Entropy Partial Transport with Tree Metrics: Theory and Practice

Tam Le[*]
RIKEN AIP

Truyen Nguyen[*]
University of Akron

Abstract

Optimal transport (OT) theory provides powerful tools to compare probability measures. However, OT is limited to nonnegative measures having the same mass, and suffers serious drawbacks about its computation and statistics. This leads to several proposals of regularized variants of OT in the recent literature. In this work, we consider an *entropy partial transport* (EPT) problem for nonnegative measures on a tree having different masses. The EPT is shown to be equivalent to a standard complete OT problem on a one-node extended tree. We derive its dual formulation, then leverage this to propose a novel regularization for EPT which admits fast computation and negative definiteness. To our knowledge, the proposed regularized EPT is the first approach that yields a *closed-form* solution among available variants of unbalanced OT for general nonnegative measures. For practical applications without prior knowledge about the tree structure for measures, we propose tree-sliced variants of the regularized EPT, computed by averaging the regularized EPT between these measures using random tree metrics, built adaptively from support data points. Exploiting the negative definiteness of our regularized EPT, we introduce a positive definite kernel, and evaluate it against other baselines on benchmark tasks such as document classification with word embedding and topological data analysis. In addition, we empirically demonstrate that our regularization also provides effective approximations.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1 Introduction

Optimal transport (OT) theory offers powerful tools to compare probability measures (Villani, 2008). OT has been applied for various tasks in machine learning (Courty et al., 2017; Bunne et al., 2019; Nadjahi et al., 2019; Peyré and Cuturi, 2019), statistics (Mena and Niles-Weed, 2019; Weed and Berthet, 2019) and computer graphics (Solomon et al., 2015; Lavenant et al., 2018). However, OT requires input measures having the same mass which may limit its applications in practice since one often needs to deal with measures of unequal masses. For instance, in natural language processing, we can view a document as a measure where each word is regarded as a point in the support with a unit mass. Thus, documents with different lengths lead to their associated measures having different masses.

To tackle the transport problem for measures having different masses, Caffarelli and McCann (2010) proposed the *partial optimal transport* (POT) where one only transports a fixed amount of mass from a measure into another. Later, Figalli (2010) extended the theory of POT, notably, about the uniqueness of solutions. A different approach is to optimize the sum of a transport functional and two convex entropy functionals which quantify the deviation of the marginals of the transport plan from the input measures (Liero et al., 2018), i.e., the *optimal entropy transport* (OET) problem. This formulation recovers many different previous works. For examples, when the entropy is equal to the total variation distance or the ℓ^2 distance, the OET is respectively equivalent to the *generalized Wasserstein distance* (Piccoli and Rossi, 2014, 2016) or the *unbalanced mass transport* (Benamou, 2003). It is worth noting that the *generalized Wasserstein distance* shares the same spirit as the *Kantorovich-Rubinstein* discrepancy (Hanin, 1992; Guittet, 2002; Lellmann et al., 2014). Another variant is the *unnormalized optimal transport* (Gangbo et al., 2019) which mixes Wasserstein distance and the ℓ^p distance. There are several applications of the transport problem for measures having different masses such as in machine learning (Frogner et al., 2015; Janati et al., 2019), deep learning (Yang and

[*]: Two authors contributed equally.

Uhler, 2019), topological data analysis (Lacombe et al., 2018), computational imaging (Lee et al., 2019), and computational biology (Schiebinger et al., 2019).

One important case for the OET problem is when the entropy is equal to the Kullback-Leibler (KL) divergence and a particular cost function is used, then OET is equivalent to the *Kantorovich-Hellinger* distance (i.e., *Wasserstein-Fisher-Rao* distance) (Chizat et al., 2018; Liero et al., 2018). In addition, one can apply the Sinkhorn-based algorithm to efficiently solve OET problem when the entropy is equal to KL divergence, i.e., *Sinkhorn-based approach for unbalanced optimal transport* (Sinkhorn-UOT) (Frogner et al., 2015; Chizat et al., 2018). Pham et al. (2020) showed that the complexity of Sinkhorn-based algorithm for Sinkhorn-UOT is quadratic which is similar to the case of entropic regularized OT (Cuturi, 2013) for probability measures. However, for large-scale applications where the supports of measures contain a large number of points, the computation of Sinkhorn-UOT becomes prohibited. Following the sliced-Wasserstein (SW) distance (Rabin et al., 2011; Bonneel et al., 2015) which projects supports into a one-dimensional space and employs the closed-form solution of the univariate optimal transport (1d-OT), Bonneel and Coeurjolly (2019) propose the *sliced partial optimal transport* (SPOT) for nonnegative measures having different masses. Unlike the standard 1d-OT, one does not have a closed-form solution for measures of unequal masses that are supported in a one-dimensional space. With an assumption of a unit mass on each support, Bonneel and Coeurjolly (2019) derived an efficient algorithm to solve the SPOT problem in quadratic complexity for the worst case. Especially, in practice, their proposed algorithm is nearly linear for computation. However, as in SW, the SPOT uses one-dimensional projection for supports which limits its capacity to capture a structure of a distribution, especially in high-dimensional settings (Le et al., 2019b; Liutkus et al., 2019).

In this work, we aim to develop an efficient and scalable approach for the transport problem when input measures have different masses. Inspired by the tree-sliced Wasserstein (TSW) distance (Le et al., 2019b) which has fast closed-form computation and remedies the curse of dimensionality for SW, we propose to consider the *entropy partial transport* (EPT) problem with tree metrics. As a high level, our main contribution is three-fold as follows:

- We establish a relationship between the EPT problem with mass constraint and a formulation with Lagrangian multiplier. Then, we employ it to transform the EPT problem to the standard complete OT problem on a suitable one-node extended tree.
- We derive a dual formulation for our EPT problem. We then leverage it to propose a novel regularization which admits a closed-form formula and negative definiteness. Consequently, we introduce positive definite kernels for our regularized EPT. We also derive tree-sliced variants of the regularized EPT for applications without prior knowledge about tree structure for measures.
- We empirically show that (i) our regularization provides both efficient approximations and fast computations, and (ii) the performances of the proposed kernels for our regularized EPT compare favorably with other baselines in applications.

The paper is organized as follow: we review tree metric and introduce important notations in §2. In §3, we develop the theory for EPT with tree metrics and derive an efficient regularization for EPT computation in practice. In §4, we distinguish our approach with other related work in the literature. Then, we evaluate our proposal on document classification with word embeddings and topological data analysis in §5, before giving a conclusion in §6. We have released code for our proposal[1].

2 Preliminaries

Let $\mathcal{T} = (V, E)$ be a tree rooting at node r and with nonnegative edge lengths $\{w_e\}_{e\in E}$, where V is the collection of nodes and E is the collection of edges. For convenience, we use $\mathcal{T}$ to denote the set of all nodes together with all points on its edges[2]. We then recall the definition of tree metric (Semple and Steel, 2003, §7, p.145–182) as follow:

Definition 2.1 (Tree metric)**.** *A metric $d_{\mathcal{T}} : \Omega \times \Omega \to [0, \infty)$ is called a tree metric on Ω if there exists tree $\mathcal{T}$ such that $\Omega \subseteq \mathcal{T}$ and for $x, y \in \Omega$, $d_{\mathcal{T}}(x, y)$ equals to the length of the (unique) path between x and y.*

Assume that V is a subset of a vector space, and let $d_{\mathcal{T}}(\cdot, \cdot)$ be the tree metric on $\mathcal{T}$. Hereafter, the unique shortest path in $\mathcal{T}$ connecting x and y is denoted by $[x, y]$. Let ω be the unique Borel measure (i.e., the length measure) on $\mathcal{T}$ satisfying $\omega([x, y]) = d_{\mathcal{T}}(x, y)$ for all $x, y \in \mathcal{T}$. Given $x \in \mathcal{T}$, the set $\Lambda(x)$ stands for the subtree below x. Precisely,

$$\Lambda(x) := \big\{y \in \mathcal{T} : x \in [r, y]\big\}. \tag{1}$$

We shall use notation $\mathcal{M}(\mathcal{T})$ to represent the set of all nonnegative Borel measures on $\mathcal{T}$ with a finite mass.

[1] https://github.com/lttam/EntropyPartialTransport

[2] Tree $\mathcal{T}$ has a finite number of nodes, but all points on edges can be considered for the tree $\mathcal{T}$ and so the tree includes an infinite number of points.

Also let $C(\mathcal{T})$ be the set of all continuous functions on $\mathcal{T}$, while $L^\infty(\mathcal{T})$ be the collection of all Borel measurable functions on $\mathcal{T}$ that are bounded ω-a.e. Then, $L^\infty(\mathcal{T})$ is a Banach space under the norm

$$\|f\|_{L^\infty(\mathcal{T})} := \inf\{a \in \mathbb{R} : |f(x)| \le a \text{ for } \omega\text{-a.e. } x \in \mathcal{T}\}.$$

3 Entropy Partial Transport (EPT) with Tree Metrics

Let $b \ge 0$ be a constant, $c : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ be a continuous cost with $c(x,x) = 0$, F_1, $F_2 : [0,\infty) \to (0,\infty)$ be entropy functions which are convex and lower semicontinuous, and let $w_1, w_2 : \mathcal{T} \to [0,\infty)$ be two nonnegative weights. For $\mu, \nu \in \mathcal{M}(\mathcal{T})$, consider the set

$$\Pi_\le(\mu,\nu) := \Big\{\gamma \in \mathcal{M}(\mathcal{T}\times\mathcal{T}) : \gamma_1 \le \mu,\, \gamma_2 \le \nu\Big\}$$

with γ_i $(i = 1,2)$ denoting the i^{th} marginal of the measure γ. For $\gamma \in \Pi_\le(\mu,\nu)$, the Radon-Nikodym derivatives of γ_1 w.r.t. μ and of γ_2 w.r.t. ν exist due to $\gamma_1 \le \mu$ and $\gamma_2 \le \nu$. From now on, we let f_1 and f_2 respectively denote these Radon-Nikodym derivatives, i.e., $\gamma_1 = f_1\mu$ and $\gamma_2 = f_2\nu$. Then $0 \le f_1 \le 1$ μ-a.e. and $0 \le f_2 \le 1$ ν-a.e. Throughout the paper, $\bar{m}$ stands for the minimum of the total masses of μ and ν. That is, $\bar{m} := \min\{\mu(\mathcal{T}), \nu(\mathcal{T})\}$. Inspired by Caffarelli and McCann (2010); Liero et al. (2018), we fix a number $m \in [0,\bar{m}]$ and consider the following EPT problem:

$$\mathcal{W}_{c,m}(\mu,\nu) := \inf_{\gamma\in\Pi_\le(\mu,\nu),\,\gamma(\mathcal{T}\times\mathcal{T})=m} \Big[\mathcal{F}_1(\gamma_1|\mu) + \mathcal{F}_2(\gamma_2|\nu) + b\int_{\mathcal{T}\times\mathcal{T}} c(x,y)\gamma(dx,dy)\Big], \quad (2)$$

where $\mathcal{F}_1(\gamma_1|\mu) := \int_{\mathcal{T}} w_1(x)F_1(f_1(x))\mu(dx)$ and $\mathcal{F}_2(\gamma_2|\nu) := \int_{\mathcal{T}} w_2(x)F_2(f_2(x))\nu(dx)$ are the weighted relative entropies. The role of the two entropies in the minimization problem is to force the marginals of γ close to μ and ν respectively. Let us introduce a Lagrange multiplier $\lambda \in \mathbb{R}$ conjugate to the constraint $\gamma(\mathcal{T}\times\mathcal{T}) = m$. As a result, we instead study the following formulation

$$\mathrm{ET}_{c,\lambda}(\mu,\nu) := \inf_{\gamma\in\Pi_\le(\mu,\nu)} \Big[\mathcal{F}_1(\gamma_1|\mu) + \mathcal{F}_2(\gamma_2|\nu) + b\int_{\mathcal{T}\times\mathcal{T}} [c(x,y)-\lambda]\gamma(dx,dy)\Big].$$

In this paper, we focus on the specific entropy functions $F_1(s) = F_2(s) = |s-1|$. Thus, the quantity of interest becomes

$$\mathrm{ET}_{c,\lambda}(\mu,\nu) = \inf_{\gamma\in\Pi_\le(\mu,\nu)} \mathcal{C}_\lambda(\gamma), \quad (3)$$

where $\mathcal{C}_\lambda(\gamma)$ is defined as follow:

$$\begin{aligned}\mathcal{C}_\lambda(\gamma) &:= \int_{\mathcal{T}} w_1[1-f_1(x)]\mu(dx) + \int_{\mathcal{T}} w_2[1-f_2(x)]\nu(dx) + b\int_{\mathcal{T}\times\mathcal{T}} [c(x,y)-\lambda]\gamma(dx,dy)\\ &= \int_{\mathcal{T}} w_1\mu(dx) + \int_{\mathcal{T}} w_2\nu(dx) - \int_{\mathcal{T}} w_1\gamma_1(dx) - \int_{\mathcal{T}} w_2\gamma_2(dx) + b\int_{\mathcal{T}\times\mathcal{T}} [c(x,y)-\lambda]\gamma(dx,dy). \quad (4)\end{aligned}$$

Notice that problem (3) is a generalization of the *generalized Wasserstein distance* $\mathcal{W}_1^{a,b}(\mu,\nu)$ introduced in (Piccoli and Rossi, 2014, 2016). We next display some relationships between problem (2) with mass constraint m and problem (3) with Lagrange multiplier λ. For this, let $\Gamma^0(\lambda)$ denote the set of all optimal plans (i.e., minimizers γ) for $\mathrm{ET}_{c,\lambda}(\mu,\nu)$. Then, since $\mathcal{C}_\lambda(\gamma)$ is an affine function of $\gamma \in \Pi_\le(\mu,\nu)$, the set $\Gamma^0(\lambda)$ is a nonempty convex set. Indeed, for any $\tilde{\gamma}, \hat{\gamma} \in \Gamma^0(\lambda)$ and for any $t \in [0,1]$ we have $(1-t)\tilde{\gamma} + t\hat{\gamma} \in \Gamma^0(\lambda)$ due to $\mathcal{C}_\lambda((1-t)\tilde{\gamma} + t\hat{\gamma}) = (1-t)\mathcal{C}_\lambda(\tilde{\gamma}) + t\mathcal{C}_\lambda(\hat{\gamma}) \le (1-t)\mathcal{C}_\lambda(\gamma) + t\mathcal{C}_\lambda(\gamma) = \mathcal{C}_\lambda(\gamma)$ for every $\gamma \in \Pi_\le(\mu,\nu)$. The following result extends (Caffarelli and McCann, 2010, Corollary 2.1) and reveals the connection between problem (2) and problem (3).

Theorem 3.1. *Let $u(\lambda) := -\mathrm{ET}_{c,\lambda}(\mu,\nu)$ for $\lambda \in \mathbb{R}$, and denote*

$$\partial u(\lambda) := \Big\{p \in \mathbb{R} : u(t) \ge u(\lambda) + p(t-\lambda), \forall t \in \mathbb{R}\Big\}$$

for the set of all subgradients of u at λ. Also, set $\partial u(\mathbb{R}) := \cup_{\lambda\in\mathbb{R}}\partial u(\lambda)$. Then, we have

i) u is a convex function on $\mathbb{R}$, and

$$\partial u(\lambda) = \big\{b\,\gamma(\mathcal{T}\times\mathcal{T}) : \gamma \in \Gamma^0(\lambda)\big\} \quad \forall\lambda \in \mathbb{R}.$$

Also if $\lambda_1 < \lambda_2$, then $m_1 \le m_2$ for every $m_1 \in \partial u(\lambda_1)$ and $m_2 \in \partial u(\lambda_2)$.

ii) u is differentiable at λ if and only if every optimal plan in $\Gamma^0(\lambda)$ has the same mass. When this happens, we in addition have $u'(\lambda) = b\,\gamma(\mathcal{T}\times\mathcal{T})$ for any $\gamma \in \Gamma^0(\lambda)$.

iii) If there exists a constant $M > 0$ such that $w_1(x) + w_2(y) \le b\,c(x,y) + M$ for all $x, y \in \mathcal{T}$, then $\partial u(\mathbb{R}) = [0, b\,\bar{m}]$. Moreover, $u(\lambda) = -\int_{\mathcal{T}} w_1\mu(dx) - \int_{\mathcal{T}} w_2\nu(dx)$ when $\lambda < -M$, and $u'(\lambda) = b\,\bar{m}$ for $\lambda > \|c\|_{L^\infty(\mathcal{T}\times\mathcal{T})}$.

Proof is placed in the Supplementary (§A.1). For any $m \in [0,\bar{m}]$, part iii) of Theorem 3.1 implies that there exists $\lambda \in \mathbb{R}$ such that $b\,m \in \partial u(\lambda)$. It then follows from part i) of this theorem that $m = \gamma^*(\mathcal{T}\times\mathcal{T})$ for

some $\gamma^* \in \Gamma^0(\lambda)$. It is also clear that this γ^* is an optimal plan for $\mathcal{W}_{c,m}(\mu,\nu)$, and

$$\mathcal{W}_{c,m}(\mu,\nu) = \mathrm{ET}_{c,\lambda}(\mu,\nu) + \lambda b\, m.$$

Thus solving the auxiliary problem (3) gives us a solution to the original problem (2). When u is differentiable, the relation between m and λ is given explicitly as $u'(\lambda) = b\,m$. Note that the above selection of λ is unique only if the function u is strictly convex. Nevertheless, it enjoys the following monotonicity regardless of the uniqueness: if $m_1 < m_2$, then $\lambda_1 \le \lambda_2$. Indeed, we have $m_1 = \gamma^1(\mathcal{T}\times\mathcal{T})$ and $m_2 = \gamma^2(\mathcal{T}\times\mathcal{T})$ for some $\gamma^1 \in \Gamma^0(\lambda_1)$ and $\gamma^2 \in \Gamma^0(\lambda_2)$. Since $\gamma^1(\mathcal{T}\times\mathcal{T}) < \gamma^2(\mathcal{T}\times\mathcal{T})$, one has $\lambda_1 \le \lambda_2$ by i) of Theorem 3.1.

To investigate problem (3), we recast it as the standard complete OT problem by using an observation in (Caffarelli and McCann, 2010). More precisely, let $\hat{s}$ be a point outside $\mathcal{T}$ and consider the set $\hat{\mathcal{T}} := \mathcal{T} \cup \{\hat{s}\}$. We next extend the cost function to $\hat{\mathcal{T}}\times\hat{\mathcal{T}}$ as follow

$$\hat{c}(x,y) := \begin{cases} b[c(x,y) - \lambda] & \text{if } x,y \in \mathcal{T},\\ w_1(x) & \text{if } x \in \mathcal{T} \text{ and } y = \hat{s},\\ w_2(y) & \text{if } x = \hat{s} \text{ and } y \in \mathcal{T},\\ 0 & \text{if } x = y = \hat{s}. \end{cases}$$

The measures μ,ν are extended accordingly by adding a Dirac mass at the isolated point $\hat{s}$: $\hat{\mu} = \mu + \nu(\mathcal{T})\delta_{\hat{s}}$ and $\hat{\nu} = \nu + \mu(\mathcal{T})\delta_{\hat{s}}$. As $\hat{\mu},\hat{\nu}$ have the same total mass on $\hat{\mathcal{T}}$, we can consider the standard complete OT problem between $\hat{\mu},\hat{\nu}$ as follow

$$\mathrm{KT}(\hat{\mu},\hat{\nu}) := \inf_{\hat{\gamma}\in\Gamma(\hat{\mu},\hat{\nu})} \int_{\hat{\mathcal{T}}\times\hat{\mathcal{T}}} \hat{c}(x,y)\hat{\gamma}(dx,dy), \qquad (5)$$

where $\Gamma(\hat{\mu},\hat{\nu}) := \Big\{\hat{\gamma} \in \mathcal{M}(\hat{\mathcal{T}}\times\hat{\mathcal{T}}) : \hat{\mu}(U) = \hat{\gamma}(U\times\hat{\mathcal{T}}),\ \hat{\nu}(U) = \hat{\gamma}(\hat{\mathcal{T}}\times U) \text{ for all Borel sets } U \subset \hat{\mathcal{T}}\Big\}$.

A one-to-one correspondence between $\gamma \in \Pi_{\le}(\mu,\nu)$ and $\hat{\gamma} \in \Gamma(\hat{\mu},\hat{\nu})$ is given by

$$\hat{\gamma} = \gamma + [(1-f_1)\mu]\otimes\delta_{\hat{s}} + \delta_{\hat{s}}\otimes[(1-f_2)\nu] + \gamma(\mathcal{T}\times\mathcal{T})\delta_{(\hat{s},\hat{s})}. \qquad (6)$$

Indeed, if $\gamma \in \Pi_{\le}(\mu,\nu)$, then it is clear that $\hat{\gamma}$ defined by (6) satisfies $\hat{\gamma} \in \Gamma(\hat{\mu},\hat{\nu})$. The converse is guaranteed by the next technical result.

Lemma 3.2. *For $\hat{\gamma} \in \Gamma(\hat{\mu},\hat{\nu})$, let γ be the restriction of $\hat{\gamma}$ to $\mathcal{T}$. Then, relation* (6) *holds and $\gamma \in \Pi_{\le}(\mu,\nu)$.*

Proof is placed in the Supplementary (§A.2).

These observations in particular display the following connection between the EPT problem and the standard complete OT problem.

Proposition 3.3 (EPT versus complete OT)**.** *For every $\mu,\nu \in \mathcal{M}(\mathcal{T})$, we have $\mathrm{ET}_{c,\lambda}(\mu,\nu) = \mathrm{KT}(\hat{\mu},\hat{\nu})$. Moreover, relation* (6) *gives a one-to-one correspondence between optimal solution γ for EPT problem* (3) *and optimal solution $\hat{\gamma}$ for standard complete OT problem* (5)*.*

Proof is placed in the Supplementary (§A.3).

3.1 Dual Formulations

The relationship given in Proposition 3.3 allows us to obtain the dual formulation of EPT in problem (3) from that of problem (5) proved in (Caffarelli and McCann, 2010, Corollary 2.6).

Theorem 3.4 (Dual formula for general cost)**.** *For any $\lambda \ge 0$ and nonnegative weights $w_1(x), w_2(x)$, we have*

$$\mathrm{ET}_{c,\lambda}(\mu,\nu) = \sup_{(u,v)\in\mathbb{K}} \Big[\int_{\mathcal{T}} u(x)\mu(dx) + \int_{\mathcal{T}} v(x)\nu(dx)\Big],$$

where $\mathbb{K} := \Big\{(u,v) :\ u \le w_1,\ -b\lambda + \inf_{x\in\mathcal{T}}[b\,c(x,y) - w_1(x)] \le v(y) \le w_2(y),\ u(x)+v(y) \le b[c(x,y)-\lambda]\Big\}$.

Proof is placed in the Supplementary (§A.4).

This dual formula is *our main theoretical result* which leads to our *novel efficient regularization* for the EPT (see §3.2), and can be rewritten more explicitly when the cost function c is the tree metric. Hereafter, we use $c(x,y) = d_{\mathcal{T}}(x,y)$. To ease the notations, we simply write $\mathrm{ET}_{\lambda}(\mu,\nu)$ for $\mathrm{ET}_{d_{\mathcal{T}},\lambda}(\mu,\nu)$.

Corollary 3.5 (Dual formula for tree metric)**.** *Assume that $\lambda \ge 0$ and the nonnegative weights w_1, w_2 are b-Lipschitz w.r.t. $d_{\mathcal{T}}$. Then, we have*

$$\mathrm{ET}_{\lambda}(\mu,\nu) = \sup\big\{\textstyle\int_{\mathcal{T}} f(\mu-\nu) :\ f \in \mathbb{L}\big\} - \tfrac{b\lambda}{2}\big[\mu(\mathcal{T}) + \nu(\mathcal{T})\big], \qquad (7)$$

where $\mathbb{L} := \Big\{f \in C(\mathcal{T}) :\ -w_2 - \frac{b\lambda}{2} \le f \le w_1 + \frac{b\lambda}{2},\ |f(x)-f(y)| \le b\,d_{\mathcal{T}}(x,y)\Big\}$.

Proof is placed in the Supplementary (§A.5).

Corollary 3.5 extends the dual formulation for the *generalized Wasserstein distance* $\mathcal{W}_1^{a,b}(\mu,\nu)$ proved in (Piccoli and Rossi, 2016, Theorem 2) and (Chung and Trinh, 2019). In the next section, we will leverage (7) to propose an effective regularization for computation in practice.

Remark 3.6. *An example of b-Lipschitz weight is $w(x) = a_1\, d_{\mathcal{T}}(x,x_0) + a_0$ for some $x_0 \in \mathcal{T}$ and for some constants $a_1 \in [0,b]$ and $a_0 \in [0,\infty)$.*

As a consequence of the dual formulation, we obtain the following geometric properties:

Proposition 3.7 (Geometric structures of metric d)**.** *Assume that $\lambda \geq 0$ and the weights w_1, w_2 are positive and b-Lipschitz w.r.t. $d_{\mathcal{T}}$. Define $d(\mu,\nu) := \mathrm{ET}_\lambda(\mu,\nu) + \frac{b\lambda}{2}\big[\mu(\mathcal{T}) + \nu(\mathcal{T})\big]$. Then, we have*

i) $d(\mu+\sigma, \nu+\sigma) = d(\mu,\nu)$, $\forall \sigma \in \mathcal{M}(\mathcal{T})$.

ii) d is a divergence and satisfies the triangle inequality $d(\mu,\nu) \leq d(\mu,\sigma) + d(\sigma,\nu)$.

iii) If in addition $w_1 = w_2$, then $(\mathcal{M}(\mathcal{T}), d)$ is a complete metric space. Moreover, it is a geodesic space in the sense that for every two points μ and ν in $\mathcal{M}(\mathcal{T})$ there exists a path $\varphi : [0,a] \to \mathcal{M}(\mathcal{T})$ with $a := d(\mu,\nu)$ such that $\varphi(0) = \mu$, $\varphi(a) = \nu$, and

$$d(\varphi(t), \varphi(s)) = |t-s| \quad \textit{for all } t, s \in [0,a].$$

Proof is placed in the Supplementary (§A.6).

Let $m \in [0, \bar{m}]$, and choose $\lambda \geq 0$ such that there exists an optimal plan γ^0 for $\mathrm{ET}_\lambda(\mu,\nu)$ with $\gamma^0(\mathcal{T}\times\mathcal{T}) = m$. As pointed out right after Theorem 3.1, this choice of λ is possible. Then, the proof of Lemma A.1 in the Supplementary (§A.6) shows that

$$\inf_{\gamma \in \Pi_\leq(\mu,\nu),\, \gamma(\mathcal{T}\times\mathcal{T}) = m} \Big[\mathcal{F}_1(\gamma_1|\mu) + \mathcal{F}_2(\gamma_2|\nu) + b\int_{\mathcal{T}\times\mathcal{T}} c(x,y)\gamma(dx,dy)\Big] \leq d(\mu,\nu).$$

Moreover, the equality happens if and only if there exists an optimal plan γ^0 for $\mathrm{ET}_\lambda(\mu,\nu)$ such that $m = \gamma^0(\mathcal{T}\times\mathcal{T}) = \frac{1}{2}[\mu(\mathcal{T}) + \nu(\mathcal{T})]$. The necessary conditions for the latter one to hold are $\mu(\mathcal{T}) = \nu(\mathcal{T})$ and $m = \bar{m}$.

3.2 An Efficient Regularization for Entropy Partial Transport with Tree Metrics

First observe that any $f \in \mathbb{L}$ can be represented by

$$f(x) = f(r) + \int_{[r,x]} g(y)\omega(dy)$$

for some function $g \in L^\infty(\mathcal{T})$ with $\|g\|_{L^\infty(\mathcal{T})} \leq b$. Note that condition $|f(x) - f(y)| \leq b\, d_{\mathcal{T}}(x,y)$ is equivalent to $\|g\|_{L^\infty(\mathcal{T})} \leq b$. It follows that $\mathbb{L} \subset \mathbb{L}_0$, where we define for $0 \leq \alpha \leq \frac{1}{2}[b\lambda + w_1(r) + w_2(r)]$ that $\mathbb{L}_\alpha$ is the collection of all functions f of the form

$$f(x) = s + \int_{[r,x]} g(y)\omega(dy),$$

with s being a constant in the interval $\Big[-w_2(r) - \frac{b\lambda}{2} + \alpha, w_1(r) + \frac{b\lambda}{2} - \alpha\Big]$ and with $\|g\|_{L^\infty(\mathcal{T})} \leq b$. This leads us to consider the following *regularization* for $\mathrm{ET}_\lambda(\mu,\nu)$:

$$\widetilde{\mathrm{ET}}^\alpha_\lambda(\mu,\nu) := \sup\big\{\textstyle\int_{\mathcal{T}} f(\mu-\nu) :\ f \in \mathbb{L}_\alpha\big\} - \tfrac{b\lambda}{2}\big[\mu(\mathcal{T}) + \nu(\mathcal{T})\big]. \quad (8)$$

Especially, when $\alpha = 0$ and notice that $\mathbb{L} \subset \mathbb{L}_0$, $\widetilde{\mathrm{ET}}^0_\lambda(\mu,\nu)$ is an upper bound of $\mathrm{ET}_\lambda(\mu,\nu)$ through the dual formulation. The next result gives a *closed-form formula* for $\widetilde{\mathrm{ET}}^\alpha_\lambda(\mu,\nu)$ and is *our main formula used for computation in practice.*

Proposition 3.8 (closed-form for regularized EPT)**.** *Assume that $\lambda, w_1(r), w_2(r)$ are nonnegative numbers. Then, for $0 \leq \alpha \leq \frac{1}{2}[b\lambda + w_1(r) + w_2(r)]$, we have*

$$\widetilde{\mathrm{ET}}^\alpha_\lambda(\mu,\nu) = \textstyle\int_{\mathcal{T}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|\,\omega(dx) - \frac{b\lambda}{2}\big[\mu(\mathcal{T}) + \nu(\mathcal{T})\big] + \big[w_i(r) + \frac{b\lambda}{2} - \alpha\big]|\mu(\mathcal{T}) - \nu(\mathcal{T})|$$

with $i := 1$ if $\mu(\mathcal{T}) \geq \nu(\mathcal{T})$ and $i := 2$ if $\mu(\mathcal{T}) < \nu(\mathcal{T})$. In particular, the map $\alpha \longmapsto \widetilde{\mathrm{ET}}^\alpha_\lambda(\mu,\nu)$ is nonincreasing and

$$|\widetilde{\mathrm{ET}}^{\alpha_1}_\lambda(\mu,\nu) - \widetilde{\mathrm{ET}}^{\alpha_2}_\lambda(\mu,\nu)| = |\alpha_1 - \alpha_2||\mu(\mathcal{T}) - \nu(\mathcal{T})|.$$

Proof is placed in the Supplementary (§A.7).

It is also possible to use $\widetilde{\mathrm{ET}}^\alpha_\lambda(\mu,\nu)$ to *upper or lower bound* the distance $\mathrm{ET}_\lambda(\mu,\nu)$ as follows:

Proposition 3.9 (Bounds for ET_λ with $\widetilde{\mathrm{ET}}^\alpha_\lambda$)**.** *Assume that $\lambda \geq 0$ and the weights w_1, w_2 are b-Lipschitz w.r.t. $d_{\mathcal{T}}$. Then,*

$$\mathrm{ET}_\lambda(\mu,\nu) \leq \widetilde{\mathrm{ET}}^0_\lambda(\mu,\nu).$$

In addition, if $[4L_{\mathcal{T}} - \lambda]b \leq w_1(r) + w_2(r)$ where $L_{\mathcal{T}} := \max_{x\in\mathcal{T}} \omega([r,x])$, then

$$\widetilde{\mathrm{ET}}^\alpha_\lambda(\mu,\nu) \leq \mathrm{ET}_\lambda(\mu,\nu),$$

for every $2bL_{\mathcal{T}} \leq \alpha \leq \frac{1}{2}[b\lambda + w_1(r) + w_2(r)]$.

Proof is placed in the Supplementary (§A.8).

Analogous to Proposition 3.7, we obtain:

Proposition 3.10 (Geometric structures of regularized metric d_α)**.** *Assume that $\lambda, w_1(r), w_2(r)$ are nonnegative numbers. For $0 \leq \alpha < \frac{b\lambda}{2} + \min\{w_1(r), w_2(r)\}$, define*

$$d_\alpha(\mu,\nu) := \widetilde{\mathrm{ET}}^\alpha_\lambda(\mu,\nu) + \frac{b\lambda}{2}\big[\mu(\mathcal{T}) + \nu(\mathcal{T})\big]. \quad (9)$$

Then, we have

i) $d_\alpha(\mu+\sigma, \nu+\sigma) = d_\alpha(\mu,\nu)$, $\forall \sigma \in \mathcal{M}(\mathcal{T})$.

ii) d_α *is a divergence and satisfies the triangle inequality* $d_\alpha(\mu,\nu) \leq d_\alpha(\mu,\sigma) + d_\alpha(\sigma,\nu)$.

iii) If in addition $w_1(r) = w_2(r)$*, then* $(\mathcal{M}(\mathcal{T}), d_\alpha)$ *is a complete metric space. Moreover, it is a geodesic space in the sense defined in part iii) of Proposition 3.7 but with* d_α *replacing* d.

Proof is placed in the Supplementary (§A.9).

Our next result about *negative definiteness* is a cornerstone to build *positive definite kernels* upon either $\widetilde{\mathrm{ET}}_\lambda^\alpha$ or d_α for kernel-dependent frameworks.

Proposition 3.11 (Negative definiteness)**.** *With the same assumptions as in Proposition 3.8 for* $\widetilde{\mathrm{ET}}_\lambda^\alpha$ *and in Proposition 3.10 for* d_α*, both* $\widetilde{\mathrm{ET}}_\lambda^\alpha$ *and* d_α *are negative definite.*

Proof is placed in the Supplementary (§A.10).

From Proposition 3.11 and following (Berg et al., 1984, Theorem 3.2.2, p.74), given $t > 0$, the kernels $k_{\widetilde{\mathrm{ET}}_\lambda^\alpha}(\mu,\nu) := \exp\left(-t\widetilde{\mathrm{ET}}_\lambda^\alpha(\mu,\nu)\right)$ and $k_{d_\alpha}(\mu,\nu) := \exp\left(-t d_\alpha(\mu,\nu)\right)$ are positive definite.

3.3 Tree-sliced Variants by Sampling Tree Metrics

In most of practical applications, we usually do not have prior knowledge about tree structure for measures. Therefore, we need to choose or sample tree metrics from support data points for a given task. We use the tree metric sampling methods in (Le et al., 2019b, §4): (i) *partition-based tree metric sampling* for a low-dimensional space, or (ii) *clustering-based tree metric sampling* for a high-dimensional space. Moreover, those tree metric sampling methods are not only fast for computation[3], but also adaptive to the distribution of supports. We further propose the tree-sliced variants of the regularized EPT, computed by averaging the regularized EPT using those randomly sampled tree metrics. One advantage is to reduce the quantization effects or cluster sensitivity problems (i.e, support data points are quantized, or clustered into an adjacent hypercube, or cluster respectively) within the tree metric sampling procedure.

Although one can leverage tree metrics to approximate arbitrary metrics (Bartal, 1996, 1998; Charikar et al., 1998; Indyk, 2001; Fakcharoenphol et al., 2004), our goal is rather to sample tree metrics and use them as ground metrics in the regularized EPT, similar to TSW. Despite the fact that one-dimensional projections do not have interesting properties in terms of distortion viewpoints, they remain useful for SPOT (or SW, sliced-Gromov-Wasserstein (Vayer et al., 2019)). In the same vein, we believe that trees with high distortion are still useful for EPT, similar as in TSW. Moreover, one may not need to spend excessive effort to optimize ET_λ (in Equation (7)) for a randomly sampled tree metric since it can lead to overfitting within the computation of the EPT itself. Therefore, the proposed efficient regularization of EPT (e.g, $\widetilde{\mathrm{ET}}_\lambda^\alpha$ in Equation (8)) is not only fast for computation (i.e., closed-form), but also gives a benefit to overcome the overfitting problem within the computation of the EPT.

[3] E.g., the complexity of the clustering-based tree metric is $\mathcal{O}(H_\mathcal{T} m \log \kappa)$ when we set κ clusters for the farthest-point clustering (Gonzalez, 1985), and $H_\mathcal{T}$ for the predefined tree deepest level for m input support data points.

4 Discussion and Related Work

One can leverage tree metrics to approximate arbitrary metrics for speeding up a computation (Bartal, 1996, 1998; Charikar et al., 1998; Indyk, 2001; Fakcharoenphol et al., 2004). For instances, (i) Indyk and Thaper (2003) applied tree metrics (e.g., quadtree) to approximate OT with Euclidean cost metric for a fast image retrieval. (ii) Sato et al. (2020) considered a generalized Kantorovich-Rubinstein discrepancy (Hanin, 1992; Guittet, 2002; Lellmann et al., 2014) with general weights for unbalanced OT, and used a quadtree as in (Indyk and Thaper, 2003) to approximate the proposed distance via a dynamic programming with infinitely many states. They then derived an efficient algorithm with a quasi-linear time complexity to speed up the dynamic programming computation by leveraging high-level programming techniques. However, such approximations following the approach of Indyk and Thaper (2003) result in large distortions in high dimensional spaces (Naor and Schechtman, 2007).

Tree metrics are also leveraged for several advanced OT problems, e.g., tree-Wasserstein barycenters (Le et al., 2019a); or a variant of Gromov-Wasserstein (a.k.a., flow-based alignment approaches) (Le et al., 2021). Additionally, ultrametric, a special case of tree metrics, is also utilized on Gromov-Wasserstein (Mémoli et al., 2021) and Gromov-Hausdoff (Mémoli et al., 2019) for metric measure spaces.

5 Experiments

In this section, we first illustrate that $\widetilde{\mathrm{ET}}_\lambda^\alpha$ (Equation (8)) is an efficient approximation for ET_λ (Equation (7)). Then, we evaluate our proposed $\widetilde{\mathrm{ET}}_\lambda^\alpha$ and d_α (Equation (9)) for comparing measures in document classification with word embedding and topological data analysis (TDA). Experiments are evaluated with Intel Xeon CPU E7-8891v3 2.80GHz and 256GB RAM.

Documents with word embedding. We consider each document as a measure where each word is regarded as a point in the support with a unit mass. Following Kusner et al. (2015); Le et al. (2019b), we applied the *word2vec* word embedding (Mikolov et al., 2013), pretrained on Google News[4] containing about 3 millions words/phrases. Each word/phrase in a document is mapped into a vector in $\mathbb{R}^{300}$. We removed all SMART stop word (Salton and Buckley, 1988), and dropped words in documents if they are not available in the pretrained *word2vec*.

Geometric structured data via persistence diagrams (PD) in TDA. TDA has recently emerged in machine learning community as a powerful tool to analyze geometric structured data such as material data, or linked twist maps (Adams et al., 2017; Lacombe et al., 2018; Le and Yamada, 2018). TDA applies algebraic topology methods (e.g., persistence homology) to extract robust topological features (e.g., connected components, rings, cavities) and output a multiset of 2-dimensional points (i.e., PD). The coordinates of a 2-dimensional point in PD are corresponding to the birth and death time of a particular topological feature. Therefore, each point in PD summarizes a life span of a topological feature. We can regard PD as measures where each 2-dimensional point is considered as a point in the support with a unit mass.

Tree metric sampling. In our experiments, we do not have prior knowledge about tree metrics for neither word embeddings in documents nor 2-dimensional points in PD. To compute the EPT (e.g., $\widetilde{\mathrm{ET}}^{\alpha}_{\lambda}$ and d_α), we considered n_s randomized tree metrics. We employed the clustering-based tree metric sampling for word embeddings in documents (i.e., high-dimensional space $\mathbb{R}^{300}$), while we used the partition-based tree metric sampling for 2-dimensional points in PD (i.e., low-dimensional space $\mathbb{R}^2$). Those tree metric sampling methods are built with a predefined deepest level $H_{\mathcal{T}}$ of tree $\mathcal{T}$ as a stopping condition as in (Le et al., 2019b).

Baselines and setup. We considered 2 typical baselines based on OT theory for measures with different masses: (i) Sinkhorn-UOT (Frogner et al., 2015; Chizat et al., 2018) (i.e., entropic regularization approach), and (ii) SPOT (Bonneel and Coeurjolly, 2019) (i.e., sliced-formula approach based on 1-dimensional projection). Following Le et al. (2019b), we apply the kernel approach in the form $\exp(-t\bar{d})$ with SVM for document classification with word embedding. Here, $\bar{d}$ is a discrepancy between measures and $t > 0$. We also employed this kernel approach for various tasks in TDA, e.g., orbit recognition and object shape classification with SVM, as well as change point detection for material data analysis with kernel Fisher discriminant ratio (KFDR) (Harchaoui et al., 2009). While kernels for $\widetilde{\mathrm{ET}}^{\alpha}_{\lambda}$ and d_α are positive definite, kernels for Sinkhorn-UOT and SPOT are empirically indefinite[5]. When kernels are indefinite, we regularized for the corresponding Gram matrices by adding a sufficiently large diagonal term as in (Cuturi, 2013; Le et al., 2019b). For SVM, we randomly split each dataset into 70%/30% for training and test with 10 repeats. Typically, we choose hyper-parameters via cross validation, choose $1/t$ from $\{q_{10}, q_{20}, q_{50}\}$ where q_s is the $s\%$ quantile of a subset of corresponding discrepancies observed on a training set, use 1-vs-1 strategy with Libsvm[6] for multi-class classification, and choose SVM regularization from $\{10^{-2:1:2}\}$. For Sinkhorn-UOT, we select the entropic regularization from $\{0.01, 0.05, 0.1, 0.5, 1\}$. Following Proposition 3.9, we take $\alpha = 0$ for $\widetilde{\mathrm{ET}}^{\alpha}_{\lambda}$ and d_α in all our experiments.

5.1 Efficient Approximation of $\widetilde{\mathrm{ET}}^{0}_{\lambda}$ for ET_λ

We randomly sample 500K pairs of documents in `TWITTER` dataset. Following Proposition 3.3, we compute ET_λ via the corresponding KT (Equation (5)). Our goal is to compare $\widetilde{\mathrm{ET}}^{0}_{\lambda}$ to ET_λ.

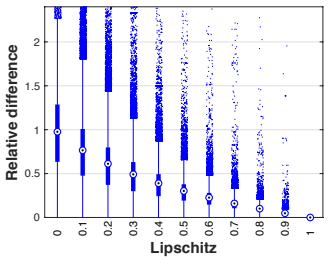

Figure 1: Relative difference between $\widetilde{\mathrm{ET}}^{0}_{\lambda}$ and ET_λ w.r.t. Lipschitz const. of w_1, w_2.

Change Lipschitz constants. We choose $w_1(x) = w_2(x) = a_1 d_{\mathcal{T}}(r, x) + a_0$, and set $\lambda = b = 1$, $a_0 = 1$. In particular, $a_1 \in [0, b]$ since w_1, w_2 are b-Lipschitz functions (see Corollary 3.5 and Remark 3.6). We illustrate the relative difference $(\widetilde{\mathrm{ET}}^{0}_{\lambda} - \mathrm{ET}_\lambda)/\mathrm{ET}_\lambda$ when a_1 is changed in $[0, b]$ in Figure 1. We observe that when a_1 is close to b (i.e., the Lipschitz constants of w_1, w_2 are close to b), $\widetilde{\mathrm{ET}}^{0}_{\lambda}$ becomes closer to ET_λ. When $a_1 = b$, the values of $\widetilde{\mathrm{ET}}^{0}_{\lambda}$ is almost identical to ET_λ.

Change λ. From the results in Figure 1, we set $a_1 = b$ to investigate the relative different between $\widetilde{\mathrm{ET}}^{0}_{\lambda}$ and ET_λ when λ is changed. As illustrated in Figure 2,

[4] https://code.google.com/p/word2vec

[5] We empirically observed negative eigenvalues in Gram matrices corresponding to kernels for Sinkhorn-UOT and SPOT.

[6] https://www.csie.ntu.edu.tw/∼cjlin/libsvm/

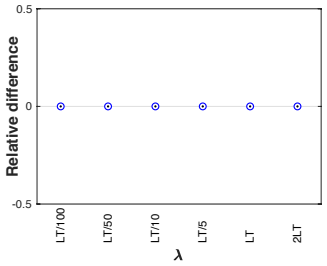

Figure 2: Relative difference between $\widetilde{\mathrm{ET}}_\lambda^0$ and ET_λ w.r.t. λ when $a_1 = b$. (LT := $L_\mathcal{T}$)

$\widetilde{\mathrm{ET}}_\lambda^0$ is almost identical to ET_λ regardless the value of λ when $a_1 = b$.

5.2 Document Classification with Word Embedding

We consider 4 datasets: `TWITTER`, `RECIPE`, `CLASSIC` and `AMAZON` for document classification with word embedding. Statistical characteristics of these datasets are summarized in Figure 3.

5.3 Topological Data Analysis (TDA)

5.3.1 Orbit Recognition

We considered a synthesized dataset, proposed by Adams et al. (2017), for link twist map which is a discrete dynamical system to model flows in DNA microarrays (Hertzsch et al., 2007). There are 5 classes of orbits. Following Le and Yamada (2018), we generated 1000 orbits for each class of orbits, and each orbit has 1000 points. We used the 1-dimensional topological features for PD extracted with Vietoris-Rips complex filtration (Edelsbrunner and Harer, 2008).

5.3.2 Object Shape Classification

We evaluated our approach for object shape classification on a subset of `MPEG7` dataset (Latecki et al., 2000) containing 10 classes where each class has 20 samples as in (Le and Yamada, 2018). For simplicity, we followed Le and Yamada (2018) to extract 1-dimensional topological features for PD with Vietoris-Rips complex filtration[7] (Edelsbrunner and Harer, 2008).

5.3.3 Change Point Detection for Material Analysis

We applied our approach on change point detection for material analysis with KFDR as a statistical score on granular packing system (GPS) (Francois et al., 2013) and SiO_2 (Nakamura et al., 2015) datasets. Statistical characteristics of these datasets are summarized in Figure 5. Following Le and Yamada (2018), we set 10^{-3} for the regularization parameter in KFDR and used the ball model filtration to extract 2-dimensional topological features for PD in GPS dataset, and 1-dimensional topological features for PD in SiO_2 dataset. Note that we omit the baseline kernel for Sinkhorn-UOT in this application since the computation with Sinkhorn-UOT is out of memory.

We illustrate the KFDR graphs in Figure 5. For GPS dataset, all kernel approaches get the change point at the index 23 which supports the observation (corresponding id = 23) in (Anonymous, 1972). For SiO_2 dataset, all kernel approaches get the change point in a supported range ($35 \leq \text{id} \leq 50$), obtained by a traditional physical approach (Elliott, 1983). The KFDR results of kernels corresponding to d_0 and $\widetilde{\mathrm{ET}}_\lambda^0$ compare favorably with those of kernel for SPOT.

5.4 Results of SVM, Time Consumption and Discussions

We illustrate the results of SVM and time consumption of kernel matrices in document classification with word embedding and TDA in Figure 3 and Figure 4 respectively. The performances of kernels for $\widetilde{\mathrm{ET}}_\lambda^0$ and d_0 outperform those of kernels for SPOT. They also outperform those of kernels for Sinkhorn-UOT on TDA, and are comparative on document classification. The fact that SPOT uses the 1-dimensional projection for support data points may limit its ability to capture high-dimensional structure in data distributions (Le et al., 2019b; Liutkus et al., 2019). The regularized EPT remedies this problem by leveraging the tree metrics which have more flexibility and degrees of freedom (e.g., choose a tree rather than a line). In addition, while kernels for $\widetilde{\mathrm{ET}}_\lambda^0$ and d_0 are positive definite, kernels for SPOT and Sinkhorn-UOT are empirically indefinite. The indefiniteness of kernels may affect their performances in some applications, e.g., kernels for Sinkhorn-UOT work well for document classification with word embedding, but perform poorly in TDA applications. There are also similar observations in (Le et al., 2019b). Additionally, we illustrate a trade-off between performances and computational time for different number of (tree) slices in `TWITTER` dataset in Figure 6. The performances are usually improved with more slices, but with a trade-off about the computational time. In applications, we observed that a good trade off is about $n_s = 10$ slices.

Tree metric sampling. Time consumption for the tree metric sampling is negligible in applications. With the predefined tree deepest level $H_\mathcal{T} = 6$ and tree branches $\kappa = 4$ as in (Le et al., 2019b), it took $1.5, 11.0, 17.5, 20.5$ seconds for `TWITTER`, `RECIPE`, `CLASSIC`, `AMAZON` datasets respectively, and $21.0, 0.1$

[7] A more complicated and advanced filtration for this task is considered in (Turner et al., 2014).

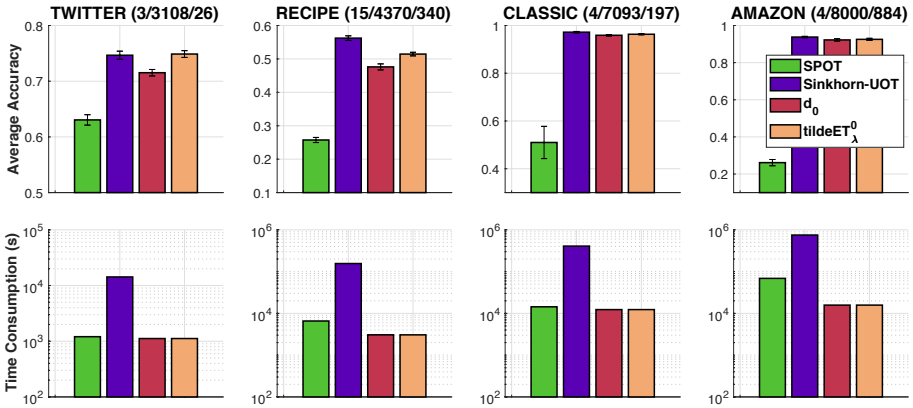

Figure 3: SVM results and time consumption of kernel matrices on document classification. For each dataset, the numbers in the parenthesis are respectively the number of classes, the number of documents, and the maximum number of unique words for each document.

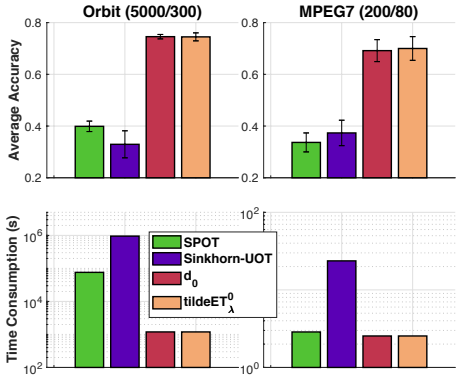

Figure 4: SVM results and time consumption of kernel matrices for TDA. For each dataset, the numbers in the parenthesis are respectively the number of PD, and the maximum number of points in PD.

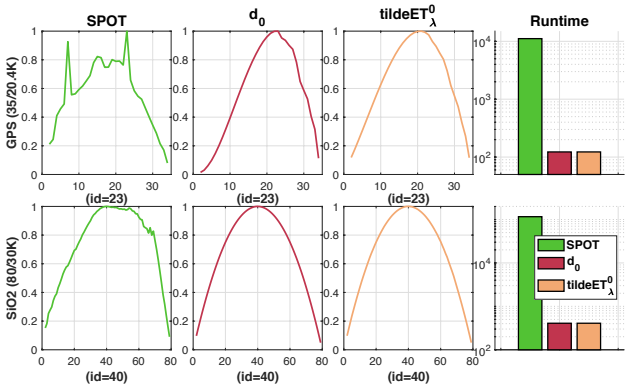

Figure 5: KFDR graphs and time consumption of kernel matrices for change point detection. For each dataset, the numbers in the parenthesis are respectively the number of PDs, and the maximum number of points in PD.

seconds for `Orbit`, `MPEG7` datasets respectively.

$\widetilde{\text{ET}}^0_\lambda$ versus ET_λ. We also compare $\widetilde{\text{ET}}^0_\lambda$ and ET_λ (or KT) in `TWITTER` dataset for document classification, and in `MPEG7` dataset for object shape recognition in TDA. The performances of $\widetilde{\text{ET}}^0_\lambda$ and ET_λ are identical (i.e., their kernel matrices are almost the same for those datasets), but $\widetilde{\text{ET}}^0_\lambda$ is faster than ET_λ about 11 times in `TWITTER` dataset, and 81 times in `MPEG7` dataset when $n_s = 10$ slices.

Further results are placed in the supplementary (§B).

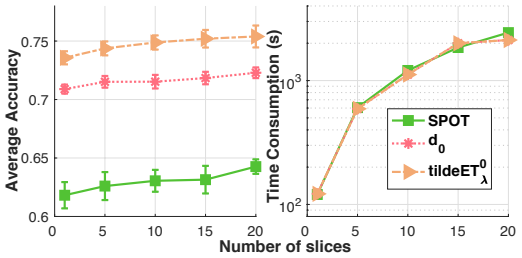

Figure 6: SVM results and time consumption for corresponding kernel matrices in `TWITTER` dataset w.r.t. the number of (tree) slices.

6 Conclusion

We have developed a rigorous theory for the entropy partial transport (EPT) problem for nonnegative measures on a tree having different masses. We show that the EPT problem is equivalent to a standard complete OT problem on a suitable one-node extended tree which allows us to develop its dual formulation. By leveraging the dual problem, we proposed efficient novel regularization for the EPT which yields *closed-form solution* for a fast computation and *negative definiteness*—an important property to build positive definite kernels required in many kernel-dependent frameworks. Moreover, our regularization also provides effective approximations in applications. We further derive tree-sliced variants of the regularized EPT for practical applications without prior knowledge about a tree structure for measures. The question about sampling optimal tree metrics for the tree-sliced variants from data points is left for future work.

Acknowledgements

We thank Nhan-Phu Chung, Nhat Ho for fruitful discussions, and anonymous reviewers for their comments. TL acknowledges the support of JSPS KAKENHI Grant number 20K19873. The research of TN is supported in part by a grant from the Simons Foundation (#318995).

References

Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., Chepushtanova, S., Hanson, E., Motta, F., and Ziegelmeier, L. (2017). Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(1):218–252.

Anonymous (1972). What is random packing? *Nature*, 239:488–489.

Bartal, Y. (1996). Probabilistic approximation of metric spaces and its algorithmic applications. In *Proceedings of 37th Conference on Foundations of Computer Science*, pages 184–193.

Bartal, Y. (1998). On approximating arbitrary metrices by tree metrics. In *ACM Symposium on Theory of Computing (STOC)*, volume 98, pages 161–168.

Benamou, J.-D. (2003). Numerical resolution of an "unbalanced" mass transport problem. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 37(5):851–868.

Berg, C., Christensen, J. P. R., and Ressel, P., editors (1984). *Harmonic analysis on semigroups*. Springer-Verglag, New York.

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

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015). Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45.

Bunne, C., Alvarez-Melis, D., Krause, A., and Jegelka, S. (2019). Learning Generative Models across Incomparable Spaces. In *International Conference on Machine Learning (ICML)*, volume 97.

Caffarelli, L. A. and McCann, R. J. (2010). Free boundaries in optimal transport and monge-ampere obstacle problems. *Annals of mathematics*, pages 673–730.

Charikar, M., Chekuri, C., Goel, A., Guha, S., and Plotkin, S. (1998). Approximating a finite metric by a small number of tree metrics. In *Proceedings 39th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 379–388.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. (2018). Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609.

Chung, N.-P. and Trinh, T.-S. (2019). Duality and quotient spaces of generalized wasserstein spaces. *arXiv preprint arXiv:1904.12461*.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 3730–3739.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300.

Edelsbrunner, H. and Harer, J. (2008). Persistent homology-a survey. *Contemporary mathematics*, 453:257–282.

Elliott, S. R. (1983). Physics of amorphous materials. *Longman Group*.

Fakcharoenphol, J., Rao, S., and Talwar, K. (2004). A tight bound on approximating arbitrary metrics by tree metrics. *Journal of Computer and System Sciences*, 69(3):485–497.

Figalli, A. (2010). The optimal partial transport problem. *Archive for rational mechanics and analysis*, 195(2):533–560.

Francois, N., Saadatfar, M., Cruikshank, R., and Sheppard, A. (2013). Geometrical frustration in amorphous and partially crystallized packings of spheres. *Physical review letters*, 111(14):148001.

Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. (2015). Learning with a wasserstein loss. In *Advances in neural information processing systems*, pages 2053–2061.

Gangbo, W., Li, W., Osher, S., and Puthawala, M. (2019). Unnormalized optimal transport. *Journal of Computational Physics*, 399:108940.

Gonzalez, T. F. (1985). Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306.

Guittet, K. (2002). Extended kantorovich norms: a tool for optimization. *INRIA report*.

Hanin, L. G. (1992). Kantorovich-rubinstein norm and its application in the theory of lipschitz spaces. *Proceedings of the American Mathematical Society*, 115(2):345–352.

Harchaoui, Z., Moulines, E., and Bach, F. R. (2009). Kernel change-point analysis. In *Advances in neural information processing systems*, pages 609–616.

Hertzsch, J.-M., Sturman, R., and Wiggins, S. (2007). Dna microarrays: design principles for maximizing ergodic, chaotic mixing. *Small*, 3(2):202–218.

Indyk, P. (2001). Algorithmic applications of low-distortion geometric embeddings. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 10–33.

Indyk, P. and Thaper, N. (2003). Fast image retrieval via embeddings. In *International workshop on statistical and computational theories of vision*, volume 2, page 5.

Janati, H., Cuturi, M., and Gramfort, A. (2019). Wasserstein regularization for sparse multi-task regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1407–1416.

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

Lacombe, T., Cuturi, M., and Oudot, S. (2018). Large scale computation of means and clusters for persistence diagrams using optimal transport. In *Advances in Neural Information Processing Systems*, pages 9770–9780.

Latecki, L. J., Lakamper, R., and Eckhardt, T. (2000). Shape descriptors for non-rigid shapes with a single closed contour. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 424–429.

Lavenant, H., Claici, S., Chien, E., and Solomon, J. (2018). Dynamical optimal transport on discrete surfaces. In *SIGGRAPH Asia 2018 Technical Papers*, page 250. ACM.

Le, T., Ho, N., and Yamada, M. (2021). Flow-based alignment approaches for probability measures in different spaces. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Le, T., Huynh, V., Ho, N., Phung, D., and Yamada, M. (2019a). On scalable variant of wasserstein barycenter. *arXiv preprint arXiv:1910.04483*.

Le, T. and Yamada, M. (2018). Persistence Fisher kernel: A Riemannian manifold kernel for persistence diagrams. In *Advances in Neural Information Processing Systems*, pages 10007–10018.

Le, T., Yamada, M., Fukumizu, K., and Cuturi, M. (2019b). Tree-sliced variants of Wasserstein distances. In *Advances in neural information processing systems*, pages 12283–12294.

Lee, J., Bertrand, N. P., and Rozell, C. J. (2019). Parallel unbalanced optimal transport regularization for large scale imaging problems. *arXiv preprint arXiv:1909.00149*.

Lellmann, J., Lorenz, D. A., Schonlieb, C., and Valkonen, T. (2014). Imaging with kantorovich–rubinstein discrepancy. *SIAM Journal on Imaging Sciences*, 7(4):2833–2859.

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

Liutkus, A., Simsekli, U., Majewski, S., Durmus, A., and Stöter, F.-R. (2019). Sliced-wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning*, pages 4104–4113.

Mémoli, F., Munk, A., Wan, Z., and Weitkamp, C. (2021). The ultrametric gromov-wasserstein distance. *arXiv preprint arXiv:2101.05756*.

Mémoli, F., Smith, Z., and Wan, Z. (2019). Gromov-hausdorff distances on p-metric spaces and ultrametric spaces. *arXiv preprint arXiv:1912.00564*.

Mena, G. and Niles-Weed, J. (2019). Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, pages 4541–4551.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Nadjahi, K., Durmus, A., Simsekli, U., and Badeau, R. (2019). Asymptotic guarantees for learning generative models with the sliced-wasserstein distance. In *Advances in Neural Information Processing Systems*, pages 250–260.

Nakamura, T., Hiraoka, Y., Hirata, A., Escolar, E. G., and Nishiura, Y. (2015). Persistent homology and many-body atomic structure for medium-range order in the glass. *Nanotechnology*, 26(30):304001.

Naor, A. and Schechtman, G. (2007). Planar Earthmover is not in L_1. *SIAM Journal on Computing*, 37(3):804–826.

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

Pham, K., Le, K., Ho, N., Pham, T., and Bui, H. (2020). On unbalanced optimal transport: An analysis of Sinkhorn algorithm. In *Proceedings of the International Conference on Machine Learning*.

Piccoli, B. and Rossi, F. (2014). Generalized wasserstein distance and its application to transport equations with source. *Archive for Rational Mechanics and Analysis*, 211(1):335–358.

Piccoli, B. and Rossi, F. (2016). On properties of the generalized wasserstein distance. *Archive for Rational Mechanics and Analysis*, 222(3):1339–1365.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2011). Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Sato, R., Yamada, M., and Kashima, H. (2020). Fast unbalanced optimal transport on tree. In *Advances in neural information processing systems*.

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

Semple, C. and Steel, M. (2003). Phylogenetics. *Oxford Lecture Series in Mathematics and its Applications*.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. (2015). Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66.

Turner, K., Mukherjee, S., and Boyer, D. M. (2014). Persistent homology transform for modeling shapes and surfaces. *Information and Inference: A Journal of the IMA*, 3(4):310–344.

Vayer, T., Flamary, R., Tavenard, R., Chapel, L., and Courty, N. (2019). Sliced Gromov-Wasserstein. *Advances in Neural Information Processing Systems*.

Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Weed, J. and Berthet, Q. (2019). Estimation of smooth densities in wasserstein distance. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, pages 3118–3119.

Yang, K. D. and Uhler, C. (2019). Scalable unbalanced optimal transport using generative adversarial networks. In *International Conference on Learning Representations*.