Scalable Unbalanced Sobolev Transport for Measures on a Graph

Tam Le [*,†,‡] **Truyen Nguyen** [*,◇] **Kenji Fukumizu** [†]

The Institute of Statistical Mathematics [†]
The University of Akron [◇]
RIKEN AIP [‡]

Abstract

Optimal transport (OT) is a popular and powerful tool for comparing probability measures. However, OT suffers a few drawbacks: (i) input measures required to have the same mass, (ii) a high computational complexity, and (iii) indefiniteness which limits its applications on kernel-dependent algorithmic approaches. To tackle issues (ii)–(iii), Le et al. (2022) recently proposed Sobolev transport for measures on a graph having the *same total mass* by leveraging the graph structure over supports. In this work, we consider measures that may have *different total mass* and are supported on a graph metric space. To alleviate the disadvantages (i)–(iii) of OT, we propose a novel and scalable approach to extend Sobolev transport for this *unbalanced* setting where measures may have different total mass. We show that the proposed *unbalanced Sobolev transport* (UST) admits a closed-form formula for fast computation, and it is also negative definite. Additionally, we derive geometric structures for the UST and establish relations between our UST and other transport distances. We further exploit the negative definiteness to design positive definite kernels and evaluate them on various simulations to illustrate their fast computation and comparable performances against other transport baselines for unbalanced measures on a graph.

1 INTRODUCTION

Optimal transport (OT) has become a popular approach and its theory lays out a compelling toolkit for data analysis on probability distributions. OT has been leveraged in several research areas such as machine learning (Peyré and Cuturi, 2019; Nadjahi et al., 2019; Titouan et al., 2019; Bunne et al., 2019, 2022; Janati et al., 2020; Muzellec et al., 2020; Paty et al., 2020; Mukherjee et al., 2021; Altschuler et al., 2021; Fatras et al., 2021; Le et al., 2021a,b; Liu et al., 2021; Nguyen et al., 2021b; Scetbon et al., 2021; Si et al., 2021; Takezawa et al., 2022; Fan et al., 2022), computer vision (Nguyen et al., 2021a; Saleh et al., 2022; Wang et al., 2022b), and statistics (Mena and Niles-Weed, 2019; Weed and Berthet, 2019; Liu et al., 2022; Nguyen et al., 2022; Nietert et al., 2022; Wang et al., 2022a) to name a few. Nevertheless, it has some fundamental disadvantages.

One drawback of OT is that it requires input measures having the *same mass* for the transportation. To address this problem, several proposals have been developed in the recent literature. For examples, the *partial optimal transport* (POT) (Caffarelli and McCann, 2010; Figalli, 2010) constraints a fixed amount of mass for transportation; the *optimal entropy transport* (OET) (Liero et al., 2018; Chizat et al., 2018b; Kondratyev et al., 2016) optimizes a sum of a transport functional and two convex entropy functionals. Additionally, there are various other approaches, e.g., the Kantorovich-Rubinstein discrepancy (Hanin, 1992; Guittet, 2002; Lellmann et al., 2014; Sato et al., 2020), the unbalanced mass transport (Benamou, 2003), the generalized Wasserstein distance (Piccoli and Rossi, 2014, 2016), the unnormalized optimal transport (Gangbo et al., 2019), and the entropy partial transport (Le and Nguyen, 2021). These approaches are either special cases of the OET (e.g., by using some specific instances of entropy functional such as the total variation distance, ℓ^2 distance), or a variant of OET (e.g., by using the ℓ^p distance, partial transport in place of the entropy functional, transport functional respectively). It is worth pointing out that the unbalanced setting for measures with unequal mass has been applied in several application domains and learning problems, e.g., color transfer and shape matching (Bonneel et al., 2015); multi-label learning (Frogner et al., 2015); positive-unlabeled learning (Chapel et al., 2020); natural language processing and topological data analysis (Le and Nguyen, 2021). In par-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

[*]: Two authors contributed equally.

ticular, the unbalanced approach becomes essential when supports of input measures are subject to noise or have outliers since such supports are not desirably aligned in the matching problem (Frogner et al., 2015; Balaji et al., 2020; Mukherjee et al., 2021).

Another drawback of standard OT is that it has a high computational complexity. This disadvantage also exists in the unbalanced optimal transport (UOT), which hinders its applications, especially for large-scale settings. For examples, let us consider the OET with Kullback-Leibler divergence for the entropy functional which is widely used in applications. For this, one can leverage the entropic regularization to derive efficient Sinkhorn-based algorithmic approach (Frogner et al., 2015; Chizat et al., 2018a; Séjourné et al., 2019) which has a quadratic complexity (Pham et al., 2020). Another popular approach to scale up UOT is to exploit geometric structures of supports, e.g., one-dimensional structure (Bonneel and Coeurjolly, 2019; Séjourné et al., 2022), tree structure (Le and Nguyen, 2021; Sato et al., 2020). More concretely, Bonneel and Coeurjolly (2019) proposed the sliced partial optimal transport (SPOT) by projecting supports into a random one-dimensional space. By assuming a unit mass on each support, they developed an efficient algorithmic approach with a quadratic complexity for the worst case. Nonetheless, SPOT suffers a curse of dimensionality since using one-dimensional projections for supports limits its ability to capture topological structures of distributions, especially in a high-dimensional space. Le and Nguyen (2021) proposed the entropy partial transport (EPT) by exploiting a tree structure to remedy the curse of dimensionality for SPOT. Moreover, EPT yields the first closed-form solution among various variants of UOT (i.e., its complexity is linear to the number of edges in a tree) for fast computation which is applicable for large-scale settings. However, tree structure may be a restricted condition which narrows down its practical usage in applications.

The aforementioned circumstances motivate us to consider measures with *unequal mass* and supported on a *graph metric* space which has more degrees of freedom (i.e., graph structure rather than tree structure) and appears more popularly in applications. Inspired by the Sobolev transport (Le et al., 2022) for probability measures on a graph, we propose a *novel and scalable* approach to leverage graph structure and extend Sobolev transport for the *unbalanced* setting. At a high level, our contributions are three-fold as follow:

- we propose a novel *p-order unbalanced Sobolev transport* (UST) ($p \geq 1$) for measures with unequal mass and supported on a graph metric space. We prove that UST admits a *closed-form formula* for a fast computation and it is *negative definite*;
- we derive geometric structures for the UST and propose *positive definite kernels* built upon the UST. Additionally, we establish relations between UST and the EPT on a *graph*;
- we empirically illustrate that UST is fast for computation (i.e., *closed-form solution* of UST). Also various simulations demonstrate that the performances of the proposed kernels for UST compare favorably with other unbalanced transport baselines for measures with unequal mass on a graph.

The paper is organized as follows: we introduce notations and the problem setup in §2. In §3, we extend and derive the EPT for unbalanced measures *on a graph*. We then present our main contribution: the UST for measures with unequal mass on a graph in §4 and derive its properties in §5. In §6, we evaluate the proposed kernel for UST against other unbalanced transport baselines for measures with unequal mass on a graph on various simulations. We conclude our work in §7. The detailed proofs for our theoretical results are placed in Appendix §A.2. Furthermore, we have released code for our proposals.[1]

2 PRELIMINARIES

In this section, we introduce our problem setting, notations, and review relevant definitions.

We consider the same graph setting $\mathbb{G} = (V, E)$ where V, E are sets of nodes and edges respectively as in (Le et al., 2022) for Sobolev transport. More precisely, $\mathbb{G}$ is an undirected, connected and physical graph in the sense that $V \subset \mathbb{R}^n$ and each edge $e \in E$ is the standard line segment in $\mathbb{R}^n$ connecting the two corresponding end-points of e. Graph $\mathbb{G}$ has positive edge lengths $\{w_e\}_{e\in E}$ and is imposed a graph metric $d_{\mathbb{G}}(\cdot, \cdot)$ which equals to the length of the shortest path on $\mathbb{G}$. Following a convention in (Le et al., 2022), by graph $\mathbb{G}$, we mean the set of all nodes in V and all points forming the edges in E, i.e., the continuous setting for graph $\mathbb{G}$. We also assume that there exists a fixed root node $z_0 \in V$ such that for every $x \in \mathbb{G}$, $d_{\mathbb{G}}(x, z_0)$ is attained by the unique shortest path connecting x and z_0, i.e., the uniqueness property of the shortest paths (Le et al., 2022).

Given a point $x \in \mathbb{G}$ (resp. an edge $e \in E$ in $\mathbb{G}$), we denote $\Lambda(x)$ (resp. γ_e) as the collection of all points $y \in \mathbb{G}$ such that the unique shortest path in $\mathbb{G}$ connecting the root node z_0 and y contains the point x (resp. the edge e). That is,

$$\Lambda(x) \triangleq \big\{y \in \mathbb{G} : \, x \in [z_0, y]\big\}, \tag{1}$$

$$\gamma_e \triangleq \big\{y \in \mathbb{G} : \, e \subset [z_0, y]\big\}, \tag{2}$$

where we write $[z_0, y]$ for the shortest path in $\mathbb{G}$ connecting the root node z_0 and y.

We denote $\mathcal{M}(\mathbb{G})$ (resp. $\mathcal{M}(\mathbb{G}\times\mathbb{G})$) as the set of all nonnegative Borel measures on $\mathbb{G}$ (resp. $\mathbb{G} \times \mathbb{G}$) with a finite mass. By continuous function f on $\mathbb{G}$, we mean that $f : \mathbb{G} \to \mathbb{R}$

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

is continuous w.r.t. the topology on $\mathbb{G}$ induced by the Euclidean distance. Similar adoption is also applied for continuous functions on $\mathbb{G}\times\mathbb{G}$. We denote $C(\mathbb{G})$ as the collection of all continuous functions on $\mathbb{G}$.

Given a scalar $b > 0$, a function $w : \mathbb{G} \to \mathbb{R}$ is called b-Lipschitz w.r.t. the graph metric $d_{\mathbb{G}}$ if

$$|w(x) - w(y)| \leq b\, d_{\mathbb{G}}(x, y), \forall x, y \in \mathbb{G}.$$

For $1 \leq p \leq \infty$, we denote p' as its conjugate, i.e., $p' \in [1, \infty]$ s.t., $\frac{1}{p} + \frac{1}{p'} = 1$. For a nonnegative Borel measure ω on $\mathbb{G}$, let $L^p(\mathbb{G}, \omega)$ denote the space of all Borel measurable functions $f : \mathbb{G} \to \mathbb{R}$ satisfying $\int_{\mathbb{G}} |f(y)|^p \omega(\mathrm{d}y) < \infty$. When $p = \infty$, we assume that f is bounded ω-a.e. instead. Functions $f_1, f_2 \in L^p(\mathbb{G}, \omega)$ are considered to be the same if $f_1(x) = f_2(x)$ for ω-a.e. $x \in \mathbb{G}$. Then, $L^p(\mathbb{G}, \omega)$ is a normed space with the norm defined by

$$\|f\|_{L^p(\mathbb{G},\omega)} \triangleq \left(\int_{\mathbb{G}} |f(y)|^p \omega(\mathrm{d}y)\right)^{\frac{1}{p}} \text{ for } 1 \leq p < \infty, \text{ and}$$

$$\|f\|_{L^\infty(\mathbb{G},\omega)} \triangleq \inf\{t \in \mathbb{R} : |f(x)| \leq t \text{ for } \omega\text{-a.e. } x \in \mathbb{G}\}.$$

Recall that Sobolev transport for probability measures on a graph is an instance of integral probability metrics (IPM) (Müller, 1997). Intuitively, the definition of Sobolev transport is based on the dual form of the 1-order Wasserstein distance, but its Lipschitz constraint for the critic function is considered in the graph-based Sobolev space (see (Le et al., 2022, §3) for the detail). As a consequence, it may *not* possible to directly leverage approaches for standard OT (e.g., partial OT, entropy (partial) transport) to extend Sobolev transport for *unbalanced* measures on a *graph*.

In this paper, we propose a *detour* to develop unbalanced Sobolev transport for measures with unequal mass on a graph. We first take a step back to leverage the EPT (for unbalanced measures on a *tree*) (Le and Nguyen, 2021) and extend it for unbalanced measures on a *graph* (§3). Although it is still a great challenge to efficiently compute the EPT for unbalanced measures on a *graph*, this novel extension (especially its dual form) plays a cornerstone in deriving a scalable approach for the proposed unbalanced Sobolev transport (UST) (§4).

3 ENTROPY PARTIAL TRANSPORT ON A GRAPH

The entropy partial transport (EPT) (Le and Nguyen, 2021) is developed for unbalanced measures on a *tree*. In this section, we propose an extension of EPT for unbalanced measures on a *graph*. Intuitively, EPT optimizes a sum of a transport function and two convex entropy functions in a similar spirit to the OET (Liero et al., 2018; Chizat et al., 2018b). We first consider the primal formulation of EPT on a *graph*. We then derive its dual formulation which is the main result of this section. This novel dual formulation paves the way for our development of the UST (§4).

Given two measures $\mu, \nu \in \mathcal{M}(\mathbb{G})$ which may have different total mass, consider the set

$$\Pi_{\leq}(\mu, \nu) \triangleq \{\gamma \in \mathcal{M}(\mathbb{G} \times \mathbb{G}) : \gamma_1 \leq \mu,\, \gamma_2 \leq \nu\}$$

where γ_1 and γ_2 respectively denote the first and second marginals of γ; by $\gamma_1 \leq \mu$, we mean that $\gamma_1(B) \leq \mu(B)$ for every Borel set $B \subset \mathbb{G}$. Similar convention is used when we write $\gamma_2 \leq \nu$.

For $\gamma \in \Pi_{\leq}(\mu, \nu)$, let f_1 and f_2 respectively be the Radon-Nikodym derivatives of γ_1 w.r.t. μ and of γ_2 w.r.t. ν, i.e., $\gamma_1 = f_1\mu$ and $\gamma_2 = f_2\nu$. Then, we have $0 \leq f_1 \leq 1$ μ-a.e., and $0 \leq f_2 \leq 1$ ν-a.e. The weighted relative entropies of γ_1 w.r.t. μ and of γ_2 w.r.t. ν are defined by

$$\mathcal{F}_1(\gamma_1|\mu) \triangleq \int_{\mathbb{G}} w_1(x) F_1(f_1(x))\mu(\mathrm{d}x),$$

$$\mathcal{F}_2(\gamma_2|\nu) \triangleq \int_{\mathbb{G}} w_2(x) F_2(f_2(x))\nu(\mathrm{d}x),$$

where $F_1,\ F_2 : [0, 1] \to (0, \infty)$ are convex and lower semi-continuous entropy functions; and $w_1, w_2 : \mathbb{G} \to [0, \infty)$ are given nonnegative weight functions.

Given a continuous cost function $c : \mathbb{G} \times \mathbb{G} \to \mathbb{R}$ with $c(x, x) = 0$, a constant $b \geq 0$ and a fixed scalar $m \in [0, \bar{m}]$ where $\bar{m} \triangleq \min\{\mu(\mathbb{G}), \nu(\mathbb{G})\}$, we consider the primal formulation of EPT problem on a *graph*:

$$\mathrm{W}_{c,m}(\mu, \nu) \triangleq \inf_{\gamma \in \Pi_{\leq}(\mu,\nu),\, \gamma(\mathbb{G}\times\mathbb{G}) = m} \Big[\mathcal{F}_1(\gamma_1|\mu) + \mathcal{F}_2(\gamma_2|\nu) + b \int_{\mathbb{G}\times\mathbb{G}} c(x, y)\gamma(\mathrm{d}x, \mathrm{d}y)\Big]. \quad (3)$$

Following (Le and Nguyen, 2021), we consider

$$F_1(s) = F_2(s) = |s - 1|$$

for the entropy functions in (3) and form a Lagrange multiplier $\lambda \in \mathbb{R}$ conjugate to the constraint $\gamma(\mathbb{G} \times \mathbb{G}) = m$. As a result, we instead study the problem

$$\mathrm{ET}_{c,\lambda}(\mu, \nu) = \inf_{\gamma \in \Pi_{\leq}(\mu,\nu)} \mathcal{C}_\lambda(\gamma), \quad (4)$$

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

$$\mathcal{C}_\lambda(\gamma) \triangleq \int_{\mathbb{G}} w_1\mu(\mathrm{d}x) + \int_{\mathbb{G}} w_2\nu(\mathrm{d}x) - \int_{\mathbb{G}} w_1\gamma_1(\mathrm{d}x) - \int_{\mathbb{G}} w_2\gamma_2(\mathrm{d}x) + b\int_{\mathbb{G}\times\mathbb{G}} [c(x, y) - \lambda]\gamma(\mathrm{d}x, \mathrm{d}y). \quad (5)$$

The connection between problem (3) with mass constraint m and problem (4) with Lagrange multiplier λ is given in Theorem A.1 (Appendix §A.1). Also, from Theorem A.1, we see that solving the auxiliary problem (4) gives us a solution to the original problem (3). We now derive a novel dual formulation for problem (4) which paves the way for our proposed UST (§4).

Theorem 3.1 (Dual formula for general cost)**.** *For $\lambda \geq 0$, nonnegative weights w_1, w_2, and two input measures $\mu, \nu \in \mathcal{M}(\mathbb{G})$, we have*

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

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

The main idea of proving this result is to attach to the graph $\mathbb{G}$ a new point $\hat{s}$, and then suitably and carefully extend the cost c and the input distributions μ, ν to the set $\hat{\mathbb{G}} \triangleq \mathbb{G} \cup \{\hat{s}\}$ inspired by an observation in (Caffarelli and McCann, 2010). The key point of this extension is to ensure that the extended input distributions on $\hat{\mathbb{G}}$ have the same total mass and the value of the new balanced OT between extended input distributions on $\hat{\mathbb{G}}$ is *equal* to that of the original EPT on graph $\mathbb{G}$ (i.e., the unbalanced setting). We then exploit the dual theory for the new balanced OT problem on $\hat{\mathbb{G}}$ to establish the dual formulation for our EPT problem on graph $\mathbb{G}$ (see Appendix §A.2 for detailed proof). When the ground cost c is the graph metric $d_{\mathbb{G}}$, the dual formula in Theorem 3.1 can be rewritten in a simpler and more symmetric form as follows.

Corollary 3.2 (Dual formula for graph metric)**.** *Assume that $\lambda \geq 0$ and the nonnegative weight functions w_1, w_2 are b-Lipschitz w.r.t. $d_{\mathbb{G}}$. For simplicity, let $\mathrm{ET}_\lambda \triangleq \mathrm{ET}_{d_{\mathbb{G}},\lambda}$. Then, we have*

$$\mathrm{ET}_\lambda(\mu,\nu) = \sup_{f\in\mathbb{U}} \int_{\mathbb{G}} f(\mu-\nu) - \frac{b\lambda}{2}\big[\mu(\mathbb{G}) + \nu(\mathbb{G})\big], \quad (6)$$

where $\mathbb{U} \triangleq \big\{f \in C(\mathbb{G}) : -w_2 - \frac{b\lambda}{2} \leq f \leq w_1 + \frac{b\lambda}{2},\ |f(x) - f(y)| \leq b\,d_{\mathbb{G}}(x,y)\big\}$.

Remark 3.3. *We remark that one cannot directly use the dual formulation in (Le and Nguyen, 2021), or that of (Piccoli and Rossi, 2014, 2016) for unbalanced measures on a* graph *since the considered problem does not satisfy the conditions imposed in these approaches for duality.*

In principal, for input unbalanced measures on a graph, it is simpler to learn the optimal f^* in dual form (6) than to learn the optimal γ^* in primal form (4). This is due to the fact that the critic f^* is a function on the lower dimensional space compared to γ^*. Moreover, the Lipschitz constraint for f^* is easier to handle than the constraint $\Pi_{\leq}(\mu,\nu)$ for γ^*. Nevertheless, it is still a challenge to effectively compute ET_λ using (6).

As illustrated in (Le et al., 2019; Le and Nguyen, 2021) for transport problems on a *tree*, the Lipschitz constraint for the critic f can be effectively optimized by leveraging the *tree structure* supports. Furthermore, the Lipschitz constraint is linked with the 1-order Wasserstein distance via the Kantorovich duality formulation. Due to the different nature of duality for p-order Wasserstein distance when $p > 1$, it is however unknown that one can extend the fast computational results in (Le et al., 2019; Le and Nguyen, 2021) to p-order Wasserstein distance with $p > 1$, even for measures on a *tree*.

To alleviate this, we propose in the next section an efficient p-order unbalanced Sobolev transport for measures with unequal mass on a *graph* for any $p \geq 1$.

4 UNBALANED SOBOLEV TRANSPORT

As pointed out in §3, it is a great challenge to efficiently compute ET_λ (i.e., the EPT problem) for unbalanced measures on a *graph* using either the primal form (4) or the dual form (6). To overcome this issue, we propose in this section an efficient variant called unbalanced Sobolev transport (UST) distance. We further derive a novel closed-form formula which allows a fast computation for the proposed transport distance, especially for large-scale settings.

Our strategy in defining the UST is based on the dual formulation (6) (in Corollary 3.2) but by simultaneously relaxing the two constraints for critic function f in the set $\mathbb{U}$. This approach is partially adopted in (Le and Nguyen, 2021) for the EPT problem for measures on a *tree*, but they only relax the first corresponding constraint for f in the set $\mathbb{U}$ (i.e., the *bounded constraint* for the critic function f). However, keeping the *Lipschitz constraint* for f limits the approach in (Le and Nguyen, 2021) to be extended to more general structures rather than tree structure (e.g., graph structure). We note that the Lipschitz constraint is about bounding the derivative of f and hence it is more fundamental and relevant than the first constraint. In this paper, we propose to also relax the Lipschitz constraint by leveraging a notion of Sobolev functions. This approach relies on the following concept of derivatives for functions on graphs introduced by Le et al. (2022), which can be viewed as a generalized version of the fundamental theorem of calculus for a *graph*.

Definition 4.1 (Graph-based Sobolev space (Le et al., 2022))**.** *Let ω be a nonnegative Borel measure on $\mathbb{G}$, and let $1 \leq p \leq \infty$. A continuous function $f : \mathbb{G} \to \mathbb{R}$ is in the Sobolev space $W^{1,p}(\mathbb{G},\omega)$ if there exists a function $h \in L^p(\mathbb{G},\omega)$ satisfying*

$$f(x) - f(z_0) = \int_{[z_0,x]} h(y)\omega(\mathrm{d}y), \forall x \in \mathbb{G}.$$

Such function h is unique in $L^p(\mathbb{G},\omega)$ and is called the graph derivative of f w.r.t. the measure ω. Hereafter, this graph derivative of f is denoted by f'.

From Definition 4.1 and the property of $L^p(\mathbb{G},\omega)$ space, we have

$$W^{1,p_2}(\mathbb{G},\omega) \subset W^{1,p_1}(\mathbb{G},\omega),$$

whenever $1 \leq p_1 \leq p_2 \leq \infty$. In particular, $W^{1,\infty}(\mathbb{G},\omega)$ is the smallest space and $W^{1,1}(\mathbb{G},\omega)$ is the largest space. Ad-

ditionally, we prove that $W^{1,\infty}(\mathbb{G},\omega^*)$ contains the space of all Lipschitz continuous functions, and both spaces coincide when $\mathbb{G}$ is a tree (see Lemma A.2 in Appendix §A.1 for the detail). Hereafter, let ω^* denote the length measure on $\mathbb{G}$ as defined in (Le et al., 2022, §4.1) (see Appendix §B.1 for a review). We propose to regularize the transport ET_λ in (6) by relaxing the constraint set $\mathbb{U}$ for critic function f in two ways:

• Firstly, we replace the *Lipschitz condition* for the critic function f in the set $\mathbb{U}$ (in Corollary 3.2) by instead considering this constraint in the graph-based Sobolev space, i.e., $f \in W^{1,p'}(\mathbb{G},\omega)$ with $\|f'\|_{L^{p'}(\mathbb{G},\omega)} \leq b$. This has the following advantages: (i) we can enlarge the constraint set on the Sobolev space $W^{1,p'}(\mathbb{G},\omega)$ by decreasing the value of parameter p'; (ii) we can vary the constraint set by choosing a suitable measure ω on $\mathbb{G}$. The measure ω can be interpreted as a cost of moving a unit mass from one location to another, and this cost is the same as the graph metric $d_\mathbb{G}$ when ω is chosen as the length measure ω^* of $\mathbb{G}$. Even when $p=1$ and $\omega=\omega^*$, this relaxation viewpoint still has the fundamental benefit: it allows us to extend most of the main results in (Le and Nguyen, 2021) for *tree* structure to *graph* structure.

We emphasize that extending the approach in (Le and Nguyen, 2021) (i.e., EPT problem for measures on a *tree*) to EPT problem for measures on a *graph* $\mathbb{G}$ is problematic. In this special case, we know from Lemma A.2 (Appendix §A.1) that our corresponding *Sobolev constraint* is equivalent the *Lipschitz constraint* when $\mathbb{G}$ is a tree. However, Lemma A.2 also implies that the Sobolev constraint set is possibly *larger* for a general graph $\mathbb{G}$. This flexibility of Sobolev functions enables us to overcome the limitation of the approach in (Le and Nguyen, 2021) (i.e., for a *tree* structure) and gives us an effective way to exploit the *graph* structure by working with critic function f of a specific form in Sobolev space (see Definition 4.1). Our obtained results in this section reveal that *critic of Sobolev type in the sense of Definition 4.1 is more suitable for EPT problem for measures on a* graph *than critic of the Lipschitz type*.

• Secondly, we relax the first condition for f in the set $\mathbb{U}$ (i.e., the *bounded constraint* for the critic function f) by using the following observation. According to Definition 4.1, any function $f \in W^{1,p'}(\mathbb{G},\omega)$ can be represented as

$$f(x) = f(z_0) + \int_{[z_0,x]} f'(y)\omega(dy).$$

If in addition $\|f'\|_{L^{p'}(\mathbb{G},\omega)} \leq b$, then by Hölder inequality, the second term on the right hand side is controlled by $b\,\omega\big([z_0,x]\big)^{\frac{1}{p}}$. Thus, instead of requiring

$$-w_2(x) - \frac{b\lambda}{2} \leq f(x) \leq w_1(x) + \frac{b\lambda}{2}, \quad \forall x \in \mathbb{G}$$

as in the definition of $\mathbb{U}$, we suggest to constrain only the first term $f(z_0)$.

Putting these two ways of regularization together, we propose to consider the following constraint set $\mathbb{U}^\alpha_{p'}$ as a relaxation of the constraint set $\mathbb{U}$ for the critic function f in Corollary 3.2. Note that the choice of $\alpha=0$ corresponds to our above discussion. Here, we generalize our theoretical development for a more general α to allow an extra degree of freedom which might be potentially useful in practical applications, e.g., by tuning α for further improvement.

Definition 4.2 (The regularized set $\mathbb{U}^\alpha_{p'}$ for critic function)**.** *For $1 \leq p \leq \infty$ and $0 \leq \alpha \leq \frac{1}{2}[b\lambda + w_1(z_0) + w_2(z_0)]$, let $\mathbb{U}^\alpha_{p'}$ be the collection of all functions $f \in W^{1,p'}(\mathbb{G},\omega)$ satisfying*

$$f(z_0) \in I_\alpha \triangleq \Big[-w_2(z_0) - \frac{b\lambda}{2} + \alpha, w_1(z_0) + \frac{b\lambda}{2} - \alpha\Big]$$

and

$$\|f'\|_{L^{p'}(\mathbb{G},\omega)} \leq b.$$

Equivalently, $\mathbb{U}^\alpha_{p'}$ is the collection of all functions f of the form

$$f(x) = s + \int_{[z_0,x]} h(y)\omega(\mathrm{d}y) \tag{7}$$

with $s \in I_\alpha$ and with $h : \mathbb{G} \to \mathbb{R}$ being some function satisfying

$$\|h\|_{L^{p'}(\mathbb{G},\omega)} \leq b.$$

It is clear from Definition 4.2 that $\mathbb{U} \subset \mathbb{U}^0_{p'}$ (see Corollary 3.2 for set $\mathbb{U}$). The requirement $\alpha \leq \frac{1}{2}[b\lambda + w_1(z_0) + w_2(z_0)]$ is to ensure that the interval I_α is nonempty. By constraining critic f to the relaxed set $\mathbb{U}^\alpha_{p'}$ and noting that the last term in (6) is simply a constant depending on the total masses of μ and ν, we propose the following regularization of the transport ET_λ in Corollary 3.2, namely *unbalanced Sobolev transport* (UST).

Definition 4.3 (Unbalanced Sobolev transport)**.** *Let ω be a nonnegative Borel measure on graph $\mathbb{G}$. Given $1 \leq p \leq \infty$ and $0 \leq \alpha \leq \frac{1}{2}[b\lambda + w_1(z_0) + w_2(z_0)]$. For $\mu,\nu \in \mathcal{M}(\mathbb{G})$, the unbalanced Sobolev transport is defined as follow*

$$\mathrm{US}^\alpha_p(\mu,\nu) \triangleq \sup_{f \in \mathbb{U}^\alpha_{p'}} \Big[\int_\mathbb{G} f(x)\mu(\mathrm{d}x) - \int_\mathbb{G} f(x)\nu(\mathrm{d}x)\Big].$$

The measure ω used for representing critic f in $U^\alpha_{p'}$ (see (7)) acts as the ground cost of moving masses on graph $\mathbb{G}$ from one location to another. Especially, when ω is chosen as the length measure ω^* of graph $\mathbb{G}$, we have $\omega([x,y]) = d_\mathbb{G}(x,y)$ (see Lemma B.2 in Appendix §B.1).

We then show the connection between 1-order UST and the dual formulation of EPT on graph $\mathbb{G}$ with the *Lipschitz constraint*, but the *bounded constraint* only applied on the critic function at root node z_0. Precisely, we obtain:

Lemma 4.4. *Recall that ω^* be the length measure of graph $\mathbb{G}$. For $\omega = \omega^*$, we have*

$$\mathrm{US}^0_1(\mu,\nu) \geq \sup\Big[\int_\mathbb{G} f(\mu-\nu) :\ f \in \mathbb{U}_0\Big] \tag{8}$$

where $\mathbb{U}_0 \triangleq \Big\{ f \in C(\mathbb{G}) : -w_2(z_0) - \frac{b\lambda}{2} \le f(z_0) \le w_1(z_0) + \frac{b\lambda}{2},\ |f(x) - f(y)| \le b\, d_{\mathbb{G}}(x,y) \Big\}$. *Moreover, the inequality in* (8) *becomes the equality if* $\mathbb{G}$ *is a tree.*

We next state our fundamental result, which demonstrates that the proposed UST (Definition 4.3) for measures with unequal mass on a *graph* is computationally effective. We in fact obtain a closed-form formula for UST in terms of an integral explicitly depending on the input measures. This yields a substantial computational advantage in comparison with the EPT approach for unbalanced measures on a graph (i.e., ET_λ) which requires to solve sophisticated optimization problems either in the primal (4) or its dual (6). To our knowledge, *the proposed UST is the* first *approach which yields a closed-form solution among available variants of unbalanced OT for measures with unequal mass on a* graph.

Proposition 4.5. *Let* ω *be a nonnegative measure on graph* $\mathbb{G}$. *Let* $1 \le p \le \infty$ *and* $0 \le \alpha \le \frac{1}{2}[b\lambda + w_1(z_0) + w_2(z_0)]$. *Then, for two input measures* $\mu, \nu \in \mathcal{M}(\mathbb{G})$, *we have*

$$\mathrm{US}_p^\alpha(\mu,\nu) = b\Big[\int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p\, \omega(\mathrm{d}x)\Big]^{\frac{1}{p}} + \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})|,$$

where $\Lambda(x)$ *is defined by* (1) *and*

$$\Theta \triangleq \begin{cases} w_1(z_0) + \frac{b\lambda}{2} - \alpha & \text{if} \quad \mu(\mathbb{G}) \ge \nu(\mathbb{G}), \\ w_2(z_0) + \frac{b\lambda}{2} - \alpha & \text{if} \quad \mu(\mathbb{G}) < \nu(\mathbb{G}). \end{cases} \tag{9}$$

The constant Θ depends on μ and ν unless $\mu(\mathbb{G}) = \nu(\mathbb{G})$ or $w_1(z_0) = w_2(z_0)$. The integral in the above expression can be computed explicitly and efficiently as in the following corollary when the two input distributions are supported on nodes of the graph (i.e., the node set V of graph $\mathbb{G}$).

Corollary 4.6. *Under the same assumptions as in Proposition 4.5 and assume in addition that* $\omega(\{x\}) = 0$ *for every* $x \in \mathbb{G}$. *Suppose that* $\mu, \nu \in \mathcal{M}(\mathbb{G})$ *are supported on nodes in* V *of graph* $\mathbb{G}$.[2] *Then, we have*

$$\mathrm{US}_p^\alpha(\mu,\nu) = b\Big(\sum\nolimits_{e\in E} w_e\, |\mu(\gamma_e) - \nu(\gamma_e)|^p\Big)^{\frac{1}{p}} + \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})|. \tag{10}$$

Remark 4.7 (UST for non-physical graph)**.** *We have assumed that* $\mathbb{G}$ *is a physical graph as in §2. However, Corollary 4.6 shows that the* p*-order unbalanced Sobolev transport* US_p^α *does not depend on this physical assumption when input measures are supported on nodes. Precisely, it only depends on the graph structure* (V, E) *and edge weights* w_e. *Thus,* US_p^α *can be applied for non-physical graph* $\mathbb{G}$.

We next describe a preprocessing step on graph $\mathbb{G}$ and analyze the time complexity in computing US_p^α .

[2]We discuss an extension for measures supported in $\mathbb{G}$ in Appendix §B.2.

Preprocessing step. To compute US_p^α, we apply a preprocessing step to form the set γ_e for each edge $e \in E$ in graph $\mathbb{G}$ by identifying shortest paths from the root node z_0 to other nodes (e.g., by Dijkstra algorithm with a complexity $\mathcal{O}(|E| + |V|\log|V|)$ where $|E|, |V|$ are the numbers of egdes and nodes of graph $\mathbb{G}$ respectively). Especially, observe that any edge e with $\gamma_e = \varnothing$ does not contribute to the computation of US_p^α. Therefore, one can remove such edge e in the summation in (10). We emphasize that this preprocessing step only involves the graph structure itself and is independent of input measures.

Computational complexity. Let $E_{\mu,\nu} \triangleq \{e \in E \mid e \subset [z_0, z] \text{ for some } z \in \mathrm{supp}(\mu) \cup \mathrm{supp}(\nu)\}$, where $\mathrm{supp}(\mu), \mathrm{supp}(\nu)$ are respectively the support of measures μ, ν. Then, the computational complexity of $\mathrm{US}_p^\alpha(\mu,\nu)$ is linear to the number of edges in $E_{\mu,\nu}$.

Related work. Beyond the pure graph of supports, the metric structure inherited from the graph metric space plays an important role in our work. More precisely, an edge weight w_e is considered as a cost to move a unit mass from one node to the other node of edge e (i.e., graph metric distance between two edge nodes). Therefore, one should distinguish our approach with the unbalanced diffusion earth mover's distance (Tong et al., 2022) which uses an affinity between two edge nodes in their graph.

• **Relation with Sobolev transport (ST) (Le et al., 2022).** We emphasize that ST is only valid for measures with *equal* mass on a graph. It *cannot* be applied for our considered problem where input measures may have *different* total mass. Even though both ST and the proposed UST are instances of integral probability metrics (IPM), it is nontrivial to effectively extend ST for unbalanced measures on a *graph* by defining a function set for the critic. The theoretical results of EPT on a *graph* in §3 play the fundamental role in developing our proposed UST.

Remark 4.8 (The special case of balanced mass)**.** *When input measures have the same mass, from Lemma A.6 of §A.1.5, the proposed unbalanced Sobolev transport (with* $b = 1$*) coincides with the balanced Sobolev transport (Le et al., 2022, Definition 3.2).*

• **Relation with EPT on a tree (Le and Nguyen, 2021).** As we discussed previously, extending the approach in (Le and Nguyen, 2021) for EPT on a *tree* to our considered problem (i.e., EPT on a *graph*) is problematic. We see from Lemma A.2 (Appendix §A.1) and Lemma 4.4 that the *Sobolev constraint* set in our approach is possibly *larger* than the *Lipschitz constraint* set for a general graph $\mathbb{G}$, but these two constraint sets coincide when $\mathbb{G}$ is a tree. Our results illustrate that it is more efficient to exploit *graph structure* for *critic of Sobolev type* (as in our approach) than *critic of the Lipschitz type* (as in EPT on a tree).

5 PROPERTIES OF UNBALANCED SOBOLEV TRANSPORT

In this section, we derive geometric structures together with bounds for UST and prove its negative definiteness. Consequently, we develop positive definite kernels upon UST, required in many kernel-dependent frameworks.

We first show that US_p^α possess the metric property. Moreover, it makes the space of measures $\mathcal{M}(\mathbb{G})$ a geodesic space. Thus, $(\mathcal{M}(\mathbb{G}), \mathrm{US}_p^\alpha)$ inherits all geometric properties of the geodesic space.

Proposition 5.1 (Geometric structures of US_p^α)**.** *Let ω be a nonnegative Borel measure on $\mathbb{G}$. Assume that $\lambda, w_1(z_0), w_2(z_0) \geq 0$. For $1 \leq p \leq \infty$ and $0 \leq \alpha < \frac{b\lambda}{2} + \min\{w_1(z_0), w_2(z_0)\}$, then we have*

i) $\mathrm{US}_p^\alpha(\mu + \sigma, \nu + \sigma) = \mathrm{US}_p^\alpha(\mu, \nu)$, $\forall \mu, \nu, \sigma \in \mathcal{M}(\mathbb{G})$.

ii) US_p^α *is a divergence*[3] *and satisfies the triangle inequality:*

$$\mathrm{US}_p^\alpha(\mu, \nu) \leq \mathrm{US}_p^\alpha(\mu, \sigma) + \mathrm{US}_p^\alpha(\sigma, \nu), \forall \mu, \nu, \sigma \in \mathcal{M}(\mathbb{G}).$$

iii) If in addition $w_1(z_0) = w_2(z_0)$, then US_p^α is a metric and $(\mathcal{M}(\mathbb{G}), \mathrm{US}_p^\alpha)$ is a complete metric space. Moreover, it is a geodesic space in the sense that for every two points μ and ν in $\mathcal{M}(\mathbb{G})$ there exists a path $\varphi : [0, a] \to \mathcal{M}(\mathbb{G})$ with $a \triangleq \mathrm{US}_p^\alpha(\mu, \nu)$ such that $\varphi(0) = \mu$, $\varphi(a) = \nu$, and

$$\mathrm{US}_p^\alpha(\varphi(t), \varphi(s)) = |t - s|, \text{ for all } t, s \in [0, a].$$

In Proposition A.4 (Appendix §A.1), we also establish a comparison between US_p^α for different exponent p. We next derive a lower bound for US_1^0 in terms of ET_λ. In fact, a more general estimate holds true for every $p \geq 1$ and is given in Proposition A.5 (Appendix §A.1). As a consequence of Corollary 3.2 and Lemma 4.4 and since $\mathbb{U} \subset \mathbb{U}_0$, we obtain:

Proposition 5.2 (Lower bound for US_1^0)**.** *Recall that ω^* is the length measure on $\mathbb{G}$. Assume that w_1, w_2 are b-Lipschitz w.r.t. $d_\mathbb{G}$. For $\omega = \omega^*$, $\mu, \nu \in \mathcal{M}(\mathbb{G})$, we have*

$$\mathrm{US}_1^0(\mu, \nu) \geq \mathrm{ET}_\lambda(\mu, \nu) + \frac{b\lambda}{2}\big[\mu(\mathbb{G}) + \nu(\mathbb{G})\big].$$

We emphasize that when $\mathbb{G}$ is a tree, our EPT on a *graph* (i.e., $\mathrm{ET}_{c,\lambda}$ and ET_λ) coincide with the ones defined in (Le and Nguyen, 2021). Furthermore, we have:

Proposition 5.3 (Lower bounds)**.** *Assume that $\mathbb{G}$ is a tree and $\omega = \omega^*$. The followings hold true:*

i) $\mathrm{US}_1^\alpha(\mu, \nu) = d_\alpha(\mu, \nu)$. *Also for $1 \leq p \leq \infty$, we have*

$$\mathrm{US}_p^\alpha(\mu, \nu) \geq \omega^*(\mathbb{G})^{-\frac{1}{p'}} d_\alpha(\mu, \nu) + \Theta\left[1 - \omega^*(\mathbb{G})^{-\frac{1}{p'}}\right] |\mu(\mathbb{G}) - \nu(\mathbb{G})|,$$

where d_α is defined in (Le and Nguyen, 2021, Eq. (9)).

ii) If $\mu(\mathbb{G}) = \nu(\mathbb{G})$, then for $1 \leq p \leq \infty$, we have

$$\mathrm{US}_p^\alpha(\mu, \nu) \geq b\,\omega^*(\mathbb{G})^{-\frac{1}{p'}} \left[\sup_{x,y \in \mathbb{G}} d_\mathbb{G}(x, y)\right]^{1-p} \mathcal{W}_p^p(\mu, \nu),$$

where $\mathcal{W}_p$ is the p-order Wasserstein distance[4] *with cost $d_\mathbb{G}^p$. Moreover, the equality is attained when $p = 1$.*

We next prove the negative definiteness for UST. This important property allows us to build positive definite kernels upon UST, required for kernel-dependent machine learning algorithmic approaches.

Proposition 5.4. *Under the same assumptions as in Corollary 4.6 and $w_1(z_0) = w_2(z_0)$. Then, US_p^α is negative definite on $\mathcal{M}(\mathbb{G})$ for any $1 \leq p \leq 2$.*

From Proposition 5.4 and by using (Berg et al., 1984, Theorem 3.2.2), we obtain that the kernel

$$k_{\mathrm{US}_p^\alpha}(\mu, \nu) \triangleq \exp(-t\mathrm{US}_p^\alpha(\mu, \nu))$$

is positive definite on $\mathcal{M}(\mathbb{G})$ for any given $t > 0$ and $1 \leq p \leq 2$.

6 EXPERIMENTS

In this section, we illustrate the fast computation (i.e., closed-form solution) of the proposed UST and comparable performances of the proposed positive definite kernel associated to UST against other popular unbalanced transport baselines and their corresponding kernels. More concretely, we evaluate for *measures with unequal mass on a given graph* under two simulations: document classification and topological data analysis (TDA).

Document classification. We consider four traditional document datasets: `TWITTER`, `RECIPE`, `CLASSIC`, and `AMAZON`. Their characteristics are summarized in Figure 1. We represent each document as a measure by considering each word in the document as its support with a unit mass. Therefore, *documents with different lengths have different total mass*. We employ the same word embedding procedure as in (Le and Nguyen, 2021) to embed words into vectors in $\mathbb{R}^{300}$.

TDA. We carry out two tasks: orbit recognition on `Orbit` dataset and object shape recognition on `MPEG7` dataset. For

[3] I.e., $\mathrm{US}_p^\alpha \geq 0$, and $\mathrm{US}_p^\alpha(\mu, \nu) = 0$ if and only if $\mu = \nu$.

[4] The definition of $\mathcal{W}_p$ is recalled in Appendix §B.1.

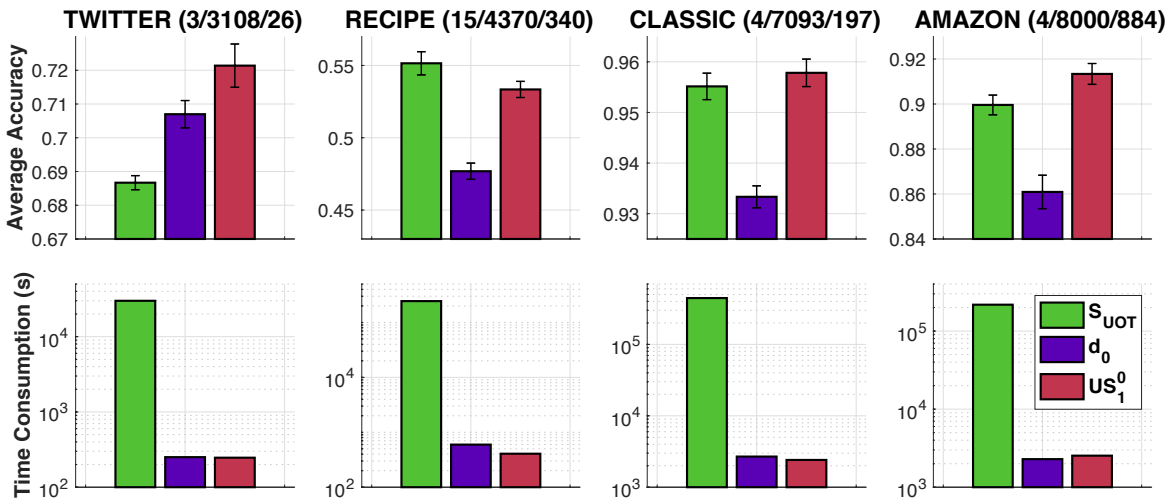

Figure 1: SVM results and time consumption for kernel matrices in document classification with graph $\mathbb{G}_{\text{Sqrt}}$. For each dataset, the numbers in the parenthesis are the number of classes; the number of documents; and the maximum number of unique words for each document respectively.

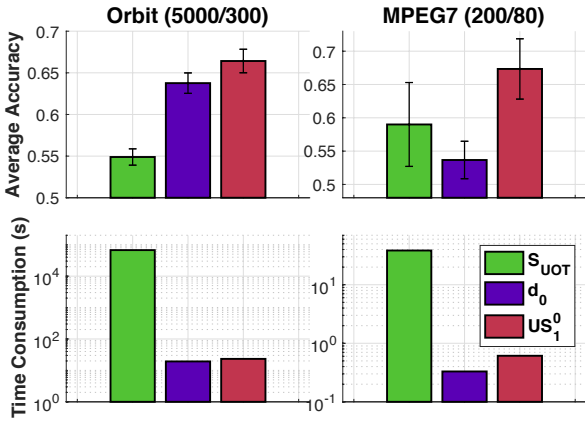

Figure 2: SVM results and time consumption for kernel matrices in TDA with graph $\mathbb{G}_{\text{Sqrt}}$. For each dataset, the numbers in the parenthesis are respectively the number of PD; and the maximum number of points in PD.

`Orbit` dataset, it is synthesized as in (Adams et al., 2017) for link twist map which are discrete dynamical systems to model flows in DNA microarrays (Hertzsch et al., 2007). There are five classes of orbits in the dataset. For each class, we generated 1000 orbits where each orbit contains 1000 points. For `MPEG7` dataset (Latecki et al., 2000), we consider its 10-class subset where each class has 20 samples as in (Le and Yamada, 2018). The characteristics of the considered `Orbit` and `MPEG7` datasets are summarized in Figure 2. We use the same procedure as in (Le and Nguyen, 2021) to extract persistence diagram (PD) for orbits and object shapes. PD are multisets of points in $\mathbb{R}^2$. Each point in PD summarizes the lifespan (i.e., birth and death time) of a topological feature (e.g., connected component, ring, cavity). We represent each PD as a measure by regarding each 2-dimensional point in PD as its support with a unit mass. Consequently, *persistence diagrams having a different number of topological features are represented as measures with different total mass*.

Notice that supports in document classification simulations are in high-dimensional spaces (i.e., in $\mathbb{R}^{300}$) while supports in TDA simulations are in low-dimensional spaces (i.e., in $\mathbb{R}^2$). Therefore, we can observe the effects of dimensions to the proposed UST and other unbalanced transport baselines from these simulations. We next describe various graph settings (i.e., the assumed graph metric spaces for measures) considered in our experiments.

Graph settings. We use the same graph settings (i.e., $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$) employed in (Le et al., 2022, §5) for our simulations on document classification and TDA. For these graphs, we consider the number of nodes: $M = 10^2, 10^3, 10^4, 4{\times}10^4$. We note that these graphs satisfy the assumptions in §2. Similar to the observations in (Le et al., 2022), each node in these graphs has a high probability to satisfy the root node condition, i.e., the uniqueness property of the shortest path (see Appendix §B.2 for a further discussion).

Root node z_0 for UST. The UST is defined over graph $\mathbb{G}$ with a root node z_0. From Definition 4.1, the root node z_0 imposes its own geometry by characterizing the graph derivative of functions on $\mathbb{G}$. To alleviate this dependency, we follow the sliced approach in (Le et al., 2022) for Sobolev transport by averaging over different choices of the root node z_0 in graph $\mathbb{G}$, which can be viewed as a sliced variant for UST.

Baselines, and experimental setup. We consider two typical UOT approaches for measures with unequal mass and

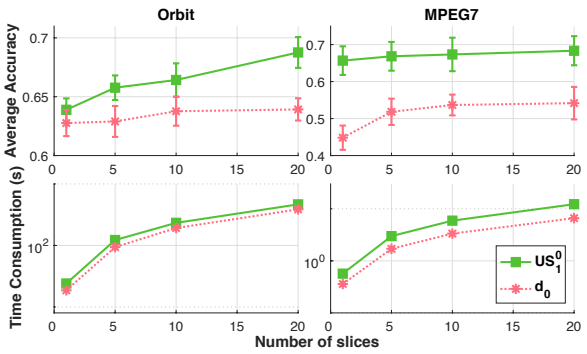

Figure 3: SVM results and time consumption for kernel matrices of slice variants in TDA with graph $\mathbb{G}_{\text{Sqrt}}$.

supported on a graph metric space as baselines: (i) the Sinkhorn-based UOT (Frogner et al., 2015; Chizat et al., 2018a) (S_{UOT})[5] with a graph metric ground cost, and (ii) the distance d_α of EPT on a *tree* (Le et al., 2022, Eq. (9)) (see Proposition 5.3 for its relation with US_p^α) where the tree structures are randomly sampled from graph $\mathbb{G}$. From results in Lemma 4.4 and Proposition 5.2 and for simplicity, we consider $\alpha = 0$ and $p = 1$ (and d_0 for EPT on a *tree* as in (Le and Nguyen, 2021))[6]. We further note that there are different approaches for simulations on document classification and TDA. However, that is *not* the purpose of our empirical simulations which compare different unbalanced transports for measures with unequal mass on a graph in the same settings.

We apply the kernel approach in the form $\exp(-t\bar{d})$, where $\bar{d}$ is a discrepancy for unbalanced measures on a *graph* and $t > 0$, with support vector machines (SVM) for the simulations on document classification and TDA. Note that kernels for US_p^α and d_α are positive definite, but kernels for S_{UOT} is empirically indefinite (see (Peyré and Cuturi, 2019, §8.3)). Similar to (Le and Nguyen, 2021), we regularized the Gram matrices for kernels with S_{UOT} by adding a sufficiently large diagonal term.

For simplicity, we employ the same setup for the EPT problem in (Le and Nguyen, 2021), i.e., using $\lambda = b = 1$ for the EPT. From Corollary 4.6 and Proposition 5.4, we consider the weight functions $w_1(x) = w_2(x) = a_1 d_{\mathbb{G}}(z_0, x) + a_0$ where $a_1 = b$ and $a_0 = 1$.

For kernel SVM, we use the same setting as in (Le and Nguyen, 2021). In each dataset, we randomly split it into $70\%/30\%$ for training and test with 10 repeats. We use 1-vs-1 strategy for SVM with multiclass data. Hyperparameters are typically chosen by cross validation. For kernel hyperparameter, we choose $1/t$ from $\{q_s, 2q_s, 5q_s\}$ with $s = 10, 20, \ldots, 90$ where q_s is the $s\%$ quantile of a random subset of corresponding distances on training data. For SVM regularization hyperparameter, we choose it from $\{0.01, 0.1, 1, 10, 100\}$. For S_{UOT}, we choose the entropic regularization from $\{0.01, 0.1, 1, 10\}$. The reported time consumption for each kernel matrices also includes the corresponding preprocessing, e.g., compute shortest paths on graph $\mathbb{G}$ for US_p^α and S_{UOT}, or sampling random tree structures from $\mathbb{G}$ for d_α of EPT on a tree.

Results of SVM, time consumption and discussions. We illustrate the SVM results and time consumption for kernel matrices for document classification and TDA in Figure 1 and Figure 2 with $M = 10^4$ for document datasets, $M = 10^3$ for `Orbit` and $M = 10^2$ for `MPEG7` for graph $\mathbb{G}_{\text{Sqrt}}$. The performances of kernels for our proposed UST compare favorably with other approaches (except S_{UOT} on `RECIPE`). Additionally, the time consumption of US_1^0 and d_0 is several-order faster than that of S_{UOT}. Recall that kernels for S_{UOT} is indefinite, which may affect performances of S_{UOT} in some datasets (e.g., `Orbit`, `TWITTER`). In Figure 3, we illustrate the effects of the number of slices (i.e., the number of root nodes used for averaging) for US_1^0 and d_0 for TDA. Generally, performances of those approaches are improved with more slices but with a trade-off on time consumption. We observe that 10 slices give a good trade-off in applications. Extensive further empirical results can be seen in Appendix §B.3, e.g., for various graph structures, graph sizes M, and different orders p of UST.

7 CONCLUSION

In this work, we proposed unbalanced Sobolev transport (UST) for measures with unequal mass on a *graph*. UST is the *first* variant of UOT having a *closed-form* formula for a fast computation. Additionally, UST is negative definite which allows to build positive definite kernels, required for kernel-dependent frameworks. Since UST exploits the graph metric structure of supports, it may restrict to applications with prior graph structures, or applications where one can build graphs from supports. On the other hand, we have not forseen any negative societal impacts of our work.

Acknowledgements

We thank anonymous reviewers and area chairs for their comments. KF has been supported in part by Grant-in-Aid for Transformative Research Areas (A) 22H05106. The research of TN is supported in part by a grant from the Simons Foundation (#318995). TL gratefully acknowledges the support of JSPS KAKENHI Grant number 20K19873. Finally, this research was enabled in part by computational support provided by Makoto Yamada.

[5] Séjourné et al. (2019) derived a debiased version for S_{UOT} which may be helpful in applications. The debiased version is also empirically indefinite and has the same complexity as S_{UOT}.

[6] One may tune these parameters for further improvements.

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.

Altschuler, J. M., Chewi, S., Gerber, P., and Stromme, A. J. (2021). Averaging on the Bures-Wasserstein manifold: Dimension-free convergence of gradient descent. *Advances in Neural Information Processing Systems*.

Balaji, Y., Chellappa, R., and Feizi, S. (2020). Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944.

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.

Bunne, C., Papaxanthos, L., Krause, A., and Cuturi, M. (2022). Proximal optimal transport modeling of population dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 6511–6528. PMLR.

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

Chapel, L., Alaya, M. Z., and Gasso, G. (2020). Partial optimal tranport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33:2903–2913.

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

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

Fan, J., Haasler, I., Karlsson, J., and Chen, Y. (2022). On the complexity of the optimal transport problem with graph-structured cost. In *International Conference on Artificial Intelligence and Statistics*, pages 9147–9165. PMLR.

Fatras, K., Séjourné, T., Flamary, R., and Courty, N. (2021). Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, pages 3186–3197. PMLR.

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

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.

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.

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

Janati, H., Muzellec, B., Peyré, G., and Cuturi, M. (2020). Entropic optimal transport between (unbalanced) gaussian measures has a closed form. In *Advances in neural information processing systems*.

Kondratyev, S., Monsaingeon, L., and Vorotnikov, D. (2016). A new optimal transport distance on the space of finite radon measures. *Advances in Differential Equations*, 21(11/12):1117–1164.

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.

Le, T., Ho, N., and Yamada, M. (2021a). Flow-based alignment approaches for probability measures in different spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 3934–3942. PMLR.

Le, T. and Nguyen, T. (2021). Entropy partial transport with tree metrics: Theory and practice. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pages 3835–3843. PMLR.

Le, T., Nguyen, T., Phung, D., and Nguyen, V. A. (2022). Sobolev transport: A scalable metric for probability measures with graph metrics. In *International Conference on Artificial Intelligence and Statistics*, pages 9844–9868.

Le, T., Nguyen, T., Yamada, M., Blanchet, J., and Nguyen, V. A. (2021b). Adversarial regression with doubly non-negative weighting matrices. *Advances in Neural Information Processing Systems*, 34.

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. (2019). Tree-sliced variants of Wasserstein distances. In *Advances in neural information processing systems*, pages 12283–12294.

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.

Liu, L., Pal, S., and Harchaoui, Z. (2022). Entropy regularized optimal transport independence criterion. In *International Conference on Artificial Intelligence and Statistics*, pages 11247–11279. PMLR.

Liu, Y., Yamada, M., Tsai, Y.-H. H., Le, T., Salakhutdinov, R., and Yang, Y. (2021). LSMI-Sinkhorn: Semi-supervised mutual information estimation with optimal transport. In *European Conference on Machine Learning and Principles & Practice of Knowledge Discovery in Databases (ECML-PKDD)*.

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.

Mukherjee, D., Guha, A., Solomon, J. M., Sun, Y., and Yurochkin, M. (2021). Outlier-robust optimal transport. In *International Conference on Machine Learning*, pages 7850–7860. PMLR.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443.

Muzellec, B., Josse, J., Boyer, C., and Cuturi, M. (2020). Missing data imputation using optimal transport. In *International Conference on Machine Learning*, pages 7130–7140. PMLR.

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.

Nguyen, T., Pham, Q.-H., Le, T., Pham, T., Ho, N., and Hua, B.-S. (2021a). Point-set distances for learning representations of 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10478–10487.

Nguyen, T. D., Trippe, B. L., and Broderick, T. (2022). Many processors, little time: Mcmc for partitions via optimal transport couplings. In *International Conference on Artificial Intelligence and Statistics*, pages 3483–3514.

Nguyen, V., Le, T., Yamada, M., and Osborne, M. A. (2021b). Optimal transport kernels for sequential and parallel neural architecture search. In *International Conference on Machine Learning*, pages 8084–8095. PMLR.

Nietert, S., Goldfeld, Z., and Cummings, R. (2022). Outlier-robust optimal transport: Duality, structure, and statistical analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 11691–11719. PMLR.

Paty, F.-P., d'Aspremont, A., and Cuturi, M. (2020). Regularity as regularization: Smooth and strongly convex Brenier potentials in optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 1222–1232. PMLR.

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.

Saleh, M., Wu, S.-C., Cosmo, L., Navab, N., Busam, B., and Tombari, F. (2022). Bending graphs: Hierarchical shape matching using gated optimal transport. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11757–11767.

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

Scetbon, M., Cuturi, M., and Peyré, G. (2021). Low-rank Sinkhorn factorization. *International Conference on Machine Learning (ICML)*.

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

Séjourné, T., Vialard, F.-X., and Peyré, G. (2022). Faster unbalanced optimal transport: Translation invariant sinkhorn and 1-d frank-wolfe. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 4995–5021. PMLR.

Si, N., Murthy, K., Blanchet, J., and Nguyen, V. A. (2021). Testing group fairness via optimal transport projections. *International Conference on Machine Learning*.

Takezawa, Y., Sato, R., Kozareva, Z., Ravi, S., and Yamada, M. (2022). Fixed support tree-sliced wasserstein barycenter. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 1120–1137. PMLR.

Titouan, V., Courty, N., Tavenard, R., and Flamary, R. (2019). Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pages 6275–6284. PMLR.

Tong, A., Huguet, G., Shung, D., Natik, A., Kuchroo, M., Lajoie, G., Wolf, G., and Krishnaswamy, S. (2022). Embedding signals on graphs with unbalanced diffusion earth mover's distance. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5647–5651.

Wang, J., Gao, R., and Xie, Y. (2022a). Two-sample test with kernel projected wasserstein distance. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 8022–8055. PMLR.

Wang, J., Zhang, Z., Chen, M., Zhang, Y., Wang, C., Sheng, B., Qu, Y., and Xie, Y. (2022b). Optimal transport for label-efficient visible-infrared person re-identification. In *European Conference on Computer Vision (ECCV)*, pages 93–109.

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.

In the appendix, we give further theoretical results and detailed proofs in §A. Additionally, we also give brief reviews about important definitions used in our work, additional discussions and further empirical results in §B.

Notations. Besides the notations in the main manuscript, we further denote $\langle x_1, x_2\rangle$ as the line segment in $\mathbb{R}^n$ connecting two points x_1, x_2 and (x_1, x_2) as the same line segment but without its two end-points.

A PROOFS AND ADDITIONAL THEORETICAL RESULTS

In this section, we give detailed proofs for the theoretical results in the main manuscript. We also provide some additional results for the unbalanced Sobolev transport (UST).

A.1 Further Theoretical Results

We include here some additional results for the transport problems and the unbalanced Sobolev transport US_p^α.

A.1.1 The Connection between Problem (3) and Problem (4)

We show the connection between problem (3) and problem (4) for EPT on a *graph* by following a similar reasoning as EPT on a *tree* (Le and Nguyen, 2021). It is a direct extension of results in (Le and Nguyen, 2021).

Theorem A.1. *Let $H(\lambda) \triangleq -\mathrm{ET}_{c,\lambda}(\mu,\nu)$ for $\lambda \in \mathbb{R}$, and denote*

$$\partial H(\lambda) \triangleq \Big\{p \in \mathbb{R} : H(t) \geq H(\lambda) + p(t-\lambda), \forall t \in \mathbb{R}\Big\}$$

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

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

$$\partial H(\lambda) = \big\{b\,\gamma(\mathbb{G}\times\mathbb{G}) : \gamma \in \Gamma^0(\lambda)\big\} \quad \forall \lambda \in \mathbb{R},$$

where we write Γ^0 for a set of all optimal plans γ. Also if $\lambda_1 < \lambda_2$, then $m_1 \leq m_2$ for every $m_1 \in \partial H(\lambda_1)$ and $m_2 \in \partial H(\lambda_2)$.

ii) H is differentiable at λ if and only if every optimal plan in $\Gamma^0(\lambda)$ has the same mass. When this happens, we also have

$$H'(\lambda) = b\,\gamma(\mathbb{G}\times\mathbb{G}),$$

for any $\gamma \in \Gamma^0(\lambda)$.

iii) If there exists a constant $M > 0$ such that

$$w_1(x) + w_2(y) \leq b\,[c(x,y) + M],$$

for all $x, y \in \mathbb{G}$, then $\partial H(\mathbb{R}) = [0, b\,\bar{m}]$. Moreover,

$$H(\lambda) = -\int_{\mathbb{G}} w_1\mu(\mathrm{d}x) - \int_{\mathbb{G}} w_2\nu(\mathrm{d}x),$$

when $\lambda < -M$, and $H'(\lambda) = b\,\bar{m}$ for $\lambda > \|c\|_{L^\infty(\mathbb{G}\times\mathbb{G})}$.

The proof is placed in §A.2.1.

For any $m \in [0, \bar{m}]$, part iii) of Theorem A.1 implies that there exists $\lambda \in \mathbb{R}$ such that $b\,m \in \partial H(\lambda)$. It then follows from part i) of this theorem that $m = \gamma^*(\mathbb{G}\times\mathbb{G})$ for some $\gamma^* \in \Gamma^0(\lambda)$. It is also clear that this γ^* is an optimal plan for $\mathrm{W}_{c,m}(\mu,\nu)$, and

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

Thus solving the auxiliary problem (4) gives us a solution to the original problem (3). When H is differentiable, the relation between m and λ is given explicitly as

$$H'(\lambda) = b\, m.$$

Note that the above selection of λ is unique only if the function H is strictly convex. Nevertheless, it enjoys the following monotonicity regardless of the uniqueness: if $m_1 < m_2$, then $\lambda_1 \leq \lambda_2$. Indeed, we have $m_1 = \gamma^1(\mathbb{G} \times \mathbb{G})$ and $m_2 = \gamma^2(\mathbb{G} \times \mathbb{G})$ for some $\gamma^1 \in \Gamma^0(\lambda_1)$ and $\gamma^2 \in \Gamma^0(\lambda_2)$. Since $\gamma^1(\mathbb{G} \times \mathbb{G}) < \gamma^2(\mathbb{G} \times \mathbb{G})$, one has $\lambda_1 \leq \lambda_2$ by i) of Theorem A.1.

A.1.2 $W^{1,\infty}(\mathbb{G}, \omega^*)$ versus Lipschitz space

We describe the connection between the Sobolev space $W^{1,\infty}(\mathbb{G}, \omega^*)$ and the space of Lipschitz continuous functions. The definition of the length measure ω^* is reviewed in §B.1.1).

Lemma A.2. *Let ω^* be the length measure on graph $\mathbb{G}$, and let $f : \mathbb{G} \to \mathbb{R}$ be a function. We have:*

i) If $|f(x) - f(y)| \leq b\, d_{\mathbb{G}}(x, y)$, $\forall x,\, y \in \mathbb{G}$, then $f \in W^{1,\infty}(\mathbb{G}, \omega^*)$ with $\|f'\|_{L^\infty(\mathbb{G},\omega^*)} \leq b$.

ii) Assume in addition that $\mathbb{G}$ is a tree. Then, $f \in W^{1,\infty}(\mathbb{G}, \omega^*)$ with $\|f'\|_{L^\infty(\mathbb{G},\omega^*)} \leq b$ implies that $|f(x) - f(y)| \leq b\, d_{\mathbb{G}}(x, y)$ for every $x,\, y \in \mathbb{G}$.

The proof is placed in §A.2.2.

Remark A.3. *Our proof for Lemma A.2 (in §A.2.2) also shows that the result in part ii) of Lemma A.2 in fact holds for every measure ω. Precisely, let ω be a nonnegative Borel measure on a tree $\mathbb{G}$. Then, we have $f \in W^{1,\infty}(\mathbb{G}, \omega)$ with $\|f'\|_{L^\infty(\mathbb{G},\omega)} \leq b$ implies that $|f(x) - f(y)| \leq b\, \omega([x, y])$ for every $x,\, y \in \mathbb{G}$.*

A.1.3 Comparison between Sobolev Spaces with Diferent Exponents

We derive a comparison between UST with different exponent p, and its proof is a direct consequence of our closed-form formula given in Proposition 4.5.

Proposition A.4 (Relation for different p)**.** *Assume that ω is a nonnegative Borel measure on $\mathbb{G}$. Then for any $1 \leq p \leq q \leq \infty$ and $\mu, \nu \in \mathcal{M}(\mathbb{G})$, we have*

$$\mathrm{US}_p^\alpha(\mu, \nu) - \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| \leq \omega(\mathbb{G})^{\frac{1}{p}-\frac{1}{q}} \Big[\mathrm{US}_q^\alpha(\mu, \nu) - \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})|\Big],$$

where Θ is the constant defined by (9)*.*

Proof of Proposition A.4. The case $p = q$ is trivial, so let us consider $1 \leq p < q \leq \infty$. Then by using Proposition 4.5 and Hölder's inequality, we obtain

$$\begin{aligned}
\mathrm{US}_p^\alpha(\mu, \nu) - \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| &= b \left(\int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p\, \omega(\mathrm{d}x)\right)^{\frac{1}{p}} \\
&\leq b\, \omega(\mathbb{G})^{\frac{1}{p}-\frac{1}{q}} \left(\int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^q\, \omega(\mathrm{d}x)\right)^{\frac{1}{q}} \\
&= \omega(\mathbb{G})^{\frac{1}{p}-\frac{1}{q}} \Big[\mathrm{US}_q^\alpha(\mu, \nu) - \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})|\Big].
\end{aligned}$$

■

A.1.4 Lower Bound for US_p^0

We derive a lower bound for US_p^0 which is a generalization of the result for $p = 1$ in Proposition 5.2.

Proposition A.5 (Lower bound for US_p^0)**.** *Let ω^* be the length measure on $\mathbb{G}$, and assume that w_1 and w_2 are b-Lipschitz w.r.t. $d_{\mathbb{G}}$. Then by taking $\omega = \omega^*$, we have for every $1 \leq p \leq \infty$ that*

$$\mathrm{US}_p^0(\mu, \nu) \geq \omega^*(\mathbb{G})^{-\frac{1}{p'}} \left\{\mathrm{ET}_\lambda(\mu, \nu) + \frac{b\lambda}{2}\big[\mu(\mathbb{G}) + \nu(\mathbb{G})\big]\right\} + \Theta[1 - \omega^*(\mathbb{G})^{-\frac{1}{p'}}]|\mu(\mathbb{G}) - \nu(\mathbb{G})|$$

for every $\mu,\, \nu \in \mathcal{M}(\mathbb{G})$. Here Θ is the constant defined by (9)*.*

Proof. This is a consequence of Corollary 3.2, Lemma 4.4, and Proposition A.4. ■

A.1.5 The Special Case of Balanced Mass

Observe that for the case $\mu(\mathbb{G}) = \nu(\mathbb{G})$, the constraint $f(z_0) \in I_\alpha$ in the definition of $\mathbb{U}_{p'}^\alpha$ is redundant. Indeed, we have:

Lemma A.6. *Let ω be a nonnegative Borel measure on $\mathbb{G}$. Assume that $\mu, \nu \in \mathcal{M}(\mathbb{G})$ satisfy $\mu(\mathbb{G}) = \nu(\mathbb{G})$. Then,*

$$\mathrm{US}_p^\alpha(\mu,\nu) = \sup\Big\{ \int_{\mathbb{G}} f(\mu-\nu) :\ f \in W^{1,p'}(\mathbb{G},\omega),\ \|f'\|_{L^{p'}(\mathbb{G},\omega)} \le b \Big\}.$$

In particular, $\mathrm{US}_p^\alpha(\mu,\nu)$ is independent of the parameters α, λ and the weights w_1, w_2.

Proof. This follows from the fact that Definition 4.3 is unchanged in the case $\mu(\mathbb{G}) = \nu(\mathbb{G})$ when the critic function f is translated by a constant. ■

From Lemma A.6, we see that for the case $\mu(\mathbb{G}) = \nu(\mathbb{G})$, our proposed unbalanced Sobolev transport US_p^α with $b = 1$ coincides with the balanced Sobolev transport $\mathcal{S}_p$ (defined in (Le et al., 2022, Definition 3.2)).

A.1.6 Infinite Divisibility for Unbalanced Sobolev Transport Kernel

Recall that given $t > 0$ and $1 \le p \le 2$, the unbalanced Sobolev transport kernel $k_{\mathrm{US}_p^\alpha}(\mu,\nu) \triangleq \exp(-t\mathrm{US}_p^\alpha(\mu,\nu))$ is positive definite (see §5 and Proposition 5.4).

For $i \in \mathbb{N}^*$, the kernel $k_{\mathrm{US}_{pi}^\alpha}(\mu,\nu) \triangleq \exp(-\frac{t}{i}\mathrm{US}_p^\alpha(\mu,\nu))$ is positive definite. Additionally, $k_{\mathrm{US}_p^\alpha}(\mu,\nu) = \Big[k_{\mathrm{US}_{pi}^\alpha}(\mu,\nu)\Big]^i$. Therefore, $k_{\mathrm{US}_p^\alpha}$ is indefinitely divisible following (Berg et al., 1984, Definition 2.6 in §3).

Hence, one does not need to recompute the Gram matrix for unbalanced Sobolev transport kernel $k_{\mathrm{US}_p^\alpha}$ for different values of t. Indeed, it is suffice to compute the Gram matrix of $k_{\mathrm{US}_p^\alpha}$ once for some fixed t and leverage its indefinite divisibility for other values of t.

A.2 Detailed Proofs

In this section, we give detailed proofs for our theoretical results.

A.2.1 Proof of Theorem A.1

Proof of Theorem A.1. We employ a similar reasoning for EPT on a tree (Le and Nguyen, 2021) to prove the relation between problem (3) and problem (4) for EPT on a *graph* as follow:

i) Note that $\lambda \mapsto \mathrm{ET}_{c,\lambda}(\mu,\nu)$ is a concave function since it is the infimum of a family of concave functions in λ. Therefore, H is convex on $\mathbb{R}$. In particular, H is differentiable almost everywhere on $\mathbb{R}$.

Let $\lambda \in \mathbb{R}$, recall the definition of $\mathcal{C}_\lambda(\gamma)$ in Equation (5). Then for any $\gamma \in \Gamma^0(\lambda)$, we have

$$\mathrm{ET}_{c,\lambda+\delta}(\mu,\nu) \le \mathcal{C}_{\lambda+\delta}(\gamma) = \mathcal{C}_\lambda(\gamma) - b\delta\gamma(\mathbb{G}\times\mathbb{G}) = \mathrm{ET}_{c,\lambda}(\mu,\nu) - b\delta\gamma(\mathbb{G}\times\mathbb{G})\ \ \forall\delta\in\mathbb{R}. \tag{11}$$

This implies that

$$\big\{b\,\gamma(\mathbb{G}\times\mathbb{G}) : \gamma \in \Gamma^0(\lambda)\big\} \subset \partial H(\lambda).$$

We next show that the opposite inclusion is also true, i.e., $\big\{b\,\gamma(\mathbb{G}\times\mathbb{G}) : \gamma \in \Gamma^0(\lambda)\big\} = \partial H(\lambda)$. This is obviously holds if $\partial H(\lambda)$ is singleton, which holds for example when H is differentiable at λ. Hence we only need to consider λ for which the convex set $\partial H(\lambda)$ has more than one element.

Let $m \in \partial H(\lambda)$, then m can be expressed as a convex combination of extreme points $m_1,\ldots,m_N$ of $\partial H(\lambda)$, i.e., $m = \sum_{i=1}^N t_i m_i$ with $0 \le t_i \le 1$ and $\sum_{i=1}^N t_i = 1$. As m_i is an extreme point of $\partial H(\lambda)$, there exists a sequence $\lambda_n \to \lambda$ such that λ_n is a differentiable point of H and $H'(\lambda_n) \to m_i$.

Let $\gamma^n \in \Gamma^0(\lambda_n)$, then $b\,\gamma^n(\mathbb{G}\times\mathbb{G}) = H'(\lambda_n) \to m_i$. By compactness, there exists a subsequence $\{\gamma^{n_k}\}$ and $\tilde\gamma^i \in \Pi_\le(\mu,\nu)$ such that $\gamma^{n_k} \to \tilde\gamma^i$ weakly. It follows that $\gamma^{n_k}(\mathbb{G}\times\mathbb{G}) \to \tilde\gamma^i(\mathbb{G}\times\mathbb{G})$, and hence we must have $b\,\tilde\gamma^i(\mathbb{G}\times\mathbb{G}) = m_i$. We have

$$\begin{aligned}\mathcal{C}_{\lambda_{n_k}}(\gamma^{\lambda_{n_k}}) = \mathcal{C}_\lambda(\gamma^{\lambda_{n_k}}) + b(\lambda-\lambda_{n_k})\gamma^{n_k}(\mathbb{G}\times\mathbb{G}) &\ge \mathrm{ET}_{c,\lambda}(\mu,\nu) + b(\lambda-\lambda_{n_k})\gamma^{n_k}(\mathbb{G}\times\mathbb{G}) \\ &\ge \mathrm{ET}_{c,\lambda}(\mu,\nu) - b\bar m|\lambda-\lambda_{n_k}|\end{aligned}$$

and for any $\gamma \in \Gamma^0(\lambda)$, there holds

$$\mathcal{C}_{\lambda_{n_k}}(\gamma^{\lambda_{n_k}}) \le \mathcal{C}_{\lambda_{n_k}}(\gamma) = \mathcal{C}_\lambda(\gamma) + b(\lambda - \lambda_{n_k})\gamma(\mathbb{G}\times\mathbb{G}) = \mathrm{ET}_{c,\lambda}(\mu,\nu) + b(\lambda-\lambda_{n_k})\gamma(\mathbb{G}\times\mathbb{G}).$$

We thus deduce that $\lim_{k\to\infty}\mathcal{C}_{\lambda_{n_k}}(\gamma^{\lambda_{n_k}}) = \mathrm{ET}_{c,\lambda}(\mu,\nu)$. These together with the lower semicontinuity of $\mathcal{C}_\lambda$ give

$$\begin{aligned}\mathrm{ET}_{c,\lambda}(\mu,\nu) = \liminf_{k\to\infty}\mathcal{C}_{\lambda_{n_k}}(\gamma^{\lambda_{n_k}}) &= \liminf_{k\to\infty}\Big[\mathcal{C}_\lambda(\gamma^{\lambda_{n_k}}) + b(\lambda-\lambda_{n_k})\gamma^{n_k}(\mathbb{G}\times\mathbb{G})\Big]\\ &= \liminf_{k\to\infty}\mathcal{C}_\lambda(\gamma^{\lambda_{n_k}}) \ge \mathcal{C}_\lambda(\tilde\gamma^i).\end{aligned}$$

Therefore, $\tilde\gamma^i \in \Gamma^0(\lambda)$ with mass $b\,\tilde\gamma^i(\mathbb{G}\times\mathbb{G}) = m_i$. Due to the convexity of $\Gamma^0(\lambda)$, we have $\bar\gamma := \sum_{i=1}^N t_i\tilde\gamma^i \in \Gamma^0(\lambda)$ with $b\,\bar\gamma(\mathbb{G}\times\mathbb{G}) = \sum_{i=1}^N t_i m_i = m$. That is,

$$\partial H(\lambda) \subset \big\{b\,\gamma(\mathbb{G}\times\mathbb{G}) : \gamma\in\Gamma^0(\lambda)\big\},$$

and we thus infer that $\big\{b\,\gamma(\mathbb{G}\times\mathbb{G}) : \gamma\in\Gamma^0(\lambda)\big\} = \partial H(\lambda)$ for all $\lambda\in\mathbb{R}$.

In order to prove the second part of i), let $\gamma\in\Gamma^0(\lambda_1)$ and $\tilde\gamma\in\Gamma^0(\lambda_2)$ be arbitrary. We have

$$\begin{aligned}\mathrm{ET}_{c,\lambda_2}(\mu,\nu) = \mathcal{C}_{\lambda_2}(\tilde\gamma) &= \mathcal{C}_{\lambda_1}(\tilde\gamma) - b(\lambda_2-\lambda_1)\tilde\gamma(\mathbb{G}\times\mathbb{G})\\ &\ge \mathrm{ET}_{c,\lambda_1}(\mu,\nu) - b(\lambda_2-\lambda_1)\tilde\gamma(\mathbb{G}\times\mathbb{G}).\end{aligned} \tag{12}$$

Hence by combining with (11), we deduce that

$$\mathrm{ET}_{c,\lambda_1}(\mu,\nu) - b(\lambda_2-\lambda_1)\tilde\gamma(\mathbb{G}\times\mathbb{G}) \le \mathrm{ET}_{c,\lambda_2}(\mu,\nu) \le \mathrm{ET}_{c,\lambda_1}(\mu,\nu) - b(\lambda_2-\lambda_1)\gamma(\mathbb{G}\times\mathbb{G}),$$

which yields $\gamma(\mathbb{G}\times\mathbb{G}) \le \tilde\gamma(\mathbb{G}\times\mathbb{G})$. This together with the above characterization of $\partial H(\lambda)$ implies the second part of i).

ii) If H is differentiable at λ, then $\partial H(\lambda)$ is a singleton set. However, as $\partial H(\lambda) = \big\{b\,\gamma(\mathbb{G}\times\mathbb{G}) : \gamma\in\Gamma^0(\lambda)\big\}$ by i), we thus infer that the mass $\gamma(\mathbb{G}\times\mathbb{G})$ must be the same for every $\gamma\in\Gamma^0(\lambda)$.

Next assume that every element in $\Gamma^0(\lambda)$ has the same mass, say m. For $\delta\neq 0$, let $\gamma^{\lambda+\delta}\in\Gamma^0(\lambda+\delta)$ and $m(\lambda+\delta) \triangleq \gamma^{\lambda+\delta}(\mathbb{G}\times\mathbb{G})$. Then, we claim that

$$\lim_{\delta\to 0} m(\lambda+\delta) = m. \tag{13}$$

Assume the claim for the moment, and let $\delta>0$. Then, as in (11)–(12), we have

$$\mathrm{ET}_{c,\lambda+\delta}(\mu,\nu) \le \mathrm{ET}_{c,\lambda}(\mu,\nu) - b\delta m \quad\text{and}\quad \mathrm{ET}_{c,\lambda+\delta}(\mu,\nu) \ge \mathrm{ET}_{c,\lambda}(\mu,\nu) - b\delta m(\lambda+\delta).$$

It follows that

$$-bm(\lambda+\delta) \le \frac{\mathrm{ET}_{c,\lambda+\delta}(\mu,\nu) - \mathrm{ET}_{c,\lambda}(\mu,\nu)}{\delta} \le -bm.$$

This together with claim (13) gives $\lim_{\delta\to 0^+}\frac{\mathrm{ET}_{c,\lambda+\delta}(\mu,\nu)-\mathrm{ET}_{c,\lambda}(\mu,\nu)}{\delta} = -bm$. By the same argument, we also have $\lim_{\delta\to 0^-}\frac{\mathrm{ET}_{c,\lambda+\delta}(\mu,\nu)-\mathrm{ET}_{c,\lambda}(\mu,\nu)}{\delta} = -bm$. Thus, we infer that H is differentiable at λ with $H'(\lambda) = bm$. Therefore, it remains to prove claim (13).

Indeed, by compactness there exists a subsequence, still labeled by $\gamma^{\lambda+\delta}$, and $\gamma\in\Pi_{\le}(\mu,\nu)$ such that $\gamma^{\lambda+\delta}\to\gamma$ weakly as $\delta\to 0$. As in i), we can show that $\gamma\in\Gamma^0(\lambda)$. Then, as the mass functional is weakly continuous, we obtain $m(\lambda+\delta) = \gamma^{\lambda+\delta}(\mathbb{G}\times\mathbb{G}) \to \gamma(\mathbb{G}\times\mathbb{G}) = m$. We in fact have shown that any subsequence of $\{m(\lambda+\delta)\}_\delta$ has a further subsequence converging to the same number m. Therefore, the full sequence $\{m(\lambda+\delta)\}_\delta$ must converge to m, and hence (13) is proved.

iii) For any $\lambda\in\mathbb{R}$, we have by i) that $\partial H(\lambda) = \big\{b\,\gamma(\mathbb{G}\times\mathbb{G}) : \gamma\in\Gamma^0(\lambda)\big\} \subset [0, b\,\bar m]$. Thus, we only need to prove $[0,b\,\bar m]\subset\partial H(\mathbb{R})$. First, note that as $\partial H(\lambda)\subset\mathbb{R}$ is a compact and convex set, it must be a finite and closed interval. Therefore, if we let

$$\gamma^\lambda_{min} := \operatorname*{arg\,min}_{\gamma\in\Gamma^0(\lambda)}\gamma(\mathbb{G}\times\mathbb{G}) \quad\text{and}\quad \gamma^\lambda_{max} := \operatorname*{arg\,max}_{\gamma\in\Gamma^0(\lambda)}\gamma(\mathbb{G}\times\mathbb{G}),$$

then it follows from ii) that $\partial H(\lambda) = \big[b\,\gamma^\lambda_{min}(\mathbb{G}\times\mathbb{G}), b\,\gamma^\lambda_{max}(\mathbb{G}\times\mathbb{G})\big]$ for every $\lambda\in\mathbb{R}$. From Equation (5), it is clear that $\partial H(\lambda) = \{0\}$ for λ negative enough. Indeed, if we take $\lambda < -M$, then as $w_1(x) + w_2(y) \le b\,[c(x,y)+M]$, we

have $0 < b\,[c(x,y) - \lambda] - w_1(x) - w_2(y)$ for all $x, y \in \mathbb{G}$. Then, we obtain from Equation (5) that $\mathcal{C}_\lambda(0) \leq \mathcal{C}_\lambda(\gamma)$ for every $\gamma \in \Pi_{\leq}(\mu,\nu)$ and the strict inequality holds if $\gamma \neq 0$. Thus, $\Gamma^0(\lambda) = \{0\}$ which gives $\partial H(\lambda) = \{0\}$ and $H(\lambda) = -\int_{\mathbb{G}} w_1 \mu(\mathrm{d}x) - \int_{\mathbb{G}} w_2 \nu(\mathrm{d}x)$.

We next show that $\partial H(\lambda) = \{b\,\bar{m}\}$ for λ positive enough. Since $c(x,y)$ is bounded due to its continuity on $\mathbb{G} \times \mathbb{G}$, we can choose $\lambda \in \mathbb{R}$ such that $c(x,y) - \lambda < 0$ for all $x, y \in \mathbb{G}$. Let $\gamma \in \Gamma^0(\lambda)$. We claim that either $\gamma_1 = \mu$ or $\gamma_2 = \nu$. Indeed, since otherwise we have $\gamma_1(A_0) < \mu(A_0)$ and $\gamma_2(B_0) < \nu(B_0)$ for some Borel sets $A_0, B_0 \subset \mathbb{G}$. Let $\tilde{\gamma} := \gamma + [(\mu - \gamma_1)\chi_{A_0}] \otimes [(\nu - \gamma_2)\chi_{B_0}]$. Then, for any Borel set $A \subset \mathbb{G}$ we have

$$\begin{aligned}\tilde{\gamma}_1(A) = \gamma_1(A) + \mu(A \cap A_0) - \gamma_1(A \cap A_0) &= \gamma_1(A \setminus A_0) + \mu(A \cap A_0) \\ &\leq \mu(A \setminus A_0) + \mu(A \cap A_0) = \mu(A).\end{aligned}$$

Likewise, $\tilde{\gamma}_2(B) \leq \nu(B)$ for any Borel set $B \subset \mathbb{G}$. Thus $\tilde{\gamma} \in \Pi_{\leq}(\mu,\nu)$. On the other hand, it is clear from (5) and the facts $\gamma_1 \leq \tilde{\gamma}_1$, $\gamma_2 \leq \tilde{\gamma}_2$, and $c - \lambda < 0$ that $\mathcal{C}_\lambda(\tilde{\gamma}) < \mathcal{C}_\lambda(\gamma)$. This is impossible and so the claim is proved. That is, either $\gamma_1 = \mu$ or $\gamma_2 = \nu$. It follows that $\gamma(\mathbb{G} \times \mathbb{G}) = \bar{m}$ for every $\gamma \in \Gamma^0(\lambda)$, and hence $\partial H(\lambda) = \{b\,\bar{m}\}$ due to i). This also means that H is differentiable at λ with $H'(\lambda) = b\,\bar{m}$.

Therefore, it remains to show that

$$(0, b\,\bar{m}) \subset \partial H(\mathbb{R}) = \bigcup_{\lambda \in \mathbb{R}} \big[b\,\gamma^\lambda_{min}(\mathbb{G} \times \mathbb{G}), b\,\gamma^\lambda_{max}(\mathbb{G} \times \mathbb{G})\big]. \tag{14}$$

Assume by contradiction that there exists $m \in (0, b\,\bar{m})$ such that $m \notin \partial H(\lambda)$ for every $\lambda \in \mathbb{R}$. For convenience, we adopt the following notation: for sets $A, B \subset \mathbb{R}$ and $r \in \mathbb{R}$, we write $A < r$ if $a < r$ for every $a \in A$, and $A < B$ if $a < b$ for every $a \in A$ and $b \in B$. Let us consider the following two sets

$$S_1 := \{\lambda : \partial H(\lambda) < m\} \quad \text{and} \quad S_2 := \{\lambda : \partial H(\lambda) > m\}.$$

Then $\lambda \in S_1$ if λ is negative enough, and $\lambda \in S_2$ if λ is positive enough. For any $\lambda_1 \in S_1$ and $\lambda_2 \in S_2$, we have $\partial H(\lambda_1) < m < \partial H(\lambda_2)$, and hence $\lambda_1 < \lambda_2$ by the monotonicity in i). That is, $S_1 < S_2$ and so we obtain

$$\lambda^* := \sup\{\lambda : \lambda \in S_1\} \leq \inf\{\lambda : \lambda \in S_2\} =: \lambda^{**}.$$

If $\lambda^* < \lambda^{**}$, then for any $\lambda \in (\lambda^*, \lambda^{**})$ we have $\lambda \notin S_1$ and $\lambda \notin S_2$. Therefore, $\partial H(\lambda) \not< m$ and $\partial H(\lambda) \not> m$. Hence, we can find $m_1, m_2 \in \partial H(\lambda)$ such that $m_1 \geq m$ and $m_2 \leq m$. Thus, $m \in [m_2, m_1] \subset \partial H(\lambda)$ due to the convexity of the set $\partial H(\lambda)$. This contradicts our hypothesis, and we conclude that $\lambda^* = \lambda^{**}$.

We next select sequences $\{\lambda^1_n\} \subset S_1$ and $\{\lambda^2_n\} \subset S_2$ such that $\lambda^1_n \to \lambda^*$ and $\lambda^2_n \to \lambda^{**} = \lambda^*$. For each n, let

$$\gamma^n_{min} := \operatorname*{arg\,min}_{\gamma \in \Gamma^0(\lambda^1_n)} \gamma(\mathbb{G} \times \mathbb{G}) \quad \text{and} \quad \gamma^n_{max} := \operatorname*{arg\,max}_{\gamma \in \Gamma^0(\lambda^2_n)} \gamma(\mathbb{G} \times \mathbb{G}).$$

By compactness, there exist subsequences, still labeled as $\{\gamma^n_{min}\}$ and $\{\gamma^n_{max}\}$, and $\gamma^*, \gamma^{**} \in \Pi_{\leq}(\mu,\nu)$ such that $\gamma^n_{min} \to \gamma^*$ weakly and $\gamma^n_{max} \to \gamma^{**}$ weakly. By arguing exactly as in i), we then obtain $\gamma^*, \gamma^{**} \in \Gamma^0(\lambda^*)$, $\gamma^n_{min}(\mathbb{G}\times\mathbb{G}) \to \gamma^*(\mathbb{G} \times \mathbb{G})$, and $\gamma^n_{max}(\mathbb{G} \times \mathbb{G}) \to \gamma^{**}(\mathbb{G} \times \mathbb{G})$. As $b\,\gamma^n_{min}(\mathbb{G} \times \mathbb{G}) < m$ due to $\lambda^1_n \in S_1$, we must have $b\,\gamma^*(\mathbb{G} \times \mathbb{G}) \leq m$. Likewise, we have $b\,\gamma^{**}(\mathbb{G} \times \mathbb{G}) \geq m$ as $b\,\gamma^n_{max}(\mathbb{G} \times \mathbb{G}) > m$ for all n. Hence, $m \in [b\,\gamma^*(\mathbb{G} \times \mathbb{G}), b\,\gamma^{**}(\mathbb{G} \times \mathbb{G})]$. Since $\gamma^*, \gamma^{**} \in \Gamma^0(\lambda^*)$, we infer that $m \in \partial H(\lambda^*)$. This is a contradiction and the proof is complete. We note that since $\lambda^1_n \leq \lambda^* \leq \lambda^2_n$, we have from the monotonicity in i) that

$$\gamma^n_{min}(\mathbb{G} \times \mathbb{G}) \leq \gamma(\mathbb{G} \times \mathbb{G}) \leq \gamma^n_{max}(\mathbb{G} \times \mathbb{G})$$

for every $\gamma \in \Gamma^0(\lambda^*)$. By sending n to infinity, it follows that $\gamma^*(\mathbb{G}\times\mathbb{G}) \leq \gamma(\mathbb{G}\times\mathbb{G}) \leq \gamma^{**}(\mathbb{G}\times\mathbb{G})$ for every $\gamma \in \Gamma^0(\lambda^*)$. That is, $\gamma^* = \gamma^{\lambda^*}_{min}$ and $\gamma^{**} = \gamma^{\lambda^*}_{max}$. ■

A.2.2 Proof of Lemma A.2

Proof of Lemma A.2. Let us define

$$A \triangleq \Big\{ f \in C(\mathbb{G}) :\, |f(x) - f(y)| \leq b\, d_{\mathbb{G}}(x,y) \Big\}.$$

and

$$B \triangleq \Big\{ f \in W^{1,\infty}(\mathbb{G}, \omega^*) : \|f'\|_{L^\infty(\mathbb{G},\omega^*)} \le b \Big\}$$

i) The statement of this part is equivalent to showing that $A \subset B$. Let $f \in A$. Then f is continuous on $\mathbb{G}$, and

$$|f(x) - f(y)| \le b\, d_{\mathbb{G}}(x, y) \quad \forall x, y \in \mathbb{G}. \tag{15}$$

On each edge e and similar to the real line, the Lipschitz condition (15) implies that there exists a function $h_e : e \to \mathbb{R}$ with the following properties: $|h_e(z)| \le b$ for ω^*-a.e. $z \in e$, and

$$f(x) = f(y) + \int_{\langle y,x\rangle} h_e(z)\, \omega^*(\mathrm{d}z) \quad \forall x, y \in e,$$

where we recall that $\langle y, x\rangle$ denotes the line segment in $\mathbb{R}^n$ connecting y and x (noting that for general graph, $\langle y, x\rangle$ might not be the same as the shortest path $[y, x]$). Let us glue them together by taking $h(z) = h_e(z)$ if z is an interior point of an edge e. Then $h : \mathbb{G} \to \mathbb{R}$ is a function satisfying: $|h(z)| \le b$ for ω^*-a.e. $z \in G$. That is, $h \in L^\infty(\mathbb{G}, \omega^*)$ with $\|h\|_{L^\infty(\mathbb{G},\omega^*)} \le b$. Moreover, for every edge e in $\mathbb{G}$ we have

$$f(x) = f(y) + \int_{\langle y,x\rangle} h(z)\, \omega^*(\mathrm{d}z) \quad \forall x, y \in e. \tag{16}$$

Now let $x \in \mathbb{G}$ be arbitrary. Let us break the unique shortest path $[z_0, x]$ connecting z_0 and x into sub line segments $\langle z_0, y_0\rangle$, $\langle y_0, y_1\rangle$, ..., $\langle y_{m-1}, y_m\rangle$, $\langle y_m, x\rangle$ such that each of them is contained in exactly one edge. Then by applying (16) to each of these sub line segments, we obtain

$$\begin{aligned} f(x) - f(z_0) &= [f(x) - f(y_m)] + [f(y_m) - f(y_{m-1})] + \cdots + [f(y_0) - f(z_0)] \\ &= \int_{\langle y_m,x\rangle} h(z)\, \omega^*(\mathrm{d}z) + \int_{\langle y_{m-1},y_m\rangle} h(z)\, \omega^*(\mathrm{d}z) + \cdots + \int_{\langle z_0,y_0\rangle} h(z)\, \omega^*(\mathrm{d}z) \\ &= \int_{[z_0,x]} h(z)\, \omega^*(\mathrm{d}z). \end{aligned}$$

Thus, we have proved that

$$f(x) = f(z_0) + \int_{[z_0,x]} h(z)\, \omega^*(\mathrm{d}z) \quad \forall x \in \mathbb{G}.$$

Therefore, according to Definition 4.1 we conclude that $f \in W^{1,\infty}(\mathbb{G}, \omega^*)$ with $\|f'\|_{L^{p'}(\mathbb{G},\omega^*)} \le b$. It then follows that $f \in B$, and hence $A \subset \mathbb{B}$ as desired.

ii) Assume that $\mathbb{G}$ is a tree. We can and will assume that z_0 is the root of this tree. We need to show that $B \subset A$. For this, let $f \in B$. Then by Definition 4.1, we have $\|f'\|_{L^\infty(\mathbb{G},\omega^*)} \le b$ and

$$f(x) = f(z_0) + \int_{[z_0,x]} f'(z)\, \omega^*(\mathrm{d}z) \quad \forall x \in \mathbb{G}.$$

Thus for any two points $x, y \in \mathbb{G}$, we obtain

$$|f(x) - f(y)| = \left| \int_{[z_0,x]} f'(z)\, \omega^*(\mathrm{d}z) - \int_{[z_0,y]} f'(z)\, \omega^*(\mathrm{d}z) \right|. \tag{17}$$

Let $\hat{z}$ be the deepest node on the tree that belongs to both path $[z_0, x]$ and path $[z_0, y]$. Due to the tree structure, the joining of path $[x, \hat{z}]$ and path $[\hat{z}, y]$ constitutes the shortest path $[x, y]$ connecting the points x and y. These together with (17) imply that

$$\begin{aligned} |f(x) - f(y)| &= \left| \int_{[\hat{z},x]} f'(z)\, \omega^*(\mathrm{d}z) - \int_{[\hat{z},y]} f'(z)\, \omega^*(\mathrm{d}z) \right| \\ &\le \int_{[x,\hat{z}]} |f'(z)|\, \omega^*(\mathrm{d}z) + \int_{[\hat{z},y]} |f'(z)|\, \omega^*(\mathrm{d}z) \\ &= \int_{[x,y]} |f'(z)|\, \omega^*(\mathrm{d}z) \le \|f'\|_{L^\infty(\mathbb{G},\omega^*)} \omega^*([x, y]) \le b\, \omega^*([x, y]). \end{aligned}$$

By the property of the length measure given in Lemma B.2, we then infer that $|f(x) - f(y)| \le b\, d_{\mathbb{G}}(x, y)$ for every $x, y \in \mathbb{G}$. It follows that $f \in A$. Therefore, we have proved that $B \subset A$ as desired. ■

A.2.3 Proof of Theorem 3.1

The proof of Theorem 3.1 is based on two auxiliary lemmas. Before stating these lemmas, let us describe the the setting and associated problem.

First, in order to investigate problem (4), 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 graph $\mathbb{G}$ and consider the set $\hat{\mathbb{G}} := \mathbb{G} \cup \{\hat{s}\}$. We next extend the cost function to $\hat{\mathbb{G}} \times \hat{\mathbb{G}}$ as follow

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

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

$$\mathrm{KT}(\hat{\mu}, \hat{\nu}) \triangleq \inf_{\hat{\gamma} \in \Gamma(\hat{\mu}, \hat{\nu})} \int_{\hat{\mathbb{G}} \times \hat{\mathbb{G}}} \hat{c}(x,y) \hat{\gamma}(\mathrm{d}x, \mathrm{d}y), \tag{18}$$

where

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

This reformulation under an observation in (Caffarelli and McCann, 2010) helps us to transform an unbalanced optimal transport (EPT) on a graph into a corresponding standard complete OT. Therefore, we can not only bypass all the issues coming from the unbalanced setting, but also rely on many results in the standard setting for OT.

We then adapt the procedure in (Caffarelli and McCann, 2010) to derive the dual formulation for the EPT on a *graph*.

Additionally, we have a one-to-one correspondence between $\gamma \in \Pi_{\leq}(\mu, \nu)$ and $\hat{\gamma} \in \Gamma(\hat{\mu}, \hat{\nu})$ as follow

$$\hat{\gamma} = \gamma + [(1 - f_1)\mu] \otimes \delta_{\hat{s}} + \delta_{\hat{s}} \otimes [(1 - f_2)\nu] + \gamma(\mathbb{G} \times \mathbb{G})\delta_{(\hat{s}, \hat{s})}. \tag{19}$$

Indeed, if $\gamma \in \Pi_{\leq}(\mu, \nu)$, then it is clear that $\hat{\gamma}$ defined by (19) satisfies $\hat{\gamma} \in \Gamma(\hat{\mu}, \hat{\nu})$. The converse is guaranteed by the next technical result.

Lemma A.7. *For $\hat{\gamma} \in \Gamma(\hat{\mu}, \hat{\nu})$, let γ be the restriction of $\hat{\gamma}$ to $\mathbb{G}$. Then, relation* (19) *holds and $\gamma \in \Pi_{\leq}(\mu, \nu)$.*

Proof. We first observe for any Borel set $A \subset \mathbb{G}$ that

$$\hat{\gamma}(A \times \{\hat{s}\}) = \hat{\gamma}(A \times \hat{\mathbb{G}}) - \hat{\gamma}(A \times \mathbb{G}) = \hat{\mu}(A) - \gamma(A \times \mathbb{G}) = \mu(A) - \gamma_1(A) = \int_A (1 - f_1)\mu(\mathrm{d}x).$$

For the same reason, we have $\hat{\gamma}(\{\hat{s}\} \times B) = \int_B (1 - f_2)\nu(dx)$ for any set Borel set $B \subset \mathbb{G}$. Also,

$$\begin{aligned} \hat{\gamma}(\{\hat{s}\} \times \{\hat{s}\}) &= \hat{\gamma}(\hat{\mathbb{G}} \times \{\hat{s}\}) - \hat{\gamma}(\mathbb{G} \times \{\hat{s}\}) \\ &= \hat{\gamma}(\hat{\mathbb{G}} \times \hat{\mathbb{G}}) - \hat{\gamma}(\hat{\mathbb{G}} \times \mathbb{G}) - \big[\hat{\gamma}(\mathbb{G} \times \hat{\mathbb{G}}) - \hat{\gamma}(\mathbb{G} \times \mathbb{G})\big] \\ &= \hat{\mu}(\hat{\mathbb{G}}) - \hat{\nu}(\mathbb{G}) - \hat{\mu}(\mathbb{G}) + \gamma(\mathbb{G} \times \mathbb{G}) = \gamma(\mathbb{G} \times \mathbb{G}). \end{aligned}$$

Since (19) is obviously true for sets of the form $A \times B$ with $A, B \subset \mathbb{G}$ being Borel sets, we only need to verify it for sets of the following three forms: $(A \cup \{\hat{s}\}) \times B$, $A \times (B \cup \{\hat{s}\})$, $(A \cup \{\hat{s}\}) \times (B \cup \{\hat{s}\})$ for Borel sets $A, B \subset \mathbb{G}$. We check it case by case as follows.

• (i) For $(A \cup \{\hat{s}\}) \times B$: Using the above observation, we have

$$\hat{\gamma}((A \cup \{\hat{s}\}) \times B) = \hat{\gamma}(A \times B) + \hat{\gamma}(\{\hat{s}\} \times B) = \gamma(A \times B) + \int_B (1 - f_2)\nu(\mathrm{d}x).$$

Therefore, (19) holds in this case.

• (ii) For $A \times (B \cup \{\hat{s}\})$): (19) is also true for this case because

$$\hat{\gamma}(A \times (B \cup \{\hat{s}\})) = \hat{\gamma}(A \times B) + \hat{\gamma}(A \times \{\hat{s}\}) = \gamma(A \times B) + \int_A (1 - f_1)\mu(\mathrm{d}x).$$

• (iii) For $(A \cup \{\hat{s}\}) \times (B \cup \{\hat{s}\})$: (19) is true as well since

$$\begin{aligned}\hat{\gamma}((A \cup \{\hat{s}\}) \times (B \cup \{\hat{s}\})) &= \hat{\gamma}(A \times B) + \hat{\gamma}(A \times \{\hat{s}\}) + \hat{\gamma}(\{\hat{s}\} \times B) + \hat{\gamma}(\{\hat{s}\} \times \{\hat{s}\}) \\ &= \gamma(A \times B) + \int_A (1 - f_1)\mu(\mathrm{d}x) + \int_B (1 - f_2)\nu(\mathrm{d}x) + \gamma(\mathbb{G} \times \mathbb{G}).\end{aligned}$$

Now as (19) holds, we obviously have $\gamma(U \times \mathbb{G}) \leq \hat{\gamma}(U \times \mathbb{G}) \leq \hat{\gamma}(U \times \hat{\mathbb{G}}) = \hat{\mu}(U) = \mu(U)$ for any Borel set $U \subset \mathbb{G}$. Likewise, $\gamma(\mathbb{G} \times U) \leq \nu(U)$ for any Borel set $U \subset \mathbb{G}$. Therefore, $\gamma \in \Pi_{\leq}(\mu, \nu)$. ■

These observations in particular display the following connection between the EPT problem on a graph (4) and the corresponding standard complete OT problem (18).

Lemma A.8 (EPT on a graph versus its corresponding 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* (19) *gives a one-to-one correspondence between optimal solution γ for EPT problem* (4) *and optimal solution $\hat{\gamma}$ for standard complete OT problem* (18)*.*

Proof. We derive two parts as follow:

• (i) We show that $\mathrm{KT}(\hat{\mu}, \hat{\nu}) \leq \mathrm{ET}_{c,\lambda}(\mu, \nu)$:

For any $\gamma \in \Pi_{\leq}(\mu, \nu)$, let $\hat{\gamma}$ be given by (19). Then, $\hat{\gamma} \in \Gamma(\hat{\mu}, \hat{\nu})$ and

$$\begin{aligned}\mathrm{KT}(\hat{\mu}, \hat{\nu}) \leq \int_{\hat{\mathbb{G}} \times \hat{\mathbb{G}}} \hat{c}(x, y)\hat{\gamma}(\mathrm{d}x, \mathrm{d}y) &= b \int_{\mathbb{G} \times \mathbb{G}} [c(x, y) - \lambda]\gamma(\mathrm{d}x, \mathrm{d}y) \\ &\quad + \int_{\mathbb{G}} w_1[1 - f_1(x)]\mu(\mathrm{d}x) + \int_{\mathbb{G}} w_2[1 - f_2(x)]\nu(\mathrm{d}x).\end{aligned}$$

It follows that $\mathrm{KT}(\hat{\mu}, \hat{\nu}) \leq \mathrm{ET}_{c,\lambda}(\mu, \nu)$.

• (ii) We show that $\mathrm{KT}(\hat{\mu}, \hat{\nu}) \geq \mathrm{ET}_{c,\lambda}(\mu, \nu)$:

To see this, for any $\hat{\gamma} \in \Gamma(\hat{\mu}, \hat{\nu})$ we let γ be the restriction of $\hat{\gamma}$ to $\mathcal{T}$. Then by Lemma A.7, we have $\gamma \in \Pi_{\leq}(\mu, \nu)$ and (19) holds. Consequently,

$$\begin{aligned}\int_{\hat{\mathbb{G}} \times \hat{\mathbb{G}}} \hat{c}(x, y)\hat{\gamma}(\mathrm{d}x, \mathrm{d}y) &= b \int_{\mathbb{G} \times \mathbb{G}} [c(x, y) - \lambda]\gamma(\mathrm{d}x, \mathrm{d}y) \\ &\quad + \int_{\mathbb{G}} w_1[1 - f_1(x)]\mu(\mathrm{d}x) + \int_{\mathbb{G}} w_2[1 - f_2(x)]\nu(\mathrm{d}x) \\ &\geq \mathrm{ET}_{c,\lambda}(\mu, \nu).\end{aligned}$$

By taking the infimum over $\hat{\gamma}$, we infer that $\mathrm{KT}(\hat{\mu}, \hat{\nu}) \geq \mathrm{ET}_{c,\lambda}(\mu, \nu)$.

Thus, from the above two parts, we obtain

$$\mathrm{KT}(\hat{\mu}, \hat{\nu}) = \mathrm{ET}_{c,\lambda}(\mu, \nu).$$

The relation about the optimal solutions also follows from the above arguments. ■

Given the above two lemmas, we are ready to present the proof of Theorem 3.1.

Proof of Theorem 3.1 . From Lemma A.8 and the dual formulation for $\mathrm{KT}(\hat{\mu}, \hat{\nu})$ proved in (Caffarelli and McCann, 2010, Corollary 2.6), we have

$$\mathrm{ET}_{c,\lambda}(\mu, \nu) = \sup_{\substack{\hat{u} \in L^1(\hat{\mu}),\, \hat{v} \in L^1(\hat{\nu}) \\ \hat{u}(x) + \hat{v}(y) \leq \hat{c}(x,y)}} \int_{\hat{\mathbb{G}}} \hat{u}(x)\hat{\mu}(\mathrm{d}x) + \int_{\hat{\mathbb{G}}} \hat{v}(x)\hat{\nu}(\mathrm{d}x) =: I.$$

Therefore, it is enough to prove that $I = J$ where

$$J \triangleq \sup_{(u,v)\in\mathbb{K}} \Big[\int_{\mathbb{G}} u(x)\mu(\mathrm{d}x) + \int_{\mathbb{G}} v(x)\nu(\mathrm{d}x) \Big].$$

For (u, v) satisfying $u \le w_1$, $v \le w_2$ and $u(x) + v(y) \le b[c(x, y) - \lambda]$, we extend it to $\hat{\mathbb{G}}$ by taking $\hat{u}(\hat{s}) = 0$ and $\hat{v}(\hat{s}) = 0$. Then, it is clear that $\hat{u}(x) + \hat{v}(y) \le \hat{c}(x, y)$ for $x, y \in \hat{\mathbb{G}}$, and

$$I \ge \int_{\hat{\mathbb{G}}} \hat{u}(x)\hat{\mu}(\mathrm{d}x) + \int_{\hat{\mathbb{G}}} \hat{v}(x)\hat{\nu}(\mathrm{d}x) = \int_{\mathbb{G}} u(x)\mu(\mathrm{d}x) + \int_{\mathbb{G}} v(x)\nu(\mathrm{d}x).$$

It follows that $I \ge J$. In order to prove the converse, let $(\hat{u}, \hat{v})$ be a maximizer for I. Then, by considering $(\hat{u}-\hat{u}(\hat{s}), \hat{v}+\hat{u}(\hat{s}))$, we can assume that $\hat{u}(\hat{s}) = 0$. Also, if we let $v(y) := \inf_{x\in\hat{\mathbb{G}}}[\hat{c}(x, y) - \hat{u}(x)]$, then $(\hat{u}, v)$ is still in the admissible class for I and $\hat{v}(y) \le v(y)$. This implies that $(\hat{u}, v)$ is also a maximizer for I. For these reasons, we can assume w.l.g. that the maximizer $(\hat{u}, \hat{v})$ has the following additional properties: $\hat{u}(\hat{s}) = 0$ and

$$\hat{v}(y) = \inf_{x\in\hat{\mathbb{G}}} [\hat{c}(x, y) - \hat{u}(x)] \quad \forall y \in \hat{\mathbb{G}}.$$

In particular, $\hat{v}(\hat{s}) = \inf_{x\in\hat{\mathbb{G}}}[\hat{c}(x, \hat{s}) - \hat{u}(x)]$. For convenience, define $w_1(\hat{s}) = 0$ and consider the following two possibilities.

• (i) For $\inf_{x\in\hat{\mathbb{G}}}[w_1(x) - \hat{u}(x)] \ge 0$:

Since $\hat{c}(\hat{s}, \hat{s}) - \hat{u}(\hat{s}) = 0$ and $\inf_{x\in\mathbb{G}}[\hat{c}(x, \hat{s}) - \hat{u}(x)] = \inf_{x\in\mathbb{G}}[w_1(x) - \hat{u}(x)] \ge 0$, we have $\hat{v}(\hat{s}) = 0$.

Also, $\hat{v}(y) \le \hat{c}(\hat{s}, y) - \hat{u}(\hat{s}) \le w_2(y)$ for all $y \in \hat{\mathbb{G}}$. For each $y \in \mathbb{G}$, by using the facts $\hat{u} \le w_1$ and $\hat{c}(\hat{s}, y) - w_1(\hat{s}) = w_2(y) \ge 0$ we get

$$\hat{v}(y) \ge \inf_{x\in\hat{\mathbb{G}}} [\hat{c}(x, y) - w_1(x)] = \inf_{x\in\mathbb{G}} \{b[c(x, y) - \lambda] - w_1(x)\} = -b\lambda + \inf_{x\in\mathbb{G}} [b\, c(x, y) - w_1(x)].$$

Thus $(\hat{u}, \hat{v}) \in \mathbb{K}$ and

$$\begin{aligned} I = \int_{\hat{\mathbb{G}}} \hat{u}(x)\hat{\mu}(\mathrm{d}x) + \int_{\hat{\mathbb{G}}} \hat{v}(x)\hat{\nu}(\mathrm{d}x) &= \int_{\mathbb{G}} \hat{u}(x)\hat{\mu}(\mathrm{d}x) + \int_{\mathbb{G}} \hat{v}(x)\hat{\nu}(\mathrm{d}x) + \hat{v}(\hat{s})\mu(\mathbb{G}) \\ &= \int_{\mathbb{G}} \hat{u}(x)\mu(\mathrm{d}x) + \int_{\mathbb{G}} \hat{v}(x)\nu(\mathrm{d}x) \le J. \end{aligned}$$

• (ii) For $\inf_{x\in\hat{\mathbb{G}}}[w_1(x) - \hat{u}(x)] < 0$:

By arguing as in the above case (i), we have $\hat{v}(\hat{s}) = \inf_{x\in\mathbb{G}}[w_1(x) - \hat{u}(x)] < 0$ and

$$I = \int_{\mathbb{G}} \hat{v}(x)\nu(\mathrm{d}x) + \int_{\mathbb{G}} \hat{u}(x)\mu(\mathrm{d}x) + \mu(\mathbb{G}) \inf_{\mathbb{G}}[w_1 - \hat{u}]. \tag{20}$$

Let $\tilde{u}(x) := \min\{\hat{u}(x), w_1(x)\}$. Then, it is obvious that $\tilde{u}(x)+\hat{v}(y) \le \hat{c}(x, y)$ and $\tilde{u}(\hat{s}) = 0$. Since $\inf_{x\in\mathbb{G}}[w_1(x)-\hat{u}(x)] < 0$, there exists $x_0 \in \mathbb{G}$ such that $w_1(x_0) < \hat{u}(x_0)$. Thus, $\tilde{u}(x_0) = w_1(x_0)$ and hence $\inf_{\mathbb{G}}[w_1 - \tilde{u}] \le 0$. As $\tilde{u} \le w_1$, we infer further that $\inf_{\mathbb{G}}[w_1 - \tilde{u}] = 0$. We also have

$$\begin{aligned} &\int_{\mathbb{G}} \hat{u}(x)\mu(\mathrm{d}x) + \mu(\mathbb{G}) \inf_{\mathbb{G}}[w_1 - \hat{u}] \\ &= \int_{\mathbb{G}} \tilde{u}(x)\mu(\mathrm{d}x) + \int_{\mathbb{G}:\hat{u}>w_1} [\hat{u}(x) - w_1(x)]\mu(\mathrm{d}x) + \mu(\mathbb{G}) \inf_{\mathbb{G}}[w_1 - \hat{u}] \le \int_{\mathbb{G}} \tilde{u}(x)\mu(\mathrm{d}x). \end{aligned}$$

This together with (20) gives

$$I \le \int_{\mathbb{G}} \tilde{u}(x)\mu(\mathrm{d}x) + \int_{\mathbb{G}} \hat{v}(x)\nu(\mathrm{d}x).$$

Now let $\tilde{v}(y) = \inf_{x\in\hat{\mathbb{G}}}[\hat{c}(x, y) - \tilde{u}(x)]$ for $y \in \mathbb{G}$. Then, $\hat{v}(y) \le \tilde{v}(y) \le \hat{c}(\hat{s}, y) - \tilde{u}(\hat{s}) = w_2(y)$ for $y \in \mathbb{G}$. For each $y \in \mathbb{G}$, by using the facts $\tilde{u} \le w_1$ and $\hat{c}(\hat{s}, y) - w_1(\hat{s}) = w_2(y) \ge 0$ we also get

$$\tilde{v}(y) \ge \inf_{x\in\hat{\mathbb{G}}} [\hat{c}(x, y) - w_1(x)] = \inf_{x\in\mathbb{G}} \{b[c(x, y) - \lambda] - w_1(x)\} = -b\lambda + \inf_{x\in\mathbb{G}} [b\, c(x, y) - w_1(x)].$$

It follows that $(\tilde{u}, \tilde{v}) \in \mathbb{K}$ and

$$I \leq \int_{\mathbb{G}} \tilde{u}(x)\mu(\mathrm{d}x) + \int_{\mathbb{G}} \tilde{v}(x)\nu(\mathrm{d}x) \leq J.$$

Thus we conclude that $I = J$ and the theorem follows. ■

A.2.4 Proof of Corollary 3.2

Proof of Corollary 3.2. Notice that as w_i $(i = 1, 2)$ is b-Lipschitz w.r.t. $d_{\mathbb{G}}$, we have for every $x \in \mathbb{G}$ that

$$-w_i(x) \leq \inf_{y\in\mathbb{G}} \big[b\, d_{\mathbb{G}}(x, y) - w_i(y)\big]. \tag{21}$$

Let $\mathbb{K}$ be the set defined in the statement of Theorem 3.1. Then for each $(u, v) \in \mathbb{K}$, let

$$v^*(x) := \inf_{y\in\mathbb{G}} \big\{b[d_{\mathbb{G}}(x, y) - \lambda] - v(y)\big\} = -b\lambda + \inf_{y\in\mathbb{G}} \big[b\, d_{\mathbb{G}}(x, y) - v(y)\big] \geq u(x),$$
$$v^{**}(y) := \inf_{x\in\mathbb{G}} \big\{b[d_{\mathbb{G}}(x, y) - \lambda] - v^*(x)\big\} = -b\lambda + \inf_{x\in\mathbb{G}} \big[b\, d_{\mathbb{G}}(x, y) - v^*(x)\big] \geq v(y).$$

By using $-b\lambda + \inf_{x\in\mathbb{G}}[b\, d_{\mathbb{G}}(x, y) - w_1(x)] \leq v(y) \leq w_2(y)$ and (21), we obtain for every $x \in \mathbb{G}$ that

$$v^*(x) \leq -b\lambda - v(x) \leq -\inf_{y\in\mathbb{G}}[b\, d_{\mathbb{G}}(x, y) - w_1(y)] \leq w_1(x),$$
$$v^*(x) \geq -b\lambda + \inf_{y\in\mathbb{G}} \big[b\, d_{\mathbb{G}}(x, y) - w_2(y)\big] \geq -b\lambda - w_2(x).$$

We also have v^* is b-Lipschitz, i.e., $|v^*(x_1) - v^*(x_2)| \leq b\, d_{\mathbb{G}}(x_1, x_2)$. Indeed, let $x_1, x_2 \in \mathbb{G}$. Then for any $\epsilon > 0$, there exists $y_1 \in \mathbb{G}$ such that

$$b\, d_{\mathbb{G}}(x_1, y_1) - v(y_1) < v^*(x_1) + b\lambda + \epsilon.$$

It follows that

$$v^*(x_2) - v^*(x_1) \leq b\, d_{\mathbb{G}}(x_2, y_1) - v(y_1) + \epsilon - [b\, d_{\mathbb{G}}(x_1, y_1) - v(y_1)] \leq b\, d_{\mathbb{G}}(x_1, x_2) + \epsilon.$$

Since this holds for every $\epsilon > 0$, we get

$$v^*(x_2) - v^*(x_1) \leq b\, d_{\mathbb{G}}(x_1, x_2).$$

By interchanging the role of x_1 and x_2, we also obtain $v^*(x_1) - v^*(x_2) \leq b\, d_{\mathbb{G}}(x_1, x_2)$. Thus,

$$|v^*(x_1) - v^*(x_2)| \leq b\, d_{\mathbb{G}}(x_1, x_2).$$

Hence, we have shown that $v^* \in \mathbb{U}^*$ with

$$\mathbb{U}^* := \Big\{f \in C(\mathbb{G}) :\, -b\lambda - w_2 \leq f \leq w_1,\, |f(x) - f(y)| \leq b\, d_{\mathbb{G}}(x, y)\Big\}.$$

We next claim $v^{**} = -b\lambda - v^*$. For this, it is clear from the definition that $v^{**}(y) \leq -b\lambda - v^*(y)$. On the other hand, from the Lipschitz property of v^* we obtain

$$-v^*(y) \leq b\, d_{\mathbb{G}}(x, y) - v^*(x) \quad \forall x \in \mathbb{G},$$

which gives $-b\lambda - v^*(y) \leq v^{**}(y)$. Thus, we conclude that $v^{**} = -b\lambda - v^*$ as claimed.

From these, we obtain that

$$\begin{aligned}\int_{\mathbb{G}} u(x)\mu(\mathrm{d}x) + \int_{\mathbb{G}} v(x)\nu(\mathrm{d}x) &\leq \int_{\mathbb{G}} v^*(x)\mu(\mathrm{d}x) + \int_{\mathbb{G}} v^{**}(x)\nu(\mathrm{d}x)\\ &= \int_{\mathbb{G}} v^*(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} v^*(x)\nu(\mathrm{d}x) - b\lambda\nu(\mathbb{G})\\ &\leq -b\lambda\nu(\mathbb{G}) + \sup\left\{\int_{\mathbb{G}} f(\mu - \nu) :\, f \in \mathbb{U}^*\right\}.\end{aligned}$$

This together with Theorem 3.1 in the main text implies that

$$\mathrm{ET}_\lambda(\mu,\nu) \leq -b\lambda\nu(\mathbb{G}) + \sup\left\{\int_{\mathbb{G}} f(\mu-\nu) :\ f \in \mathbb{U}^*\right\}.$$

To prove the converse, let $f \in \mathbb{U}^*$. Define $u := f$ and $v := -b\lambda - f$. Then, we have

$$u(x) \leq w_1(x),$$

$$v(x) \leq -b\lambda - [-b\lambda - w_2(x)] = w_2(x),$$

and

$$v(x) \geq -b\lambda - w_1(x) \geq -b\lambda + \inf_{y\in\mathbb{G}}[b\, d_{\mathbb{G}}(x,y) - w_1(y)].$$

Also, the Lipschitz property of f gives

$$u(x) + v(y) = -b\lambda + f(x) - f(y) \leq b[d_{\mathbb{G}}(x,y) - \lambda] \quad \forall x, y \in \mathbb{G}.$$

Thus $(u, v) \in \mathbb{K}$, and hence we obtain from Theorem 3.1 in the main text that

$$-b\lambda\nu(\mathbb{G}) + \int_{\mathbb{G}} f(\mu-\nu) = \int_{\mathbb{G}} u(x)\mu(\mathrm{d}x) + \int_{\mathbb{G}} v(x)\nu(\mathrm{d}x) \leq \mathrm{ET}_\lambda(\mu,\nu).$$

As this holds for every $f \in \mathbb{U}^*$, we get

$$-b\lambda\nu(\mathbb{G}) + \sup\left\{\int_{\mathbb{G}} f(\mu-\nu) :\ f \in \mathbb{U}^*\right\} \leq \mathrm{ET}_\lambda(\mu,\nu).$$

Thus, we have shown that

$$\mathrm{ET}_\lambda(\mu,\nu) = -b\lambda\nu(\mathbb{G}) + \sup\left\{\int_{\mathbb{G}} f(\mu-\nu) :\ f \in \mathbb{U}^*\right\}. \tag{22}$$

Now consider $f = \tilde{f} - \frac{b\lambda}{2}$. Then, $f \in \mathbb{U}^*$ if and only if $\tilde{f} \in \mathbb{U}$. Moreover,

$$\int_{\mathbb{G}} f(\mu-\nu) = -\frac{b\lambda}{2}\big[\mu(\mathbb{G}) - \nu(\mathbb{G})\big] + \int_{\mathbb{G}} \tilde{f}(\mu-\nu).$$

Therefore, the conclusion of the corollary follows from (22). ■

A.2.5 Proof of Lemma 4.4

Proof of Lemma 4.4. By using part i) of Lemma A.2, we see that

$$\mathbb{U}_0 \subset \left\{f \in W^{1,\infty}(\mathbb{G},\omega^*) :\ -w_2(z_0) - \frac{b\lambda}{2} \leq f(z_0) \leq w_1(z_0) + \frac{b\lambda}{2},\ \|f'\|_{L^\infty(\mathbb{G},\omega^*)} \leq b\right\} = \mathbb{U}_\infty^0. \tag{23}$$

As a consequence, we obtain

$$\mathrm{US}_1^0(\mu,\nu) = \sup\Big[\int_{\mathbb{G}} f(\mu-\nu) :\ f \in \mathbb{U}_\infty^0\Big] \geq \sup\Big[\int_{\mathbb{G}} f(\mu-\nu) :\ f \in \mathbb{U}_0\Big].$$

Thus the first statement of the lemma is proved. Now if $\mathbb{G}$ is a tree. Then Lemma A.2 implies that the inclusion in (23) is actually the equality. That is, $\mathbb{U}_0 = \mathbb{U}_\infty^0$. Therefore, we get the desired identity

$$\mathrm{US}_1^0(\mu,\nu) = \sup\Big[\int_{\mathbb{G}} f(\mu-\nu) :\ f \in \mathbb{U}_0\Big].$$

■

A.2.6 Proof of Proposition 4.5

Proof of Proposition 4.5. It follows from Definition 4.3 and the representation (7) for f that

$$\mathrm{US}_p^\alpha(\mu,\nu) = \sup\Big\{s[\mu(\mathbb{G})-\nu(\mathbb{G})]:\ s\in\big[-\frac{b\lambda}{2}-w_2(z_0)+\alpha, w_1(z_0)+\frac{b\lambda}{2}-\alpha\big]\Big\} + \sup\left\{\int_{\mathbb{G}}\Big[\int_{[z_0,x]}h(y)\omega(\mathrm{d}y)\Big](\mu-\nu)(dx):\ \|h\|_{L^{p'}(\mathbb{G},\omega)}\le b\right\}.$$

The first supremum equals to $[w_1(z_0)+\frac{b\lambda}{2}-\alpha][\mu(\mathbb{G})-\nu(\mathbb{G})]$ if $\mu(\mathbb{G})\ge\nu(\mathbb{G})$ and equals to $-[w_2(z_0)+\frac{b\lambda}{2}-\alpha][\mu(\mathbb{G})-\nu(\mathbb{G})]$ if $\mu(\mathbb{G})<\nu(\mathbb{G})$.

On the other hand, by the same arguments as in the proof of (Le et al., 2022, Proposition 3.5) we see that the second supremum equals to $b\left(\int_{\mathbb{G}}|\mu(\Lambda(x))-\nu(\Lambda(x))|^p\,\omega(\mathrm{d}x)\right)^{\frac{1}{p}}$. Putting them together, we obtain the desired formula for $\mathrm{US}_p^\alpha(\mu,\nu)$. ■

A.2.7 Proof of Corollary 4.6

Proof of Corollary 4.6. We first recall that $\langle u,v\rangle$ denotes the line segment in $\mathbb{R}^n$ connecting two points u,v, while (u,v) means the same line segment but without its two end-points. Then as $\omega(\{x\})=0$ for every $x\in\mathbb{G}$, we have

$$\int_{\mathbb{G}}|\mu(\Lambda(x))-\nu(\Lambda(x))|^p\,\omega(\mathrm{d}x) = \sum_{e=\langle u,v\rangle\in E}\int_{(u,v)}|\mu(\Lambda(x))-\nu(\Lambda(x))|^p\,\omega(\mathrm{d}x).$$

Since μ and ν are supported on nodes, we can rewrite the above identity as

$$\int_{\mathbb{G}}|\mu(\Lambda(x))-\nu(\Lambda(x))|^p\,\omega(\mathrm{d}x) = \sum_{e=\langle u,v\rangle\in E}\int_{(u,v)}|\mu(\Lambda(x)\setminus(u,v))-\nu(\Lambda(x)\setminus(u,v))|^p\,\omega(\mathrm{d}x).$$

For $e=\langle u,v\rangle$ and $x\in(u,v)$, we observe that $y\in\mathbb{G}\setminus(u,v)$ belongs to $\Lambda(x)$ if and only if $y\in\gamma_e$. It follows that $\Lambda(x)\setminus(u,v)=\gamma_e$, and thus we deduce from the above identity that

$$\begin{aligned}\int_{\mathbb{G}}|\mu(\Lambda(x))-\nu(\Lambda(x))|^p\,\omega(\mathrm{d}x) &= \sum_{e=\langle u,v\rangle\in E}\int_{(u,v)}|\mu(\gamma_e)-\nu(\gamma_e)|^p\,\omega(\mathrm{d}x)\\ &= \sum_{e\in E}\big|\mu(\gamma_e)-\nu(\gamma_e)\big|^p\omega(e).\end{aligned}$$

This together with Proposition 4.5 yields the postulated result. ■

A.2.8 Proof of Proposition 5.1

We begin with the following auxiliary result.

Lemma A.9. *Let $\mu,\nu\in\mathcal{M}(\mathbb{G})$. Then, $\mu=\nu$ if and only if $\mu(\Lambda(x))=\nu(\Lambda(x))$ for every x in $\mathbb{G}$.*

Proof. It is obvious that $\mu=\nu$ implies that $\mu(\Lambda(x))=\nu(\Lambda(x))$ for every x in $\mathbb{G}$. Now assume that $\mu(\Lambda(x))=\nu(\Lambda(x))$ for every x in $\mathbb{G}$. We first claim that $\mu(\{a\})=\nu(\{a\})$ for any $a\in\mathbb{G}$. Let $a\in\mathbb{G}$ be arbitray. Then there are two possibility for a: either a is a node or a is an interior point of an edge. We consider these two cases saperately.

• (i) a is an interior point of an edge $e\in E$ (i.e. a is not a node):

Let $\{a_n\}_{n=1}^\infty$ be a sequence of distinct points on the same edge e as a such that $d_{\mathbb{G}}(a_n,z_0)>d_{\mathbb{G}}(a,z_0)$ for every $n\ge1$ and $a_n\to a$ as $n\to\infty$. It follows that $\Lambda(a_n)\subset\Lambda(a)$ and $\Lambda(a)\setminus\Lambda(a_n)\downarrow\{a\}$ as $n\to\infty$. As a consequence, we have

$$\mu(\{a\}) = \lim_{n\to\infty}\mu(\Lambda(a)\setminus\Lambda(a_n)) = \lim_{n\to\infty}\big[\mu(\Lambda(a))-\mu(\Lambda(a_n))\big].$$

But as $\mu(\Lambda(x))=\nu(\Lambda(x))$ for every x in $\mathbb{G}$, we thus obtain

$$\mu(\{a\}) = \lim_{n\to\infty}\big[\nu(\Lambda(a))-\nu(\Lambda(a_n))\big] = \lim_{n\to\infty}\nu(\Lambda(a)\setminus\Lambda(a_n)) = \nu(\{a\})$$

as claimed.

• (ii) a is a node:

We can assume that a is a common node for edges $e_1, ..., e_k$. Then for each $i \in \{1, ..., k\}$, let $\{a_n^i\}_{n=1}^{\infty}$ be a sequence of distinct points on edge e_i such that $d_{\mathbb{G}}(a_n^i, z_0) > d_{\mathbb{G}}(a, z_0)$ for every $n \geq 1$ and $a_n^i \to a$ as $n \to \infty$. These choices yield $\Lambda(a_n^i) \subset \Lambda(a)$ and $\Lambda(a) \setminus \cup_{i=1}^k \Lambda(a_n^i) \downarrow \{a\}$ as $n \to \infty$. Using this and the assumption $\mu(\Lambda(x)) = \nu(\Lambda(x))$ for every x in $\mathbb{G}$, we obtain

$$\mu(\{a\}) = \lim_{n\to\infty} \big[\mu(\Lambda(a)) - \sum_{i=1}^{k} \mu(\Lambda(a_n^i))\big] = \lim_{n\to\infty} \big[\nu(\Lambda(a)) - \sum_{i=1}^{k} \nu(\Lambda(a_n^i))\big] = \nu(\{a\}).$$

Thus, we have proved the claim that $\mu(\{a\}) = \nu(\{a\})$ for every $a \in \mathbb{G}$.

On the other hand, for any points x, y belonging to the same edge

$$\mu(\langle x, y)) = \mu(\Lambda(x)) - \mu(\Lambda(y)) = \nu(\Lambda(x)) - \nu(\Lambda(y)) = \nu(\langle x, y)),$$

where $\langle x, y)$ denotes the line segment in $\mathbb{R}^n$ connecting two points x, y but without its right end-point x (while $\langle x, y\rangle$ include both end-points).

Thus, by combining them, we infer further that $\mu(\langle x, y\rangle) = \nu(\langle x, y\rangle)$ for any $x, y \in e$ and for any edge $e \in E$. It follows that $\mu = \nu$, and the proof is complete. ■

Proof of Proposition 5.1. We note first that the quantity US_p^α depends only on the values of the weights at the root z_0 of the graph. This comes from the fact that only $w_1(z_0)$ and $w_2(z_0)$ are used in the definition of $\mathbb{U}_{p'}^\alpha$.

i) This follows immediately from Proposition 4.5 in the main text.

ii) It follows from Definition 4.3 that $\mathrm{US}_p^\alpha(\mu, \mu) = 0$ and US_p^α satisfies the triangle inequality. As the constant function $f = 0$ belongs to the constraint set $\mathbb{U}_{p'}^\alpha$, we also have $\mathrm{US}_p^\alpha(\mu, \nu) \geq 0$. Next, assume that $\mathrm{US}_p^\alpha(\mu, \nu) = 0$. Then by Proposition 4.5 in the main text, we get

$$b\left(\int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p \, \omega(\mathrm{d}x)\right)^{\frac{1}{p}} + \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| = 0.$$

As $\Theta > 0$ by our assumption of α, we must have

$$\mu(\mathbb{G}) = \nu(\mathbb{G}) \quad \text{and} \quad \int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p \, \omega(\mathrm{d}x) = 0.$$

Therefore, $\mu(\Lambda(x)) = \nu(\Lambda(x))$ for every $x \in \mathbb{G}$. By using Lemma A.9, we then conclude that $\mu = \nu$.

iii) Due to the assumption $w_1(z_0) = w_2(z_0)$ we have $f \in \mathbb{U}_{p'}^\alpha$ if and only if $-f \in \mathbb{U}_{p'}^\alpha$. Hence we obtain from Definition 4.3 that $\mathrm{US}_p^\alpha(\mu, \nu) = \mathrm{US}_p^\alpha(\nu, \mu)$. This together with ii) implies that $(\mathcal{M}(\mathbb{G}), \mathrm{US}_p^\alpha)$ is a metric space. Its completeness follows from (Piccoli and Rossi, 2014, Proposition 4). As a complete metric space, it is well known that $(\mathcal{M}(\mathbb{G}), \mathrm{US}_p^\alpha)$ is a geodesic space if and only if for every $\mu, \nu \in \mathcal{M}(\mathbb{G})$ there exists $\sigma \in \mathcal{M}(\mathbb{G})$ such that

$$\mathrm{US}_p^\alpha(\mu, \sigma) = \mathrm{US}_p^\alpha(\nu, \sigma) = \frac{1}{2}\mathrm{US}_p^\alpha(\mu, \nu).$$

To verify the latter, take $\sigma := \frac{\mu+\nu}{2}$. Then using Definition 4.3 in the main text, we obtain

$$\mathrm{US}_p^\alpha(\mu, \sigma) = \frac{1}{2} \sup_{f \in \mathbb{U}_{p'}^\alpha} \int_{\mathbb{G}} f(\mu - \nu) = \frac{1}{2}\mathrm{US}_p^\alpha(\mu, \nu)$$

and

$$\mathrm{US}_p^\alpha(\nu, \sigma) = \frac{1}{2} \sup_{f \in \mathbb{U}_{p'}^\alpha} \int_{\mathbb{G}} f(\nu - \mu) = \frac{1}{2}\mathrm{US}_p^\alpha(\nu, \mu) = \frac{1}{2}\mathrm{US}_p^\alpha(\mu, \nu).$$

■

A.2.9 Proof of Proposition 5.3

Proof of Proposition 5.3. i) From its definition, we have $\mathbb{U}_\infty^\alpha = \mathbb{L}_\alpha$ with $\mathbb{L}_\alpha$ being the set defined in (Le and Nguyen, 2021, Section 3.2). As a consequence, we obtain $\mathrm{US}_1^\alpha(\mu,\nu) = d_\alpha(\mu,\nu)$. On the other hand, Proposition A.4 yields for any $1 \le p \le \infty$ that

$$\mathrm{US}_1^\alpha(\mu,\nu) - \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| \le \omega^*(\mathbb{G})^{\frac{1}{p'}} \Big[\mathrm{US}_p^\alpha(\mu,\nu) - \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})|\Big].$$

Therefore, we conclude that

$$\omega^*(\mathbb{G})^{-\frac{1}{p'}} \Big[d_\alpha(\mu,\nu) - \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})|\Big] \le \mathrm{US}_p^\alpha(\mu,\nu) - \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})|.$$

By moving and combining terms we arrive at

$$\mathrm{US}_p^\alpha(\mu,\nu) \ge \omega^*(\mathbb{G})^{-\frac{1}{p'}} d_\alpha(\mu,\nu) + \Theta[1 - \omega^*(\mathbb{G})^{-\frac{1}{p'}}]|\mu(\mathbb{G}) - \nu(\mathbb{G})|.$$

ii) Let $\bar{m} \triangleq \mu(\mathbb{G}) = \nu(\mathbb{G})$. From the definition of the p-Wasserstein distance, we have

$$\begin{aligned}
\mathcal{W}_p(\mu,\nu)^p &= \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{G}\times\mathbb{G}} d_{\mathbb{G}}(x,y)^p \gamma(\mathrm{d}x,\mathrm{d}y) \\
&\le \Big[\sup_{x,y\in\mathbb{G}} d_{\mathbb{G}}(x,y)\Big]^{p-1} \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{G}\times\mathbb{G}} d_{\mathbb{G}}(x,y)\gamma(\mathrm{d}x,\mathrm{d}y) \\
&= \Big[\sup_{x,y\in\mathbb{G}} d_{\mathbb{G}}(x,y)\Big]^{p-1} \mathcal{W}_1(\mu,\nu),
\end{aligned}$$

where

$$\Pi(\mu,\nu) \triangleq \Big\{\gamma \in \mathcal{M}(\mathbb{G}\times\mathbb{G}) : \gamma(\mathbb{G}\times\mathbb{G}) = \bar{m},\, \gamma_1 = \mu,\, \gamma_2 = \nu\Big\}.$$

Therefore, the first statement will follow if we can show that

$$\mathrm{US}_p^\alpha(\mu,\nu) \ge b\, \mathcal{W}_1(\mu,\nu). \tag{24}$$

Since $\mu(\mathbb{G}) = \nu(\mathbb{G})$, we have from Lemma A.6 that

$$\mathrm{US}_p^\alpha(\mu,\nu) = \sup\Big\{\int_{\mathbb{G}} f(\mu-\nu) :\, f \in W^{1,p'}(\mathbb{G},\omega),\, \|f'\|_{L^{p'}(\mathbb{G},\omega)} \le b\Big\}.$$

Hence by taking $g \triangleq f/b$, we can rewrite this identity as

$$\begin{aligned}
\mathrm{US}_p^\alpha(\mu,\nu) &= b \sup\Big\{\int_{\mathbb{G}} g(\mu-\nu) :\, g \in W^{1,p'}(\mathbb{G},\omega),\, \|g'\|_{L^{p'}(\mathbb{G},\omega)} \le 1\Big\} \\
&= b\, \mathcal{S}_p(\mu,\nu),
\end{aligned}$$

where $\mathcal{S}_p$ is the balanced Sobolev transport distance defined in (Le et al., 2022, Definition 3.2). On the other hand, we have $\mathcal{S}_p(\mu,\nu) \ge \omega^*(\mathbb{G})^{-\frac{1}{p'}} \mathcal{W}_1(\mu,\nu)$ by (Le et al., 2022, Lemma 4.3). Therefore, we obtain (24) as desired.

Alternatively, we can derive (24) as follows. By using $\mathbb{U}_\infty^\alpha = \mathbb{L}_\alpha$ as in the proof of part i) and the observation about the translation invariant in the proof of Lemma A.6, we see that

$$\begin{aligned}
d_\alpha(\mu,\nu) &= \sup\Big\{\int_{\mathbb{G}} f(\mu-\nu) :\, f \in \mathbb{U}_\infty^\alpha\Big\} \\
&= \sup\Big\{\int_{\mathbb{G}} f(\mu-\nu) :\, f \in W^{1,\infty}(\mathbb{G},\omega^*),\, \|f'\|_{L^\infty(\mathbb{G},\omega^*)} \le b\Big\}.
\end{aligned}$$

Then due to Lemma A.2, we can further rewrite as

$$\begin{aligned}
d_\alpha(\mu,\nu) &= \sup\Big\{\int_{\mathbb{G}} f(\mu-\nu) :\, f \in C(\mathbb{G}),\, |f(x) - f(y)| \le b\, d_{\mathbb{G}}(x,y)\Big\} \\
&= b \sup\Big\{\int_{\mathbb{G}} g(\mu-\nu) :\, g \in C(\mathbb{G}),\, |g(x) - g(y)| \le 1\, d_{\mathbb{G}}(x,y)\Big\} \\
&= b\, \mathcal{W}_1(\mu,\nu).
\end{aligned}$$

On the other hand, part i) above gives

$$\mathrm{US}_p^\alpha(\mu,\nu) \geq \omega^*(\mathbb{G})^{-\frac{1}{p'}} d_\alpha(\mu,\nu).$$

Therefore, we obtain

$$\mathrm{US}_p^\alpha(\mu,\nu) \geq b\,\omega^*(\mathbb{G})^{-\frac{1}{p'}} \mathcal{W}_1(\mu,\nu),$$

for every $1 \leq p \leq \infty$.

For $p = 1$, the equality happens since $p' = \infty$ and

$$\mathrm{US}_1^\alpha(\mu,\nu) = \sup\Big\{ \int_{\mathbb{G}} f(\mu-\nu) :\ f \in \mathbb{U}_\infty^\alpha \Big\} = b\,\mathcal{W}_1(\mu,\nu).$$

Thus, the second statement follows.

■

A.2.10 Proof of Proposition 5.4

Proof of Proposition 5.4. We first prove that ℓ_p distance is negative definite for $1 \leq p \leq 2$, where

$$\ell_p(x,z) \triangleq \left(\sum_{i=1}^m \left| x_{(i)} - z_{(i)} \right|^p \right)^{1/p} \quad \text{for} \quad x, z \in \mathbb{R}^m.$$

It is easy to see that the function $(u,v) \mapsto (u-v)^2$ is negative definite for $u, v \in \mathbb{R}$. Using this and by applying (Berg et al., 1984, Corollary 2.10), the function $(u,v) \mapsto (u-v)^p$ is negative definite for $1 \leq p \leq 2$.

Therefore, for $1 \leq p \leq 2$, the function ℓ_p^p is negative definite since it is a sum of negative definite functions. Using this and by applying (Berg et al., 1984, Corollary 2.10), we have ℓ_p is negative definite for $1 \leq p \leq 2$.

We are now ready to prove the Proposition 5.4. From Proposition 4.5, we have

$$\mathrm{US}_p^\alpha(\mu,\nu) = b\Big(\sum_{e\in E} w_e\, |\mu(\gamma_e) - \nu(\gamma_e)|^p \Big)^{\frac{1}{p}} + \Theta |\mu(\mathbb{G}) - \nu(\mathbb{G})|.$$

Let $m = |E|$. Then, $\mu \mapsto \left\{ w_e^{\frac{1}{p}} \mu(\gamma_e) \right\}_{e\in E}$ can be regarded as a feature map for measure μ onto $\mathbb{R}_+^m$. Therefore, the first term of US_p^α is equivalent to b times the ℓ_p distance between two feature maps of measures μ, ν on $\mathbb{R}_+^m$ respectively. Recall that $b \geq 0$. Thus, the first term of US_p^α is negative definite for $1 \leq p \leq 2$.

Additionally, the second term of US_p^α is Θ times the ℓ_1 distance between $\mu(\mathbb{G})$ and $\nu(\mathbb{G})$. Since $w_1(z_0) = w_2(z_0)$ and $\alpha \leq \frac{b\lambda}{2} + w_1(z_0)$, we also have from (9) that $\Theta = w_1(z_0) + \frac{b\lambda}{2} - \alpha \geq 0$. Therefore, the second term of US_p^α is also negative definite.

Hence, US_p^α is negative definite for any $1 \leq p \leq 2$. ■

B FURTHER RESULTS AND DISCUSSIONS

B.1 Brief Reviews

We give brief reviews for some definitions used in our work.

B.1.1 Length Measure on Graphs

We recall the definition and properties in (Le et al., 2022, §4.1) about the length measure on graphs.

Definition B.1 (Length measure)**.** *Let ω^* be the unique Borel measure on $\mathbb{G}$ such that the restriction of ω^* on any edge is the length measure of that edge. That is, ω^* satisfies:*

i) For any edge e connecting two nodes u and v, we have $\omega^*(\langle x, y\rangle) = (t-s)w_e$ whenever $x = (1-s)u + sv$ and $y = (1-t)u + tv$ for $s, t \in [0, 1)$ with $s \leq t$. Here, $\langle x, y\rangle$ is the line segment in e connecting x and y.

ii) For any Borel set $F \subset \mathbb{G}$, we have

$$\omega^*(F) = \sum_{e \in E} \omega^*(F \cap e).$$

The next lemma asserts that ω^* is closely connected to the graph metric $d_{\mathbb{G}}$, and thus justifies the terminology of a length measure.

Lemma B.2 (ω^* is the length measure on graph)**.** *Suppose that $\mathbb{G}$ has no short cuts, namely, any edge e is a shortest path connecting its two end-points. Then, ω^* is a length measure in the sense that*

$$\omega^*([x, y]) = d_{\mathbb{G}}(x, y)$$

for any shortest path $[x, y]$ connecting x and y. In particular, ω^* has no atom in the sense that $\omega^*(\{x\}) = 0$ for every x in $\mathbb{G}$.

B.1.2 Wasserstein distances

We recall here the definition of the p-Wasserstein distances with graph metric ground cost on $\mathbb{G}$.

Definition B.3. *Let $1 \leq p < \infty$. Suppose that μ and ν are two nonnegative Borel measures on $\mathbb{G}$ satisfying $\mu(\mathbb{G}) = \nu(\mathbb{G})$. Then the p-Wasserstein distance between μ and ν is defined by*

$$\mathcal{W}_p(\mu, \nu)^p = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{G}\times\mathbb{G}} d_{\mathbb{G}}(x, y)^p \gamma(\mathrm{d}x, \mathrm{d}y),$$

where

$$\Pi(\mu, \nu) \triangleq \left\{\gamma \in \mathcal{M}(\mathbb{G} \times \mathbb{G}) : \gamma(\mathbb{G} \times \mathbb{G}) = \bar{m},\, \gamma_1 = \mu,\, \gamma_2 = \nu\right\}$$

with $\bar{m} \triangleq \mu(\mathbb{G}) = \nu(\mathbb{G})$.

B.1.3 Kernels

We review some important definitions and theorems/corollaries about kernels that are used in our work.

- **Positive Definite Kernels (Berg et al., 1984, pp. 66–67).** A kernel function $k : \Omega \times \Omega \to \mathbb{R}$ is called positive definite if for every positive integer m and every points $x_1, x_2, ..., x_m \in \Omega$, we have

$$\sum_{i,j=1}^{m} c_i c_j k(x_i, x_j) \geq 0 \qquad \text{for every} \quad c_1, ..., c_m \in \mathbb{R}.$$

- **Negative Definite Kernels (Berg et al., 1984, pp. 66–67).** A kernel function $k : \Omega \times \Omega \to \mathbb{R}$ is called negative definite if for every integer $m \geq 2$ and every points $x_1, x_2, ..., x_m \in \Omega$, we have

$$\sum_{i,j=1}^{m} c_i c_j k(x_i, x_j) \leq 0, \qquad \text{for every} \quad c_1, ..., c_m \in \mathbb{R} \;\text{ s.t. }\; \sum_{i=1}^{m} c_i = 0.$$

- **Theorem 3.2.2 in (Berg et al., 1984, pp. 74).** Let κ be a *negative definite* kernel. Then for every $t > 0$, the kernel

$$k_t(x, z) \triangleq \exp\left(-t\kappa(x, z)\right)$$

  is positive definite.

- **Definition 2.6 in (Berg et al., 1984, pp. 76).** A positive definite kernel κ is called *infinitely divisible* if for each $n \in \mathbb{N}^*$, there exists a positive definite kernel κ_n such that

$$\kappa = (\kappa_n)^n.$$

- **Corollary 2.10 in (Berg et al., 1984, pp. 78).** Let κ be a *negative definite* kernel. Then for $0 < t < 1$, the kernel

$$k_t(x, z) \triangleq [\kappa(x, z)]^t$$

  is negative definite.

B.2 Further Discussions

In this subsection, we discuss some extension for our work and describe more details for some parts in the main manuscript.

Path length for points in $\mathbb{G}$. We can canonically measure a path length connecting any two points $x, y \in \mathbb{G}$ where x, y are not necessary to be nodes in V. Indeed, for two points $x, y \in \mathbb{R}^n$ belonging to the same edge $e = \langle u, v\rangle$ which connects two nodes u and v in V, then we have

$$x = (1-s)u + sv,$$
$$y = (1-t)u + tv,$$

for some numbers $t, s \in [0, 1]$. Therefore, the length of the path connecting x and y along the edge e (i.e., the line segment $\langle x, y\rangle$) is defined by $|t - s|w_e$. Hence, the length for an arbitrary path in $\mathbb{G}$ can be similarly defined by breaking down into pieces over edges and summing over their corresponding lengths (Le et al., 2022).

Lipschitz nonnegative weight function on graph $\mathbb{G}$. An example of b-Lipschitz nonegative weight function on $\mathbb{G}$ is

$$w(x) = a_1 d_{\mathbb{G}}(z_0, x) + a_0,$$

for some constants $a_1 \in [0, b]$ and $a_0 \in [0, \infty)$.

Extension to measures supported on $\mathbb{G}$. The closed-form formula for US_p^α in (10) can be extended for measures with finite supports on $\mathbb{G}$ (i.e., measures which may have supports on edges) by using the same strategy to measure a path length connecting z_0 and y for any $z_0, y \in \mathbb{G}$ (see §2). More precisely, we break down edges containing supports into pieces and sum over their corresponding values instead of the sum over edges for US_p^α in (10).

About the assumption of uniqueness property of the shortest paths on $\mathbb{G}$. As discussed in the supplementary of (Le et al., 2022), since $w_e \in \mathbb{R}$ for any edge $e \in E$ of graph $\mathbb{G}$., it is almost surely that every node in the graph can be regarded as unique-path root node (with a high probability, lengths of paths connecting any two nodes in graph $\mathbb{G}$ are different). Additionally, for some special graph, e.g., a grid of nodes, there is no unique-path root node for such graph. However, by perturbing each node of such graph (or lengths of edges in $\mathbb{G}$ in case $\mathbb{G}$ is a non-physical graph, i.e., w_e) with a small deviation ε, we can obtain a graph satisfying the unique-path root node assumption.

About the unbalanced Sobolev transport. Similar to the work (Le et al., 2022), we assume that we know the graph metric space (i.e., the graph structure) where supports of measures are belongs to. Giving such graph, we define the unbalanced Sobolev transport for measures which may have different total mass and are supported on that graph metric space. We leave a question to learn an optimal graph metric structure from data (i.e., supports of measures) for unbalanced Sobolev transport for future work.

About graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ (Le et al., 2022). First, we use a clustering method, e.g., the farthest-point clustering, to partition supports of measures into at most M clusters.[7] Then, let V denote the set of centroids of these clusters. For edges, in graph $\mathbb{G}_{\text{Log}}$, we randomly choose $M\log(M)$ edges; and $M^{3/2}$ edges for graph $\mathbb{G}_{\text{Sqrt}}$, we also denote the set of those sampled edges as $\tilde{E}$.

For each edge e, its corresponding weight w_e is computed by the Euclidean distance between the two corresponding nodes of e. Let n_c be the number of connected components in the graph $\tilde{\mathbb{G}}(V, \tilde{E})$, we then randomly add $(n_c - 1)$ more edges between these n_c connected components to construct a connected graph $\mathbb{G}$ from $\tilde{\mathbb{G}}$.Let E_c be the set of these $(n_c - 1)$ added edges and denote set $E = \tilde{E} \cup E_c$, then $\mathbb{G}(V, E)$ is the considered graph.

Datasets and Computational Devices. For document dataset (i.e., `TWITTER, RECIPE, CLASSIC, AMAZON`), orbit dataset (`Orbit`) and a 10-class subset of `MPEG7` dataset, one can contact the authors of (Le et al., 2022) to access to these datasets. For computational devices, we run all of our experiments on commodity hardware.

B.3 Further Empirical Results

In this subsection, we provide further empirical results for our work.

[7] M is the input number of clusters for the clustering method. Therefore, the result has at most M clusters depending on input data.

B.3.1 Extended Empirical Results for the Main Text

Similar to Figure 3 in the main text for TDA, we illustrate the effect of the number of slices for document classification with graph $\mathbb{G}_{\text{Sqrt}}$ in Figure 4.

We also consider a graph $\mathbb{G}$ with a different setting:$\mathbb{G}_{\text{Log}}$. Recall that for Figure 1, Figure 2, Figure 3 in the main text and Figure 4, results are for graph $\mathbb{G}_{\text{Sqrt}}$ where $M = 10^4$ for document datasets, $M = 10^3$ for `MPEG7` dataset and $M = 10^2$ for `Orbit` dataset.[8] We illustrate corresponding results for graph $\mathbb{G}_{\text{Log}}$ in Figure 5, Figure 6, Figure 7, and Figure 8 respectively.

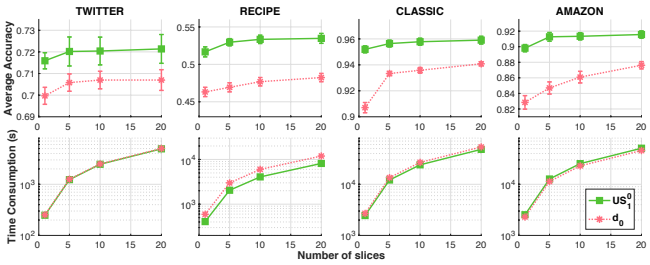

Figure 4: SVM results and time consumption for kernel matrices of slice variants for UST and EPT on a tree in document classification with graph $\mathbb{G}_{\text{Sqrt}}$.

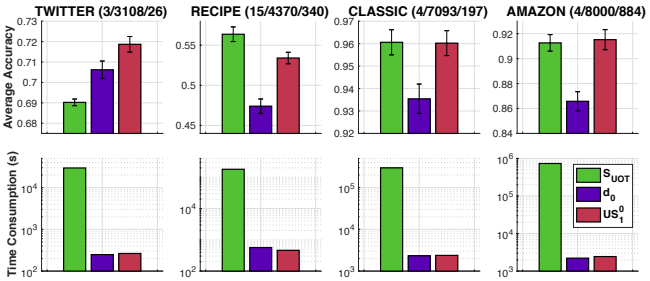

Figure 5: SVM results and time consumption for kernel matrices in document classification with graph $\mathbb{G}_{\text{Log}}$. For each dataset, the numbers in the parenthesis are the number of classes; the number of documents; and the maximum number of unique words for each document respectively.

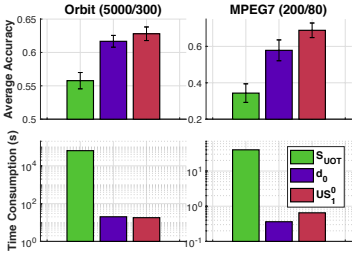

Figure 6: SVM results and time consumption for kernel matrices in TDA with graph $\mathbb{G}_{\text{Log}}$. For each dataset, the numbers in the parenthesis are respectively the number of PD; and the maximum number of points in PD.

B.3.2 Further Empirical Results

We also provides further results for document classification and TDA as follow:

For document classification.

- For $M = 10^2$, we illustrate the SVM results and time consumption for kernels matrices and the effect of the number of slices for graph $\mathbb{G}_{\text{Sqrt}}$ in Figure 9 and Figure 10 respectively. The corresponding results for graph $\mathbb{G}_{\text{Log}}$ are in Figure 11 and Figure 12.

[8] There is a typo in the main text (§6): It should be $M = 10^3$ is for `MPEG7` and $M = 10^2$ is for `Orbit`.

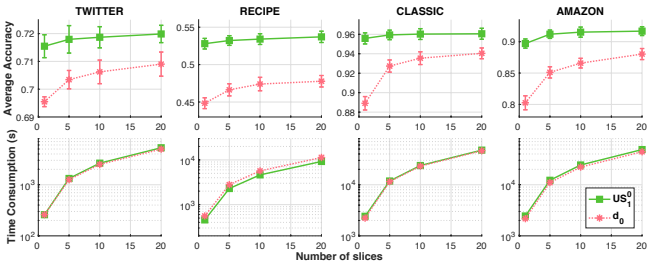

Figure 7: SVM results and time consumption for kernel matrices of slice variants for UST and EPT on a tree in document classification with graph $\mathbb{G}_{\text{Log}}$.

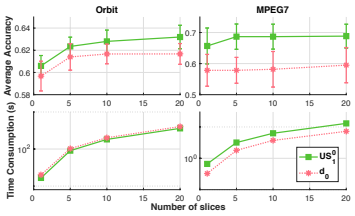

Figure 8: SVM results and time consumption for kernel matrices of slice variants for UST and EPT on a tree in TDA with graph $\mathbb{G}_{\text{Log}}$.

- For $M = 10^3$, we illustrate the SVM results and time consumption for kernels matrices and the effect of the number of slices for graph $\mathbb{G}_{\text{Sqrt}}$ in Figure 13 and Figure 14 respectively. The corresponding results for graph $\mathbb{G}_{\text{Log}}$ are in Figure 15 and Figure 16.

- For $M = 10^4$, we illustrate the SVM results and time consumption for kernels matrices and the effect of the number of slices for graph $\mathbb{G}_{\text{Sqrt}}$ in Figure 17 and Figure 18 respectively. The corresponding results for graph $\mathbb{G}_{\text{Log}}$ are in Figure 19 and Figure 20.

- For $M = 4 \times 10^4$, we illustrate the SVM results and time consumption for kernels matrices and the effect of the number of slices for graph $\mathbb{G}_{\text{Sqrt}}$ in Figure 21 and Figure 22 respectively. The corresponding results for graph $\mathbb{G}_{\text{Log}}$ are in Figure 23 and Figure 24.

For TDA.

- For $M = 10^2$, we illustrate the SVM results and time consumption for kernels matrices and the effect of the number of slices for graph $\mathbb{G}_{\text{Sqrt}}$ in Figure 25 and Figure 26 respectively. The corresponding results for graph $\mathbb{G}_{\text{Log}}$ are in Figure 27 and Figure 28.

- For $M = 10^3$, we illustrate the SVM results and time consumption for kernels matrices and the effect of the number of slices for graph $\mathbb{G}_{\text{Sqrt}}$ in Figure 29 and Figure 30 respectively. The corresponding results for graph $\mathbb{G}_{\text{Log}}$ are in Figure 31 and Figure 32.

- For $M = 10^4$ on `Orbit` dataset and $M = 10^3$ on `MPEG7` dataset (due to the same size of `MPEG7` dataset), we illustrate the SVM results and time consumption for kernels matrices and the effect of the number of slices for graph $\mathbb{G}_{\text{Sqrt}}$ in Figure 33 and Figure 34 respectively. The corresponding results for graph $\mathbb{G}_{\text{Log}}$ are in Figure 35 and Figure 36.

With different exponent p for UST. We also carry out experiments for different p in unbalanced Sobolev transport using the same setting for M in the main text (i.e., $M = 10^4$ for document datasets, $M = 10^3$ for `MPEG7` dataset and $M = 10^2$ for `Orbit` dataset) on graph $\mathbb{G}_{\text{Sqrt}}$ and graph $\mathbb{G}_{\text{Log}}$. Figure 37 and Figure 38 illustrate performances on document classification and TDA respectively with graph $\mathbb{G}_{\text{Sqrt}}$. For graph $\mathbb{G}_{\text{Log}}$, the corresponding results are shown in Figure 39 and Figure 40.[9]

[9]We skip plots about time consumption since the time consumption of UST for $p = 1$ and $p = 2$ are almost identical. Please refer to other Figures where we illustrate the time consumption of UST for $p = 1$.

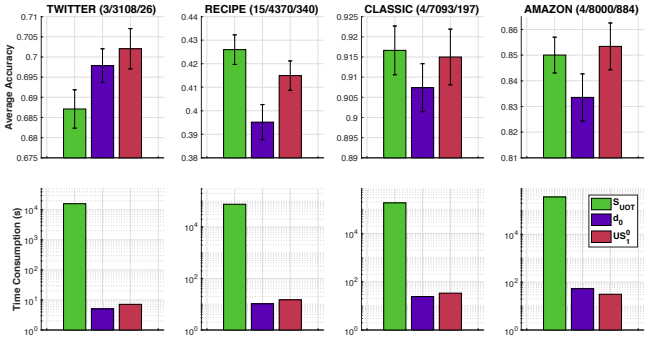

Figure 9: SVM results and time consumption for kernel matrices in document classification with graph $\mathbb{G}_{\text{Sqrt}}$ with $M = 10^2$.

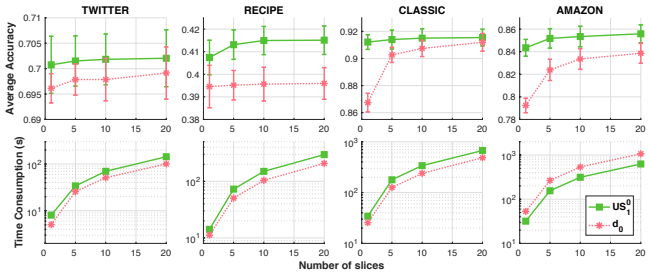

Figure 10: SVM results and time consumption for kernel matrices of slice variants for UST and EPT on a tree in document classification with graph $\mathbb{G}_{\text{Sqrt}}$ with $M = 10^2$.

With Sinkhorn divergence-based approach for UOT (Séjourné et al., 2019) as an extra baseline. Furthermore, we also consider Sinkhorn divergence-based approach for UOT (SD_{UOT}) (Séjourné et al., 2019) as an extra baseline. As we noted in the main manuscript, SD_{UOT} is the debiased version of Sinkhorn-based approach for UOT (S_{UOT}) which may be helpful for applications. Both SD_{UOT} and S_{UOT} are empirically indefinite and they have the same computational complexity.

We illustrate SVM results for document classification and TDA with the extra baseline SD_{UOT} for both graph $\mathbb{G}_{\text{Sqrt}}$ and $\mathbb{G}_{\text{Log}}$ corresponding to Figure 1 (in the main text), Figure 2 (in the main text), Figure 5, and Figure 6 in Figure 41, Figure 42, Figure 43, Figure 44 respectively.

B.3.3 Further Discussions on Empirical Results

The unbalanced Sobolev transport (UST) US_p^α versus d_α of entropy partial transport (EPT) on a tree. Overall, performances of the UST compare favorably with those of d_α of EPT on a tree. Moreover, time consumption of UST is comparable to that of d_α of EPT on trees. So, by exploiting the full graph structure, UST improves performances of d_α of EPT on a tree and still keeps the advantage about the computational complexity.

The unbalanced Sobolev transport (UST) versus Sinkhorn-based unbalanced optimal transport (UOT). The performances of UST is comparable to those of Sinkhorn-based UOT. Recall that kernels for UST are positive definite while kernels for Sinkhorn-based UOT are empirically indefinite. This indefiniteness may affect performances of Sinkhorn-UOT in some settings (e.g., datasets or graph structure). It is worth noting that the UST is several order faster than Sinkhorn-based UOT. Therefore, it is prohibited to apply Sinkhorn-based UOT for large-scale settings while our proposed approach (UST) is scalable to such settings.

The effects of the number of slices (i.e., the number of root nodes used for averaging). In general, when one increases the number of slices for the UST (and d_α of EPT on a tree), their corresponding performances are also increased but it comes with a trade-off about time consumption (i.e., linear to the number of slices). We observe that 10 slices seems a good trade-off between performances and time consumption, similar to observations in (Le and Nguyen, 2021).

Unbalanced Sobolev transport with different p. In our experiments on document classification and TDA, we observe that $p = 1$ for UST consistently gives better performances than $p = 2$ for UST.[10] Generally, one may turn parameter p to

[10]Recall that UST with $p = 1$ has a stronger connection to EPT on graphs thatn UST with $p = 2$ as illustrated in Lemma A.2.

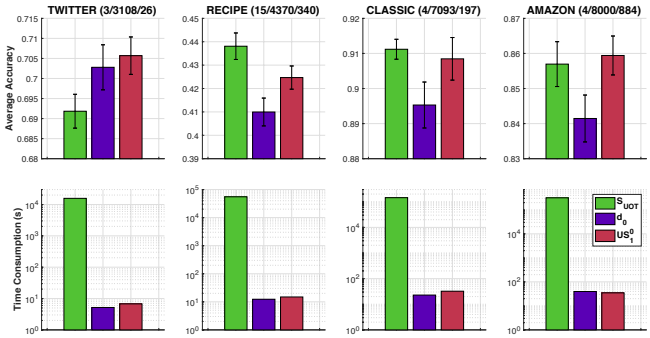

Figure 11: SVM results and time consumption for kernel matrices in document classification with graph $\mathbb{G}_{\text{Log}}$ with $M = 10^2$.

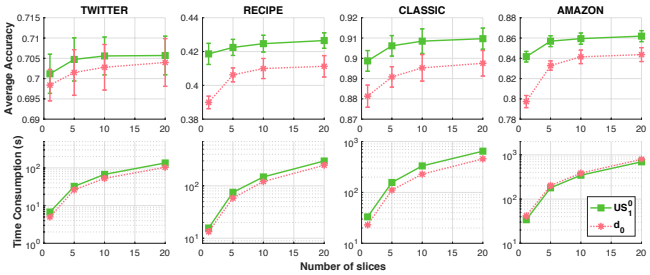

Figure 12: SVM results and time consumption for kernel matrices of slice variants for UST and EPT on a tree in document classification with graph $\mathbb{G}_{\text{Log}}$ with $M = 10^2$.

improve performances of UST in applications.

The extra baseline: Sinkhorn divergence-based approach for UOT. In our experiments, the performances of the extra baseline SD_{UOT} are relative with those of S_{UOT} when comparing with performances of d_α (EPT on a tree) and our proposed UST. The debias property of SD_{UOT} improves performances of S_{UOT} in some datasets, especially for datasets in TDA tasks (`Orbit` and `MPEG7`). For document datasets, performances of SD_{UOT} and S_{UOT} are comparative (the role of debias property is not clear).

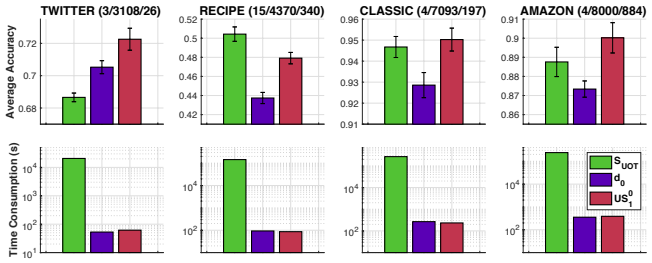

Figure 13: SVM results and time consumption for kernel matrices in document classification with graph $\mathbb{G}_{\text{Sqrt}}$ with $M = 10^3$.

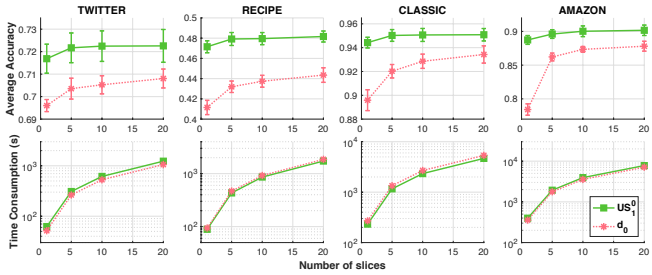

Figure 14: SVM results and time consumption for kernel matrices of slice variants for UST and EPT on a tree in document classification with graph $\mathbb{G}_{\text{Sqrt}}$ with $M = 10^3$.

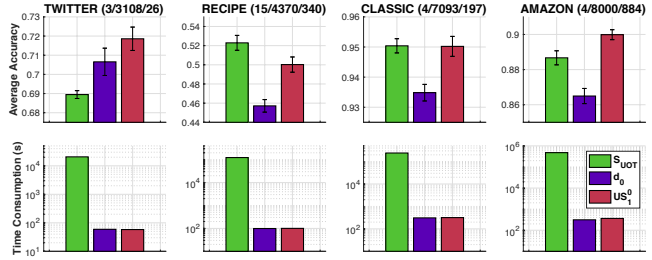

Figure 15: SVM results and time consumption for kernel matrices in document classification with graph $\mathbb{G}_{\text{Log}}$ with $M = 10^3$.

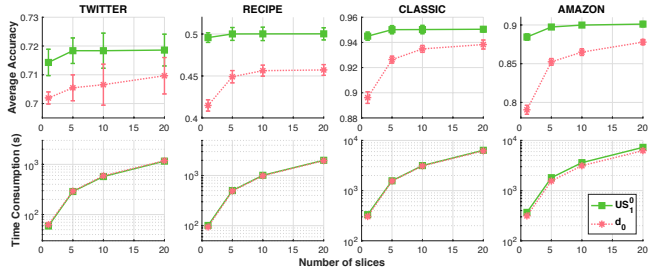

Figure 16: SVM results and time consumption for kernel matrices of slice variants for UST and EPT on a tree in document classification with graph $\mathbb{G}_{\text{Log}}$ with $M = 10^3$.

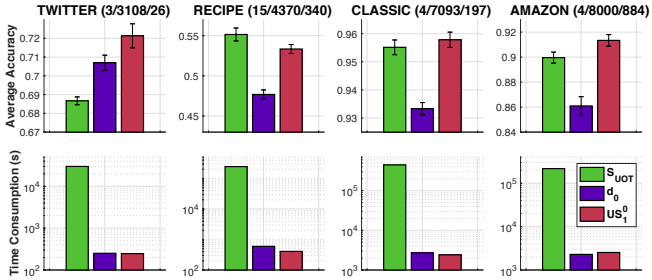

Figure 17: SVM results and time consumption for kernel matrices in document classification with graph $\mathbb{G}_{\text{Sqrt}}$ with $M = 10^4$.

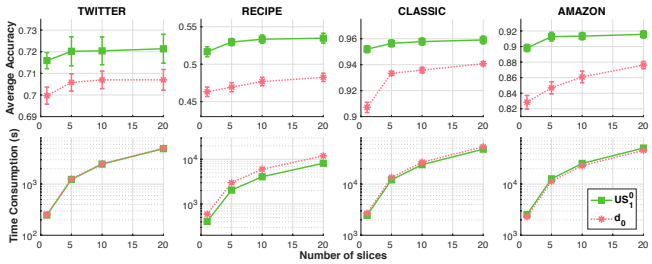

Figure 18: SVM results and time consumption for kernel matrices of slice variants for UST and EPT on a tree in document classification with graph $\mathbb{G}_{\text{Sqrt}}$ with $M = 10^4$.

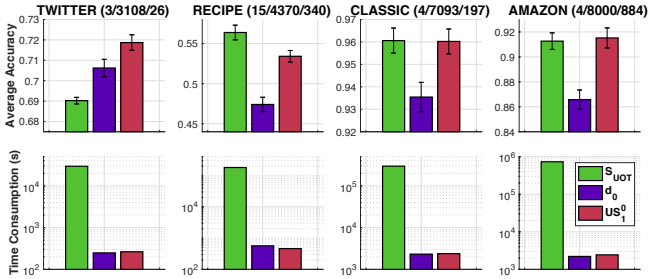

Figure 19: SVM results and time consumption for kernel matrices in document classification with graph $\mathbb{G}_{\text{Log}}$ with $M = 10^4$.

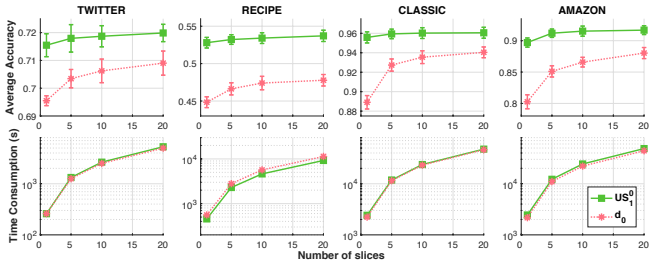

Figure 20: SVM results and time consumption for kernel matrices of slice variants for UST and EPT on a tree in document classification with graph $\mathbb{G}_{\text{Log}}$ with $M = 10^4$.

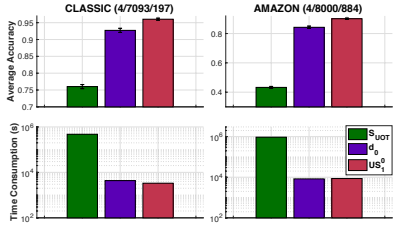

Figure 21: SVM results and time consumption for kernel matrices in document classification with graph $\mathbb{G}_{\text{Sqrt}}$ with $M = 4 \times 10^4$.

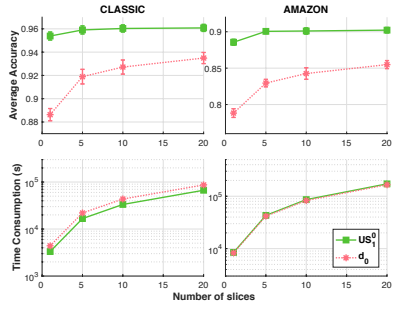

Figure 22: SVM results and time consumption for kernel matrices of slice variants for UST and EPT on a tree in document classification with graph $\mathbb{G}_{\text{Sqrt}}$ with $M = 4 \times 10^4$.

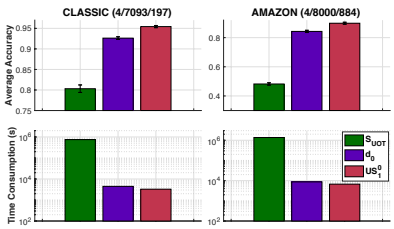

Figure 23: SVM results and time consumption for kernel matrices in document classification with graph $\mathbb{G}_{\text{Log}}$ with $M = 4 \times 10^4$.

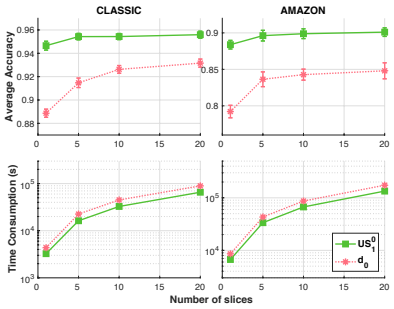

Figure 24: SVM results and time consumption for kernel matrices of slice variants for UST and EPT on a tree in document classification with graph $\mathbb{G}_{\text{Log}}$ with $M = 4 \times 10^4$.

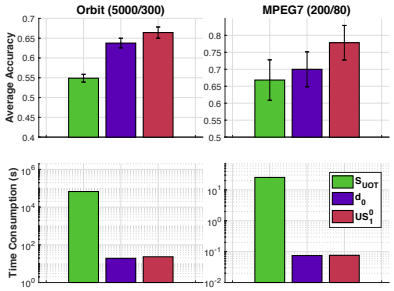

Figure 25: SVM results and time consumption for kernel matrices in TDA with graph $\mathbb{G}_{\text{Sqrt}}$ with $M = 10^2$.

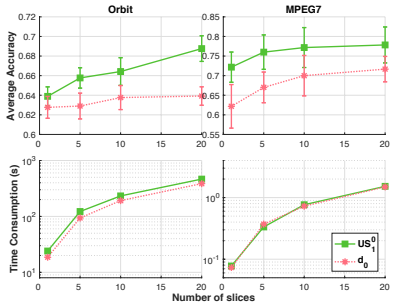

Figure 26: SVM results and time consumption for kernel matrices of slice variants for UST and EPT on a tree in TDA with graph $\mathbb{G}_{\text{Sqrt}}$ with $M = 10^2$.

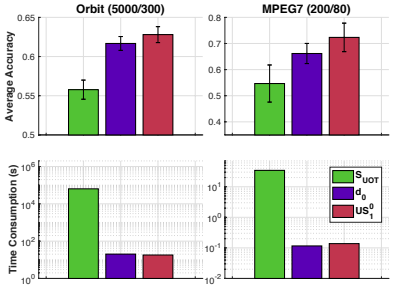

Figure 27: SVM results and time consumption for kernel matrices in TDA with graph $\mathbb{G}_{\text{Log}}$ with $M = 10^2$.

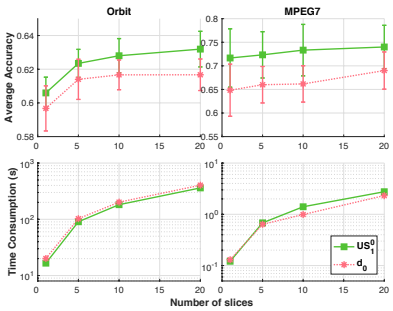

Figure 28: SVM results and time consumption for kernel matrices of slice variants for UST and EPT on a tree in TDA with graph $\mathbb{G}_{\text{Log}}$ with $M = 10^2$.

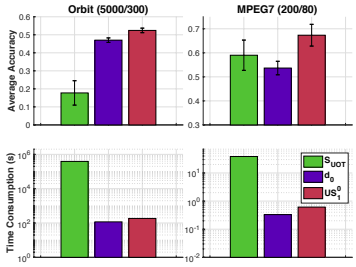

Figure 29: SVM results and time consumption for kernel matrices in TDA with graph $\mathbb{G}_{\text{Sqrt}}$ with $M = 10^3$.

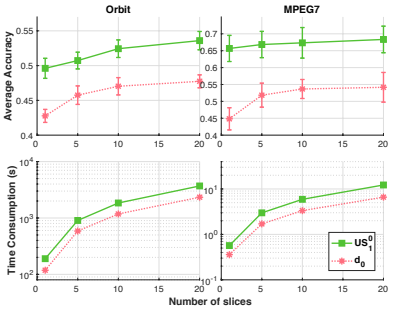

Figure 30: SVM results and time consumption for kernel matrices of slice variants for UST and EPT on a tree in TDA with graph $\mathbb{G}_{\text{Sqrt}}$ with $M = 10^3$.

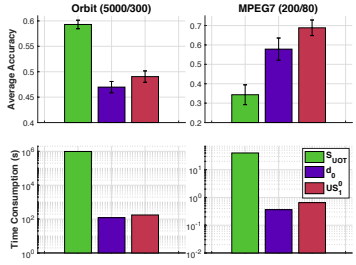

Figure 31: SVM results and time consumption for kernel matrices in TDA with graph $\mathbb{G}_{\text{Log}}$ with $M = 10^3$.

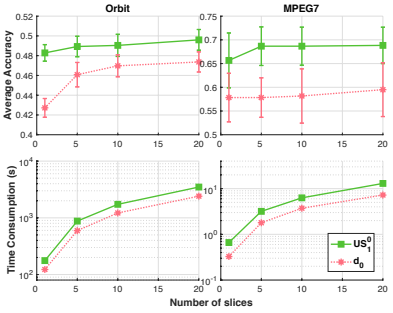

Figure 32: SVM results and time consumption for kernel matrices of slice variants for UST and EPT on a tree in TDA with graph $\mathbb{G}_{\text{Log}}$ with $M = 10^3$.

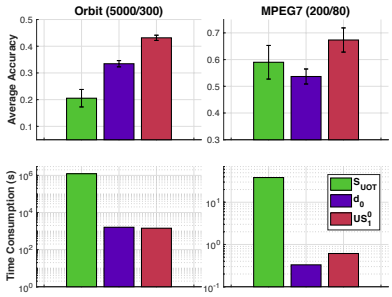

Figure 33: SVM results and time consumption for kernel matrices in TDA with graph $\mathbb{G}_{\text{Sqrt}}$ with $M = 10^4$ for `Orbit` and with $M = 10^3$ for `MPEG7`.

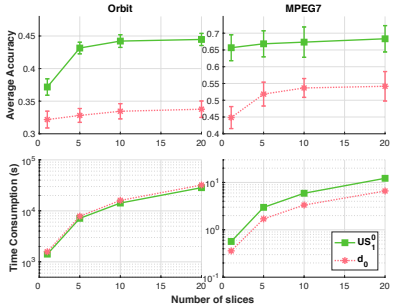

Figure 34: SVM results and time consumption for kernel matrices of slice variants for UST and EPT on a tree in TDA with graph $\mathbb{G}_{\text{Sqrt}}$ with $M = 10^4$ for `Orbit` and with $M = 10^3$ for `MPEG7`.

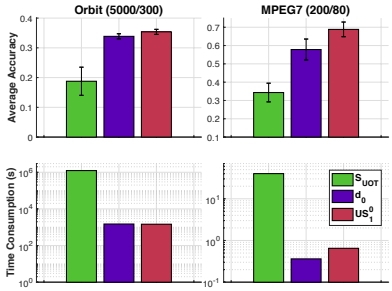

Figure 35: SVM results and time consumption for kernel matrices in TDA with graph $\mathbb{G}_{\text{Log}}$ with $M = 10^4$ for `Orbit` and with $M = 10^3$ for `MPEG7`.

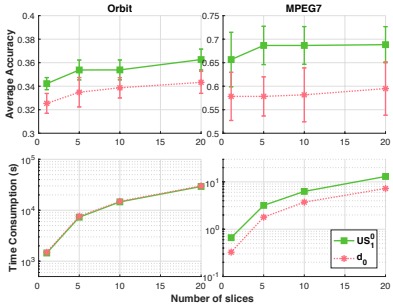

Figure 36: SVM results and time consumption for kernel matrices of slice variants for UST and EPT on a tree in TDA with graph $\mathbb{G}_{\text{Log}}$ with $M = 10^4$ for `Orbit` and with $M = 10^3$ for `MPEG7`.

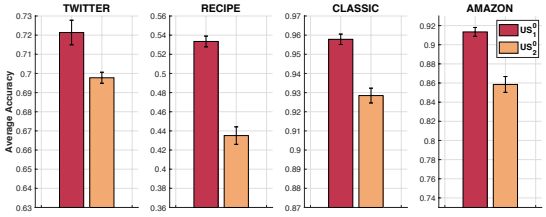

Figure 37: SVM results and time consumption for kernel matrices in document classification with graph $\mathbb{G}_{\text{Sqrt}}$ with $M = 10^4$.

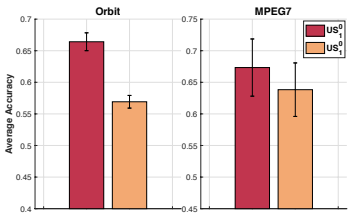

Figure 38: SVM results and time consumption for kernel matrices in TDA with graph $\mathbb{G}_{\text{Sqrt}}$ with $M = 10^2$ for `Orbit` and with $M = 10^3$ for `MPEG7`.

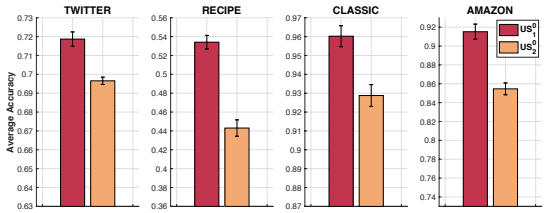

Figure 39: SVM results and time consumption for kernel matrices in document classification with graph $\mathbb{G}_{\text{Log}}$ with $M = 10^4$.

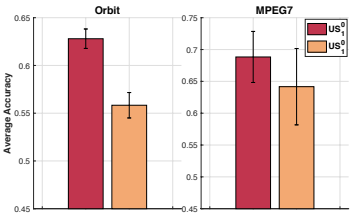

Figure 40: SVM results and time consumption for kernel matrices in TDA with graph $\mathbb{G}_{\text{Log}}$ with $M = 10^2$ for `Orbit` and with $M = 10^3$ for `MPEG7`.

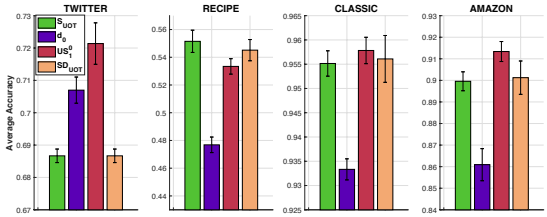

Figure 41: SVM results for document classification with graph $\mathbb{G}_{\text{Sqrt}}$ with an extra baseline (SD_{UOT}).

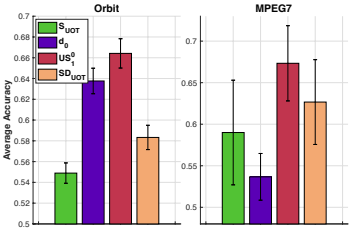

Figure 42: SVM results for TDA with graph $\mathbb{G}_{\text{Sqrt}}$ with an extra baseline (SD_{UOT}).

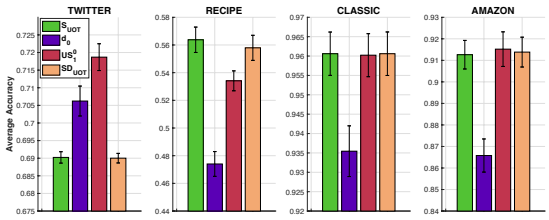

Figure 43: SVM results for document classification with graph $\mathbb{G}_{\text{Log}}$ with an extra baseline (SD_{UOT}).

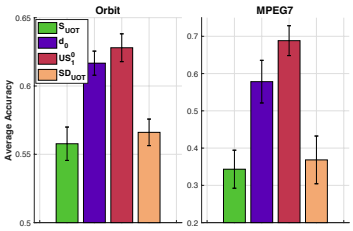

Figure 44: SVM results for TDA with graph $\mathbb{G}_{\mathrm{Log}}$ with an extra baseline ($\mathrm{SD}_{\mathrm{UOT}}$).