minimizing a sum of positive terms. SOT is symmetric, thus when $\varphi_1=\varphi_2$, we obtain symmetry of the functional w.r.t. (α,β). Assume D_φ is definite, *i.e.* $\mathrm{D}_\varphi(\alpha|\beta)=0$ implies $\alpha=\beta$. Assume now that $\mathrm{USOT}(\alpha,\beta)=0$, and denote by (π_1,π_2) the optimal marginals attaining the infimum in (6). $\mathrm{USOT}(\alpha,\beta)=0$ implies that $\mathrm{SOT}(\pi_1,\pi_2)=0$, $\mathrm{D}_\varphi(\pi_1|\alpha)=0$ and $\mathrm{D}_\varphi(\pi_2|\beta)=0$. These three terms are definite, which yields $\alpha=\pi_1=\pi_2=\beta$, hence the definiteness of USOT.

□

A.3. Comparison of SUOT, USOT, SOT, and proof of Theorem 3.3

In this section, we establish several bounds to compare SUOT, USOT and SOT on the space of compactly-supported measures. We provide the detailed derivations and auxiliary lemmas needed for the proofs. Note that Theorem 3.3 is a direct consequence from Theorems A.2 to A.4.

Theorem A.2. *Let X be a compact subset of $\mathbb{R}^d$ with radius R and consider $\alpha, \beta \in \mathcal{M}_+(\mathsf{X})$. Then,* $\mathrm{SUOT}(\alpha, \beta) \leq \mathrm{USOT}(\alpha, \beta)$.

Proof. To show that $\mathrm{SUOT}(\alpha, \beta) \leq \mathrm{USOT}(\alpha, \beta)$, we use a sub-optimality argument. Let π be the solution $\mathrm{USOT}(\alpha, \beta)$ and denote by (π_1, π_2) the marginals of π. For any $\theta \in \mathbb{S}^{d-1}$, denote by π_θ the solution of $\mathrm{OT}(\theta^\star_\sharp \pi_1, \theta^\star_\sharp \pi_2)$. By definition of USOT, the marginals of π_θ are given by $(\theta^\star_\sharp \pi_1, \theta^\star_\sharp \pi_2)$. Since the sequence $(\pi_\theta)_\theta$ is suboptimal for the problem $\mathrm{SUOT}(\alpha, \beta)$, one has

$$\mathrm{SUOT}(\alpha, \beta) \leq \int_{\mathbb{S}^{d-1}} \Big\{ \int \mathrm{C}_1 \mathrm{d}\pi_\theta + \mathrm{D}_{\varphi_1}(\theta^\star_\sharp \pi_1 | \theta^\star_\sharp \alpha) + \mathrm{D}_{\varphi_2}(\theta^\star_\sharp \pi_2 | \theta^\star_\sharp \beta) \Big\} \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{11}$$

$$\leq \int_{\mathbb{S}^{d-1}} \int \mathrm{C}_1 \mathrm{d}\pi_\theta \mathrm{d}\boldsymbol{\sigma}(\theta) + \mathrm{D}_{\varphi_1}(\pi_1 | \alpha) + \mathrm{D}_{\varphi_2}(\pi_2 | \beta) \tag{12}$$

$$= \mathrm{USOT}(\alpha, \beta), \tag{13}$$

where the second inequality results from Lemma A.5, and the last equality follows from the definition of $\mathrm{USOT}(\alpha, \beta)$. □

Theorem A.3. *Let X be a compact subset of $\mathbb{R}^d$ with radius R and consider $\alpha, \beta \in \mathcal{M}_+(\mathsf{X})$. Additionally, let $p \in [1, +\infty)$ and assume $\mathrm{C}_1(x, y) = |x - y|^p$ for $(x, y) \in \mathbb{R} \times \mathbb{R}$ and $\mathrm{C}_d(x, y) = \|x - y\|^p$ for $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$. Then,* $\mathrm{USOT}(\alpha, \beta) \leq \mathrm{UOT}(\alpha, \beta)$.

Proof. By (Bonnotte, 2013, Proposition 5.1.3), $\mathrm{SOT}(\mu, \nu) \leq K\ \mathrm{OT}(\mu, \nu)$ with $K \leq 1$. Let π be the solution of $\mathrm{UOT}(\alpha, \beta)$ with marginals (π_1, π_2). These marginals are sub-optimal for $\mathrm{USOT}(\alpha, \beta)$, we have

$$\mathrm{USOT}(\alpha, \beta) \leq \mathrm{SOT}(\pi_1, \pi_2) + \mathrm{D}_{\varphi_1}(\pi_1 | \alpha) + \mathrm{D}_{\varphi_2}(\pi_2 | \beta)\,, \tag{14}$$

$$\leq \mathrm{OT}(\pi_1, \pi_2) + \mathrm{D}_{\varphi_1}(\pi_1 | \alpha) + \mathrm{D}_{\varphi_2}(\pi_2 | \beta)\,, \tag{15}$$

$$= \mathrm{UOT}(\alpha, \beta)\,, \tag{16}$$

where the last equality is obtained because π is optimal in $\mathrm{UOT}(\alpha, \beta)$. □

Theorem A.4. *Let X be a compact subset of $\mathbb{R}^d$ with radius R and consider $\alpha, \beta \in \mathcal{M}_+(\mathsf{X})$. Additionally, let $p \in [1, +\infty)$ and assume $\mathrm{C}_1(x, y) = |x - y|^p$ for $(x, y) \in \mathbb{R}$ and $\mathrm{C}_d(x, y) = \|x - y\|^p$ for $(x, y) \in \mathbb{R}^d$. Let $\rho > 0$ and assume $D_{\varphi_1} = D_{\varphi_2} = \rho\mathrm{KL}$. Then,* $\mathrm{UOT}(\alpha, \beta) \leq c\,\mathrm{SUOT}(\alpha, \beta)^{1/(d+1)}$*, where $c = c(m(\alpha), m(\beta), \rho, R)$ is a non-decreasing function of $m(\alpha)$ and $m(\beta)$.*

Proof. We adapt the proof of (Bonnotte, 2013, Lemma 5.1.4), which establishes a bound between OT and SOT. The first step consists in bounding from above the distance between two regularized measures.

Let $\psi : \mathbb{R}^d \to \mathbb{R}_+$ be a smooth and radial function verifying $\mathrm{supp}(\psi) \subseteq B_d(\mathbf{0}, 1)$ and $\int_{\mathbb{R}^d} \psi(x) \mathrm{dLeb}(x) = 1$. Let $\psi_\lambda(x) = \lambda^{-d} \psi(x/\lambda) / \mathcal{A}(\mathbb{S}^{d-1})$ where $\mathcal{A}(\mathbb{S}^{d-1})$ is the surface area of $\mathbb{S}^{d-1}$, *i.e.* $\mathcal{A}(\mathbb{S}^{d-1}) = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ with Γ the gamma function. For any function f defined on $\mathbb{R}^s$ ($s \geq 1$), denote by $\mathcal{F}[f]$ the Fourier transform of f defined for $x \in \mathbb{R}^s$ as $\mathcal{F}[f](x) = \int_{\mathbb{R}^s} f(w) e^{-\mathrm{i}\langle w, x \rangle} \mathrm{d}w$. Let $\alpha_\lambda = \alpha * \varphi_\lambda$ and $\beta_\lambda = \beta * \varphi_\lambda$ where $*$ is the convolution operator. Let (f, g) such that $f \oplus g \leq \mathrm{C}_d$. By using the isometry properties of the Fourier transform and the definition of ψ_λ, then representing the variables with polar coordinates, we have

$$\int_{\mathbb{R}^d} \varphi^\circ(f(x)) \mathrm{d}\alpha_\lambda(x) = \int_{\mathbb{R}^d} \mathcal{F}[\varphi^\circ \circ f](w) \mathcal{F}[\alpha](w) \mathcal{F}[\psi](\lambda w) \mathrm{d}w \tag{17}$$

$$= \int_{\mathbb{S}^{d-1}} \int_0^{+\infty} \mathcal{F}[\varphi^\circ \circ f](r\theta) \mathcal{F}[\alpha](r\theta) \mathcal{F}[\psi](\lambda r) r^{d-1} \mathrm{d}r \mathrm{d}\boldsymbol{\sigma}(\theta)\,. \tag{18}$$

Since $\varphi^\circ \circ f$ is a real-valued function, $\mathcal{F}[\varphi^\circ \circ f]$ is an even function, then

$$\int_{\mathbb{R}^d} \varphi^\circ(f(x)) \mathrm{d}\alpha_\lambda(x) \tag{19}$$

$$= \frac{1}{2} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \mathcal{F}[\varphi^\circ \circ f](r\theta) \mathcal{F}[\alpha](r\theta) \mathcal{F}[\psi](\lambda r) |r|^{d-1} \mathrm{d}r \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{20}$$

$$= \frac{1}{2} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \mathcal{F}[\varphi^\circ \circ f](r\theta) \mathcal{F}[\theta^\star_\sharp \alpha](r) \mathcal{F}[\psi](\lambda r) |r|^{d-1} \mathrm{d}r \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{21}$$

$$= \frac{1}{2} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \mathcal{F}[\varphi^\circ \circ f](r\theta) \left(\int_{-R}^{R} e^{-\mathrm{i}ru} \mathrm{d}\theta^\star_\sharp \alpha(u) \right) \mathcal{F}[\psi](\lambda r) \, |r|^{d-1} \, \mathrm{d}r \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{22}$$

$$= \frac{1}{2} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \left(\int_{\mathbb{R}^d} \int_{-R}^{R} \varphi^\circ(f(x)) e^{-\mathrm{i}r(u + \langle \theta, x \rangle)} \mathrm{d}\theta^\star_\sharp \alpha(u) \right) \mathcal{F}[\psi](\lambda r) \, |r|^{d-1} \, \mathrm{d}x \mathrm{d}r \mathrm{d}\boldsymbol{\sigma}(\theta) \, . \tag{23}$$

Equation (21) follows from the property of push-forward measures, (22) results from the definition of the Fourier transform and $u \in [-R, R]$, and (23) results from the definition of the Fourier transform and Fubini's theorem. By making a change of variables (x becomes $x - u\theta$), we obtain

$$\int_{\mathbb{R}^d} \varphi^\circ(f(x)) \mathrm{d}\alpha_\lambda(x) \tag{24}$$

$$= \frac{1}{2} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \int_{\mathbb{R}^d} \int_{-R}^{R} \varphi^\circ(f(x - u\theta)) e^{-\mathrm{i}r\langle \theta, x \rangle} \mathrm{d}\theta^\star_\sharp \alpha(u) \mathcal{F}[\psi](\lambda r) \, |r|^{d-1} \, \mathrm{d}x \mathrm{d}r \mathrm{d}\boldsymbol{\sigma}(\theta) \tag{25}$$

$$= \frac{1}{2} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \int_{B_d(\mathbf{0}, 2R+\lambda)} \int_{-R}^{R} \varphi^\circ(f(x - u\theta)) e^{-\mathrm{i}r\langle \theta, x \rangle} \mathrm{d}\theta^\star_\sharp \alpha(u) \mathcal{F}[\psi](\lambda r) \, |r|^{d-1} \, \mathrm{d}x \mathrm{d}r \mathrm{d}\boldsymbol{\sigma}(\theta) \, , \tag{26}$$

where (26) follows from the assumption that $\mathrm{supp}(\alpha) \subseteq B_d(\mathbf{0}, R)$. Indeed, this implies that $\mathrm{supp}(\alpha_\lambda) \subseteq B_d(\mathbf{0}, R + \lambda)$, thus the domain of $x \mapsto \varphi^\circ \circ f(x - u\theta)$ is contained in $B_d(\mathbf{0}, 2R + \lambda)$.

Similarly, one can show that

$$\int_{\mathbb{R}^d} \varphi^\circ(g(y)) \mathrm{d}\beta_\lambda(y) \tag{27}$$

$$= \frac{1}{2} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \int_{B_d(\mathbf{0}, 2R+\lambda)} \int_{-R}^{R} \varphi^\circ(g(y - u\theta)) e^{-\mathrm{i}r\langle \theta, y \rangle} \mathrm{d}\theta^\star_\sharp \beta(u) \mathcal{F}[\psi](\lambda r) \, |r|^{d-1} \, \mathrm{d}y \mathrm{d}r \mathrm{d}\boldsymbol{\sigma}(\theta) \, . \tag{28}$$

By (26) and (28), and applying Fubini's theorem, we obtain

$$\int_{\mathbb{R}^d} \varphi^\circ(f(x)) \mathrm{d}\alpha_\lambda(x) + \int_{\mathbb{R}^d} \varphi^\circ(g(y)) \mathrm{d}\beta_\lambda(y) \tag{29}$$

$$\leq \frac{1}{2} \int_{\mathbb{R}} \int_{B_d(\mathbf{0}, 2R+\lambda)} \int_{\mathbb{S}^{d-1}} \Big\{ \int_{-R}^{R} \varphi^\circ(f(x - u\theta)) \mathrm{d}\theta^\star_\sharp \alpha(u) + \int_{-R}^{R} \varphi^\circ(g(x - u\theta)) \mathrm{d}\theta^\star_\sharp \beta(u) \Big\} e^{-\mathrm{i}r\langle \theta, x \rangle} \mathcal{F}[\psi](\lambda r) \, |r|^{d-1} \, \mathrm{d}\boldsymbol{\sigma}(\theta) \mathrm{d}x \mathrm{d}r \tag{30}$$

$$\leq c_1 (2R + \lambda)^d \int_{\mathbb{S}^{d-1}} \mathrm{UOT}(\theta^\star_\sharp \alpha, \theta^\star_\sharp \beta) \mathrm{d}\boldsymbol{\sigma}(\theta) \int_{\mathbb{R}} \lambda^{-d} \big| \mathcal{F}[\psi](r) \, |r|^{d-1} \, \big| \mathrm{d}r \tag{31}$$

$$\leq c_2 (2R + \lambda)^d \lambda^{-d} \mathrm{SUOT}(\alpha, \beta) \tag{32}$$

where $c_1 > 0$ is independent from α and β, and $c_2 = c_1 \int_{\mathbb{R}} \left| \mathcal{F}[\psi](r) \right| |r|^{d-1} \, \mathrm{d}r$. Equation (32) is obtained by taking the supremum of (30) over the set of potentials $(\tilde{f}, \tilde{g})$ such that for $u \in [-R, R]$, $\exists (x, \theta) \in B_d(\mathbf{0}, 2R + \lambda) \times \mathbb{S}^{d-1}$, $\tilde{f}(u) = f(x - u\theta)$, $\tilde{g}(u) = g(x - u\theta)$, which is included in the set of potentials (f', g') s.t. $f' : \mathbb{R} \to \mathbb{R}$, $g' : \mathbb{R} \to \mathbb{R}$ and $f' \oplus g' \leq \mathrm{C}_1$.

We deduce from the dual formulation of UOT (3) and (32) that,

$$\mathrm{UOT}(\alpha_\lambda, \beta_\lambda) \leq c_2 (2R+\lambda)^d \lambda^{-d} \mathrm{SUOT}(\alpha, \beta)\,. \tag{33}$$

The last step of the proof consists in relating $\mathrm{UOT}(\alpha_\lambda, \beta_\lambda)$ with $\mathrm{UOT}(\alpha, \beta)$. For any (f, g) such that $f \oplus g \leq \mathrm{C}_d$, we have

$$\int_{\mathbb{R}^d} \varphi^\circ(f(x)) \mathrm{d}\alpha(x) + \int_{\mathbb{R}^d} \varphi^\circ(g(y)) \mathrm{d}\beta(y) - \mathrm{UOT}(\alpha_\lambda, \beta_\lambda) \tag{34}$$

$$\leq \int_{\mathbb{R}^d} \varphi^\circ(f(x)) \mathrm{d}\alpha(x) + \int_{\mathbb{R}^d} \varphi^\circ(g(x)) \mathrm{d}\beta(x) - \int_{\mathbb{R}^d} \varphi^\circ(f(x)) \mathrm{d}\alpha_\lambda(x) - \int_{\mathbb{R}^d} \varphi^\circ(g(y)) \mathrm{d}\beta_\lambda(y) \tag{35}$$

$$\leq \int_{\mathbb{R}^d} \{\varphi^\circ(f(x)) - \psi_\lambda * \varphi^\circ(f(x))\} \mathrm{d}\alpha(x) + \int_{\mathbb{R}^d} \{\varphi^\circ(g(y)) - \psi_\lambda * \varphi^\circ(g(y))\} \mathrm{d}\beta(y)\,. \tag{36}$$

For $x \in \mathbb{R}^d$,

$$\varphi^\circ(f(x)) - \psi_\lambda * \varphi^\circ(f(x)) = \frac{\lambda^{-d}}{\mathcal{A}(\mathbb{S}^{d-1})} \int_{\mathbb{R}^d} \left(\varphi^\circ(f(x)) - \varphi^\circ(f(y))\right) \psi\left(\frac{x-y}{\lambda}\right) \mathrm{d}y \tag{37}$$

$$\leq \frac{\lambda^{-d}}{\mathcal{A}(\mathbb{S}^{d-1})} \int_{\mathbb{R}^d} |\varphi^\circ(f(x)) - \varphi^\circ(f(y))| \psi\left(\frac{x-y}{\lambda}\right) \mathrm{d}y\,, \tag{38}$$

Since $\mathrm{D}_\varphi = \rho \mathrm{KL}$, then for $z \in \mathbb{R}$, $\varphi^\circ(z) = \rho(1 - e^{-z/\rho})$, so for $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\varphi^\circ(f(x)) - \varphi^\circ(f(y)) = \rho(e^{-f(y)/\rho} - e^{-f(x)/\rho}) \tag{39}$$

By Lemma A.8, the potentials (f, g) are bounded by constants depending on $m(\alpha), m(\beta)$, thus we can bound (39) as follows.

$$|\varphi^\circ(f(x)) - \varphi^\circ(f(y))| \leq \rho e^{-\lambda^\star/\rho} \left(1 - e^{-R/\rho}\right)\,, \tag{40}$$

with $\lambda^\star \in [-R + \frac{\rho}{2} \log \frac{m(\alpha)}{m(\beta)}, \frac{R}{2} + \frac{\rho}{2} \log \frac{m(\alpha)}{m(\beta)}]$. We thus derive the following upper-bound on (38).

$$\varphi^\circ(f(x)) - \psi_\lambda * \varphi^\circ(f(x)) \leq \frac{\lambda^{-d}}{\mathcal{A}(\mathbb{S}^{d-1})} \rho e^{-\lambda^\star/\rho} \left(1 - e^{-R/\rho}\right) \int_{\mathbb{R}^d} \psi\left(\frac{x-y}{\lambda}\right) \mathrm{d}y \tag{41}$$

$$\leq \frac{\lambda^{-d+1}}{\mathcal{A}(\mathbb{S}^{d-1})} \rho e^{-\lambda^\star/\rho} \left(1 - e^{-R/\rho}\right) \int_{\mathbb{R}^d} \frac{1}{\lambda} \psi\left(\frac{x-y}{\lambda}\right) \mathrm{d}y \tag{42}$$

$$\leq \frac{\lambda^{-d+1}}{\mathcal{A}(\mathbb{S}^{d-1})} \sqrt{\frac{m(\beta)}{m(\alpha)}} \rho e^{R/\rho} \left(1 - e^{-R/\rho}\right) \int_{\mathbb{R}^d} \frac{1}{\lambda} \psi\left(\frac{x-y}{\lambda}\right) \mathrm{d}y \tag{43}$$

By doing the change of variables $z = (y - x)/\lambda$ and using the fact that ψ is a radial function and $\int_{\mathbb{R}^d} \psi(z) \mathrm{dLeb}(z) = 1$, we obtain $\int_{\mathbb{R}^d} \frac{1}{\lambda} \psi\left(\frac{x-y}{\lambda}\right) \mathrm{d}y = 1$. Therefore,

$$\varphi^\circ(f(x)) - \psi_\lambda * \varphi^\circ(f(x)) \leq \frac{\lambda^{-d+1}}{\mathcal{A}(\mathbb{S}^{d-1})} \sqrt{\frac{m(\beta)}{m(\alpha)}} \rho e^{R/\rho} \left(1 - e^{-R/\rho}\right) \tag{44}$$

$$\leq \frac{\lambda}{\mathcal{A}(\mathbb{S}^{d-1})} \sqrt{\frac{m(\beta)}{m(\alpha)}} \rho e^{R/\rho} \left(1 - e^{-R/\rho}\right)\,. \tag{45}$$

Similarly, using the bounds on g in Lemma A.8, one can show that

$$|\varphi^\circ(g(x)) - \varphi^\circ(g(y))| \leq \rho e^{\lambda^\star/\rho} \left(e^{R/\rho} - e^{-R/\rho}\right) \leq \rho \sqrt{\frac{m(\alpha)}{m(\beta)}} e^{R/2\rho} \left(e^{R/\rho} - e^{-R/\rho}\right)\,, \tag{46}$$

therefore,

$$\varphi^{\circ}(g(x)) - \psi_\lambda * \varphi^{\circ}(g(x)) \leq \frac{\lambda}{\mathcal{A}(\mathbb{S}^{d-1})}\sqrt{\frac{m(\alpha)}{m(\beta)}}\rho e^{R/2\rho}\left(e^{R/\rho} - e^{-R/\rho}\right). \tag{47}$$

We conclude that,

$$\int_{\mathbb{R}^d}\varphi^{\circ}(f(x))\mathrm{d}\alpha(x) + \int_{\mathbb{R}^d}\varphi^{\circ}(g(y))\mathrm{d}\beta(y) - \mathrm{UOT}(\alpha_\lambda,\beta_\lambda) \tag{48}$$

$$\leq \frac{\lambda\rho}{\mathcal{A}(\mathbb{S}^{d-1})}\left\{m(\alpha)e^{-\lambda^\star/\rho}\left(1-e^{-R/\rho}\right) + m(\beta)e^{\lambda^\star/\rho}\left(e^{R/\rho}-e^{-R/\rho}\right)\right\} \tag{49}$$

$$\leq \frac{\lambda\rho}{\mathcal{A}(\mathbb{S}^{d-1})}\sqrt{m(\alpha)m(\beta)}\left\{e^{R/\rho}\left(1-e^{-R/\rho}\right) + e^{R/2\rho}\left(e^{R/\rho}-e^{-R/\rho}\right)\right\} \tag{50}$$

Taking the supremum on both sides over (f,g) such that $f\oplus g \leq \mathrm{C}_d$ yields,

$$\mathrm{UOT}(\alpha,\beta) - \mathrm{UOT}(\alpha_\lambda,\beta_\lambda) \tag{51}$$

$$\leq \frac{\lambda\rho}{\mathcal{A}(\mathbb{S}^{d-1})}\left\{m(\alpha)e^{-\lambda^\star/\rho}\left(1-e^{-R/\rho}\right) + m(\beta)e^{\lambda^\star/\rho}\left(e^{R/\rho}-e^{-R/\rho}\right)\right\} \tag{52}$$

$$\leq \frac{\lambda\rho}{\mathcal{A}(\mathbb{S}^{d-1})}\sqrt{m(\alpha)m(\beta)}\left\{e^{R/\rho}\left(1-e^{-R/\rho}\right) + e^{R/2\rho}\left(e^{R/\rho}-e^{-R/\rho}\right)\right\}. \tag{53}$$

Finally, by combining (33) with the above inequality, we obtain

$$\mathrm{UOT}(\alpha,\beta) \tag{54}$$

$$\leq \frac{\lambda\rho}{\mathcal{A}(\mathbb{S}^{d-1})}\sqrt{m(\alpha)m(\beta)}\left\{e^{R/\rho}\left(1-e^{-R/\rho}\right) + e^{R/2\rho}\left(e^{R/\rho}-e^{-R/\rho}\right)\right\} \tag{55}$$

$$+ c_2(2R+\lambda)^d\lambda^{-d}\mathrm{SUOT}(\alpha,\beta) \tag{56}$$

$$\leq c\lambda\big(1+(2R+\lambda)^d\lambda^{-(d+1)}\mathrm{SUOT}(\alpha,\beta)\big), \tag{57}$$

where c is a constant satisfying $c \geq c_2$ and

$$c \geq \rho\sqrt{m(\alpha)m(\beta)}\left\{e^{R/\rho}\left(1-e^{-R/\rho}\right) + e^{R/2\rho}\left(e^{R/\rho}-e^{-R/\rho}\right)\right\}/\mathcal{A}(\mathbb{S}^{d-1}). \tag{58}$$

We conclude the proof by plugging $\lambda = R^{d/(d+1)}\mathrm{SUOT}(\alpha,\beta)^{1/(d+1)}$ in (57) and using the fact that $\mathrm{SUOT}(\alpha,\beta)$ is bounded from above: $\mathrm{SUOT}(\alpha,\beta) \leq \rho(m(\alpha)+m(\beta))$ since on the one hand, π is suboptimal in (3) thus $\mathrm{UOT}(\alpha,\beta) \leq \rho(m(\alpha)+m(\beta))$, and on the other hand, $m(\alpha) = m(\theta^\star_\sharp\alpha)$ for any $\theta \in \mathbb{S}^{d-1}$. □

Lemma A.5. *For any $\theta \in \mathbb{S}^{d-1}$ and $\alpha,\beta \in \mathcal{M}_+(\mathbb{R}^d)$, $D_\varphi(\theta^\star_\sharp\alpha|\theta^\star_\sharp\beta) \leq D_\varphi(\alpha|\beta)$.*

Proof. For $\alpha,\beta \in \mathcal{M}_+(\mathbb{R}^s)$ with $s \geq 1$, the dual characterization of φ-divergences reads (Liero et al., 2018, Theorem 2.7)

$$\mathrm{D}_\varphi(\alpha|\beta) = \sup_{f\in\mathcal{E}(\mathbb{R}^s)}\int_{\mathbb{R}^s}\varphi^{\circ}(f(x))\mathrm{d}\beta(x) - \int_{\mathbb{R}^s}f(x)\mathrm{d}\alpha(x),$$

where $\mathcal{E}(\mathbb{R}^s)$ denotes the space of lower semi-continuous functions from $\mathbb{R}^s$ to $\mathbb{R}\cup\{+\infty\}$. Therefore, for any $\theta \in \mathbb{S}^{d-1}$ and $\alpha,\beta \in \mathcal{M}_+(\mathbb{R}^d)$,

$$\mathrm{D}_\varphi(\theta^\star_\sharp\alpha|\theta^\star_\sharp\beta) = \sup_{f\in\mathcal{E}(\mathbb{R})}\int_{\mathbb{R}}\varphi^{\circ}(f(t))\mathrm{d}(\theta^\star_\sharp\beta)(t) - \int_{\mathbb{R}}f(t)\mathrm{d}(\theta^\star_\sharp\alpha)(t) \tag{59}$$

$$= \sup_{g:\mathbb{R}^d\to\mathbb{R}\ s.t.\ \exists f\in\mathcal{E}(\mathbb{R}),\, g=f\circ\theta^\star}\int_{\mathbb{R}^d}\varphi^{\circ}(g(x))\mathrm{d}\beta(x) - \int_{\mathbb{R}^d}g(x)\mathrm{d}\alpha(x) \tag{60}$$

where (60) results from the definition of push-forward measures. We conclude the proof by observing that the supremum in (60) is taken over a subset of $\mathcal{E}(\mathbb{R}^d)$. □

Lemma A.6. *(Santambrogio, 2015, Proposition 1.11) Let $p \in [1, +\infty)$ and assume $\mathrm{C}_d(x, y) = \|x - y\|^p$. Let α, β with compact support, such that $\mathrm{C}_d(x, y) \leq R^p$ for $(x, y) \in \mathrm{supp}(\alpha) \times \mathrm{supp}(\beta)$. Then without loss of generality the dual potentials (f, g) of $\mathrm{UOT}(\alpha, \beta)$ satisfy $f(x) \in [0, R]$ and $g(y) \in [-R, R]$.*

Lemma A.7. *(Séjourné et al., 2022b, Proposition 2) Define the translation-invariant dual formulation*

$$\mathrm{UOT}(\alpha, \beta) = \sup_{f \oplus g \leq \mathrm{C}_d} \sup_{\lambda \in \mathbb{R}} \int \varphi_1^\circ(f + \lambda) \mathrm{d}\alpha + \int \varphi_2^\circ(g - \lambda) \mathrm{d}\beta. \tag{61}$$

Let $\rho > 0$ and assume $D_{\varphi_1} = D_{\varphi_2} = \rho \mathrm{KL}$. Take optimal potentials (f, g) in (61). Then optimal potentials in (3) are given by $(f + \lambda^\star(f, g), g - \lambda^\star(f, g))$, where the optimal translation $\lambda^\star$ reads

$$\lambda^\star(f, g) \triangleq \frac{1}{2}\left[\mathrm{S}_\rho^\beta(g) - \mathrm{S}_\rho^\alpha(f)\right], \quad \mathrm{S}_\rho^\alpha(f) \triangleq -\rho \log \int e^{-f/\rho} \mathrm{d}\alpha,$$

and we call $\mathrm{S}_\rho^\alpha(f)$ the soft-minimum of f. When $m(\alpha) = 1$ and $m \leq f(x) \leq M$, then $m \leq \mathrm{S}_\rho^\alpha(f) \leq M$.

Lemma A.8. *Assume (α, β) have compact support such that, for $(x, y) \in \mathrm{supp}(\alpha) \times \mathrm{supp}(\beta)$, $C(x, y) \leq R$. Then, without loss of generality, one can restrict the optimization of the dual formulation (3) of $\mathrm{UOT}(\alpha, \beta)$ over the set of potentials satisfying for $(x, y) \in \mathrm{supp}(\alpha) \times \mathrm{supp}(\beta)$,*

$$f(x) \in [\lambda^\star, \lambda^\star + R], \quad g(y) \in [-\lambda^\star - R, -\lambda^\star + R],$$

where $\lambda^\star \in [-R + \frac{\rho}{2} \log \frac{m(\alpha)}{m(\beta)}, \frac{R}{2} + \frac{\rho}{2} \log \frac{m(\alpha)}{m(\beta)}]$. In particular, one has

$$f(x) \in [-R + \tfrac{\rho}{2} \log \tfrac{m(\alpha)}{m(\beta)}, \tfrac{3R}{2} + \tfrac{\rho}{2} \log \tfrac{m(\alpha)}{m(\beta)}], \quad g(y) \in [-\tfrac{3R}{2} - \tfrac{\rho}{2} \log \tfrac{m(\alpha)}{m(\beta)}, 2R - \tfrac{\rho}{2} \log \tfrac{m(\alpha)}{m(\beta)}]$$

Proof. Consider the translation-invariant dual formulation (61): if (f, g) are optimal, then for any $\lambda \in \mathbb{R}$, $(f + \lambda, g - \lambda)$ are also optimal. We leverage the structure of the dual constraint $f \oplus g \leq \mathrm{C}_d$ with Lemma A.6. Since for $(x, y) \in \mathrm{supp}(\alpha) \times \mathrm{supp}(\beta)$, $\mathrm{C}_d(x, y) \leq R$, then without loss of generality, $f(x) \in [0, R]$ and $g(y) \in [-R, R]$. The potentials (f, g) are optimal for the translation-invariant dual energy, and we need a bound for the original dual functional (3). To this end, we leverage Lemma A.7 to compute the optimal translation, such that $(f, g) = (f + \lambda^\star(f, g), g - \lambda^\star(f, g))$. Let $\bar{\alpha} = \alpha / m(\alpha)$ and $\bar{\beta} = \beta / m(\beta)$ be the normalized probability measures. The translation can be written as,

$$\lambda^\star(f, g) = \frac{1}{2}\left[\mathrm{S}_\rho^{\bar{\beta}}(g) - \mathrm{S}_\rho^{\bar{\alpha}}(f)\right] + \frac{\rho}{2} \log \frac{m(\alpha)}{m(\beta)}, \tag{62}$$

where the functional S_ρ^α is defined in Lemma A.7. Since $\bar{\alpha}$ and $\bar{\beta}$ are probability measures, then by (Genevay et al., 2019, Proposition 1), $f(x) \in [0, R]$ and $g(x) \in [-R, R]$ respectively imply $\mathrm{S}_\rho^{\bar{\alpha}}(f) \in [0, R]$ and $\mathrm{S}_\rho^{\bar{\beta}}(g) \in [-R, R]$. Combining these bounds on $\mathrm{S}_\rho^{\bar{\alpha}}(f)$, $\mathrm{S}_\rho^{\bar{\beta}}(g)$ with the expression of $\lambda^\star(f, g)$ (62) yields the desired bounds on the optimal potentials (f, g) of the dual formulation (3). □

A.4. Metrizing weak* convergence: Proof of Theorem 3.4

Proof. Let (α_n) be a sequence of measures in $\mathcal{M}_+(\mathsf{X})$ and $\alpha \in \mathcal{M}_+(\mathsf{X})$, where $\mathsf{X} \subset \mathbb{R}^d$ is compact with radius $R > 0$. First, we assume that $\alpha_n \rightharpoonup \alpha$. Then, by (Liero et al., 2018, Theorem 2.25), under our assumptions, $\alpha_n \rightharpoonup \alpha$ is equivalent to $\lim_{n \to +\infty} \mathrm{UOT}(\alpha_n, \alpha) = 0$. This implies that $\lim_{n \to +\infty} \mathrm{SUOT}(\alpha_n, \alpha) = 0$ and $\lim_{n \to +\infty} \mathrm{USOT}(\alpha_n, \alpha) = 0$, since by Theorem 3.3 and non-negativity of SUOT (Proposition 3.2),

$$0 \leq \mathrm{SUOT}(\alpha_n, \alpha) \leq \mathrm{USOT}(\alpha_n, \alpha) \leq \mathrm{UOT}(\alpha_n, \alpha).$$

Conversely, assume either that $\lim_{n \to +\infty} \mathrm{SUOT}(\alpha_n, \alpha) = 0$ or $\lim_{n \to +\infty} \mathrm{USOT}(\alpha_n, \alpha) = 0$. First assume there exists $M > 0$ such that for large enough $n \in \mathbb{N}^*, m(\alpha_n) \leq M$, then by Theorem 3.3, there exists $c > 0$ such that $\mathrm{UOT}(\alpha_n, \alpha) \leq c\big(\mathrm{SUOT}(\alpha_n, \alpha)\big)^{1/(d+1)}$. Since c is doesn't depend on the masses $(m(\alpha_n), m(\alpha))$, it does not depend on n. By Theorem 3.3, it yields metric equivalence between SUOT, USOT and UOT, thus $\lim_{n \to +\infty} \mathrm{UOT}(\alpha_n, \alpha) = 0$. By (Liero et al., 2018, Theorem 2.25), we eventually obtain $\alpha_n \rightharpoonup \alpha$, which is the desired result.

The remaining step thus consists in proving that the sequence of masses $(m(\alpha_n))_{n\in\mathbb{N}^*}$ is indeed uniformly bounded by $M > 0$ for large enough n. Note that for any $(\alpha, \beta) \in \mathcal{M}_+(\mathbb{R}^d)$, one has $\mathrm{UOT}(\alpha, \beta) \geq \rho(\sqrt{m(\alpha)} - \sqrt{m(\beta)})^2$. Indeed one has $\mathrm{UOT}(\alpha, \beta) \geq \mathcal{D}(\lambda, -\lambda)$, where $\mathcal{D}$ denotes the dual functional (3) and $\lambda = \frac{\rho}{2} \log \frac{m(\alpha)}{m(\beta)}$. Note that the pair $(\lambda, -\lambda)$ are feasible dual potentials for the constraint $f \oplus g \leq \mathrm{C}_d$, because the cost C_d is positive in our setting. The property of push-forwards measures means that for any $\theta \in \mathbb{S}^{d-1}$, one has $m(\theta^\star_\sharp \alpha) = m(\alpha)$. Therefore, we obtain the following bounds for n large enough.

$$\begin{aligned}\mathrm{USOT}(\alpha_n, \alpha) \geq \mathrm{SUOT}(\alpha_n, \alpha) &\geq \int_{\mathbb{S}^{d-1}} \rho\Big(\sqrt{m(\theta^\star_\sharp \alpha_n)} - \sqrt{m(\theta^\star_\sharp \alpha)}\Big)^2 \mathrm{d}\boldsymbol{\sigma}(\theta), \\ &= \rho(\sqrt{m(\alpha_n)} - \sqrt{m(\alpha)})^2.\end{aligned}$$

Hence, $\lim_{n\to+\infty} \mathrm{SUOT}(\alpha_n, \alpha) = 0$ or $\lim_{n\to+\infty} \mathrm{USOT}(\alpha_n, \alpha) = 0$ implies $\lim_{n\to+\infty} m(\alpha_n) = m(\alpha)$. In other terms the mass of sequence converges and is thus uniformly bounded for large enough n. Since we proved that $m(\alpha_n) < M$ and $m(\alpha)$ is finite, it ends the proof. □

A.5. Application to sliced partial OT: Proof of Theorem 3.5

The proof of Theorem 3.5 relies on a formulation for SUOT and USOT when $\mathrm{D}_{\varphi_1} = \mathrm{D}_{\varphi_2} = \rho\mathrm{TV}$, which we prove below. Equation (63) is proved in (Piccoli & Rossi, 2014), and can then be applied to SUOT. We include it for completeness. Equation (64) is our contribution and is specific to USOT.

Lemma A.9. *Let $\rho > 0$ and assume $D_{\varphi_1} = D_{\varphi_2} = \rho\mathrm{TV}$ and $\mathrm{C}_d(x, y) = ||x - y||$. Then, for any $(\alpha, \beta) \in \mathcal{M}_+(\mathbb{R}^d)$,*

$$\mathrm{UOT}(\alpha, \beta) = \sup_{f\in\mathcal{E}} \int f(x)\mathrm{d}(\alpha - \beta)(x), \tag{63}$$

where

$$\mathcal{E} = \{f : \mathbb{R}^d \to \mathbb{R},\ ||f||_{Lip} \leq 1,\ ||f||_\infty \leq \rho\},$$

and $||f||_\infty \triangleq \sup_{x\in\mathbb{R}^d} |f(x)|$ and $||f||_{Lip} \triangleq \sup_{(x,y)\in\mathbb{R}^d} \frac{|f(x)-f(y)|}{\mathrm{C}_d(x,y)}$.

Furthermore, for $\mathrm{C}_1(x, y) = |x - y|$ and an empirical approximation $\hat{\boldsymbol{\sigma}}_N = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i}$ of $\boldsymbol{\sigma}$, one has

$$\mathrm{USOT}(\alpha, \beta) = \sup_{(f_\theta)\in\mathcal{E}} \int_{\mathbb{R}^d} \left(\int_{\mathbb{S}^{d-1}} f_\theta(\theta^\star(x))\mathrm{d}\hat{\boldsymbol{\sigma}}_N(\theta)\right)\mathrm{d}(\alpha - \beta)(x)\,, \tag{64}$$

where

$$\mathcal{E} = \{\forall\theta \in \mathrm{supp}(\hat{\boldsymbol{\sigma}}_N),\ f_\theta : \mathbb{R} \to \mathbb{R},\ ||f_\theta||_{Lip} \leq 1,\ ||\int_{\mathbb{S}^{d-1}} f_\theta \circ \theta^\star \mathrm{d}\hat{\boldsymbol{\sigma}}_N(\theta)||_\infty \leq \rho\},$$

and the Lipschitz norm here is defined w.r.t. C_1 as $||f||_{Lip} \triangleq \sup_{(x,y)\in\mathbb{R}^d} \frac{|f(x)-f(y)|}{\mathrm{C}_1(x,y)}$

Proof. We start with the formulation of Equation 3 and Theorem 3.7. For USOT one has

$$\begin{aligned}\mathrm{USOT}(\alpha, \beta) = \sup_{f_\theta(\cdot)\oplus g_\theta(\cdot)\leq \mathrm{C}_1} &\int \varphi_1^\circ\Big(\int_{\mathbb{S}^{d-1}} f_\theta(\theta^\star(x))\mathrm{d}\sigma_N(\theta)\Big)\mathrm{d}\alpha(x) \\ &+ \int \varphi_2^\circ\Big(\int_{\mathbb{S}^{d-1}} g_\theta(\theta^\star(y))\mathrm{d}\sigma_N(\theta)\Big)\mathrm{d}\beta(y)\,.\end{aligned}$$

When $\mathrm{D}_\varphi = \rho\mathrm{TV}$, the function φ° reads $\varphi^\circ(x) = x$ for $x \in [-\rho, \rho]$, $\varphi^\circ(x) = \rho$ when $x \geq \rho$, and $\varphi^\circ(x) = -\infty$ otherwise. Noting $f_{avg}(x) = \int_{\mathbb{S}^{d-1}} f_\theta(\theta^\star(x))\mathrm{d}\sigma_N(\theta)$ and $g_{avg}(x) = \int_{\mathbb{S}^{d-1}} g_\theta(\theta^\star(x))\mathrm{d}\sigma_N(\theta)$. This formula on φ° imposes $f_{avg}(x) \geq -\rho$ and $g_{avg}(x) \geq -\rho$. Furthermore, since we perform a supremum w.r.t. (f_{avg}, g_{avg}) where φ° attains a plateau, then without loss of generality, we can impose the constraint $f_{avg}(x) \leq \rho$ and $g_{avg}(x) \geq \rho$, as it will have no impact on the optimal dual functional value. Thus we have that $||f_{avg}||_\infty \leq \rho$ and $||g_{avg}||_\infty \leq \rho$. To obtain the Lipschitz property, we use the constraint that $f_\theta(\cdot) \oplus g_\theta(\cdot) \leq \mathrm{C}_1$ for any $\theta \in \mathrm{supp}(\sigma_N)$, as well as (Santambrogio, 2015, Proposition 3.1). Thus by using c-transform for the cost $\mathrm{C}_1(x,y) = |x-y|$, we can take w.l.o.g $f_\theta(\cdot) = -g_\theta(\cdot)$ with $f_\theta(\cdot)$ a 1-Lipschitz function. Thus w.l.o.g we can perform the supremum over $(f_\theta)_\theta \in \mathcal{E}$, and rephrase the functional as desired, since we have that $\varphi^\circ(f_{avg}) = f_{avg}$.

The proof for UOT is exactly the same, except that our inputs are (f, g) instead of (f_θ, g_θ). □

We can now prove Theorem 3.5.

Proof of Theorem 3.5. First we prove that in that setting USOT is a metric. Reusing Lemma A.9, we have that for any measures (α, β, γ)

$$\begin{aligned}
\mathrm{USOT}(\alpha,\gamma) &= \sup_{(f_\theta)_\theta \in \mathcal{E}} \int \left(\int_{\mathbb{S}^{d-1}} f_\theta(\theta^\star(x)) \mathrm{d}\sigma_N \right) \mathrm{d}(\alpha - \gamma)(x) \\
&= \sup_{(f_\theta)_\theta \in \mathcal{E}} \int \left(\int_{\mathbb{S}^{d-1}} f_\theta(\theta^\star(x)) \mathrm{d}\sigma_N \right) \mathrm{d}(\alpha - \beta + \beta - \gamma)(x) \\
&\leq \sup_{(f_\theta)_\theta \in \mathcal{E}} \int \left(\int_{\mathbb{S}^{d-1}} f_\theta(\theta^\star(x)) \mathrm{d}\sigma_N \right) \mathrm{d}(\alpha - \beta)(x) \\
&\quad + \sup_{(f_\theta)_\theta \in \mathcal{E}} \int \left(\int_{\mathbb{S}^{d-1}} f_\theta(\theta^\star(x)) \mathrm{d}\sigma_N \right) \mathrm{d}(\beta - \gamma)(x) \\
&= \mathrm{USOT}(\alpha,\beta) + \mathrm{USOT}(\beta,\gamma).
\end{aligned}$$

Note that reusing Lemma A.9, we have that SUOT is a sliced integral probability metric over the space of bounded and Lipschitz functions. More precisely, we satisfy the assumptions of (Nadjahi et al., 2020, Theorem 3), so that one has $\mathrm{UOT}(\alpha,\beta) \leq c(\rho,R)(\mathrm{SUOT}(\alpha,\beta))^{1/(d+1)}$.

To prove that USOT and SUOT metrize the weak* convergence, the proof is very similar to that of Theorem 3.4 detailed above. Assuming that $\alpha_n \rightharpoonup \alpha$ implies $\mathrm{SUOT}(\alpha_n, \alpha) \to 0$ and $\mathrm{USOT}(\alpha_n, \alpha) \to 0$ is already proved in Appendix A.4. To prove the converse, the proof is also the same, i.e. we use the property that SUOT, USOT and UOT are equivalent metrics, which holds as we assumed that supports of (α, β) are compact in a ball of radius R. Note that since the bound $\mathrm{UOT}(\alpha,\beta) \leq c(\rho,R)(\mathrm{SUOT}(\alpha,\beta))^{1/(d+1)}$ holds independently of the measure's masses, we do not need to uniformly bound $m(\alpha_n)$, compared to the KL setting of Theorem 3.4. □

A.6. Sample complexity: Proof of Theorem 3.6

Theorem 3.6 is obtained by adapting (Nadjahi et al., 2020, Theorems 4 and 5). We provide the detailed derivations below.

Proof of Theorem 3.6. Let α, β in $\mathcal{M}_+(\mathbb{R}^d)$ with respective empirical approximations $\hat{\alpha}_n, \hat{\beta}_n$ over n samples. By using the definition of SUOT, the triangle inequality and the assumed sample complexity of UOT for univariate measures, we show that

$$\begin{aligned}
&\mathbb{E}\left|\mathrm{SUOT}(\alpha,\beta) - \mathrm{SUOT}(\hat{\alpha}_n, \hat{\beta}_n)\right| && (65)\\
&= \mathbb{E}\left|\int_{\mathbb{S}^{d-1}} \left\{\mathrm{UOT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta) - \mathrm{UOT}(\theta^\star_\sharp\hat{\alpha}_n, \theta^\star_\sharp\hat{\beta}_n)\right\} \mathrm{d}\boldsymbol{\sigma}(\theta)\right| && (66)\\
&\leq \mathbb{E}\left\{\int_{\mathbb{S}^{d-1}} \left|\mathrm{UOT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta) - \mathrm{UOT}(\theta^\star_\sharp\hat{\alpha}_n, \theta^\star_\sharp\hat{\beta}_n)\right| \mathrm{d}\boldsymbol{\sigma}(\theta)\right\} && (67)\\
&\leq \int_{\mathbb{S}^{d-1}} \mathbb{E}\left|\mathrm{UOT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta) - \mathrm{UOT}(\theta^\star_\sharp\hat{\alpha}_n, \theta^\star_\sharp\hat{\beta}_n)\right| \mathrm{d}\boldsymbol{\sigma}(\theta) && (68)\\
&\leq \int_{\mathbb{S}^{d-1}} \kappa(n) \mathrm{d}\boldsymbol{\sigma}(\theta) = \kappa(n)\,, && (69)
\end{aligned}$$

which completes the proof for the first setting.

Next, let $\alpha \in \mathcal{M}_+(\mathbb{R}^d)$ with corresponding empirical approximation $\hat{\alpha}_n$. Then, using the definition of SUOT, the triangle inequality (w.r.t. integral) and the assumed convergence rate in UOT,

$$\mathbb{E}\,|\mathrm{SUOT}(\hat{\alpha}_n, \alpha)| \tag{70}$$
$$= \mathbb{E}\left|\int_{\mathbb{S}^{d-1}} \mathrm{UOT}(\theta^\star_\sharp \hat{\alpha}_n, \theta^\star_\sharp \alpha)\mathrm{d}\boldsymbol{\sigma}(\theta)\right| \leq \mathbb{E}\left\{\int_{\mathbb{S}^{d-1}} \left|\mathrm{UOT}(\theta^\star_\sharp \hat{\alpha}_n, \theta^\star_\sharp \alpha)\right| \mathrm{d}\boldsymbol{\sigma}(\theta)\right\} \tag{71}$$
$$\leq \int_{\mathbb{S}^{d-1}} \mathbb{E}\left|\mathrm{UOT}(\theta^\star_\sharp \hat{\alpha}_n, \theta^\star_\sharp \alpha)\right| \mathrm{d}\boldsymbol{\sigma}(\theta) \leq \int_{\mathbb{S}^{d-1}} \xi(n)\mathrm{d}\boldsymbol{\sigma}(\theta) = \xi(n)\,. \tag{72}$$

Additionally, if we assume that $\mathrm{UOT}^{1/p}$ satisfies non-negativity, symmetry and the triangle inequality on $\mathcal{M}_+(\mathbb{R}) \times \mathcal{M}_+(\mathbb{R})$, then by Proposition 3.2, $\mathrm{SUOT}^{1/p}$ verifies these three metric properties on $\mathcal{M}_+(\mathbb{R}^d) \times \mathcal{M}_+(\mathbb{R}^d)$, and we can derive its sample complexity as follows. For any α, β in $\mathcal{M}_+(\mathbb{R}^d)$ with respective empirical approximations $\hat{\alpha}_n, \hat{\beta}_n$, applying the triangle inequality yields for $p \in [1, +\infty)$,

$$\left|\mathrm{UOT}^{1/p}(\alpha, \beta) - \mathrm{UOT}^{1/p}(\hat{\alpha}_n, \hat{\beta}_n)\right| \leq \mathrm{UOT}^{1/p}(\hat{\alpha}_n, \alpha) + \mathrm{UOT}^{1/p}(\hat{\beta}_n, \beta)\,. \tag{73}$$

Taking the expectation of (73) with respect to $\hat{\alpha}_n, \hat{\beta}_n$ gives,

$$\mathbb{E}\left|\mathrm{SUOT}^{1/p}(\alpha, \beta) - \mathrm{SUOT}^{1/p}(\hat{\alpha}_n, \hat{\beta}_n)\right| \leq \mathbb{E}|\mathrm{SUOT}^{1/p}(\hat{\alpha}_n, \alpha)| + \mathbb{E}|\mathrm{SUOT}^{1/p}(\hat{\beta}_n, \beta)| \tag{74}$$
$$\leq \{\mathbb{E}\,|\mathrm{SUOT}(\hat{\alpha}_n, \alpha)|\}^{1/p} + \{\mathbb{E}|\mathrm{SUOT}(\hat{\beta}_n, \beta)|\}^{1/p} \tag{75}$$
$$\leq \xi(n)^{1/p} + \xi(n)^{1/p} = 2\xi(n)^{1/p}\,, \tag{76}$$

where (75) is immediate if $p = 1$, and results from applying Hölder's inequality on $\mathbb{S}^{d-1}$ if $p > 1$, and (76) follows from (72). □

A.7. Strong duality: Proof of Theorem 3.7

Proof of Theorem 3.7. Note that the result for SUOT is already proved in Lemma A.12. Thus we focus on the proof of duality for USOT. We start from the definition of USOT, reformulate it to apply the strong duality result of Proposition A.10 and obtain our reformulation. We first have that

$$\begin{aligned}
\mathrm{USOT}(\alpha, \beta) = &\inf_{(\pi_1,\pi_2)\in\mathcal{M}_+(\mathbb{R}^d)^2} \left\{\mathrm{SOT}(\pi_1, \pi_2) + \mathrm{D}_{\varphi_1}(\pi_1|\alpha) + \mathrm{D}_{\varphi_2}(\pi_2|\beta)\right\},\\
= &\inf_{(\pi_1,\pi_2)\in\mathcal{M}_+(\mathbb{R}^d)^2} \Bigg\{\int_{\mathbb{S}^{d-1}} \left[\sup_{f_\theta \oplus g_\theta \leq \mathrm{C}_1} \int f_\theta \mathrm{d}(\theta^\star_\sharp \pi_1) + \int g_\theta \mathrm{d}(\theta^\star_\sharp \pi_2)\right] \mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta)\\
&+ \sup_{\tilde{f}\in\mathcal{E}(\mathbb{R}^d)} \int \varphi_1^\circ(\tilde{f}(x))\mathrm{d}\alpha(x) - \int \tilde{f}(x)\mathrm{d}\pi_1(x)\\
&+ \sup_{\tilde{g}\in\mathcal{E}(\mathbb{R}^d)} \int \varphi_2^\circ(\tilde{g}(y))\mathrm{d}\beta(y) - \int \tilde{g}(y)\mathrm{d}\pi_2(y)\Bigg\},\\
= &\inf_{(\pi_1,\pi_2)\in\mathcal{M}_+(\mathbb{R}^d)^2} \Bigg\{\sup_{f_\theta \oplus g_\theta \leq \mathrm{C}_1} \int_{\mathbb{S}^{d-1}} \left[\int f_\theta \mathrm{d}(\theta^\star_\sharp \pi_1) + \int g_\theta \mathrm{d}(\theta^\star_\sharp \pi_2)\right] \mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta)\\
&+ \sup_{\tilde{f}\in\mathcal{E}(\mathbb{R}^d)} \int \varphi_1^\circ(\tilde{f}(x))\mathrm{d}\alpha(x) - \int \tilde{f}(x)\mathrm{d}\pi_1(x)\\
&+ \sup_{\tilde{g}\in\mathcal{E}(\mathbb{R}^d)} \int \varphi_2^\circ(\tilde{g}(y))\mathrm{d}\beta(y) - \int \tilde{g}(y)\mathrm{d}\pi_2(y)\Bigg\},
\end{aligned}$$

where $\mathcal{E}(\mathbb{R}^d)$ denotes a set of lower-semicontinuous functions, and the last equality holds thanks to Lemma A.11.

We focus now on verifying that Proposition A.10 holds, so that we can swap the infimum and the supremum. Define the functional

$$\begin{aligned}\mathcal{L}((\pi_1,\pi_2),((f_\theta)_\theta,(g_\theta)_\theta,\tilde{f},\tilde{g})) \triangleq \int_{\mathbb{S}^{d-1}} \left[\int f_\theta \mathrm{d}(\theta^\star_\sharp \pi_1) + \int g_\theta \mathrm{d}(\theta^\star_\sharp \pi_2) \right] \mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta) \\ + \int \varphi_1^\circ(\tilde{f}(x))\mathrm{d}\alpha(x) - \int \tilde{f}(x)\mathrm{d}\pi_1(x) \\ + \int \varphi_2^\circ(\tilde{g}(y))\mathrm{d}\beta(y) - \int \tilde{g}(y)\mathrm{d}\pi_2(y)\,.\end{aligned}$$

One has that,

- For any $((f_\theta)_\theta,(g_\theta)_\theta,\tilde{f},\tilde{g})$, $\mathcal{L}$ is linear (thus convex) and lower-semicontinuous.
- For any (π_1,π_2), $\mathcal{L}$ is concave in $((f_\theta)_\theta,(g_\theta)_\theta,\tilde{f},\tilde{g})$ because φ_i° is concave and thus $\mathcal{L}$ is a sum of linear or concave functions.

Furthermore, since we assumed e.g. that $0 \in \mathrm{dom}(\varphi)$, then

$$\sup_{((f_\theta)_\theta,(g_\theta)_\theta,\tilde{f},\tilde{g})} \inf_{(\pi_1,\pi_2)\in\mathcal{M}_+(\mathbb{R}^d)^2} \mathcal{L} \leq \mathrm{USOT}(\alpha,\beta) \leq \varphi_1(0)m(\alpha) + \varphi_2(0)m(\beta)\,,$$

because the marginals $(\pi_1,\pi_2)=(0,0)$ are admissible and suboptimal. If we consider instead that $(m(\alpha),m(\beta)) \in \mathrm{dom}(\varphi)$, then we take the marginals $\pi_1 = \alpha/m(\alpha)$ and $\pi_2 = \beta/m(\beta)$, which yields an upper-bound by $m(\alpha)\varphi_1(\frac{1}{m(\alpha)}) + m(\beta)\varphi_2(\frac{1}{m(\beta)})$. Then we consider an anchor dual point $b^\star = ((f_\theta)_\theta,(g_\theta)_\theta,\tilde{f},\tilde{g})$ to bound $\mathcal{L}$ over a compact set. We take $f_\theta = 0$, $g_\theta = 0$, which are always admissible since we take $\mathrm{C}_1(x,y) \geq 0$. Then, since we assume there exists $p_i \leq 0$ in $\mathrm{dom}(\varphi_i^*)$, we take $\tilde{f} = p_1$ and $\tilde{g} = p_2$. For these potentials one has:

$$\mathcal{L}((\pi_1,\pi_2),b^\star) = \varphi_1^\circ(p_1)m(\alpha) - p_1 m(\pi_1) + \varphi_2^\circ(p_2)m(\alpha) - p_2 m(\pi_2).$$

Note that the functional at this point only depends on the masses of the marginals (π_1,π_2). Since $(p_1,p_2) \geq 0$ the set of (π_1,π_2) such that $\mathcal{L}((\pi_1,\pi_2),b^\star) \leq \varphi_1(0)m(\alpha) + \varphi_2(0)m(\beta)$ is non-empty (at least in a neighbourhood of $(\pi_1,\pi_2)=(0,0)$, and that $(m(\pi_1),m(\pi_2))$ are uniformly bounded by some constant $M > 0$. By the Banach-Alaoglu theorem, such set of measures is compact for the weak* topology.

Therefore, Proposition A.10 holds and we have strong duality, *i.e.*

$$\mathrm{USOT}(\alpha,\beta) = \sup_{\substack{f_\theta \oplus g_\theta \leq \mathrm{C}_1 \\ (\tilde{f},\tilde{g}) \in \mathcal{E}(\mathbb{R}^d)}} \inf_{(\pi_1,\pi_2)\in\mathcal{M}_+(\mathbb{R}^d)^2} \mathcal{L}((\pi_1,\pi_2),((f_\theta)_\theta,(g_\theta)_\theta,\tilde{f},\tilde{g})).$$

To achieve the proof, note that taking the infimum in (π_1,π_2) (for fixed dual variables) reads

$$\begin{aligned}\inf_{\pi_1,\pi_2\geq 0} \int \left(\int_{\mathbb{S}^{d-1}} f_\theta(\theta^\star(x))\mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta) \right)\mathrm{d}\pi_1(x) - \int \tilde{f}(x)\mathrm{d}\pi_1(x) \\ + \int \left(\int_{\mathbb{S}^{d-1}} g_\theta(\theta^\star(y))\mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta) \right)\mathrm{d}\pi_2(y) - \int \tilde{g}(y)\mathrm{d}\pi_2(y).\end{aligned}$$

Note that we applied Fubini's theorem here, which holds here because all measures have compact support, thus all quantities are finite. It allows to rephrase the minimization over $\pi_1,\pi_2 \geq 0$ as the following constraint

$$\int_{\mathbb{S}^{d-1}} f_\theta(\theta^\star(x))\mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta) \geq \tilde{f}(x), \qquad \int_{\mathbb{S}^{d-1}} g_\theta(\theta^\star(y))\mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta) \geq \tilde{g}(y),$$

otherwise the infimum is $-\infty$. However, the function φ° is non-decreasing (see (Séjourné et al., 2019, Proposition 2)). Thus the maximization in $(\tilde{f}, \tilde{g})$ is optimal when the above inequality is actually an equality, i.e.

$$\int_{\mathbb{S}^{d-1}} f_\theta(\theta^\star(x))\mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta) = \tilde{f}(x), \qquad \int_{\mathbb{S}^{d-1}} g_\theta(\theta^\star(y))\mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta) = \tilde{g}(y).$$

Plugging the above relation in the functional $\mathcal{L}$ yields the desired result on the dual of USOT and ends the proof. □

We mention a strong duality result which is very general and which we use in the proof of 3.7. This result is taken from (Liero et al., 2018, Theorem 2.4) which itself takes it from (Simons, 2006).

Proposition A.10. *(Liero et al., 2018, Theorem 2.4) Consider two sets A and B be nonempty convex sets of some vector spaces. Assume A is endowed with a Hausdorff topology. Let $L : A \times B \to \mathbb{R}$ be a function such that*

1. *$a \mapsto L(a, b)$ is convex and lower-semicontinuous on A, for every $b \in B$*
2. *$b \mapsto L(a, b)$ is concave on B, for every $a \in A$.*

If there exists $b_\star \in B$ and $\kappa > \sup_{b\in B} \inf_{a\in A} L(a, b)$ such that the set $\{a \in A,\ L(a, b_\star) < \kappa\}$ is compact in A, then

$$\inf_{a\in A} \sup_{b\in B} L(a, b) = \sup_{b\in B} \inf_{a\in A} L(a, b)$$

We also consider the following to swap the supremum in the integral which defines sliced-UOT (and in particular sliced-OT). In what follows we note sliced potentials as functions $f_\theta(z)$ with $(\theta, z) \in \mathbb{S}^{d-1} \times \mathbb{R}$, such that

$$\mathrm{SUOT}(\alpha, \beta) = \int_{\mathbb{S}^{d-1}} \Big[\sup_{f_\theta \oplus g_\theta \leq \mathrm{C}_1} \int \varphi^\circ \circ f_\theta \mathrm{d}(\theta^\star_\sharp \alpha) + \int \varphi^\circ \circ g_\theta \mathrm{d}(\theta^\star_\sharp \beta)\Big] \mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta).$$

Note that with the above definition, $z \mapsto f_\theta(z)$ is continuous for any θ, but $\theta \mapsto f_\theta(z)$ is only $\hat{\boldsymbol{\sigma}}_K$-measurable.

Lemma A.11. *Consider two sets X and Y, a measure σ such that $\sigma(X) < +\infty$. Assume Y is compact. Consider a function $\mathcal{F} : X \times Y \to \mathbb{R}$. Assume there exists a sequence (y_n) in Y such that $\mathcal{F}(\cdot, y_n) \to \sup_{y\in Y} \mathcal{F}(\cdot, y)$ uniformly. Then one has*

$$\sup_{y\in Y} \int_X \mathcal{F}(x, y)\mathrm{d}\sigma(x) = \int_X \sup_{y\in Y} \mathcal{F}(x, y)\mathrm{d}\sigma(x).$$

Proof. Define $\mathcal{G}(x) = \sup_{y\in Y} \mathcal{F}(x, y)$ and $\mathcal{H}(x, y) \triangleq \mathcal{G}(x) - \mathcal{F}(x, y)$. One has $\mathcal{H} \geq 0$ by definition, and the desired equality can be rewritten as

$$\begin{aligned} &\sup_{y\in Y} \int_X \mathcal{F}(x, y)\mathrm{d}\sigma(x) = \int_X \sup_{y\in Y} \mathcal{F}(x, y)\mathrm{d}\sigma(x) \\ \Leftrightarrow\ &\inf_{y\in Y} \int_X \mathcal{H}(x, y)\mathrm{d}\sigma(x) = 0. \end{aligned}$$

Since the integral involving $\mathcal{H}$ is non-negative, the infimum is zero if and only if we have a sequence (y_n) such that $\int_X \mathcal{H}(\cdot, y_n)\mathrm{d}\sigma \to 0$. By assumption, one has $\mathcal{F}(\cdot, y_n) \to \sup_{y\in Y} \mathcal{F}(\cdot, y)$ uniformly, i.e. $||\mathcal{H}(\cdot, y_n)||_\infty \to 0$. This implies thanks to Holder's inequality that

$$0 \leq \int_X \mathcal{H}(\cdot, y_n)\mathrm{d}\sigma \leq \sigma(X)||\mathcal{H}(\cdot, y_n)||_\infty$$

Thus by assumption one has $\int_X \mathcal{F}(\cdot, y_n)\mathrm{d}\sigma \to \int_X \mathcal{G}\mathrm{d}\sigma$, which indeed means that we have the desired permutation between supremum and integral. □

Lemma A.12. *Let $p \in [1, +\infty)$ and assume that $\mathrm{C}_1(x, y) = |x - y|^p$. Consider two positive measures (α, β) with compact support. Assume that the measure $\hat{\boldsymbol{\sigma}}_K$ is discrete, i.e. $\hat{\boldsymbol{\sigma}}_K = \frac{1}{K}\sum_{i=1}^K \delta_{\theta_i}$ with $\theta_i \in \mathbb{S}^{d-1}$, $i = 1, \ldots, n$. Then, one can swap the integral over the sphere and the supremum in the dual formulation of* SUOT*, such that*

$$\mathrm{SUOT}(\alpha, \beta) = \sup_{f_\theta \oplus g_\theta \leq \mathrm{C}_1} \int_{\mathbb{S}^{d-1}} \Big[\int \varphi^\circ \circ f_\theta \mathrm{d}(\theta^\star_\sharp \alpha) + \int \varphi^\circ \circ g_\theta \mathrm{d}(\theta^\star_\sharp \beta)\Big] \mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta).$$

In particular, this result is valid for SOT.

Proof. The proof consists in applying Lemma A.11 for (X, Y) chosen as $X = \mathrm{supp}(\hat{\boldsymbol{\sigma}}_K) \subset \mathbb{S}^{d-1}$ and

$$Y = \big\{\forall \theta \in \mathrm{supp}(\hat{\boldsymbol{\sigma}}_K),\ f_\theta : \mathbb{R} \to \mathbb{R},\ g_\theta : \mathbb{R} \to \mathbb{R},\ f_\theta(x) + g_\theta(y) \leq \mathrm{C}_1(x, y)\big\}.$$

The functions in Y are dual potentials, and by definition are continuous for any θ. Let $\mathcal{F} : X \times Y \to \mathbb{R}$ be the functional defined as

$$\mathcal{F} : (\theta, (f_\theta)_\theta, (g_\theta)_\theta) \mapsto \int f_\theta \mathrm{d}(\theta^\star_\sharp \alpha) + \int g_\theta \mathrm{d}(\theta^\star_\sharp \beta)\,.$$

Since the measures (α, β) have compact support, then by Lemma A.13, the supremum is attained over a subset of dual potentials of Y such that for any fixed $\theta \in X$, (f_θ, g_θ) are Lipschitz-continuous and bounded, thus uniformly equicontinuous functions (with constants independent of θ). By the Ascoli-Arzela theorem, the set of uniformly equicontinuous functions is compact for the uniform convergence. Hence, for any $\theta \in X$, there exists a sequence of dual potentials $(f_{\theta,n}, g_{\theta,n})$ which uniformly converges to optimal dual potentials (f_θ, g_θ) (up to extraction of subsequence). Besides, we have $\mathrm{OT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta) = \mathcal{F}(\theta, f_\theta, g_\theta)$ and $\mathcal{F}(\theta, (f_{\theta,n})_\theta, (g_{\theta,n})_\theta) \to \mathrm{OT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta)$ as $n \to +\infty$. Denote $\mathcal{F}_n(\theta) \triangleq \mathcal{F}(\theta, (f_{\theta,n})_\theta, (g_{\theta,n})_\theta)$ and $\mathrm{OT}(\theta) \triangleq \mathrm{OT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta)$. In order to apply Lemma A.11, we need to prove that the convergence of $(\mathcal{F}_n(\theta))_{n\in\mathbb{N}^*}$ to $\mathrm{OT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta)$ is uniform w.r.t. θ, *i.e.* $\sup_{\theta \in X} |\mathcal{F}_n(\theta) - \mathrm{OT}(\theta)| \to 0$ as $n \to +\infty$.

First, note that for any $\theta \in X$,

$$|\mathcal{F}_n(\theta) - \mathrm{OT}(\theta)| \leq m(\alpha)\|f_{\theta,n} - f_\theta\|_\infty + m(\beta)\|g_{\theta,n} - g_\theta\|_\infty.$$

Since for a fixed $\theta \in X$, $(f_{\theta,n}, g_{\theta,n})_{n\in\mathbb{N}^*}$ uniformly converge to (f_θ, g_θ), this means that

$$\forall \theta \in X, \forall \varepsilon > 0,\ \exists N(\varepsilon, \theta), \forall n \geq N(\varepsilon, \theta),\ m(\alpha)\|f_{\theta,n} - f_\theta\|_\infty + m(\beta)\|g_{\theta,n} - g_\theta\|_\infty < \varepsilon.$$

Since we assume that σ is supported on a discrete set, then the cardinal of X is finite and one can define $N(\varepsilon) \triangleq \max_{\theta\in X} N(\varepsilon, \theta)$. This yields,

$$\forall \varepsilon > 0,\ \exists N(\varepsilon), \forall n \geq N(\varepsilon), \sup_{\theta \in X} |\mathcal{F}_n(\theta) - \mathrm{OT}(\theta)| < \varepsilon.$$

which means that $\sup_{\theta\in X} |\mathcal{F}_n(\theta) - \mathrm{OT}(\theta)| \to 0$, thus concludes the proof. □

Lemma A.13. *Let $p \in [1, +\infty)$ and $\mathrm{C}_1(x, y) = |x - y|^p$. Consider two positive measures $(\alpha, \beta) \in \mathcal{M}_+(\mathbb{R}^d)$ whose support is such that $\mathrm{C}_d(x, y) = ||x - y||^p \leq R$. Then for any $\theta \in \mathbb{S}^{d-1}$, one can restrict without loss of generality the problem* $\mathrm{UOT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta)$ *as a supremum over dual potentials satisfying $f_\theta(x) + g_\theta(y) \leq \mathrm{C}_1(x, y)$, uniformly bounded by M and uniformly L-Lipschitz, where M and L do not depend on θ.*

Proof. We adapt the proof of (Santambrogio, 2015, Proposition 1.11), and focus on showing that the uniform boundedness and Lipschitz constant are independent of $\theta \in \mathbb{S}^{d-1}$ in this setting. Here we consider the translation-invariant formulation of UOT from (Séjourné et al., 2022b), i.e. $\mathrm{UOT}(\alpha, \beta) = \sup_{f\oplus g\leq \mathrm{C}_d} \mathcal{H}(f, g)$, where $\mathcal{H}(f, g) = \sup_{\lambda\in\mathbb{R}} \mathcal{D}(f + \lambda, g - \lambda)$. It is proved in (Séjourné et al., 2022b, Proposition 9) that the above problem has the same primal and is thus equivalent to optimize $\mathcal{D}$. By definition one has $\mathcal{H}(f, g) = \mathcal{H}(f + \lambda, g - \lambda)$ for any $\lambda \in \mathbb{R}$, i.e. this formulation shares the same invariance as Balanced OT. Thus we can reuse all arguments from (Santambrogio, 2015, Proposition 1.11), such that for

$\mathrm{UOT}(\alpha, \beta)$, one can use the constraint $f(x) + g(y) \leq \mathrm{C}_d(x, y)$ and the assumption $\mathrm{C}_d(x, y) \leq R$ to prove that without loss of generality, on can restrict to potentials such that $f(x) \in [0, R]$ and $g(y) \in [-R, R]$. Furthermore if the cost satisfies in $\mathbb{R}^d$

$$|\mathrm{C}_d(x, y) - \mathrm{C}_d(x', y')| \leq L(||x - x'|| + ||y - y'||),$$

then one can also restrict w.l.o.g. to potentials which are L-Lipschitz. For the cost $\mathrm{C}_d(x, y) = ||x - y||^p$ with $p \geq 1$, this holds with constant $L = pR^{p-1}$ because the support is bounded and the gradient of C_d is radially non-decreasing.

Regarding $\mathrm{OT}(\theta^\star_\sharp \alpha, \theta^\star_\sharp \beta)$, the bounds (M_θ, L_θ) could be refined by considering the dependence in $\theta \in \mathbb{S}^{d-1}$. However we prove now these constants can be upper-bounded by a finite constant independent of θ. In this setting we consider the cost

$$\mathrm{C}_1(\theta^\star(x), \theta^\star(y)) = |\langle \theta, x - y \rangle|^p \leq ||\theta||^p ||x - y||^p \leq ||x - y||^p,$$

by Cauchy-Schwarz inequality. Therefore, if (α, β) have supports such that $||x - y||^p \leq R$, then $(\theta^\star_\sharp \alpha, \theta^\star_\sharp \beta)$ also have supports bounded by R in $\mathbb{R}$. Similarly note that the derivative of $h(x) = x^p$ is non-decreasing for $p \geq 1$. Hence the cost $\mathrm{C}_1(\theta^\star(x), \theta^\star(y))$ has a bounded derivative, which reads

$$p|\langle \theta, x - y \rangle|^{p-1} \leq p||\theta||^{p-1}||x - y||^{p-1} \leq p|x - y||^{p-1} \leq pR^{p-1}.$$

Thus on the supports of $(\theta^\star_\sharp \alpha, \theta^\star_\sharp \beta)$ one can also bound the Lipschitz constant of the cost $\mathrm{C}_1(x, y) = |x - y|^p$ by the same constant L. □

Remark: Extending Theorem 3.7. We conjecture that Theorem 3.7 also holds when $\boldsymbol{\sigma}$ is the uniform measures over $\mathbb{S}^{d-1}$, since the above holds for any $N \in \mathbb{N}^*$ and $\hat{\boldsymbol{\sigma}}_N$ converges weakly* to $\boldsymbol{\sigma}$. Proving this result would require that potentials (f_θ, g_θ) are also regular (*i.e.*, Lipschitz and bounded) w.r.t $\theta \in \mathbb{S}^{d-1}$. This regularity is proved in (Xi & Niles-Weed, 2022) assuming (α, β) have densities, but remains unknown for discrete measures. Since discretizing $\boldsymbol{\sigma}$ corresponds to the computational approach, we assume it to be discrete, so that no additional assumption than boundedness on (α, β) is required. For instance, such result remains valid for semi-discrete UOT computation.

B. Additional details for Section 4

B.1. Frank-Wolfe methodology for computing UOT

Background: FW for UOT. Our approach to compute SUOT and USOT builds upon the construction of (Séjourné et al., 2022b). It consists in applying a Frank-Wolfe (FW) procedure over the dual formulation of UOT. Such approach is equivalent to solve a sequence of balanced OT problems between measures $(\tilde{\alpha}, \tilde{\beta})$ which are iterative renormalizations of (α, β). While the idea holds in wide generality, it is especially efficient in 1D where OT has low algorithmic complexity, and we reuse it in our sliced setting.

FW algorithm consists in optimizing a functional $\mathcal{H}$ over a compact, convex set $\mathcal{C}$ by optimizing its linearization $\nabla\mathcal{H}$. Given a current iterate x^t of FW algorithm, one computes $r^{t+1} \in \arg\max_{r \in \mathcal{C}} \langle \nabla\mathcal{H}(x^t), r \rangle$, and performs a convex update $x^{t+1} = (1 - \gamma_{t+1})x^t + \gamma_{t+1} r^{t+1}$. One typically chooses the learning rate $\gamma_t = \frac{2}{2+t}$. This yields the routine `FWStep` of Section 4 which is detailed below.

Algorithm 3 – $\texttt{FWStep}(f, g, r, s, \gamma)$

Input: $\alpha, \beta, f, g, \gamma$
Output: Normalized measures (α, β) as in Equation (80)
  $f(x) \leftarrow (1 - \gamma) f(x) + \gamma r(x)$
  $g(y) \leftarrow (1 - \gamma) g(y) + \gamma s(y)$
  Return (f, g)

In the setting of UOT, one would take $\mathcal{C} = \{f \oplus g \leq \mathrm{C}_d\}$. However, this set is not compact as it contains $(\lambda, -\lambda)$ for any $\lambda \in \mathbb{R}$. Thus, (Séjourné et al., 2022b) propose to optimise a *translation-invariant* dual functional $\mathcal{H}(f, g) \triangleq \sup_{\lambda \in \mathbb{R}} \mathcal{D}(f + \lambda, g - \lambda)$, with $\mathcal{D}$ defined Equation (3). Similar to the balanced OT dual, one has $\mathcal{H}(f + \lambda, g - \lambda) = \mathcal{H}(f, g)$, thus one can apply (Santambrogio, 2015, Proposition 1.11) to assume w.l.o.g. that e.g. $f(0) = 0$ and restrict to a compact set of functions. We emphasize that FW algorithm is well-posed to optimize $\mathcal{H}$, but not $\mathcal{D}$.

Note that once we have the dual variables (f,g) maximizing $\mathcal{H}$, we retrieve optimal dual variables maximizing $\mathcal{D}$ as $(f+\lambda^\star(f,g), g-\lambda^\star(f,g))$, where $\lambda^\star(f,g) \triangleq \arg\max_{\lambda\in\mathbb{R}} \mathcal{D}(f+\lambda, g-\lambda)$. The KL setting where $\mathrm{D}_{\varphi_1}=\rho_1\mathrm{KL}$ and $\mathrm{D}_{\varphi_2}=\rho_2\mathrm{KL}$ is especially convenient, because $\lambda^\star(f,g)$ admits a closed form, which avoids iterative subroutines to compute it. In that case, it reads

$$\lambda^\star(f,g) = \frac{\rho_1\rho_2}{\rho_1+\rho_2}\log\left(\frac{\int e^{-f(x)/\rho_1}\mathrm{d}\alpha(x)}{\int e^{-g(y)/\rho_2}\mathrm{d}\beta(y)}\right). \tag{77}$$

We summarize the FW algorithm for UOT in the proposition below. We refer to (Séjourné et al., 2022b) for more details on the algorithm and pseudo-code. We adapt this approach and result for SUOT and USOT.

Proposition B.1. *(Séjourné et al., 2022b) Assume φ° is smooth. Given current iterates $(f^{(t)}, g^{(t)})$, the linear FW oracle of* $\mathrm{UOT}(\alpha,\beta)$ *is* $\mathrm{OT}(\bar\alpha^{(t)}, \bar\beta^{(t)})$, *where $\bar\alpha^{(t)} = \nabla\varphi^\circ(f^{(t)}+\lambda^\star(f^{(t)},g^{(t)}))\alpha$ and $\bar\beta^{(t)} = \nabla\varphi^\circ(g^{(t)}-\lambda^\star(f^{(t)},g^{(t)}))\beta$. In particular, one has $m(\bar\alpha^{(t)}) = m(\bar\beta^{(t)})$, thus the balanced OT problem always has finite value. More precisely, the FW update reads*

$$(f^{(t+1)}, g^{(t+1)}) = (1-\gamma^{(t+1)})(f^{(t)}, g^{(t)}) + \gamma^{(t+1)}(r^{(t+1)}, s^{(t+1)}), \tag{78}$$

$$\textit{where} \qquad (r^{(t+1)}, s^{(t+1)}) \in \arg\max_{r\oplus s\leq \mathrm{C}_d} \int r(x)\mathrm{d}\bar\alpha^{(t)}(x) + \int s(y)\mathrm{d}\bar\beta^{(t)}(y). \tag{79}$$

Recall that the in KL setting one has $\varphi_i^\circ(x) = \rho_i(1-e^{-x/\rho_i})$, thus $\nabla\varphi_i^\circ(x) = e^{-x/\rho_i}$. Thus in that case one normalizes the measures as

$$\bar\alpha = \exp\left(-\frac{f+\lambda^\star(f,g)}{\rho_1}\right)\alpha, \qquad \bar\beta = \exp\left(-\frac{g-\lambda^\star(f,g)}{\rho_2}\right)\beta, \tag{80}$$

where $\lambda^\star$ is defined in (77).

This defines the `Norm` routine in Section 4, which we detail below.

Algorithm 4 – $\mathtt{Norm}(\alpha,\beta,f,g,\rho_1,\rho_2)$

Input: $\alpha, \beta, f, g, \boldsymbol{\rho}=(\rho_1,\rho_2)$
Output: Normalized measures (α,β) as in eq. (80)
Compute $\lambda^\star = \lambda^\star(f,g)$ as in eq. (77)
$\bar\alpha(x) \leftarrow \exp\left(-\frac{f(x)+\lambda^\star}{\rho_1}\right)\alpha(x)$
$\bar\beta(y) \leftarrow \exp\left(-\frac{g(y)-\lambda^\star}{\rho_2}\right)\beta(y)$
Return (α,β)

B.2. Frank-Wolfe methodology for computing SUOT

Proposition B.2. *Given current iterates (f_θ, g_θ), the linear Frank-Wolfe oracle of* $\mathrm{USOT}(\alpha,\beta)$ *is* $\int_{\mathbb{S}^{d-1}} \mathrm{OT}(\theta^\star_\sharp\alpha^\theta, \theta^\star_\sharp\beta^\theta)\mathrm{d}\boldsymbol{\sigma}(\theta)$, *where*

$$\alpha^\theta = \nabla\varphi^\circ\Big(f_\theta + \lambda^\star(f_\theta, g_\theta)\Big)\alpha, \qquad \beta^\theta = \nabla\varphi^\circ\Big(g_\theta - \lambda^\star(f_\theta, g_\theta)\Big)\beta.$$

As a consequence, given dual sliced potentials (r_θ, s_θ) solving $\mathrm{OT}(\theta^\star_\sharp\alpha^\theta, \theta^\star_\sharp\beta^\theta)$, *one can perform Frank-Wolfe updates (78) on (f_θ, g_θ).*

Proof. Our goal is to compute the first order variation of the SUOT functional. Given that $\mathrm{SUOT}(\alpha,\beta) = \int_{\mathbb{S}^{d-1}} \mathrm{UOT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta)\mathrm{d}\boldsymbol{\sigma}(\theta)$, one can apply Proposition B.1 slice-wise. Since measures are assumed to have compact support, one can apply the dominated convergence theorem and differentiate under the integral sign. Furthermore, the translation-invariant formulation in the setting of SUOT reads

$$\mathrm{SUOT}(\alpha,\beta) = \int_{\mathbb{S}^{d-1}} \sup_{f_\theta \oplus g_\theta \leq \mathrm{C}_1} \Bigg[\sup_{\lambda_\theta \in \mathbb{R}} \int \varphi^\circ\Big(f_\theta(\cdot) + \lambda_\theta\Big) \mathrm{d}\theta^\star_\sharp \alpha \tag{81}$$

$$+ \int \varphi^\circ\Big(g_\theta(\cdot) - \lambda_\theta\Big) \mathrm{d}\theta^\star_\sharp \beta \Bigg], \tag{82}$$

In the setting where φ° is smooth and strictly concave (such as $\mathrm{D}_\varphi = \rho\mathrm{KL}$), there always exists a unique optimal $\lambda^\star_\theta$. Furthermore, one can apply the envelope theorem such that the Fréchet differential w.r.t. to a perturbation (r_θ, s_θ) of (f_θ, g_θ) reads

$$\int_{\mathbb{S}^{d-1}} \Bigg[\int r_\theta(\cdot) \times \nabla\varphi^\circ\Big(f_\theta(\cdot) + \lambda^\star_\theta(f_\theta, g_\theta)\Big) \mathrm{d}\theta^\star_\sharp \alpha \tag{83}$$

$$+ \int s_\theta(\cdot) \times \nabla\varphi^\circ\Big(g_\theta(\cdot) - \lambda^\star_\theta(f_\theta, g_\theta)\Big) \mathrm{d}\theta^\star_\sharp \beta \Bigg] \tag{84}$$

Setting

$$\alpha_\theta = \nabla\varphi^\circ\Big(f_\theta(\cdot) + \lambda^\star(f_\theta, g_\theta)\Big)\alpha, \qquad \beta_\theta = \nabla\varphi^\circ\Big(g_\theta(\cdot) - \lambda^\star(f_\theta, g_\theta)\Big)\beta\,,$$

yields the desired result, *i.e.* the first order variation is

$$\int_{\mathbb{S}^{d-1}} \Bigg[\int r_\theta(\cdot) \mathrm{d}(\theta^\star_\sharp \alpha_\theta) + \int s_\theta(\cdot) \mathrm{d}(\theta^\star_\sharp \beta_\theta) \Bigg]. \tag{85}$$

□

B.3. Frank-Wolfe methodology for computing USOT

To compute USOT, we leverage Theorem 3.7 and derive the linear Frank-Wolfe oracle based on its translation-invariant formulation. We state the associated FW updates in the following proposition.

Proposition B.3. *Given current iterates (f_θ, g_θ), the linear Frank-Wolfe oracle of* $\mathrm{USOT}(\alpha, \beta)$ *is* $\mathrm{SOT}(\bar\alpha, \bar\beta)$*, where*

$$\bar\alpha = \nabla\varphi^\circ(f_{avg} + \lambda^\star(f_{avg}, g_{avg}))\alpha, \qquad \bar\beta = \nabla\varphi^\circ(g_{avg} - \lambda^\star(f_{avg}, g_{avg}))\beta,$$

$$f_{avg}(x) = \int_{\mathbb{S}^{d-1}} f_\theta(\theta^\star(x)) \mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta), \qquad g_{avg}(y) = \int_{\mathbb{S}^{d-1}} g_\theta(\theta^\star(y)) \mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta)$$

Thus given dual sliced potentials $(r_\theta(\cdot), s_\theta(\cdot))$ which solve $\mathrm{SOT}(\bar\alpha, \bar\beta)$*, one can then perform Frank-Wolfe updates (78) on (f_θ, g_θ) and thus (f_{avg}, g_{avg}).*

Proof. Our goal is to compute the first order variation of the USOT functional. First, we leverage Theorem 3.7 such that USOT reads

$$\mathrm{USOT}(\alpha,\beta) = \sup_{f_\theta(\cdot) \oplus g_\theta(\cdot) \leq \mathrm{C}_1} \int \varphi^\circ_1\Big(\int_{\mathbb{S}^{d-1}} f_\theta(\theta^\star(x)) \mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta)\Big) \mathrm{d}\alpha(x) \tag{86}$$

$$+ \int \varphi^\circ_2\Big(\int_{\mathbb{S}^{d-1}} g_\theta(\theta^\star(y)) \mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta)\Big) \mathrm{d}\beta(y) \tag{87}$$

$$= \sup_{f_\theta(\cdot) \oplus g_\theta(\cdot) \leq \mathrm{C}_1} \int \varphi^\circ_1\Big(f_{avg}(x)\Big) \mathrm{d}\alpha(x) + \int \varphi^\circ_2\Big(g_{avg}(y)\Big) \mathrm{d}\beta(y), \tag{88}$$

where

$$f_{avg}(x) = \int_{\mathbb{S}^{d-1}} f_\theta(\theta^\star(x))\mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta), \qquad g_{avg}(y) = \int_{\mathbb{S}^{d-1}} g_\theta(\theta^\star(y))\mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta).$$

From this, we derive the translation-invariant formulation as follows.

$$\mathrm{USOT}(\alpha,\beta) = \sup_{f_\theta(\cdot)\oplus g_\theta(\cdot)\leq \mathrm{C}_1} \sup_{\lambda\in\mathbb{R}} \int \varphi_1^\circ\Big(f_{avg}(x)+\lambda\Big)\mathrm{d}\alpha(x) \tag{89}$$

$$+ \int \varphi_2^\circ\Big(g_{avg}(y)-\lambda\Big)\mathrm{d}\beta(y), \tag{90}$$

For smooth and strictly concave φ°, there exists a unique $\lambda^\star(f_{avg}, g_{avg})$ attaining the supremum. Furthermore, one can apply the enveloppe theorem and differentiate under the integral sign (since the support is compact). Consider perturbations $(r_\theta(\cdot), s_\theta(\cdot))$ of $(f_\theta(\cdot), g_\theta(\cdot))$. Write

$$r_{avg}(x) = \int_{\mathbb{S}^{d-1}} r_\theta(\theta^\star(x))\mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta), \qquad s_{avg}(y) = \int_{\mathbb{S}^{d-1}} s_\theta(\theta^\star(y))\mathrm{d}\hat{\boldsymbol{\sigma}}_K(\theta).$$

Given that $\varphi_1^\circ(f_{avg}+r_{avg}) = \varphi_1^\circ(f_{avg}) + r_{avg}\nabla\varphi_1^\circ(f_{avg}) + o(||r_{avg}||_\infty)$, the first order variation reads

$$\int r_{avg}(x)\nabla\varphi_1^\circ\Big(f_{avg}(x)+\lambda^\star(f_{avg},g_{avg})\Big)\mathrm{d}\alpha(x) \tag{91}$$

$$+\int s_{avg}(y)\nabla\varphi_2^\circ\Big(g_{avg}(y)-\lambda^\star(f_{avg},g_{avg})\Big)\mathrm{d}\beta(y). \tag{92}$$

Then we define

$$\bar{\alpha} = \nabla\varphi_1^\circ(f_{avg}+\lambda^\star(f_{avg},g_{avg}))\alpha, \qquad \bar{\beta} = \nabla\varphi_2^\circ(g_{avg}-\lambda^\star(f_{avg},g_{avg}))\beta,$$

such that the first order variation reads

$$\int r_{avg}(x)\mathrm{d}\bar{\alpha}(x) + \int s_{avg}(y)\mathrm{d}\bar{\beta}(y). \tag{93}$$

One can then explicit the definition of (r_{avg}, s_{avg}), such that it reads

$$\int_{\mathbb{S}^{d-1}}\int r_\theta(\theta^\star(x))\mathrm{d}\bar{\alpha}(x) + \int_{\mathbb{S}^{d-1}}\int s_\theta(\theta^\star(y))\mathrm{d}\bar{\beta}(y) \tag{94}$$

$$= \int_{\mathbb{S}^{d-1}}\int r_\theta \mathrm{d}\theta^\star_\sharp\bar{\alpha}(x) + \int_{\mathbb{S}^{d-1}}\int s_\theta \mathrm{d}\theta^\star_\sharp\bar{\beta}(y). \tag{95}$$

By optimizing the above over the constraint set $\{r_\theta \oplus s_\theta \leq \mathrm{C}_1\}$, we identify the computation of $\mathrm{SOT}(\bar{\alpha},\bar{\beta})$, which concludes the proof. □

Since Proposition B.3 involves potentials averaged over $\boldsymbol{\sigma}$, we thus need to define the `AvgPot` routine detailed below.

Algorithm 5 – `AvgPot`(f_θ)

Input: sliced potentials (f_θ) with $(\theta_k)_k^K$
Output: Averaged potential f_{avg} as in Proposition B.3
Average $f_{avg} = \frac{1}{K}\sum_{k=1}^K f_\theta$

B.4. Implementation of Sliced OT to return dual potentials

Recall from Section 4, Algorithms 1 and 2 and more precisely, Propositions B.2 and B.3, that FW linear oracle is a sliced OT program, *i.e.* a set of OT problems computed between univariate distributions of $\mathcal{M}_+(\mathbb{R})$. Therefore, a key building block of our algorithm is to compute the loss and dual variables of these univariate OT problems. We explain below how one can compute the sliced OT loss and dual potentials. The computation of the loss consists in implementing closed formulas of OT between univariate distributions, as detailed in (Santambrogio, 2015, Proposition 2.17). More precisely, when $\mathrm{C}_1(x, y) = |x - y|^p$ and $(\mu, \nu) \in \mathcal{M}_+(\mathbb{R})$, then

$$\mathrm{OT}(\mu, \nu) = \int_0^1 |F_\mu^{[-1]}(t) - F_\nu^{[-1]}(t)|^p \mathrm{d}t, \tag{96}$$

where $F_\mu^{[-1]}$ denotes the inverse cumulative distribution function (ICDF) of μ.

Algorithm 6 – `SlicedOTLoss`$(\alpha, \beta, \{\theta\}, p)$

Input: α, β, projections $\{\theta\}$, exponent p
Output: $\mathrm{OT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta)$ as in eq. (96)
for $\theta \in \{\theta\}$ **do**
  Project support of $\theta^\star_\sharp\alpha$ and $\theta^\star_\sharp\beta$
  Sort weights of $(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta)$ and support $(\theta^\star(x))$, $(\theta^\star(y))$ s.t. support is non-decreasing
  Compute ICDF of $\theta^\star_\sharp\alpha$ and $\theta^\star_\sharp\beta$
  Compute $\mathrm{OT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta)$ as in eq. (96) with exponent p
end for

To compute dual potentials using backpropagation, one computes the sliced OT losses (using Algorithm 6) then calls the backpropagation w.r.t to inputs (α, β), because their gradients are optimal dual potentials (Santambrogio, 2015, Proposition 7.17). We describe this procedure in Algorithm 7.

Algorithm 7 – `SlicedOTPotentialsBackprop`$(\alpha, \beta, \{\theta\}, p)$

Input: α, β, projections $\{\theta\}$, exponent p
Output: Dual potentials (f_θ, g_θ) solving $\mathrm{OT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta)$
  Enable gradients w.r.t. $(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta)$
  Call `SlicedOTLoss`$(\alpha, \beta, \{\theta\}, p)$
  Sum (but do not average) losses $\mathcal{L} = \sum_\theta \mathrm{OT}(\theta^\star_\sharp\alpha, \theta^\star_\sharp\beta)$.
  Backpropagate $\mathcal{L}$ w.r.t. (α, β)
  Return (f_θ, g_θ) as gradients of $\mathcal{L}$ w.r.t. (α, β).

The implementation of the dual potentials using 1D closed forms relies on the north-west corner rule principle, which can be vectorized in PyTorch in order to be computed in parallel. The contribution of our implementation thus consists in making such algorithm GPU-compatible and allowing for a parallel computation for every slice simultaneously. We stress that this constitutes a non-trivial piece of code, and we refer the interested reader to the code in our supplementary material for more details on the implementation.

B.5. Output optimal sliced marginals

In all our algorithms, we focus on dual formulations of SUOT and USOT, which optimize the dual potentials. However, one might want the output variables of the primal formulation (See Definition 3.1). In particular, the marginals of optimal transport plans are interested because they are interpreted as normalized versions of inputs (α, β) where geometric outliers have been removed. We detail where this interpretation comes from in the setting of UOT, and then give how it is adapted to SUOT and USOT. In particular, we justify that the `Norm` routine suffices to compute them.

Case of UOT. We focus on the $\mathrm{D}_{\varphi_i} = \rho_i\mathrm{KL}$. As per (Liero et al., 2018, Equation 4.21), we have at optimality that the optimal transport $\pi^\star$ plan solving $\mathrm{UOT}(\alpha, \beta)$ as in Equation (2) has marginals $(\pi_1^\star, \pi_2^\star)$ which read $\pi_1^\star = e^{-f^\star/\rho_1}\alpha$ and $\pi_2^\star = e^{-g^\star/\rho_2}\beta$, where $(f^\star, g^\star)$ are the optimal dual potentials solving Equation (3). Since on $\mathrm{supp}(\pi^\star)$ one also has $f^\star(x) + g^\star(y) = \mathrm{C}_d(x, y)$, if the transportation cost $\mathrm{C}_d(x, y)$ is large (i.e. we are matching a geometric outlier), so are $f^\star(x)$

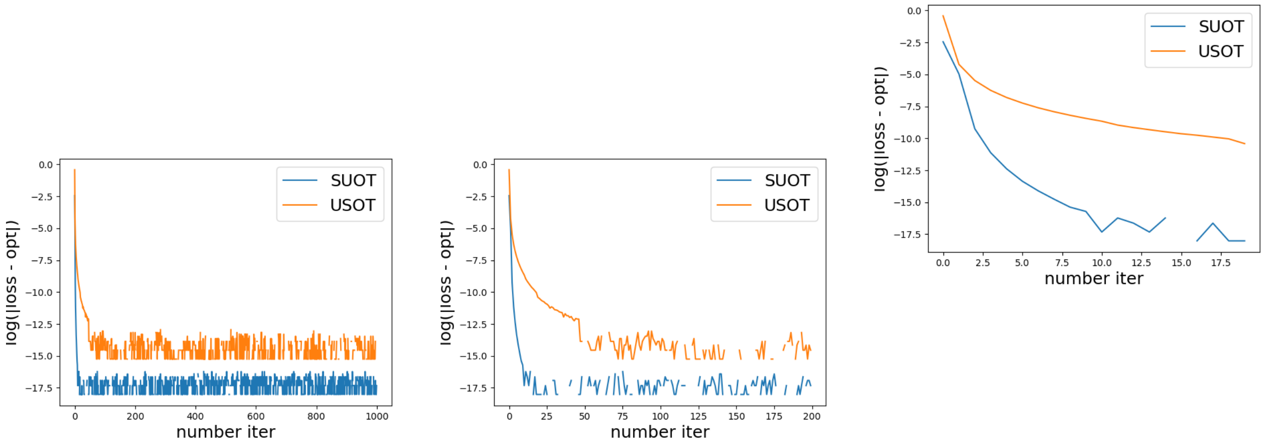


Figure 5: $|\mathrm{SUOT}(\alpha, \beta) - \widehat{\mathrm{SUOT}}_t|$ and $|\mathrm{USOT}(\alpha, \beta) - \widehat{\mathrm{USOT}}_t|$ against iteration t, where $\widehat{\mathrm{SUOT}}_t, \widehat{\mathrm{USOT}}_t$ are the estimated SUOT, USOT using t FW iterations. Plots are in log-scale. All figures are issued from the same run, but zoomed on a subset of first iterations: *(left)* 1000 iterations of FW, *(middle)* 200 iterations, *(right)* 20 iterations.

and $g^\star(y)$, and eventually the weights $\pi_1^\star(x)$ and $\pi_2^\star(y)$ are small, hence the interpretation of the geometric normalization of the measures. Note that in that case, one obtain $(\pi_1^\star, \pi_2^\star)$ by calling $\mathtt{Norm}(\alpha, \beta, f^\star, g^\star, \rho_1, \rho_2)$.

Case of SUOT. Since $\mathrm{SUOT}(\alpha, \beta)$ consists in integrating $\mathrm{UOT}(\theta^\star_\sharp \alpha, \theta^\star_\sharp \beta)$ w.r.t. $\boldsymbol{\sigma}$, it shares many similarities with UOT. For any θ, we consider π_θ and (f_θ, g_θ) solving the primal and dual formulation of $\mathrm{UOT}(\theta^\star_\sharp \alpha, \theta^\star_\sharp \beta)$. The marginals of π_θ are thus given by $(e^{-f_\theta/\rho_1}\alpha, e^{-g_\theta/\rho_2}\beta)$. In particular, we retrieve the observation made in Figure 1 that the optimal marginals change for each θ. In that case we call for each θ the routine $\mathtt{Norm}(\alpha, \beta, f_\theta, g_\theta, \rho_1, \rho_2)$.

Case of USOT. Recall that the optimal marginals (π_1, π_2) in $\mathrm{USOT}(\alpha, \beta)$ do not depend on θ, contrary to $\mathrm{SUOT}(\alpha, \beta)$. Leveraging the dual formulation of Theorem 3.7, and looking at the Lagrangian which is defined in the proof of Theorem 3.7 (see Appendix A.7), we have the optimality condition that $\pi_1 = e^{-f_{avg}/\rho_1}\alpha$ and $\pi_2 = e^{-g_{avg}/\rho_2}\beta$. Thus in that case, calling $\mathtt{Norm}(\alpha, \beta, f_{avg}, g_{avg}, \rho_1, \rho_2)$ yields the desired marginals.

B.6. Convergence of Frank-Wolfe iterations: Empirical analysis

We display below an experiment on synthetic dataset to illustrate the convergence of Frank-Wolfe iterations. We also provide insights on the number of iterations that yields a reasonable approximation: a few iterations suffices in our practical settings, typically $F = 20$.

The results are displayed in Figure 5. We consider the empirical distributions (α, β) computed over respectively, $N = 400$ and $M = 500$ samples over the unit hypercube $[0, 1]^d$, $d = 10$. Moreover, β is slightly shifted by a vector of uniform coordinates $0.5 \times \mathbf{1}_d$. We choose $\rho = 1$ and report the estimation of $\mathrm{SUOT}(\alpha, \beta)$ and $\mathrm{USOT}(\alpha, \beta)$ through Frank-Wolfe iterations. We estimate the true values by running $F = 5000$ iterations, and display the difference between the estimated score and the 'true' values. Appendix B.6 shows that numerical precision is reached in a few tens of iterations. As learning tasks do not usually require an estimation of losses up to numerical precision, we think that it is hence reasonable to take $F \approx 20$ in numerical applications.

C. Additional details on Section 5

C.1. Document classification: Technical details and additional results

C.1.1. Datasets

We sum up the statistics of the different datasets in Table 2.

Table 2: Dataset characteristics.

	BBCSport	Movies	Goodreads genre	Goodreads like
Doc	737	2000	1003	1003
Train	517	1500	752	752
Test	220	500	251	251
Classes	5	2	8	2
Mean words by doc	116 ± 54	182 ± 65	1491 ± 538	1491 ± 538
Median words by doc	104	175	1518	1518
Max words by doc	469	577	3499	3499

BBCSport. The BBCSport dataset contains articles between 2004 and 2005, and is composed of 5 classes. We average over the 5 same train/test split of (Kusner et al., 2015). The dataset can be found in https://github.com/mkusner/wmd/tree/master.

Movie Reviews. The movie reviews dataset is composed of 1000 positive and 1000 negative reviews. We take five different random 75/25 train/test split. The data can be found in http://www.cs.cornell.edu/people/pabo/movie-review-data/.

Goodreads. This dataset, proposed in (Maharjan et al., 2017), and which can be found at https://ritual.uh.edu/multi_task_book_success_2017/, is composed of 1003 books from 8 genres. A first possible classification task is to predict the genre. A second task is to predict the likability, which is a binary task where a book is said to have success if it has an average rating ≥ 3.5 on the website Goodreads (https://www.goodreads.com). The five train/test split are randomly drawn with 75/25 proportions.

C.1.2. Technical Details

All documents are embedded with the Word2Vec model (Mikolov et al., 2013) in dimension $d = 300$. The embedding can be found in https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/view?resourcekey=0-wjGZdNAUop6WykTtMip30g.

In this experiment, we report the results averaged over 5 random train/test split. For discrepancies which are approximated using random projections, we additionally average the results over 3 different computations, and we report this standard deviation in Table 1. Furthermore, we always use 500 projections to approximate the sliced discrepancies. For Frank-Wolfe based methods, we use 10 iterations, which we found to be enough to have a good accuracy. We added an ablation of these two hyperparameters in Figure 7. We report the results obtained with the best ρ for USOT and SUOT computed among a grid $\rho \in \{10^{-4}, 5 \cdot 10^{-4}, 10^{-3}, 5 \cdot 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. For USOT, the best ρ is consistently $5 \cdot 10^{-3}$ for the Movies and Goodreads datasets, and $5 \cdot 10^{-4}$ for the BBCSport dataset. For SUOT, the best ρ obtained was 0.01 for the BBCSport dataset, 1.0 for the movies dataset and 0.5 for the goodreads dataset. For UOT, we used $\rho = 1.0$ on the BBCSport dataset. For the movies dataset, the best ρ obtained on a subset was 50, but it took an unreasonable amount of time to run on the full dataset as the runtime increases with ρ (see (Chapel et al., 2021, Figure 3)). On the goodreads dataset, it took too much memory on the GPU. For Sinkhorn UOT, we used $\varepsilon = 0.001$ and $\rho = 0.1$ on the BBCSport and Goodreads datasets, and $\varepsilon = 0.01$ on the Movies dataset. For each method, the number of neighbors used for the k-NN method is obtained via cross-validation.

C.1.3. Additional experiments

Runtime. We report in Figure 6 the runtime of computing the different discrepancies between each pair of documents. On the BBCSport dataset, the documents have in average 116 words, thus the main bottleneck is the projection step for sliced OT methods. Hence, we observe that OT runs slightly faster than SOT and the sliced unbalanced counterparts. Goodreads is a dataset with larger documents, with on average 1491 words by document. Therefore, as OT scales cubically with the number of samples, we observe here that all sliced methods run faster than OT, which confirms that sliced methods scale better w.r.t. the number of samples. In this setting, we were not able to compute UOT with the POT implementation in a reasonable time. Computations have been performed with a NVIDIA A100 GPU.

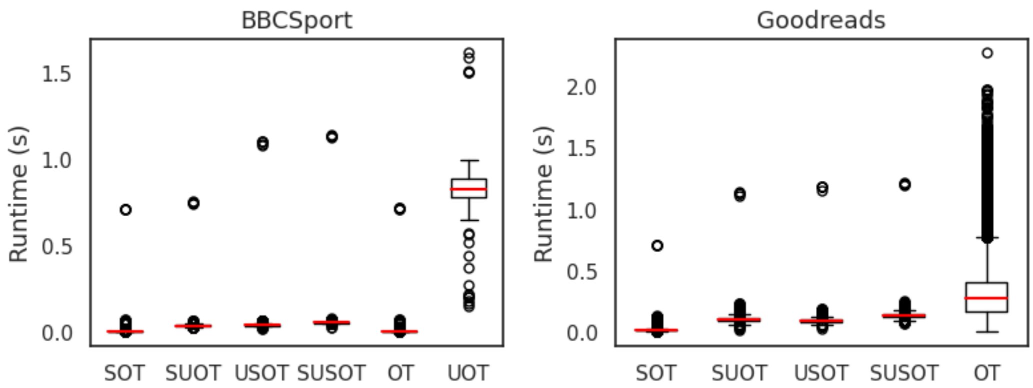


Figure 6: Runtime on the BBCSport dataset *(left)* and on the Goodreads dataset *(right)*.

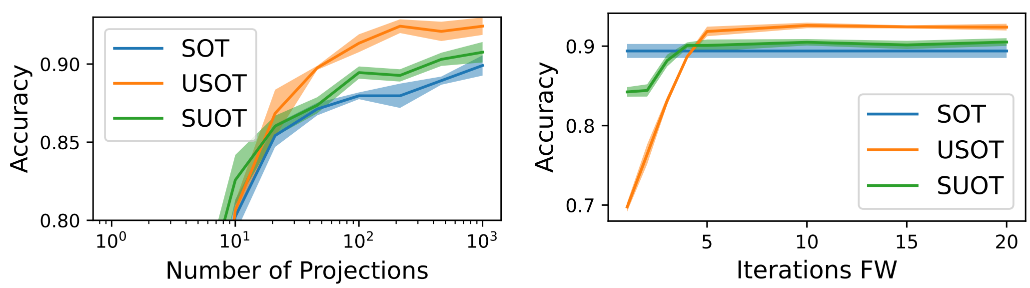


Figure 7: Ablation on BBCSport of the number of projections *(left)* and of the number of Frank-Wolfe iterations *(right)*.

Ablations. We plot in Figure 7 accuracy as a function of the number of projections and the number of iterations of the Frank-Wolfe algorithm. We averaged the accuracy obtained with the same setting described in Appendix C.1.2, with varying number of projections $K \in \{4, 10, 21, 46, 100, 215, 464, 1000\}$ and number of FW iterations $F \in \{1, 2, 3, 4, 5, 10, 15, 20\}$. Regarding the hyperparameter ρ, we selected the one returning the best accuracy, *i.e.* $\rho = 5 \cdot 10^{-4}$ for USOT and $\rho = 10^{-2}$ for SUOT.

C.2. Unbalanced sliced Wasserstein barycenters

We define below the formulation of the USOT barycenter which was used in the experiments of Figure 4 to average predictions of geophysical data. We then detail how we computed it.

Definition C.1. Consider a set of measures $(\alpha_1, \dots, \alpha_B) \in \mathcal{M}_+(\mathbb{R}^d)^B$, and a set of non-negative coefficients $(\omega_1, \dots, \omega_B) \geq 0$ such that $\sum_{b=1}^B \omega_b = 1$. We define the barycenter problem (in the KL setting) as

$$\mathcal{B}((\alpha_b)_b, (\omega_b)_b) \triangleq \inf_{\beta \in \mathcal{P}(\mathbb{R}^d)} \sum_{b=1}^B \omega_b \text{USOT}(\alpha_b, \beta), \tag{97}$$

$$= \inf_{\beta \in \mathcal{P}(\mathbb{R}^d)} \sum_{b=1}^B \inf_{(\pi_{b,1}, \pi_{b,2})} \text{SOT}(\pi_{b,1}, \pi_{b,2}) + \rho_1 \text{KL}(\pi_{b,1}|\alpha_b) + \rho_2 \text{KL}(\pi_{b,2}|\beta), \tag{98}$$

where $\mathcal{P}(\mathbb{R}^d)$ denotes the set of probability measures.

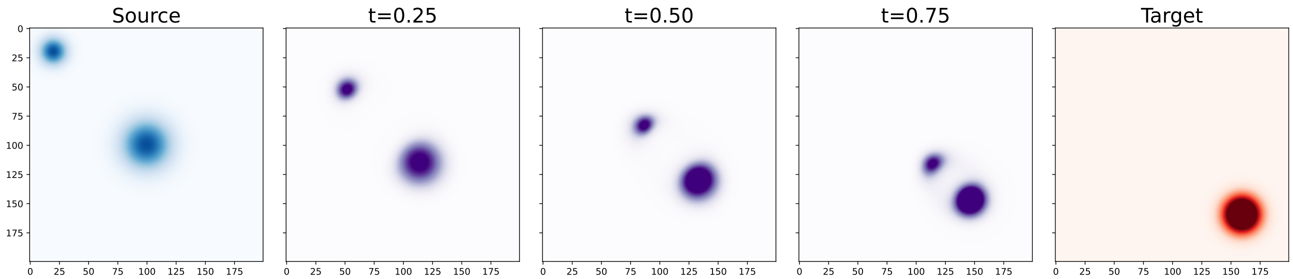
To compute the barycenter, we aggregate several building blocks. First, since we consider that the barycenter $\beta \in \mathcal{P}(\mathbb{R}^d)$ is a probability, we perform mirror descent as in (Beck & Teboulle, 2003; Cuturi & Doucet, 2014b). More precisely, we use a Nesterov accelerated version of mirror descent. We also tried projected gradient descent, but it did not yield consistent outputs (due to convergence speed (Beck & Teboulle, 2003)). Second, we use a Stochastic-USOT version (see Section 4), *i.e.* we sample new projections at each iteration of the barycenter update (but not a each iteration of the FW subroutines in Algorithm 2). This procedure is described in Algorithm 8.

Algorithm 8 – $\texttt{Barycenter}((\alpha_b)_b, (\omega_b)_b, \rho_1, \rho_2, lr)$

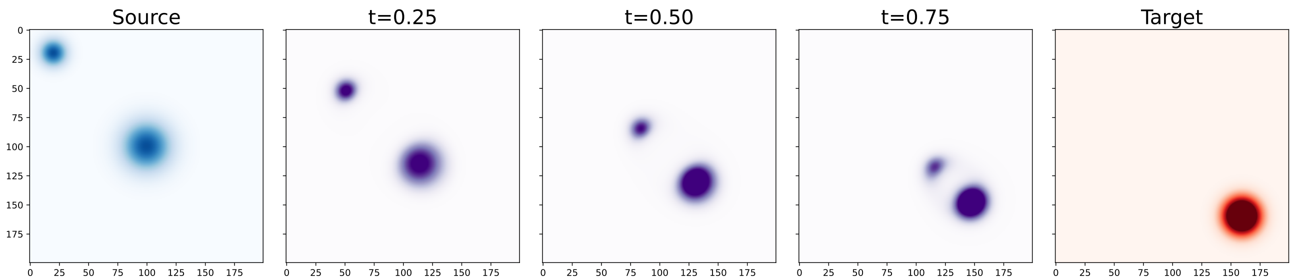
Input: measures $(\alpha_b)_b$, weights $(\omega_b)_b$, ρ_1, ρ_2, learning rate lr, FW iter F
Output: Optimal barycenter β of Equation (97)

```
t ← 1
Init (β, β̃, β̂) as uniform distribution over a grid
while not converged do do
    γ ← 2/(t+1),
    β → (1 − γ)β̂ + γβ̃
    Sample projections (θ_k)_{k=1}^K
    Compute B((α_b)_b, (ω_b)_b) by calling USOT(α_b, β, F, (θ_k)_{k=1}^K, ρ1, ρ2) in Algorithm 2 for each b
    Compute g as the gradient of B((α_b)_b, (ω_b)_b) w.r.t. variable β
    β̃ ← exp(−lr × γ^{−1} × g)β
    β̃ ← β̃/m(β̃)
    β̂ ← (1 − γ)β̂ + γβ̃
    t ← t + 1
end while
```

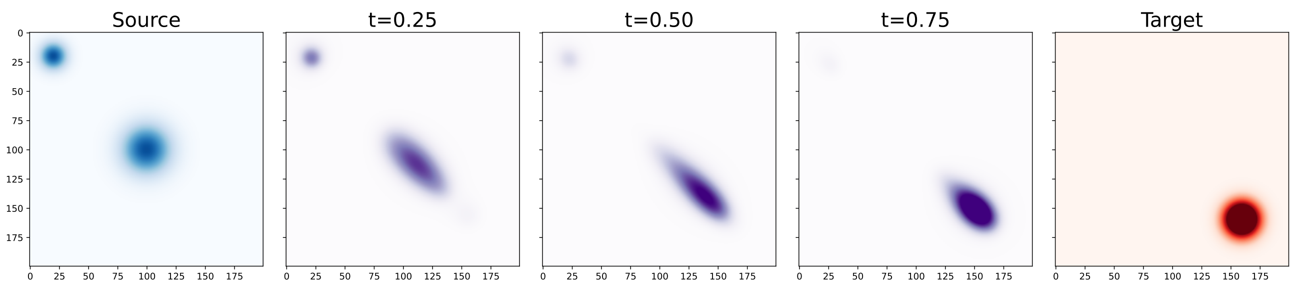
We illustrate this algorithm with several examples of interpolation in Figure 8. We propose to compute an interpolation between two measures located on a fixed grid of size 200×200 with different values of ρ_i in $\mathrm{D}_{\varphi_i} = \rho_i \mathrm{KL}$. For illustration purposes, we construct the *source* distribution as a mixture of two Gaussians with a small and a larger mode, and the *target* distribution as a single Gaussian. Those distributions are normalized over the grid such that both total norms are equal to one (which is not required by our unbalanced sliced variants but grants more interpretability and possible comparisons with SOT). Figure 8a shows the result of the interpolation at three timestamps ($t = 0.25,\ 0.5$ and 0.75) of a SOT interpolation (within this setting, $\omega_1 = 1 - t$ and $\omega_2 = t$). As expected, the two modes of the source distribution are transported over the target one. We verify in Figure 8b that for a large value of $\rho_1 = \rho_2 = 100$, the USOT interpolation behaves similarly as SOT, as expected from the theory. When $\rho_1 = \rho_2 = 0.01$, the smaller mode is not moved during the interpolation, whereas the larger one is stretched toward the target (Figure 8c). Finally, in Figure 8d, an asymmetric configuration of $\rho_1 = 0.01$ and $\rho_2 = 100$ allows to get an interpolation when only the big mode of the source distribution is displaced toward the target. In all those cases, the mirror-descent algorithm 8 is run for 500 iterations. Even for a large grid of 200×200, those different results are obtained in a $2 - 3$ minutes on a commodity GPU, while the OT or UOT barycenters are untractable with a limited computational budget.



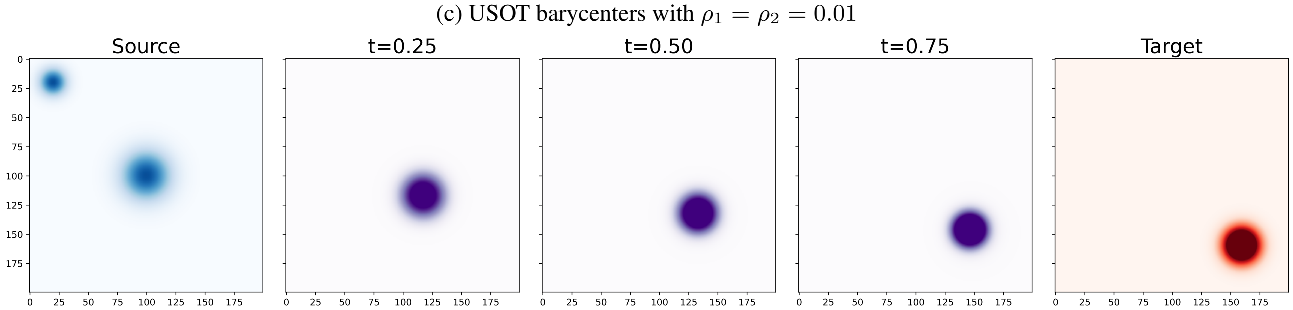
(a) SOT barycenters



(b) USOT barycenters with $\rho_1 = \rho_2 = 100$



(c) USOT barycenters with $\rho_1 = \rho_2 = 0.01$



(d) USOT barycenters with $\rho_1 = 0.01$ and $\rho_2 = 100$

Figure 8: **Interpolation with USOT as a barycenter computation**. We compare different interpolations using SOT or USOT with different settings for the ρ values