An Efficient Orlicz-Sobolev Approach for Transporting Unbalanced Measures on a Graph

Tam Le[*]
Institute of Statistical Mathematics
tam@ism.ac.jp

Truyen Nguyen[*]
University of Akron
truyennguyen@meta.com

Hideitsu Hino
Institute of Statistical Mathematics
hino@ism.ac.jp

Kenji Fukumizu
Institute of Statistical Mathematics
fukumizu@ism.ac.jp

Abstract

We investigate optimal transport (OT) for measures on graph metric spaces with different total masses. To mitigate the limitations of traditional L^p geometry, Orlicz-Wasserstein (OW) and generalized Sobolev transport (GST) employ *Orlicz geometric structure*, leveraging convex functions to capture nuanced geometric relationships and remarkably contribute to advance certain machine learning approaches. However, both OW and GST are restricted to measures with equal total mass, limiting their applicability to real-world scenarios where mass variation is common, and input measures may have noisy supports, or outliers. To address unbalanced measures, OW can either incorporate mass constraints or marginal discrepancy penalization, but this leads to a more complex two-level optimization problem. Additionally, GST provides a scalable yet rigid framework, which poses significant challenges to extend GST to accommodate nonnegative measures. To tackle these challenges, in this work we revisit the entropy partial transport (EPT) problem. By exploiting Caffarelli & McCann [12]'s insights, we develop a novel variant of EPT endowed with Orlicz geometric structure, called *Orlicz-EPT*. We establish theoretical background to solve Orlicz-EPT using a binary search algorithmic approach. Especially, by leveraging the dual EPT and the underlying graph structure, we formulate a novel regularization approach that leads to the proposed *Orlicz-Sobolev transport* (OST). Notably, we demonstrate that OST can be efficiently computed by simply solving a univariate optimization problem, in stark contrast to the intensive computation needed for Orlicz-EPT. Building on this, we derive geometric structures for OST and draw its connections to other transport distances. We empirically illustrate that OST is several-order faster than Orlicz-EPT. Furthermore, we show initial evidence on the advantages of OST for measures on a graph in document classification and topological data analysis.

1 Introduction

Orlicz-Wasserstein (OW) extends L^p geometry by leveraging a specific class of convex functions for Orlicz geometric structure. Intuitively, OW is an instance of optimal transport (OT), which utilizes Orlicz metric as its ground cost [67, 32, 28, 3, 41]. Building on this foundation, OW has proven instrumental in advancing certain machine learning approaches. For example, recent works have leveraged OW to tackle challenging problems: Altschuler & Chewi [3] use OW as a metric

[*]: equal contribution. Correspondence to: Tam Le <tam@ism.ac.jp>.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

shift for Rényi divergence, enabling novel differential-privacy-inspired techniques to overcome longstanding challenges for fast convergence of hypocoercive differential equations, while Guha et al. [28] employ OW metric to significantly improve Bayesian contraction rates in hierarchical Bayesian nonparametric models by overcoming limitations raised from the usage of traditional OT with Euclidean ground cost. However, OW's high computational complexity, stemming from its two-level optimization formula, poses a significant limitation. To address this challenge, Le et al. [41] introduce generalized Sobolev transport (GST), a scalable variant of OW suitable for practical application domains, especially for large-scale settings. Moreover, Orlicz geometric structure has been successfully applied to various machine learning problems, including linear regression [4, 66], scalable approaches [21] for reinforcement learning, kernelized support vector machines, and clustering. Additionally, Orlicz metrics play a crucial role in deriving deviation bounds for polynomial-growth functions to approximate kernel derivatives [13], and have been used as regularization in OT problems [46]. For in-depth studies on Orlicz functions, see [2, 60].

When dealing with input measures having different total masses, various approaches have been proposed in the literature to address this challenge [30, 29, 6, 12, 25, 44, 57, 58, 26, 33, 45, 18, 8, 27, 63, 64, 56, 62, 15, 5, 48, 36, 24, 16, 65, 40, 52, 7, 14, 72]. These approaches for unbalanced measures have proven effective in various domains, including color transfer [8], shape matching [8], image-to-image translation [72], multi-label learning [26], positive-unlabeled learning [15], point-cloud gradient flow [72], natural language processing [36, 40], topological data analysis (TDA) [36, 40], generative modeling [5, 72], domain adaptation [5], and robust approaches for handling noisy supports, outliers [26, 5, 48], or noisy ground cost [54, 42].

In this work, we focus on the OT problem with Orlicz geometric structure for unbalanced measures supported on a graph metric space. On one hand, OW naturally extends OT's flexibility to handle unbalanced measures by incorporating either mass constraints or marginal difference penalization, formulated as partial OT (POT) or unbalanced OT (UOT) respectively. However, these approaches result in a more complex two-level optimization problem, analogous to POT/UOT with Orlicz metric cost, which poses significant computational challenges. On the other hand, although GST provides a scalable alternative to the computationally intensive OW, it still assumes equal-mass input measures. Moreover, due to GST's definition as an optimization over the critic function, extending it to unbalanced measures is nontrivial. To address these limitations, we revisit the entropy partial transport (EPT) problem [36, 40, 72] and leverage insights from Caffarelli & McCann [12] to reformulate EPT as a standard complete OT problem. Then by carefully calibrating the corresponding ground cost for its nonnegativity, we propose Orlicz-EPT and establish a theoretical foundation for solving it by a binary search algorithmic approach. Furthermore, by exploiting the dual EPT and underlying graph structure, we introduce a novel regularization approach, leading to Orlicz-Sobolev transport (OST), which scales Orlicz-EPT for practical applications.

Contribution. In summary, our contributions are two-fold as follows:

- We revisit the EPT problem and leverage Caffarelli & McCann [12]'s insights to reformulate EPT as a standard complete OT, leading to the development of the proposed Orlicz-EPT. We establish its theoretical foundations, enabling a binary search algorithmic approach for its computation. Additionally, we develop a novel regularization approach, resulting in the proposed OST. We show that OST can be efficiently computed by simply solving a univariate optimization problem, unlike the computationally intensive Orlicz-EPT.
- We derive geometric structures for OST and establish its connections to other transport distances. Our empirical results demonstrate that OST is several-order faster than Orlicz-EPT. We also provide initial evidence on the advantages of OST for document classification and TDA.

Organization. In §2, we briefly review relevant background and notions. We revisit EPT problem and propose Orlicz-EPT in §3. In §4, we introduce the computationally efficient OST. Then we derive geometric structures for OST and draw its connections to other transport distances in §5. In §6, we discuss related work. Empirical results are presented in §7, followed by concluding remarks in §8. Proofs of key theoretical results and additional materials are deferred to the Appendices. Furthermore, we have released code for our proposed approaches.[2]

[2]The code repository is on `https://github.com/lttam/OST_OrliczEPT`.

2 Preliminaries

In this section, we introduce notations, and briefly review graph, and Orlicz functions.

Graph. We follow the graph setting as in [39]. Let V, E be the sets of nodes and edges respectively. We consider a connected, undirected, and physical[3] graph $\mathbb{G} = (V, E)$ with positive edge lengths $\{w_e\}_{e\in E}$. For continuous graph setting, we regard $\mathbb{G}$ as the set of all nodes in V and all points forming the edges in E. We equip $\mathbb{G}$ with graph metric $d_{\mathbb{G}}(x, y)$ which equals to the length of the shortest path between x and y in $\mathbb{G}$. Additionally, we assume that there exists a fixed root node $z_0 \in V$ such that the shortest path connecting z_0 and x is unique for any $x \in \mathbb{G}$, i.e., the uniqueness property of the shortest paths. We denote $\mathcal{P}(\mathbb{G})$ (resp. $\mathcal{P}(\mathbb{G}\times\mathbb{G})$) as the set of all nonnegative Borel measures on $\mathbb{G}$ (resp. $\mathbb{G}\times\mathbb{G}$) with a finite mass. Let $[x, z]$ be the shortest path connecting x and z in $\mathbb{G}$. For $x \in \mathbb{G}$, edge $e \in E$, define the sets $\Lambda(x)$ and γ_e as follows:

$$\Lambda(x) := \big\{y \in \mathbb{G} :\, x \in [z_0, y]\big\}, \qquad \gamma_e := \big\{y \in \mathbb{G} :\, e \subset [z_0, y]\big\}. \tag{1}$$

Functions on graph. By a continuous function f on $\mathbb{G}$, we mean that $f : \mathbb{G} \to \mathbb{R}$ is continuous w.r.t. the topology on $\mathbb{G}$ induced by the Euclidean distance. Henceforth, $C(\mathbb{G})$ denotes the set of all continuous functions on $\mathbb{G}$. Similar notation is used for continuous functions on $\mathbb{G}\times\mathbb{G}$. Given a positive scalar $b > 0$, then a function $f : \mathbb{G} \to \mathbb{R}$ is called b-Lipschitz w.r.t. $d_{\mathbb{G}}$ if $|f(x) - f(y)| \le b\, d_{\mathbb{G}}(x, y)$ for every x and y in $\mathbb{G}$.

A family of convex functions. We consider the collection of N-functions [2, §8.2] which are special convex functions on $\mathbb{R}_+$. Hereafter, a strictly increasing and convex function $\Phi : [0,\infty) \to [0,\infty)$ is called an N-function if $\lim_{t\to 0} \frac{\Phi(t)}{t} = 0$ and $\lim_{t\to+\infty} \frac{\Phi(t)}{t} = +\infty$.

Orlicz functional space. Given an N-function Φ and a nonnegative Borel measure ω on $\mathbb{G}$, let $L_\Phi(\mathbb{G}, \omega)$ be the linear hull of the set of all Borel measurable functions $f : \mathbb{G} \to \mathbb{R}$ satisfying $\int_{\mathbb{G}} \Phi(|f(x)|)\omega(\mathrm{d}x) < \infty$. Then, $L_\Phi(\mathbb{G}, \omega)$ is a normed space with the Luxemburg norm, defined as

$$\|f\|_{L_\Phi} := \inf\left\{t > 0 \mid \int_{\mathbb{G}} \Phi\left(\frac{|f(x)|}{t}\right)\omega(\mathrm{d}x) \le 1\right\}. \tag{2}$$

3 Orlicz-EPT: Entropy Partial Transport with Orlicz Geometric Structure

In this section, we revisit the entropy partial transport (EPT) problem [36, 40, 72], then develop Orlicz-EPT as a variant of EPT endowed with Orlicz geometric structure.

3.1 Entropy Partial Transport (EPT)

Let γ_1, γ_2 be the first and second marginals of $\gamma \in \mathcal{P}(\mathbb{G}\times\mathbb{G})$ respectively. For unbalanced measures $\mu, \nu \in \mathcal{P}(\mathbb{G})$, we consider the set $\Pi_{\le}(\mu, \nu) := \{\gamma :\, \gamma_1 \le \mu,\, \gamma_2 \le \nu\}$.[4] Additionally, let f_1, f_2 be the Radon-Nikodym derivatives of γ_1 w.r.t. μ and of γ_2 w.r.t. ν respectively, i.e., $\gamma_1 = f_1\mu$ ($0 \le f_1 \le 1$, μ-a.e.) and $\gamma_2 = f_2\nu$ ($0 \le f_2 \le 1$, ν-a.e.).

For convex and lower semicontinuous entropy functions $F_1,\, F_2 : [0, 1] \to (0,\infty)$, and nonnegative weight functions $w_1, w_2 : \mathbb{G} \to [0,\infty)$, we consider the weighted relative entropies $\mathcal{F}_1(\gamma_1|\mu) := \int_{\mathbb{G}} w_1(x)F_1(f_1(x))\mu(\mathrm{d}x)$, and $\mathcal{F}_2(\gamma_2|\nu) := \int_{\mathbb{G}} w_2(x)F_2(f_2(x))\nu(\mathrm{d}x)$. For scalar $b > 0$, scalar $m \in [0, \bar{m}]$ with $\bar{m} := \min\{\mu(\mathbb{G}), \nu(\mathbb{G})\}$, and graph metric $d_{\mathbb{G}}$ as ground cost, the EPT problem is

$$\mathrm{W}_m(\mu, \nu) := \inf_{\gamma\in\Pi_{\le}(\mu,\nu),\, \gamma(\mathbb{G}\times\mathbb{G})=m} \Big[\mathcal{F}_1(\gamma_1|\mu) + \mathcal{F}_2(\gamma_2|\nu) + b\int_{\mathbb{G}\times\mathbb{G}} d_{\mathbb{G}}(x, y)\gamma(\mathrm{d}x, \mathrm{d}y)\Big]. \tag{3}$$

Following [40, §3], by using entropy functions $F_1(s) = F_2(s) := |s - 1|$ and considering a Lagrange multiplier $\lambda \in \mathbb{R}$ conjugate to the constraint $\gamma(\mathbb{G}\times\mathbb{G}) = m$, we instead study the problem

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

[3] In the sense that V is a subset of Euclidean space $\mathbb{R}^n$, and each edge $e \in E$ is the standard line segment in $\mathbb{R}^n$ connecting the two vertices of the edge e.

[4] $\gamma_1 \le \mu$ means that $\gamma_1(B) \le \mu(B)$ for all Borel set $B \subset \mathbb{G}$, similarly for $\gamma_2 \le \nu$.

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

EPT as a standard OT. Following Caffarelli & McCann [12]'s insights, we can reformulate problem (4) as the standard complete OT problem. However, it is not guarantee that the corresponding standard OT has a nonnegative ground cost, e.g., see [12, 36, 40, 72]. Therefore, such OT reformulation may not be applicable to derive corresponding OW as in [67, 32, 28, 3, 41] since N-function is only defined for nonnegative domain (§2). Therefore, it is essential to carefully calibrate the ground cost of the corresponding standard OT of EPT to ensure its nonnegativity.

Precisely, following [40, Theorem 3.1], we henceforth consider $\lambda \geq 0$.[6] Then let $\hat{s}$ be a point outside graph $\mathbb{G}$, i.e., $\hat{s} \notin \mathbb{G}$, and extend graph metric cost $d_{\mathbb{G}}$ on $\mathbb{G}$ to a new *nonnegative* cost function $\hat{c}$ with $b\lambda$-deviation on $\hat{\mathbb{G}} := \mathbb{G} \cup \{\hat{s}\}$ as follows:

$$\hat{c}(x,y) := \begin{cases} b\, d_{\mathbb{G}}(x,y) & \text{if } x,y \in \mathbb{G}, \\ w_1(x) + b\lambda & \text{if } x \in \mathbb{G} \text{ and } y = \hat{s}, \\ w_2(y) + b\lambda & \text{if } x = \hat{s} \text{ and } y \in \mathbb{G}, \\ b\lambda & \text{if } x = y = \hat{s}. \end{cases} \tag{5}$$

For unbalanced measures μ, ν, we construct corresponding probability (balanced) measures $\hat{\mu} = \frac{\mu + \nu(\mathbb{G})\delta_{\hat{s}}}{\mu(\mathbb{G})+\nu(\mathbb{G})}$ and $\hat{\nu} = \frac{\nu + \mu(\mathbb{G})\delta_{\hat{s}}}{\mu(\mathbb{G})+\nu(\mathbb{G})}$. Let $\Pi(\hat{\mu},\hat{\nu}) := \Big\{\hat{\gamma} \in \mathcal{P}(\hat{\mathbb{G}}\times\hat{\mathbb{G}}) : \hat{\mu}(U) = \hat{\gamma}(U\times\hat{\mathbb{G}}),\, \hat{\nu}(U) = \hat{\gamma}(\hat{\mathbb{G}}\times U)$ for all Borel sets $U \subset \hat{\mathbb{G}}\Big\}$, then one can recast EPT (4) as a standard OT with cost $\hat{c}$.

Proposition 3.1. *Consider the standard OT $\mathcal{W}_{\hat{c}}$ between probability measures $\hat{\mu}, \hat{\nu}$ with cost $\hat{c}$,*

$$\mathcal{W}_{\hat{c}}(\hat{\mu},\hat{\nu}) := \inf_{\hat{\gamma}\in\Pi(\hat{\mu},\hat{\nu})} \int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \hat{c}(x,y)\hat{\gamma}(\mathrm{d}x,\mathrm{d}y), \tag{6}$$

then we have

$$\mathrm{KT}(\mu,\nu) := (\mu(\mathbb{G}) + \nu(\mathbb{G}))\,(\mathcal{W}_{\hat{c}}(\hat{\mu},\hat{\nu}) - b\lambda) = \mathrm{ET}_\lambda(\mu,\nu). \tag{7}$$

The proof is placed in Appendix §A.2.1.

Thus, we have reformulated EPT (4) for unbalanced measures as a corresponding standard complete OT (7) with *nonnegative* ground cost. Consequently, we bypass the technical challenges inherent in unbalanced settings and can leverage abundant existing results and approaches in the standard balanced setting for OT problems with unbalanced measures on a graph.

Remark 3.2 (Nonnegativity). Unlike existing approaches, e.g., as in [12, 36, 40, 72], the new ground cost $\hat{c}$ of the corresponding standard OT problem (7) for EPT is *guaranteed* to be nonnegative. Our calibration is essential for developing the associated OW from its standard OT as in [67, 32, 28, 17].

3.2 Orlicz-EPT

Following the approaches in [67, 32, 28, 17], we define *Orlicz-EPT*, which is EPT endowed with an Orlicz geometric structure, based on the standard OT problem (7) as follows:

$$\mathcal{OE}_\Phi(\mu,\nu) := (\mu(\mathbb{G}) + \nu(\mathbb{G}))\,(\mathcal{W}_\Phi(\hat{\mu},\hat{\nu}) - b\lambda), \tag{8}$$

where $\mathcal{W}_\Phi(\hat{\mu},\hat{\nu}) := \inf_{\tilde{\gamma}\in\Pi(\hat{\mu},\hat{\nu})} \inf\Big[t > 0 : \int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \Phi\left(\frac{\hat{c}(x,y)}{t}\right) \mathrm{d}\tilde{\gamma}(x,y) \leq 1\Big]$.

It should be noted that, similar to OW, Orlicz-EPT (8) is derived from the standard OT problem (7), which circumvents all challenges coming from the setting of unbalanced measures.

We next show that the objective function of Orlicz-EPT is monotone non-increasing w.r.t. t.

Proposition 3.3 (Monotonicity)**.** *Let Φ be an N-function, and let $\hat{c}$ be the cost given by* (5)*. For any probability measures $\hat{\mu}$ and $\hat{\nu}$ on $\hat{\mathbb{G}}$, define*

$$\mathcal{A}(t;\hat{\mu},\hat{\nu}) := \inf_{\tilde{\gamma}\in\Pi(\hat{\mu},\hat{\nu})} \int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \Phi\left(\frac{\hat{c}(x,y)}{t}\right) \mathrm{d}\tilde{\gamma}(x,y) \quad \textit{for} \quad t > 0. \tag{9}$$

Then the function $t \in (0,+\infty) \longmapsto \mathcal{A}(t;\hat{\mu},\hat{\nu})$ is monotone non-increasing.

The proof is placed in Appendix §A.2.2.

[5] The relationship between Problem (3) and Problem (4) is established in [40, Theorem A.1].

[6] The dual EPT result is the foundation for developing Orlicz-Sobolev transport in §4, where λ is nonnegative.

Computation. Observe that for a fixed t, $\mathcal{A}$ is a standard OT problem between $\hat{\mu}$ and $\hat{\nu}$ with the cost function $\Phi\left(\frac{\hat{c}(\cdot,\cdot)}{t}\right)$. For computational efficiency,[7] we consider its corresponding entropic regularization [20], and show that the monotonicity is preserved.

Proposition 3.4 (Entropic regularization)**.** *Define the entropic regularization of $\mathcal{A}$ as*

$$\mathcal{A}_\varepsilon(t;\hat{\mu},\hat{\nu}) := \inf_{\tilde{\gamma}\in\Pi(\hat{\mu},\hat{\nu})}\left[\int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}}\Phi\left(\frac{\hat{c}(x,y)}{t}\right)\mathrm{d}\tilde{\gamma}(x,y)-\varepsilon H(\tilde{\gamma})\right], \tag{10}$$

where $\varepsilon \geq 0$ and H is Shannon entropy defined by $H(\tilde{\gamma}) := -\int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}}(\log\tilde{\gamma}(x,y)-1)\mathrm{d}\tilde{\gamma}(x,y)$. Then the function $t\in(0,+\infty)\longmapsto\mathcal{A}_\varepsilon(t;\hat{\mu},\hat{\nu})$ is monotone non-increasing.

The proof is placed in Appendix §A.2.3.

In addition, we obtain the following upper and lower bounds for $\mathcal{A}_\varepsilon$.

Proposition 3.5 (Bounds)**.** *Let* $\mathrm{supp}(\cdot)$ *be a set of supports of a measure, then we have*

$$\mathcal{A}_\varepsilon\left(\frac{\mathcal{W}_{\hat{c}}(\hat{\mu},\hat{\nu})}{\Phi^{-1}(1+\varepsilon\left[H(\hat{\mu})+H(\hat{\nu})-1\right])};\hat{\mu},\hat{\nu}\right)\geq 1,\quad \textit{and}\quad \mathcal{A}_\varepsilon\left(\frac{L_{\hat{\mu},\hat{\nu}}}{\Phi^{-1}(1+\varepsilon)};\hat{\mu},\hat{\nu}\right)\leq 1,$$

where $L_{\hat{\mu},\hat{\nu}} := \max_{x\in\mathrm{supp}(\hat{\mu}),y\in\mathrm{supp}(\hat{\nu})}\hat{c}(x,y)$.

The proof is placed in Appendix §A.2.4.

Thanks to the monotonicity of $\mathcal{A}_\varepsilon$ in Proposition 3.4 and the limits of $\mathcal{A}_\varepsilon$ in Proposition 3.5, we can leverage the binary search approach to compute the entropic regularized Orlicz-EPT, which corresponds to the original Orlicz-EPT (8). Precisely, this entropic regularization is defined as

$$\mathcal{OE}_{\Phi,\varepsilon}(\mu,\nu) := \left(\mu(\mathbb{G})+\nu(\mathbb{G})\right)\left(\mathcal{W}_{\Phi,\varepsilon}(\hat{\mu},\hat{\nu})-b\lambda\right), \tag{11}$$

where $\mathcal{W}_{\Phi,\varepsilon}(\hat{\mu},\hat{\nu}) := \inf_{\tilde{\gamma}\in\Pi(\hat{\mu},\hat{\nu})}\inf\left[t>0:\int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}}\Phi\left(\frac{\hat{c}(x,y)}{t}\right)\mathrm{d}\tilde{\gamma}(x,y)-\varepsilon H(\tilde{\gamma})\leq 1\right]$.

Discussions. Orlicz-EPT is a novel variant of EPT that incorporates Orlicz geometric structure. Leveraging Caffarelli & McCann [12]'s insights and carefully calibrating the ground cost of the corresponding standard OT to ensure its nonnegativity, we are able to bypass all challenges of unbalanced measures and derive the proposed Orlicz-EPT from the standard OT, similar to OW [67]. We note that $\mathcal{OE}_{\Phi,\varepsilon}$ (11) performs binary search with quadratic complexity $\mathcal{A}_\varepsilon$ (10), instead of dealing with super-cubic complexity $\mathcal{A}$ (9) in $\mathcal{OE}_\Phi$ (8). Unfortunately, the two-level optimization structure of $\mathcal{OE}_{\Phi,\varepsilon}$ still retains significant complexity, severely limiting its practical applications, particularly in large-scale settings. To address this computational challenge, in the next section we exploit the dual EPT and graph structure to develop a *novel regularization* approach, resulting in the proposed Orlicz-Sobolev transport. This approach adopts the Orlicz geometric structure used in Orlicz-EPT, but offers a much more efficient computation.

4 Orlicz-Sobolev Transport: A Scalable Variant of Orlicz-EPT

In this section, we leverage the dual EPT and underlying graph structure to develop a novel regularization approach, resulting in the proposed *Orlicz-Sobolev transport* (OST).

Dual EPT. For b-Lipschitz w_1, w_2 (w.r.t. $d_{\mathbb{G}}$), from [40, Corollary 3.2], the dual EPT is

$$\mathrm{ET}_\lambda(\mu,\nu) = \sup_{f\in\mathbb{U}}\int_{\mathbb{G}} f(\mu-\nu)-\frac{b\lambda}{2}\big[\mu(\mathbb{G})+\nu(\mathbb{G})\big], \tag{12}$$

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

Let Ψ be the complement N-function of Φ and ω be a nonnegative Borel measure on $\mathbb{G}$. Then let $WL_\Psi(\mathbb{G},\omega)$ be the graph-based Orlicz-Sobolev space [41, Definition 3.1] associated to Ψ and ω. Inspired by the approach of GST [41], we consider the critic function $f\in\mathbb{U}$ within $WL_\Psi(\mathbb{G},\omega)$. Consequently, the b-Lipschitz constraint on the critic function $f\in\mathbb{U}$ is replaced by $\|f'\|_{L_\Psi}\leq b$.

[7] Entropic regularized OT reduces the computational cost of OT from super-cubic to quadratic complexity [20].

For $f \in WL_\Psi(\mathbb{G}, \omega)$, we have $f(x) = f(z_0) + \int_{[z_0,x]} f'(y)\omega(\mathrm{d}y), \forall x \in \mathbb{G}$. Let $\mathbf{1}$ be the indicator function. Then by using the generalized Hölder inequality [2, §8.11] and $\|f'\|_{L_\Psi} \le b$, we can control the integral part over the generalized graph derivative f' for $f(x)$ as follows:

$$\int_{[z_0,x]} f'(y)\omega(\mathrm{d}y) \le 2\|f'\|_{L_\Psi} \left\|\mathbf{1}_{[z_0,x]}\right\|_{L_\Phi} \le 2b \left\|\mathbf{1}_{[z_0,x]}\right\|_{L_\Phi} \le \frac{2b}{\Phi^{-1}(1/\omega(\mathbb{G}))}, \tag{13}$$

where the last inequality is due to the increasing property of N-function Φ. Therefore, instead of the bounded constraint on the critic function f in $\mathbb{U}$, we constraint only on $f(z_0)$.

Definition 4.1 (Orlicz-Sobolev transport (OST))**.** For $\alpha \in [0, \frac{1}{2}(b\lambda + w_1(z_0) + w_2(z_0))]$, let $\mathcal{I}_\alpha := \left[-w_2(z_0) - \frac{b\lambda}{2} + \alpha, w_1(z_0) + \frac{b\lambda}{2} - \alpha\right]$. The Orlicz-Sobolev transport for $\mu, \nu \in \mathcal{P}(\mathbb{G})$ is defined

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) := \sup_{f \in \mathbb{U}_{\Psi,\alpha}} \left[\int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(x)\nu(\mathrm{d}x)\right], \tag{14}$$

where $\mathbb{U}_{\Psi,\alpha} := \left\{f \in WL_\Psi(\mathbb{G},\omega) : \|f'\|_{L_\Psi} \le b, f(z_0) \in \mathcal{I}_\alpha\right\}$.

Intuitively, $\mathbb{U}_{\Psi,\alpha}$ is the collection of all functions f expressed by $f(x) = s + \int_{[z_0,x]} h(y)\omega(\mathrm{d}y), \forall x \in \mathbb{G}$, where $s \in \mathcal{I}_\alpha$, and $\|h\|_{L_\Psi} \le b$. The upper bound constraint on α is to ensure that $\mathcal{I}_\alpha$ is nonempty. When $\alpha = 0$, $\mathcal{I}_\alpha$ is the largest interval. Also, OST is an instance of the integral probability metric [49].

Computation. Given unbalanced measures $\mu, \nu \in \mathcal{P}(\mathbb{G})$, for brevity let us define

$$\Theta := \begin{cases} w_1(z_0) + \frac{b\lambda}{2} - \alpha & \text{if} \quad \mu(\mathbb{G}) \ge \nu(\mathbb{G}), \\ w_2(z_0) + \frac{b\lambda}{2} - \alpha & \text{if} \quad \mu(\mathbb{G}) < \nu(\mathbb{G}). \end{cases} \tag{15}$$

Theorem 4.2 (Univariate optimization problem for OST)**.** *OST can be computed as follows*

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) = \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| + \inf_{k>0} \frac{1}{k}\left(1 + \int_{\mathbb{G}} \Phi\left(kb\,|\mu(\Lambda(x)) - \nu(\Lambda(x))|\right)\omega(dx)\right). \tag{16}$$

The proof is placed in Appendix §A.2.5.

We derive the discrete case for OST which provides an explicit expression for the integral in (63).

Corollary 4.3 (Discrete case)**.** *Let ω be the length measure of graph $\mathbb{G}$, and assume that input measures μ, ν are supported on nodes in V of graph $\mathbb{G}$.*[8] *Then, we have*

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) = \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| + \inf_{k>0} \frac{1}{k}\left(1 + \sum_{e \in E} w_e \Phi(kb\,|\mu(\gamma_e) - \nu(\gamma_e)|)\right). \tag{17}$$

The proof is placed in Appendix §A.2.6.

Therefore, OST can be efficiently computed by simply solving the *univariate* optimization problem (17), thanks to the proposed novel regularization for critic functions in $\mathbb{U}_{\Psi,\alpha}$.

Remark 4.4 (Non-physical graph)*.* In §2, $\mathbb{G}$ is assumed to be a physical graph. Corollary 4.3 implies that OST only depends on graph structure (V, E) and edge weights w_e when input measures are supported on nodes in V of $\mathbb{G}$. Hence, OST is applicable for non-physical graph $\mathbb{G}$ for such cases.

Remark 4.5 (Complementary pairs of N-functions)*.* Corollary 4.3 also implies that one can compute OST with N-function Φ without involving its complementary N-function Ψ (17), unlike its definition (14). The univariate optimization formula (17)) for OST requires that Ψ is finite-valued, which is satisfied for any N-function Φ as it grows faster than linear.

Preprocessing for γ_e. Similar to the GST computation [41], we precompute set γ_e (1) for all edge e in $\mathbb{G}$. More concretely, we apply the Dijkstra algorithm to recompute the shortest paths from z_0 to all other vertices in V with complexity $\mathcal{O}(|E| + |V|\log|V|)$, where $|\cdot|$ denotes the set cardinality.

Sparsity. Observe that for every $x \in \mathrm{supp}(\mu)$, its mass is gathered into $\mu(\gamma_e)$ if and only if $e \subset [z_0, x]$ [41]. Let $E_{\mu,\nu} := \{e \in E \mid \exists z \in (\mathrm{supp}(\mu) \cup \mathrm{supp}(\nu)), e \subset [z_0, z]\} \subset E$. Then it suffices to compute the summation only over edges $e \in E_{\mu,\nu}$ in (17) for OST, i.e., screen out all edges $e \in E \setminus E_{\mu,\nu}$.

[8] It can be extended for measures supported in $\mathbb{G}$ (see §B.2).

5 Theoretical Properties

In this section, we leverage the computational efficiency of OST to derive its geometric structure and explore its connections to other transport distances.

Geometric structures of OST.

Proposition 5.1 (Geometric structure)**.** *Let* $0 \leq \alpha < \frac{b\lambda}{2} + \min\{w_1(z_0), w_2(z_0)\}$ *and* $\mu, \nu, \sigma \in \mathcal{P}(\mathbb{G})$.

i) $\mathcal{OS}_{\Phi,\alpha}(\mu + \sigma, \nu + \sigma) = \mathcal{OS}_{\Phi,\alpha}(\mu, \nu)$.

ii) $\mathcal{OS}_{\Phi,\alpha}$ *is a divergence,*[9] *and* $\mathcal{OS}_{\Phi,\alpha}(\mu, \nu) \leq \mathcal{OS}_{\Phi,\alpha}(\mu, \sigma) + \mathcal{OS}_{\Phi,\alpha}(\sigma, \nu)$.

iii) With an additional assumption $w_1(z_0) = w_2(z_0)$*, then* $\mathcal{OS}_{\Phi,\alpha}$ *is a metric.*

We next establish connections of OST with other transport distances, including GST [41], Sobolev transport (ST) [39], unbalanced Sobolev transport (UST) [40].

Connection of OST with GST. Denote $\mathcal{GS}_\Phi$ for the GST with N-function Φ.

Proposition 5.2. *For* $\mu(\mathbb{G}) = \nu(\mathbb{G})$*,* $b = 1$*, then* $\mathcal{OS}_{\Phi,\alpha}(\mu, \nu) = \mathcal{GS}_\Phi(\mu, \nu)$.

Connection of OST with ST. Denote $\mathcal{S}_p$ for the p-order ST, for $1 < p < \infty$.

Proposition 5.3. *For* $\mu(\mathbb{G}) = \nu(\mathbb{G})$*,* $b = 1$*, and* $\Phi(t) = \frac{(p-1)^{p-1}}{p^p} t^p$*, then* $\mathcal{OS}_{\Phi,\alpha}(\mu, \nu) = \mathcal{S}_p(\mu, \nu)$.

Connection of OST with UST. Denote $\mathcal{US}_{p,\alpha}$ for UST, for $1 < p < \infty$.

Proposition 5.4. *For* N*-function* $\Phi(t) = \frac{(p-1)^{p-1}}{p^p} t^p$*, then* $\mathcal{OS}_{\Phi,\alpha}(\mu, \nu) = \mathcal{US}_{p,\alpha}(\mu, \nu)$.

Additionally, we investigate the limit case for N-function, i.e., $\Phi(t) = t$,[10] for OST and Orlicz-EPT.

Proposition 5.5 (Limit case for OST)**.** *For* $\Phi(t) = t$*, and with the same assumptions as in Corollary 4.3, then OST yields a closed-form expression as follows:*

$$\mathcal{OS}_{\Phi,\alpha}(\mu, \nu) = b \sum_{e \in E} w_e \, |\mu(\gamma_e) - \nu(\gamma_e)| + \Theta |\mu(\mathbb{G}) - \nu(\mathbb{G})|. \tag{18}$$

Proposition 5.6 (Limit case for Orlicz-EPT)**.** *For* $\Phi(t) = t$*, then we have* $\mathcal{OE}_\Phi(\mu, \nu) = \mathrm{KT}(\mu, \nu)$ *for every* $\mu, \nu \in \mathcal{P}(\mathbb{G})$.

Proposition 5.7 (Relation of OST and Orlicz-EPT)**.** *For* $\Phi(t) = t$*, length measure* ω *on* $\mathbb{G}$*,* b*-Lipschitz* w_1, w_2 *(w.r.t.* $d_\mathbb{G}$*),* $\alpha = 0$*, and* $p = 1$*, then* $\mathcal{OS}_{\Phi,\alpha}(\mu, \nu) \geq \mathcal{OE}_\Phi(\mu, \nu) + \frac{b\lambda}{2}(\mu(\mathbb{G}) + \nu(\mathbb{G}))$.

The proofs for these theoretical results (in §5) are respectively placed in §A.2.7–§A.2.13.

6 Related Works and Discussions

In this section, we discuss relations between our proposals with related works in the literature.

GST [41]. Proposition 5.2 shows that OST provably generalizes GST [41] for unbalanced measures. We emphasize that GST is restricted for balanced measures and is defined as an optimization over the critic function, making it nontrivial to directly extend it to accommodate unbalanced measures.

EPT [36, 40, 72]. Orlicz-EPT and OST are developed from the primal and dual EPT respectively. Notably, the corresponding standard OT following [12] is not guarantee nonnegativity for ground cost, see [12, 36, 40]. Additionally, Caffarelli & McCann [12]'s insights may not be applicable to certain other UOT formulations, such as those proposed in [6, 26, 18, 63, 64, 27, 5, 48, 52]. The calibration is essential to guarantee nonnegativity for ground cost of the corresponding standard OT for EPT, paving ways to develop Orlicz-EPT. Similar to OW, Orlicz-EPT is derived from standard OT, thereby circumventing the challenges associated with unbalanced measures. Furthermore, we derive a novel regularization, resulting in the proposed OST with an efficient computation.

[9] $\mathcal{OS}_{\Phi,\alpha}(\mu, \nu) \geq 0$, and $\mathcal{OS}_{\Phi,\alpha}(\mu, \nu) = 0$ if and only if $\mu = \nu$.

[10] Notice that $\Phi(t) = t$ is not an N-function due to its linear growth. It can be considered as the limit $p \to 1^+$ of the N-function $\Phi(t) = t^p$ with $p > 1$.

UST [40] and ST [39] Proposition 5.4 shows that OST provably generalizes UST to a more general collection of N-functions. Consequently, OST also provably generalizes ST to unbalanced measures, and to a more general set of N-functions (see Proposition 5.3).

Measures on a graph. We study OT problem between *two unbalanced measures* supported on the *same* graph, a setting also explored in [40]. One should distinguish our considered problem with the research lines on computing either distances/discrepancies [55, 74, 73, 22, 11, 47] or kernels [10, 34, 53, 61] between *two (different) input graphs*.

7 Experiments

In this section, we illustrate that the computation of Orlicz-EPT is costly. Especially, OST is several-order faster than Orlicz-EPT. Following the problem setups in [40], we evaluate OST for *unbalanced measures supported a given graph*,[11] and show initial evidences on its advantages for document classification and topological data analysis (TDA).

Document classification. We use 4 real-world document datasets: `TWITTER`, `RECIPE`, `CLASSIC`, and `AMAZON` as in [40], and summarize their characteristics in Figure 2. By regarding each word in a document as a support with a unit mass, we represent each document as a nonnegative measure. Consequently, the representations of documents with different lengths are *measures with different total mass*. We apply the same word embedding procedure in [40] to map words into vectors in $\mathbb{R}^{300}$.

TDA. We consider orbit recognition on `Orbit` dataset [1], and object shape classification on `MPEG7` dataset [35] as in [40]. We summarize these dataset characteristics in Figure 3. We use persistence diagrams (PD), a multiset of 2-dimensional data points summarized topological features, to represent objects of interest. We then consider each data point in PD as a support with a unit-mass, and represent PD as nonnegative measures. As a result, PD having different numbers of topological features are presented as *measures with different total mass*.[12]

Graph. Following [40], we use the graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ [39, §5] for our experiments,[13] which empirically satisfy the assumptions in §2. Additionally, we set $M = 10^4$ for the number of nodes for these graphs, except experiments on `MPEG7` dataset with $M = 10^3$ due to its small size.

N-function. Following [41], we consider two N-functions: $\Phi_1(t) = \exp(t) - t - 1$, and $\Phi_2(t) = \exp(t^2) - 1$, and the limit case of N-functions, i.e., $\Phi_0(t) = t$.

Parameters. For simplicity, we follow the experimental setup in [40]. We set $\lambda = b = 1$, $\alpha = 0$, and 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 $a_0 = 1$. The entropic regularization ε is chosen from $\{0.01, 0.1, 1, 10\}$, typically via cross validation.

Optimization algorithm. For OST, we use `fmincon` MATLAB solver with trust-region-reflective algorithm, for solving the *univariate* optimization problem (17).

SVM classification. For document classification and TDA, we use support vector machine (SVM) with kernels $\exp(-\bar{t}\bar{d}(\cdot,\cdot))$, where $\bar{d}$ is a distance/discrepancy (e.g., OST, Orlicz-EPT) for unbalanced measures on a graph, and $\bar{t} > 0$. We regularize Gram matrices by adding a sufficiently large diagonal term for indefinite kernels [20]. Additionally, we note that there are more than 29M pairs for `AMAZON` which we need to evaluate distances/discrepancies for SVM in each run to illustrate the experiment scale.[14]

Set up. We randomly split each dataset into $70\%/30\%$ for training and test, and use 10 repeats. We basically choose hyper-parameters via cross validation. More concretely, we choose kernel hyperparameter from $\{1/q_s, 0.5/q_s, 0.2/q_s\}$ with $s = 10, 20, \ldots, 90$, where q_s is the $s\%$ quantile of a subset of distances observed on a training set; SVM regularization hyperparameter from $\{0.01, 0.1, 1, 10\}$; root node z_0 from a random 10-root-node subset of V in graph $\mathbb{G}$. Note that reported time consumption includes all preprocessing procedures, e.g., preprocessing for γ_e for OST.

[11] One should distinguish the considered problem, i.e., compare two *input unbalanced measures* supported in the *same* graph, with either OT or Gromov-Wasserstein problem between *two different input graphs* (§6).

[12] We distinguish our problem setup with [41], where objects are represented as *probability measures* instead.

[13] Due to the space limitation, corresponding experimental results for graph $\mathbb{G}_{\text{Log}}$ are placed in §B.3.

[14] See Table 1 in §B.2 for the details.

7.1 Computation

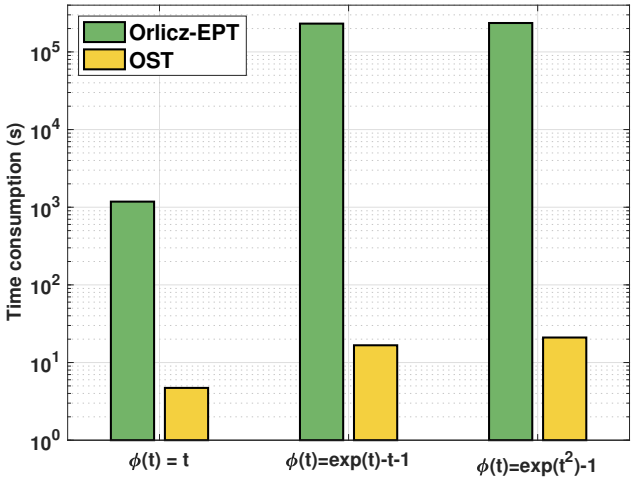

Figure 1: Time consumption.

We compare the time consumption of OST and Orlicz-EPT with Φ_1, Φ_2, and with the limit case Φ_0.

Set up. We randomly sample 10^4 pairs of nonnegative measures on `AMAZON` dataset for evaluation. We consider $M = 10^3$ for graphs, and $\varepsilon = 0.1$ for Orlicz-EPT.

Results. We illustrate the time consumptions on $\mathbb{G}_{\text{Sqrt}}$ in Figure 1. OST is several-order faster than Orlicz-EPT, i.e., at least $250\times, 13800\times, 11200\times$ for Φ_0, Φ_1, Φ_2 respectively. Notably, for N-functions Φ_1, Φ_2, Orlicz-EPT takes at least 2.6 *days*, while OST takes less than 21 *seconds*. Note that for the limit case Φ_0, Orlicz-EPT is equal to EPT on a graph (Proposition 5.6), and OST admits a closed-form expression (Proposition 5.5) for a fast computation. Consequently, Orlicz-EPT and OST with Φ_0 is more computationally efficient than those with Φ_1, Φ_2.

7.2 Document Classification

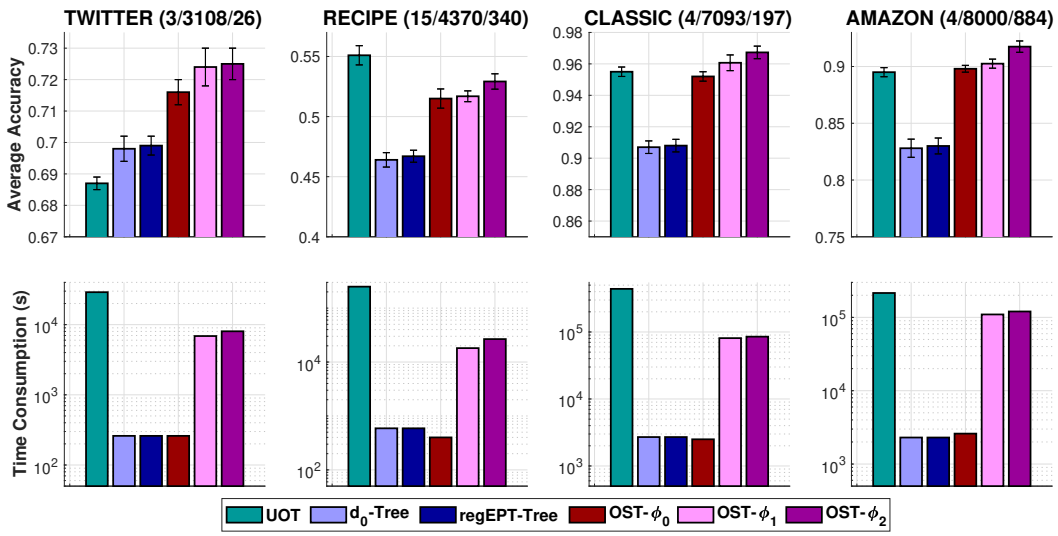

Figure 2: Document classification on graph $\mathbb{G}_{\text{Sqrt}}$. For each dataset, the numbers in the parenthesis are respectively the number of classes; the number of documents; and the maximum number of unique words for each document.

Set up. We evaluate OST with Φ_0, Φ_1, Φ_2 (§7.1), denote them as OST-Φ_i for $i = 0, 1, 2$. We exclude Orlicz-EPT due to their heavy computations (§7.1). Additionally, following [40], we consider

UOT [26, 63] with ground cost $d_{\mathbb{G}}$,[15] and special cases with tree-structure graph. More concretely, we randomly sample a tree from the given graph $\mathbb{G}$, then consider the regularized EPT and d_0, denoted as d_0-Tree and regEPT-Tree [36, Proposition 3.8, Equation (9)].

Results. We show SVM results and time consumptions of kernels on $\mathbb{G}_{\text{Sqrt}}$ in Figure 2. The performances of OST with all Φ functions are comparable to UOT, but the computation of UOT is more costly than OST. Additionally, OST outperforms d_0-Tree and regEPT-Tree. However, the computations of OST-Φ_1, OST-Φ_2 are more expensive while the computation of OST-Φ_0 is comparative to those fast-computational variants of UOT on tree (i.e., d_0-Tree and regEPT-Tree). Moreover, OST-Φ_1 and OST-Φ_2 improve performances of OST-Φ_0, but their computational time is several-order higher, which may imply that Orlicz geometric structure in OST may be helpful for document classification. The performances of UOT also agree with observations in [40].

7.3 Topological Data Analysis (TDA)

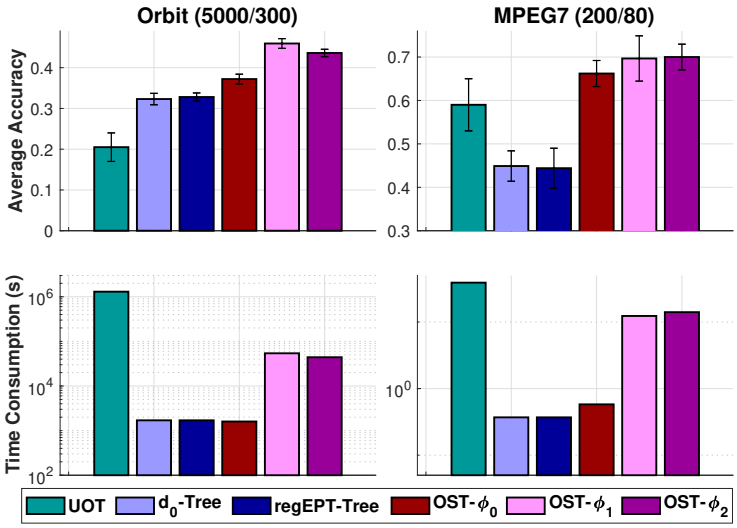

Figure 3: TDA on graph $\mathbb{G}_{\text{Sqrt}}$. For each dataset, the numbers in the parenthesis are respectively the number of PD; and the maximum number of points in PD.

Set up. Similarly, we evaluate OST-Φ_0, OST-Φ_1, OST-Φ_2, UOT, d_0-Tree, and regEPT-Tree for TDA.

Results. We illustrate SVM results and time consumptions of kernels on $\mathbb{G}_{\text{Sqrt}}$ in Figure 3. The performances of OST with all Φ functions compare favorably with other transport distance approaches. Especially, the performances of OST-Φ_1 and OST-Φ_2 compare favorably with those of OST-Φ_0, but it comes with higher computational cost (i.e., OST-Φ_0 has a closed-form expression (Proposition 5.5)), which may imply that Orlicz geometric structure may be also helpful for TDA tasks.

8 Conclusion

In this work, we propose novel approaches to extend OW/GST for unbalanced measures on a graph. Building on the EPT problem and leveraging Caffarelli & McCann [12]'s insights, we derive Orlicz-EPT by recasting it as a standard OT with a carefully calibrated ground cost, thereby bypassing challenges raised from unbalanced measures. Furthermore, by exploiting dual EPT and the underlying geometric structure, we formulate a novel regularization, resulting in the proposed OST, which is efficient in computation. It provably suffices to compute OST by simply solving a univariate optimization problem, unlike the computationally intensive Orlicz-EPT. Moreover, we illustrate empirical evidence on the advantages of OST in document classification and topological data analysis.

[15] Séjourné et al. [63, 65] derived a debiased version for (Sinkhorn-based) UOT which may be helpful in applications [40, §B.3.3]. It has the same computational complexity as UOT, and is also empirically indefinite.

Acknowledgments and Disclosure of Funding

We thank the area chairs and anonymous reviewers for their comments. KF has been supported in part by Grant-in-Aid for Transformative Research Areas (A) 22H05106 and JST CREST JPMJCR2015. HH acknowledges the support of JSPS KAKENHI JP25H01494 and JP23K24909. TL gratefully acknowledges the support of JSPS KAKENHI Grant number 23K11243, and Mitsui Knowledge Industry Co., Ltd. grant.

References

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

[2] Adams, R. A. and Fournier, J. J. *Sobolev spaces*. Elsevier, 2003.

[3] Altschuler, J. M. and Chewi, S. Faster high-accuracy log-concave sampling via algorithmic warm starts. *Journal of the ACM*, 71(3):1–55, 2024.

[4] Andoni, A., Lin, C., Sheng, Y., Zhong, P., and Zhong, R. Subspace embedding and linear regression with Orlicz norm. In *International Conference on Machine Learning*, pp. 224–233. PMLR, 2018.

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

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

[7] Bonet, C., Nadjahi, K., Sejourne, T., Fatras, K., and Courty, N. Slicing unbalanced optimal transport. *Transactions on Machine Learning Research*, 2024.

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

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

[10] Borgwardt, K., Ghisu, E., Llinares-López, F., O'Bray, L., Rieck, B., et al. Graph kernels: State-of-the-art and future challenges. *Foundations and Trends® in Machine Learning*, 13(5-6): 531–712, 2020.

[11] Brogat-Motte, L., Flamary, R., Brouard, C., Rousu, J., and d'Alché Buc, F. Learning to predict graphs with fused gromov-wasserstein barycenters. In *International Conference on Machine Learning*, pp. 2321–2335. PMLR, 2022.

[12] Caffarelli, L. A. and McCann, R. J. Free boundaries in optimal transport and Monge-Ampere obstacle problems. *Annals of mathematics*, pp. 673–730, 2010.

[13] Chamakh, L., Gobet, E., and Szabó, Z. Orlicz random Fourier features. *The Journal of Machine Learning Research*, 21(1):5739–5775, 2020.

[14] Chapel, L. and Tavenard, R. One for all and all for one: Efficient computation of partial Wasserstein distances on the line. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

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

[17] Chewi, S. *An optimization perspective on log-concave sampling and beyond*. PhD thesis, Massachusetts Institute of Technology, 2023.

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

[19] Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 1999.

[20] Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

[21] Deng, Y., Song, Z., Weinstein, O., and Zhang, R. Fast distance oracles for any symmetric norm. *Advances in Neural Information Processing Systems*, 35:7304–7317, 2022.

[22] Dong, Y. and Sawin, W. COPT: Coordinated optimal transport on graphs. *Advances in Neural Information Processing Systems*, 33:19327–19338, 2020.

[23] Edelsbrunner, H. and Harer, J. Persistent homology - A survey. *Contemporary Mathematics*, 453:257–282, 2008.

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

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

[26] Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. Learning with a Wasserstein loss. In *Advances in neural information processing systems*, pp. 2053–2061, 2015.

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

[28] Guha, A., Ho, N., and Nguyen, X. On excess mass behavior in Gaussian mixture models with Orlicz-Wasserstein distances. In *International Conference on Machine Learning, ICML*, volume 202, pp. 11847–11870. PMLR, 2023.

[29] Guittet, K. Extended Kantorovich norms: a tool for optimization. *INRIA report*, 2002.

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

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

[32] Kell, M. On interpolation and curvature via Wasserstein geodesics. *Advances in Calculus of Variations*, 10(2):125–167, 2017.

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

[34] Kriege, N. M., Johansson, F. D., and Morris, C. A survey on graph kernels. *Applied Network Science*, 5:1–42, 2020.

[35] Latecki, L. J., Lakamper, R., and Eckhardt, T. 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, pp. 424–429, 2000.

[36] Le, T. and Nguyen, T. Entropy partial transport with tree metrics: Theory and practice. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3835–3843, 2021.

[37] Le, T., Yamada, M., Fukumizu, K., and Cuturi, M. Tree-sliced variants of Wasserstein distances. In *Advances in neural information processing systems*, pp. 12283–12294, 2019.

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

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

[40] Le, T., Nguyen, T., and Fukumizu, K. Scalable unbalanced Sobolev transport for measures on a graph. In *International Conference on Artificial Intelligence and Statistics*, pp. 8521–8560, 2023.

[41] Le, T., Nguyen, T., and Fukumizu, K. Generalized Sobolev transport for probability measures on a graph. In *Forty-first International Conference on Machine Learning*, 2024.

[42] Le, T., Nguyen, T., and Fukumizu, K. Optimal transport for measures with noisy tree metric. In *International Conference on Artificial Intelligence and Statistics*, pp. 3115–3123, 2024.

[43] Le, T., Nguyen, T., Hino, H., and Fukumizu, K. Scalable Sobolev IPM for probability measures on a graph. In *Forty-second International Conference on Machine Learning*, 2025.

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

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

[46] Lorenz, D. and Mahler, H. Orlicz space regularization of continuous optimal transport problems. *Applied Mathematics & Optimization*, 85(2):14, 2022.

[47] Ma, X., Chu, X., Wang, Y., Lin, Y., Zhao, J., Ma, L., and Zhu, W. Fused Gromov-Wasserstein graph mixup for graph-level classifications. *Advances in Neural Information Processing Systems*, 36:15252–15276, 2023.

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

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

[50] Musielak, J. *Orlicz spaces and modular spaces*, volume 1034. Springer, 2006.

[51] Nguyen, K., Zhang, S., Le, T., and Ho, N. Sliced Wasserstein with random-path projecting directions. In *Forty-first International Conference on Machine Learning*, 2024.

[52] Nguyen, Q. M., Nguyen, H. H., Zhou, Y., and Nguyen, L. M. On unbalanced optimal transport: Gradient methods, sparsity and approximation error. *The Journal of Machine Learning Research*, 2023.

[53] Nikolentzos, G., Siglidis, G., and Vazirgiannis, M. Graph kernels: A survey. *Journal of Artificial Intelligence Research*, 72:943–1027, 2021.

[54] Paty, F.-P. and Cuturi, M. Subspace robust Wasserstein distances. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5072–5081, 2019.

[55] Petric Maretic, H., El Gheche, M., Chierchia, G., and Frossard, P. GOT: An optimal transport framework for graph comparison. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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

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

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

[60] Rao, M. M. and Ren, Z. D. Theory of Orlicz spaces. *Marcel Dekker*, 1991.

[61] Sando, K., Le, T., and Hino, H. Tree structure for the categorical Wasserstein Weisfeiler-Lehman graph kernel. *Transactions on Machine Learning Research (TMLR)*, 2025.

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

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

[64] Séjourné, T., Vialard, F.-X., and Peyré, G. 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, pp. 4995–5021. PMLR, 2022.

[65] Séjourné, T., Peyré, G., and Vialard, F.-X. Unbalanced optimal transport, from theory to numerics. *Handbook of Numerical Analysis*, 24:407–471, 2023.

[66] Song, Z., Wang, R., Yang, L., Zhang, H., and Zhong, P. Efficient symmetric norm regression via linear sketching. *Advances in Neural Information Processing Systems*, 32, 2019.

[67] Sturm, K.-T. Generalized Orlicz spaces and Wasserstein distances for convex–concave scale functions. *Bulletin des sciences mathématiques*, 135(6-7):795–802, 2011.

[68] Tran, T., Tran, V.-H., Chu, T., Pham, T., Ghaoui, L. E., Le, T., and Nguyen, T. Tree-sliced Wasserstein distance with nonlinear projection. In *Forty-second International Conference on Machine Learning*, 2025.

[69] Tran, V.-H., Chu, T., Nguyen, K., Pham, T., Le, T., and Nguyen, T. Spherical tree-sliced Wasserstein distance. In *The Thirteenth International Conference on Learning Representations*, 2025.

[70] Tran, V.-H., Nguyen, K., Pham, T., Chu, T., Le, T., and Nguyen, T. Distance-based tree-sliced Wasserstein distance. In *The Thirteenth International Conference on Learning Representations*, 2025.

[71] Tran, V.-H., Pham, T., Tran, T., Nguyen, K., Chu, T., Le, T., and Nguyen, T. Tree-sliced Wasserstein distance: A geometric perspective. In *Forty-second International Conference on Machine Learning*, 2025.

[72] Tran, V.-H., Tran, T., Chu, T., Le, T., and Nguyen, T. Tree-sliced Entropy Partial Transport. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[73] Xu, H., Luo, D., and Carin, L. Scalable Gromov-Wasserstein learning for graph partitioning and matching. *Advances in neural information processing systems*, 32, 2019.

[74] Xu, H., Luo, D., Zha, H., and Duke, L. C. Gromov-Wasserstein learning for graph matching and node embedding. In *International conference on machine learning*, pp. 6932–6941. PMLR, 2019.

Supplement to "An Efficient Orlicz-Sobolev Approach for Transporting Unbalanced Measures on a Graph"

In this appendix, we provide further theoretical results and detailed proofs in §A. Additionally, we give brief reviews on related notions used in our work, together with further discussions, and empirical results in §B.

A Detailed Proofs and Further Theoretical Results

In this section, we provide further theoretical results, and detailed proofs for all the theoretical results.

A.1 Further Theoretical Results

We investigate special cases for OST, alternative upper limit for $\mathcal{A}_\varepsilon$, and the limit case of N-function for entropic regularized Orlicz-EPT.

A.1.1 Special Cases for OST

We exam the special cases of OST when graph $\mathbb{G}$ is a tree.

Proposition A.1 (Relation of OST and a variant of regularized EPT)**.** *Under the same assumptions as in Proposition 5.5, and assume in addition that graph $\mathbb{G}$ is a tree, then*

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) = d_\alpha(\mu,\nu),$$

where d_α is a variant of the regularized EPT in [36, Equation (9)].

The proof is placed in Appendix §A.2.14.

Proposition A.2 (Relation of OST and standard OT)**.** *Under the same assumptions as in Proposition A.1, and assume in addition that $\mu(\mathbb{G}) = \nu(\mathbb{G})$ and $b = 1$, then*

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) = \mathcal{W}_{d_\mathbb{G}}(\mu,\nu),$$

where $\mathcal{W}_{d_\mathbb{G}}$ is the standard OT with graph metric ground cost $d_\mathbb{G}$.

The proof is placed in Appendix §A.2.15.

A.1.2 Upper Limit of $\mathcal{A}_\varepsilon$ w.r.t. Entropic Regularized OT

With a technical assumption that entropic regularized input is nonnegative for N-function Φ,[16] we derive an alternative upper limit of $\mathcal{A}_\varepsilon$ w.r.t. entropic regularized OT as summarized in the following proposition.

Proposition A.3 (Upper bound w.r.t. entropic regularized OT)**.** *We have*

$$\mathcal{A}_\varepsilon\left(\frac{\mathcal{W}_{\hat{c},\varepsilon}(\hat{\mu},\hat{\nu}) + \frac{\varepsilon}{2}\left(H(\hat{\mu}) + H(\hat{\nu})\right)}{\Phi^{-1}(1+\varepsilon\left[H(\hat{\mu}) + H(\hat{\nu}) - 1\right])}; \hat{\mu}, \hat{\nu}\right) \geq 1,$$

where $\mathcal{W}_{\hat{c},\varepsilon}(\hat{\mu},\hat{\nu})$ is the entropic regularized OT between probability measures $\hat{\mu},\hat{\nu}$ with ground cost $\hat{c}$, defined as

$$\mathcal{W}_{\hat{c},\varepsilon}(\hat{\mu},\hat{\nu}) := \inf_{\tilde{\gamma}\in\Pi(\hat{\mu},\hat{\nu})}\left[\int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \hat{c}(x,y)\tilde{\gamma}(\mathrm{d}x,\mathrm{d}y) - \varepsilon H(\tilde{\gamma})\right]. \tag{19}$$

The proof is placed in Appendix §A.2.16.

A.1.3 Limit Case for Entropic Regularized Orlicz-EPT

We consider the limit case for N-function, i.e., $\Phi(t) = t$, for $\mathcal{OE}_{\Phi,\varepsilon}$, similar to Proposition 5.6 for the original Orlicz-EPT.

[16] The technical assumption is specified in the proof (§A.2.16).

Proposition A.4 (Limit case for entropic regularized Orlicz-EPT)**.** *For $\Phi(t) = t$, and $\mu, \nu \in \mathcal{P}(\mathbb{G})$, we have*

$$\mathcal{OE}_{\Phi,\varepsilon}(\mu,\nu) = (\mu(\mathbb{G}) + \nu(\mathbb{G}))\,(\mathcal{W}_{\hat{c},\varepsilon}(\hat{\mu},\hat{\nu}) - b\lambda), \tag{20}$$

where $\mathcal{W}_{\hat{c},\varepsilon}$ is the entropic regularized optimal transport (see Equation (19)*).*

The proof is placed in Appendix §A.2.17.

A.2 Detailed Proofs

A.2.1 Proof for Proposition 3.1

Proof. Consider the cost function $\tilde{c}$ on $\hat{\mathbb{G}}$ as follows:

$$\tilde{c}(x,y) := \begin{cases} b(d_{\mathbb{G}}(x,y) - \lambda) & \text{if } x,y \in \mathbb{G}, \\ w_1(x) & \text{if } x \in \mathbb{G} \text{ and } y = \hat{s}, \\ w_2(y) & \text{if } x = \hat{s} \text{ and } y \in \mathbb{G}, \\ 0 & \text{if } x = y = \hat{s}. \end{cases} \tag{21}$$

Additionally, for unbalanced measures μ, ν, we construct corresponding balanced measures $\tilde{\mu} := \mu + \nu(\mathbb{G})\delta_{\hat{s}}$ and $\tilde{\nu} := \nu + \mu(\mathbb{G})\delta_{\hat{s}}$ where measures $\tilde{\mu}, \tilde{\nu}$ have the same total mass $(\mu(\mathbb{G})+\nu(\mathbb{G}))$. Let $\widetilde{\Pi}(\tilde{\mu},\tilde{\nu}) := \left\{\tilde{\gamma} \in \mathcal{P}(\hat{G}\times\hat{G}) : \tilde{\mu}(U) = \tilde{\gamma}(U\times\hat{\mathbb{G}}),\ \tilde{\nu}(U) = \tilde{\gamma}(\hat{\mathbb{G}}\times U) \text{ for all Borel sets } U \subset \hat{\mathbb{G}}\right\}$, then following [40, Lemma A.8], we have

$$\mathrm{ET}_\lambda(\mu,\nu) = \mathcal{W}_{\tilde{c}}(\tilde{\mu},\tilde{\nu}), \tag{22}$$

where $\mathcal{W}_{\tilde{c}}(\tilde{\mu},\tilde{\nu})$ is a standard complete OT between two balanced measures $\tilde{\mu}$ and $\tilde{\nu}$ (i.e., having the same total mass $\mu(\mathbb{G}) + \nu(\mathbb{G})$), with cost $\tilde{c}$, defined as

$$\mathcal{W}_{\tilde{c}}(\tilde{\mu},\tilde{\nu}) := \inf_{\tilde{\gamma}\in\widetilde{\Pi}(\tilde{\mu},\tilde{\nu})} \int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \tilde{c}(x,y)\tilde{\gamma}(\mathrm{d}x,\mathrm{d}y).$$

Moreover, from Equation (22), we have

$$\begin{aligned} \mathrm{ET}_\lambda(\mu,\nu) &= \mathcal{W}_{\tilde{c}}(\tilde{\mu},\tilde{\nu}) \\ &= (\mu(\mathbb{G}) + \nu(\mathbb{G}))\,\mathcal{W}_{\tilde{c}}(\hat{\mu},\hat{\nu}) && (23) \\ &= (\mu(\mathbb{G}) + \nu(\mathbb{G}))\,(\mathcal{W}_{\hat{c}}(\hat{\mu},\hat{\nu} - b\lambda) && (24) \\ &= \mathrm{KT}(\mu,\nu), \end{aligned}$$

where the equality in Equation (24) is due to $\tilde{c}(x,y) = \hat{c}(x,y) - b\lambda$ for all $x, y \in \hat{\mathbb{G}}$.

Hence, the proof is completed. ■

A.2.2 Proof for Proposition 3.3

Proof. From the definition, we have

$$\mathcal{A}(t;\hat{\mu},\hat{\nu}) := \inf_{\tilde{\gamma}\in\Pi(\hat{\mu},\hat{\nu})} \int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \Phi\left(\frac{\hat{c}(x,y)}{t}\right) \mathrm{d}\tilde{\gamma}(x,y). \tag{25}$$

Let $0 < t_1 \le t_2 < \infty$, denote $\widetilde{\gamma^*_{t_1}}, \widetilde{\gamma^*_{t_2}}$ as the optimal transport plans of $\mathcal{A}(t_1;\hat{\mu},\hat{\nu}), \mathcal{A}(t_2;\hat{\mu},\hat{\nu})$ respectively. Then, we have

$$\begin{aligned} \mathcal{A}(t_2;\hat{\mu},\hat{\nu}) &= \int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \Phi\left(\frac{\hat{c}(x,y)}{t_2}\right) \mathrm{d}\widetilde{\gamma^*_{t_2}}(x,y) \\ &\le \int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \Phi\left(\frac{\hat{c}(x,y)}{t_2}\right) \mathrm{d}\widetilde{\gamma^*_{t_1}}(x,y) \\ &\le \int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \Phi\left(\frac{\hat{c}(x,y)}{t_1}\right) \mathrm{d}\widetilde{\gamma^*_{t_1}}(x,y) \\ &= \mathcal{A}(t_1;\hat{\mu},\hat{\nu}), \end{aligned}$$

where the second inequality is due to the strictly increasing property of the N-function Φ.

Hence, the proof is completed. ■

A.2.3 Proof for Proposition 3.4

Proof. The result is followed by the same reasoning as in the proof for Proposition 3.3 where we leverage the strictly increasing property of the N-function Φ and the optimal transport plans for $\mathcal{A}_\varepsilon$.

More concretely, let $0 < t_1 \leq t_2 < \infty$, denote $\widetilde{\gamma_{t_1}^*}, \widetilde{\gamma_{t_2}^*}$ as the optimal transport plans of $\mathcal{A}_\varepsilon(t_1; \hat{\mu}, \hat{\nu}), \mathcal{A}_\varepsilon(t_2; \hat{\mu}, \hat{\nu})$ respectively. Then, we have

$$\begin{aligned}
\mathcal{A}_\varepsilon(t_2; \hat{\mu}, \hat{\nu}) &= \int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \Phi\left(\frac{\hat{c}(x,y)}{t_2}\right) \mathrm{d}\widetilde{\gamma_{t_2}^*}(x,y) - \varepsilon H(\widetilde{\gamma_{t_2}^*}) \\
&\leq \int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \Phi\left(\frac{\hat{c}(x,y)}{t_2}\right) \mathrm{d}\widetilde{\gamma_{t_1}^*}(x,y) - \varepsilon H(\widetilde{\gamma_{t_1}^*}) \\
&\leq \int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \Phi\left(\frac{\hat{c}(x,y)}{t_1}\right) \mathrm{d}\widetilde{\gamma_{t_1}^*}(x,y) - \varepsilon H(\widetilde{\gamma_{t_1}^*}) \\
&= \mathcal{A}_\varepsilon(t_1; \hat{\mu}, \hat{\nu}).
\end{aligned}$$

Hence, the proof is completed. ■

A.2.4 Proof for Proposition 3.5

Proof. We provide the proof for the lower and upper limits for $\mathcal{A}_\varepsilon$ as follows:

For lower limit. From the definition in Equation (10), we have

$$\mathcal{A}_\varepsilon\left(\frac{L_{\hat{\mu},\hat{\nu}}}{\Phi^{-1}(1+\varepsilon)}; \hat{\mu}, \hat{\nu}\right) = \inf_{\tilde{\gamma}\in\Pi(\hat{\mu},\hat{\nu})} \left[\int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \Phi\left(\frac{\hat{c}(x,y)}{\frac{L_{\hat{\mu},\hat{\nu}}}{\Phi^{-1}(1+\varepsilon)}}\right) \mathrm{d}\tilde{\gamma}(x,y) - \varepsilon H(\tilde{\gamma})\right].$$

Additionally, since N-function Φ is strictly increasing, we have

$$\Phi\left(\frac{\hat{c}(x,y)}{\frac{L_{\hat{\mu},\hat{\nu}}}{\Phi^{-1}(1+\varepsilon)}}\right) \leq \Phi(\Phi^{-1}(1+\varepsilon)) = 1+\varepsilon.$$

For convenience, given any $\tilde{\gamma} \in \Pi(\hat{\mu}, \hat{\nu})$, we define

$$\bar{\mathcal{H}}(\gamma) := -\int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \log\tilde{\gamma}(x,y)\mathrm{d}\tilde{\gamma}(x,y). \tag{26}$$

From the definition of H in Proposition 3.4, for any $\tilde{\gamma} \in \Pi(\hat{\mu}, \hat{\nu})$, we have

$$H(\tilde{\gamma}) = -\int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \log\tilde{\gamma}(x,y)\mathrm{d}\tilde{\gamma}(x,y) + 1 \geq 1,$$

where the inequality is followed by using [19, Lemma 2.1.1] (i.e., $\bar{\mathcal{H}}(\gamma) \geq 0$).

Thus, we have

$$\mathcal{A}_\varepsilon\left(\frac{L_{\hat{\mu},\hat{\nu}}}{\Phi^{-1}(1+\varepsilon)}; \hat{\mu}, \hat{\nu}\right) \leq (1+\varepsilon) - \varepsilon \leq 1. \tag{27}$$

The proof for the lower limit is completed.

For upper limit. For any $\tilde{\gamma} \in \Pi(\hat{\mu}, \hat{\nu})$, we have

$$\begin{aligned}
\mathcal{T} &:= \int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \Phi\left(\frac{\hat{c}(x,y)}{t}\right) \mathrm{d}\tilde{\gamma}(x,y) - \varepsilon H(\tilde{\gamma}) \\
&\geq \Phi\left(\int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \left(\frac{\hat{c}(x,y)}{t}\right) \mathrm{d}\tilde{\gamma}(x,y)\right) - \varepsilon H(\tilde{\gamma}) \\
&= \Phi\left(\frac{1}{t}\int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \hat{c}(x,y)\mathrm{d}\tilde{\gamma}(x,y)\right) - \varepsilon H(\tilde{\gamma}),
\end{aligned}$$

where we use the Jensen's inequality for the second row.

Additionally, for any $\tilde{\gamma} \in \Pi(\hat{\mu}, \hat{\nu})$, we have

$$\begin{aligned} H(\tilde{\gamma}) &= \bar{\mathcal{H}}(\tilde{\gamma}) + 1 \\ &\leq \bar{\mathcal{H}}(\hat{\mu}) + \bar{\mathcal{H}}(\hat{\nu}) + 1 \\ &= H(\hat{\mu}) + H(\hat{\nu}) - 1, \end{aligned}$$

where we apply [19, Theorem 2.2.1 and Theorem 2.6.5] for the inequality in the second row.

Thus, we have

$$\mathcal{T} \geq \Phi\left(\frac{1}{t}\int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \hat{c}(x,y)\mathrm{d}\tilde{\gamma}(x,y)\right) - \varepsilon\left(H(\hat{\mu}) + H(\hat{\nu}) - 1\right) \tag{28}$$

Taking the infimum of $\tilde{\gamma}$ in $\Pi(\hat{\mu}, \hat{\nu})$, we obtain

$$\mathcal{A}_\varepsilon\left(t; \hat{\mu}, \hat{\nu}\right) \geq \Phi\left(\frac{1}{t}\mathcal{W}_{\hat{c}}(\hat{\mu}, \hat{\nu})\right) - \varepsilon\left(H(\hat{\mu}) + H(\hat{\nu}) - 1\right) \tag{29}$$

Therefore, by choosing $t = \frac{\mathcal{W}_{\hat{c}}(\hat{\mu},\hat{\nu})}{\Phi^{-1}(1+\varepsilon[H(\hat{\mu})+H(\hat{\nu})-1])}$, then we have

$$\mathcal{A}_\varepsilon\left(\frac{\mathcal{W}_{\hat{c}}(\hat{\mu}, \hat{\nu})}{\Phi^{-1}(1 + \varepsilon\left[H(\hat{\mu}) + H(\hat{\nu}) - 1\right])}; \hat{\mu}, \hat{\nu}\right) \geq 1.$$

The proof for the upper limit is completed.

■

A.2.5 Proof for Theorem 4.2

Proof. For $f \in WL_\Phi(\mathbb{G}, \omega)$, as in Equation (63), we have

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

Thus, following the Definition 4.1, we have

$$\begin{aligned} \mathcal{OS}_{\Phi,\alpha}(\mu,\nu) = \sup_{f(z_0)\in\mathcal{I}_\alpha} f(z_0)(\mu(\mathbb{G}) - \nu(\mathbb{G})) + \\ \sup_{f\in WL_\Psi(\mathbb{G},\omega), \|f'\|_{L_\Psi}\leq b} \int_{\mathbb{G}} \left(\int_{[z_0,x]} f'(y)\omega(\mathrm{d}y)\right)(\mu(x) - \nu(x))\,\mathrm{d}x \end{aligned}$$

Thus, we can rewrite $\mathcal{OS}_{\Phi,\alpha}(\mu,\nu)$ as follows:

$$\begin{aligned} \mathcal{OS}_{\Phi,\alpha}(\mu,\nu) &= \sup_{f\in WL_\Psi(\mathbb{G},\omega), \|f'\|_{L_\Psi}\leq b} \int_{\mathbb{G}} \left(\int_{[z_0,x]} f'(y)\omega(\mathrm{d}y)\right)(\mu(x) - \nu(x))\,\mathrm{d}x + \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| \\ &= \sup_{f\in WL_\Psi(\mathbb{G},\omega), \|f'\|_{L_\Psi}\leq b} \int_{\mathbb{G}} \left(\int_{[z_0,x]} f'(y)\omega(\mathrm{d}y)\right)(\mu(x) - \nu(x))\,\mathrm{d}x + \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})|, \end{aligned} \tag{30}$$

where recall that Θ is defined in Equation (15).

Additionally, recall that the indicator function of the shortest path $[z_0, x]$ is as follows:

$$\mathbf{1}_{[z_0,x]}(y) = \begin{cases} 1 & \text{if } y \in [z_0, x] \\ 0 & \text{otherwise.} \end{cases} \tag{31}$$

We rewrite the objective function for the first term of $\mathcal{OS}_{\Phi,\alpha}(\mu,\nu)$ in Equation (30) as follows:

$$\begin{aligned} \int_{\mathbb{G}} \left(\int_{[z_0,x]} f'(y)\omega(\mathrm{d}y)\right)(\mu(x) - \nu(x))\,\mathrm{d}x &= \int_{\mathbb{G}}\int_{\mathbb{G}} \mathbf{1}_{[z_0,x]}(y)\, f'(y)\,(\mu(x) - \nu(x))\,\omega(\mathrm{d}y)\mathrm{d}x \\ &= \int_{\mathbb{G}} \left[\int_{\mathbb{G}} \mathbf{1}_{[z_0,x]}(y)\ (\mu(x) - \nu(x))\,\mathrm{d}x\right] f'(y)\omega(\mathrm{d}y) \qquad (32) \\ &= \int_{\mathbb{G}} \left[\mu(\Lambda(y)) - \nu(\Lambda(y))\right] f'(y)\omega(\mathrm{d}y), \qquad (33) \end{aligned}$$

where we apply the Fubini's theorem to interchange the order of integration for the second row, and use the definition of Λ in Equation (1) for the last row. Consequently, following [60, Proposition 10, pp.81] and notice that $\|bf'\|_{L_\Psi} = b\,\|f'\|_{L_\Psi}$ for $b > 0$, we have

$$\begin{aligned}
&\sup_{f\in WL_\Psi(\mathbb{G},\omega),\|f'\|_{L_\Psi}\le b}\int_{\mathbb{G}}\left(\int_{[z_0,x]} f'(y)\omega(\mathrm{d}y)\right)(\mu(x)-\nu(x))\,\mathrm{d}x \\
&\qquad = \sup_{f\in WL_\Psi(\mathbb{G},\omega),\left\|\frac{1}{b}f'\right\|_{L_\Psi}\le 1}\int_{\mathbb{G}} b\left[\mu(\Lambda(y))-\nu(\Lambda(y))\right]\left[\frac{1}{b}f'(y)\right]\omega(\mathrm{d}y) \\
&\qquad = \left\|\tilde{f}\right\|_{\Phi}, \qquad (34)
\end{aligned}$$

where $\tilde{f}(x) := b\left(\mu(\Lambda(x))-\nu(\Lambda(x))\right), \forall x\in\mathbb{G}$, and we write $\left\|\tilde{f}\right\|_{\Phi}$ for the Orlicz norm of $\tilde{f}$ with N-function Φ [60, Definition 2, pp.58] (i.e., see a review in Equation (58) in §B.1.3).

Moreover, following [60, Theorem 13, pp.69], we also have

$$\|\tilde{f}\|_{\Phi} = \inf_{k>0}\frac{1}{k}\left(1+\int_{\mathbb{G}}\Phi\left(k\left|\tilde{f}(x)\right|\right)\omega(\mathrm{d}x)\right). \qquad (35)$$

Hence, putting these Equations (30), (34), (35) together, we have

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) = \inf_{k>0}\frac{1}{k}\left(1+\int_{\mathbb{G}}\Phi\left(kb\left|\mu(\Lambda(x))-\nu(\Lambda(x))\right|\right)\omega(\mathrm{d}x)\right)+\Theta|\mu(\mathbb{G})-\nu(\mathbb{G})|. \qquad (36)$$

The proof is completed.

■

A.2.6 Proof for Corollary 4.3

Proof. Following Theorem 4.2, we have

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) = \Theta|\mu(\mathbb{G})-\nu(\mathbb{G})| + \inf_{k>0}\frac{1}{k}\left(1+\int_{\mathbb{G}}\Phi\left(kb\left|\mu(\Lambda(x))-\nu(\Lambda(x))\right|\right)\omega(\mathrm{d}x)\right), \qquad (37)$$

We next follow the same reasoning as in [41, Corollary 3.4] to compute the integral in (37) by an explicit expression.

For an edge e between two nodes $u, v \in V$ of graph $\mathbb{G}$, then u, v are also two data points in $\mathbb{R}^n$ as $\mathbb{G}$ is a physical graph. For convenience, denote $\langle u, v\rangle$ as the line segment in $\mathbb{R}^n$ connecting the two data points u, v, and (u, v) as the same line segment but without its two end-points. Therefore, we have $e = \langle u, v\rangle$.

Additionally, for any $x \in (u, v)$, we have $y \in \mathbb{G}\setminus(u, v)$ belongs to $\Lambda(x)$ if and only if $y \in \gamma_e$ (see Equation (1) for the definitions of $\Lambda(x)$ and γ_e). Thus, we have

$$\Lambda(x)\setminus(u, v) = \gamma_e. \qquad (38)$$

Consider the case where ω is the length measure of graph $\mathbb{G}$, we have $\omega(\{x\}) = 0$ for every $x \in \mathbb{G}$. Consequently,

$$\int_{\mathbb{G}}\Phi\left(kb\left|\mu(\Lambda(x))-\nu(\Lambda(x))\right|\right)\omega(\mathrm{d}x) = \sum_{e=\langle u,v\rangle\in E}\int_{(u,v)}\Phi\left(kb\left|\mu(\Lambda(x))-\nu(\Lambda(x))\right|\right)\omega(\mathrm{d}x). \qquad (39)$$

Additionally, for measures μ, ν supported on nodes V of $\mathbb{G}$, and using Equation (38), then we have

$$|\mu(\Lambda(x))-\nu(\Lambda(x))| = |\mu(\Lambda(x)\setminus(u, v))-\nu(\Lambda(x)\setminus(u, v))| = |\mu(\gamma_e)-\nu(\gamma_e)|,$$

for every edge $e = \langle u, v\rangle \in E$ of graph $\mathbb{G}$.

Therefore, we can rewrite the identity (39) as follows:

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

By combining it with (37), we obtain

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu)=\Theta|\mu(\mathbb{G})-\nu(\mathbb{G})|+\sum_{e\in E} w_e\,\Phi\left(kb\left|\mu(\gamma_e)-\nu(\gamma_e)\right|\right).$$

Hence, the proof is completed. ■

A.2.7 Proof for Proposition 5.1

Proof. The proof for each property on geometric structure of OST is as follows:

i) The result is directly followed from Equation (14) in Theorem 4.2 with the observation that

$$|\mu(\mathbb{G})-\nu(\mathbb{G})|=|(\mu+\sigma)(\mathbb{G})-(\nu+\sigma)(\mathbb{G})|$$

and

$$|\mu(\Lambda(x))-\nu(\Lambda(x))|=|(\mu+\sigma)(\Lambda(x))-(\nu+\sigma)(\Lambda(x))|.$$

ii) From Definition 4.1, choosing $f=0$, then $f\in\mathbb{U}_{\Psi,\alpha}$, and for any $\mu,\nu\in\mathcal{P}(\mathbb{G})$, we have that

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu)\geq 0.$$

Assume that $\mathcal{OS}_{\Phi,\alpha}(\mu,\nu)(\mu,\nu)=0$. Then, from Theorem 4.2, we obtain

$$\Theta|\mu(\mathbb{G})-\nu(\mathbb{G})|+\inf_{k>0}\frac{1}{k}\left(1+\int_{\mathbb{G}}\Phi\left(kb\left|\mu(\Lambda(x))-\nu(\Lambda(x))\right|\right)\omega(\mathrm{d}x)\right)=0.$$

Additionally, for $0\leq\alpha<\frac{b\lambda}{2}+\min\{w_1(z_0),w_2(z_0)\}$, we have $\Theta>0$. Consequently, we must have

$$\mu(\mathbb{G})=\nu(\mathbb{G})\quad\text{and}\quad\inf_{k>0}\frac{1}{k}\left(1+\int_{\mathbb{G}}\Phi\left(kb\left|\mu(\Lambda(x))-\nu(\Lambda(x))\right|\right)\omega(\mathrm{d}x)\right)=0.$$

Thus, $\mu(\Lambda(x))=\nu(\Lambda(x)),\forall x\in\mathbb{G}$.

By applying [40, Lemma A.9],[17] it leads to $\mu=\nu$.

Moreover, from Definition 4.1, we also have $\mathcal{OS}_{\Phi,\alpha}(\mu,\mu)=0$.

Furthermore, for any feasible function $f\in\mathbb{U}_{\Psi,\alpha}$, we have

$$\begin{aligned}\int_{\mathbb{G}} f(x)\mu(\mathrm{d}x)-\int_{\mathbb{G}} f(x)\nu(\mathrm{d}x) &= \Big[\int_{\mathbb{G}} f(x)\mu(\mathrm{d}x)-\int_{\mathbb{G}} f(x)\sigma(\mathrm{d}x)\Big]+\\ &\qquad\Big[\int_{\mathbb{G}} f(x)\sigma(\mathrm{d}x)-\int_{\mathbb{G}} f(x)\nu(\mathrm{d}x)\Big]\\ &\leq \mathcal{OS}_{\Phi,\alpha}(\mu,\sigma)+\mathcal{OS}_{\Phi,\alpha}(\sigma,\nu).\end{aligned}$$

Therefore, by taking the infimum for $f\in\mathbb{U}_{\Psi,\alpha}$, it implies that

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu)\leq\mathcal{OS}_{\Phi,\alpha}(\mu,\sigma)+\mathcal{OS}_{\Phi,\alpha}(\sigma,\nu).$$

Hence, $\mathcal{OS}_{\Phi,\alpha}$ satisfies the triangle inequality.

iii) With an additional assumption $w_1(z_0)=w_2(z_0)$, then for any function $f\in\mathbb{U}_{\Psi,\alpha}$, we also have $(-f)\in\mathbb{U}_{\Psi,\alpha}$.

Therefore, from Definition 4.1, we obtain $\mathcal{OS}_{\Phi,\alpha}(\mu,\nu)=\mathcal{OS}_{\Phi,\alpha}(\nu,\mu)$.

Thus, together with results in ii), we have $\mathcal{OS}_{\Phi,\alpha}$ is a metric.

■

[17] In §B.1.6 (Lemma B.3), we review the Lemma A.9 in Le et al. [40].

A.2.8 Proof for Proposition 5.2

Proof. For $\mu(\mathbb{G}) = \nu(\mathbb{G})$ and $b = 1$, then following Theorem 4.2 for OST and [41, Theorem 3.3] for GST, we have

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) = \inf_{k>0} \frac{1}{k}\left(1 + \int_{\mathbb{G}} \Phi\left(k\left|\mu(\Lambda(x)) - \nu(\Lambda(x))\right|\right)\omega(\mathrm{d}x)\right) = \mathcal{GS}_{\Phi}(\mu,\nu).$$

■

The proof is completed.

A.2.9 Proof for Proposition 5.3

Proof. For $\mu(\mathbb{G}) = \nu(\mathbb{G})$, $b = 1$, by applying Proposition 5.2, we have

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) = \mathcal{GS}_{\Phi}(\mu,\nu), \tag{40}$$

where we recall that $\mathcal{GS}_\Phi$ is the GST for balanced measures on a graph.

Additionally, for $1 < p < \infty$ and N-function $\Phi(t) = \frac{(p-1)^{p-1}}{p^p}t^p$, by leveraging [41, Proposition 4.4] for the connection between GST and ST, we have

$$\mathcal{GS}_{\Phi}(\mu,\nu) = \mathcal{S}_p(\mu,\nu). \tag{41}$$

Therefore, by combining Equations (40) and (41), we obtain

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) = \mathcal{S}_p(\mu,\nu).$$

The proof is completed. ■

A.2.10 Proof for Proposition 5.4

Proof. For N-function $\Phi(t) = \frac{(p-1)^{p-1}}{p^p}t^p$ with $1 < p < \infty$, from Theorem 4.2, we have

$$\begin{aligned}\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) = \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| + \\ \inf_{k>0} \tfrac{1}{k}\left(1 + \int_{\mathbb{G}} \tfrac{(p-1)^{p-1}}{p^p}k^p b^p \left|\mu(\Lambda(x)) - \nu(\Lambda(x))\right|^p \omega(\mathrm{d}x)\right).\end{aligned} \tag{42}$$

For convenience, for $k > 0$, let

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

i.e., the objective function of the univariate optimization problem for $\mathcal{OS}_{\Phi,\alpha}$.

We next consider two cases:

Case 1: $\int_{\mathbb{G}} \left|\mu(\Lambda(x)) - \nu(\Lambda(x))\right|^p \omega(\mathrm{d}x) = 0$. Then, we have

$$\inf_{k>0} T(k) = \inf_{k>0} \frac{1}{k} = 0.$$

Consequently, from Equation (42), we have

$$\begin{aligned}\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) &= \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| = b\left[\int_{\mathbb{G}} \left|\mu(\Lambda(x)) - \nu(\Lambda(x))\right|^p \omega(\mathrm{d}x)\right]^{\frac{1}{p}} + \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| \\ &= \mathcal{US}_{p,\alpha}(\mu,\nu).\end{aligned}$$

Case 2: $\int_{\mathbb{G}} |h(x)|^p \omega(\mathrm{d}x) \neq 0$. Then, we have

$$\lim_{k\to 0^+} T(k) = \lim_{k\to+\infty} T(k) = +\infty.$$

Therefore, Equation (42), we have

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) = \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| + T(k_0), \tag{43}$$

for some finite number $k_0 \in (0, +\infty)$ satisfying $T'(k_0) = 0$.

Additionally, we have

$$T'(k) = -\frac{1}{k^2} + \left(\frac{p-1}{p}\right)^p k^{p-2} b^p \int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p \, \omega(\mathrm{d}x).$$

Consequently, by solving the equation $T'(k_0) = 0$ w.r.t. k_0, we obtain

$$k_0 = \frac{1}{\frac{p-1}{p} b \left(\int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p \, \omega(\mathrm{d}x)\right)^{\frac{1}{p}}}.$$

Therefore, by plugging this value of k_0 into T, we have

$$\begin{aligned}
T(k_0) &= \frac{1}{k_0}\left(1 + \frac{(p-1)^{p-1}}{p^p} k_0^p b^p \int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p \, \omega(\mathrm{d}x)\right) \\
&= \frac{p-1}{p} b \left(\int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p \, \omega(\mathrm{d}x)\right)^{\frac{1}{p}} \times \\
&\left(1 + \frac{(p-1)^{p-1}}{p^p} \frac{1}{\frac{(p-1)^p}{p^p} b^p \left(\int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p \, \omega(\mathrm{d}x)\right)} b^p \int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p \, \omega(\mathrm{d}x)\right) \\
&= b \left(\int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p \, \omega(\mathrm{d}x)\right)^{\frac{1}{p}}.
\end{aligned}$$

Thus, by plugging this value of $T(k_0)$ into Equation (43), we obtain

$$\begin{aligned}
\mathcal{OS}_{\Phi,\alpha}(\mu, \nu) &= \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| + b \left(\int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p \, \omega(\mathrm{d}x)\right)^{\frac{1}{p}} \\
&= \mathcal{US}_{p,\alpha}(\mu, \nu).
\end{aligned}$$

Hence, we have shown that $\mathcal{OS}_{\Phi,\alpha}(\mu, \nu) = \mathcal{US}_{p,\alpha}(\mu, \nu)$ in both cases.

The proof is completed. ■

A.2.11 Proof for Proposition 5.5

Proof. Following Corollary 4.3, we have

$$\mathcal{OS}_{\Phi,\alpha}(\mu, \nu) = \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| + \inf_{k>0} \frac{1}{k} \left(1 + \sum_{e \in E} w_e \Phi(kb \, |\mu(\gamma_e) - \nu(\gamma_e)|)\right).$$

For $\Phi(t) = t$, then we have

$$\begin{aligned}
\mathcal{OS}_{\Phi,\alpha}(\mu, \nu) &= \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| + \inf_{k>0} \frac{1}{k} \left(1 + \sum_{e \in E} w_e kb \, |\mu(\gamma_e) - \nu(\gamma_e)|\right) \\
&= \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| + \inf_{k>0} \frac{1}{k} + \sum_{e \in E} w_e b \, |\mu(\gamma_e) - \nu(\gamma_e)| \\
&= \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| + b \sum_{e \in E} w_e \, |\mu(\gamma_e) - \nu(\gamma_e)|.
\end{aligned}$$

Hence, the proof is completed. ■

A.2.12 Proof for Proposition 5.6

Proof. From Equation (8), we have

$$\mathcal{OE}_{\Phi}(\mu,\nu) = (\mu(\mathbb{G}) + \nu(\mathbb{G}))\,(\mathcal{W}_{\Phi}(\hat{\mu},\hat{\nu}) - b\lambda).$$

For $\Phi(t) = t$, we further have

$$\mathcal{W}_{\Phi}(\hat{\mu},\hat{\nu}) = \inf_{\tilde{\gamma}\in\Pi(\hat{\mu},\hat{\nu})} \inf\left[t > 0 : \int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \left(\frac{\hat{c}(x,y)}{t}\right) \mathrm{d}\tilde{\gamma}(x,y) \leq 1\right]$$

Then, the infimum $(t^*, \tilde{\gamma}^*)$ satisfies

$$\int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \left(\frac{\hat{c}(x,y)}{t^*}\right) \mathrm{d}\tilde{\gamma}^*(x,y) = 1.$$

Therefore, we obtain $t^* = \int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \hat{c}(x,y)\mathrm{d}\tilde{\gamma}^*(x,y) = \mathcal{W}_{\hat{c}}(\hat{\mu},\hat{\nu})$.

Hence, we have

$$\begin{aligned}\mathcal{OE}_{\Phi}(\mu,\nu) &= (\mu(\mathbb{G}) + \nu(\mathbb{G}))\,(\mathcal{W}_{\hat{c}}(\hat{\mu},\hat{\nu}) - b\lambda)\\ &= \mathrm{KT}(\mu,\nu).\end{aligned}$$

The proof is completed. ∎

A.2.13 Proof for Proposition 5.7

Proof. For $\Phi(t) = t$, $p = 1$, from Theorem 4.2, we have

$$\begin{aligned}\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) &= \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| + \inf_{k>0} \frac{1}{k}\left(1 + \int_{\mathbb{G}} kb\,|\mu(\Lambda(x)) - \nu(\Lambda(x))|\,\omega(\mathrm{d}x)\right)\\ &= \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| + \inf_{k>0} \frac{1}{k} + b\int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|\,\omega(\mathrm{d}x)\\ &= \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| + b\int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|\,\omega(\mathrm{d}x)\\ &= \mathcal{US}_{p,\alpha}(\mu,\nu).\end{aligned}$$

Additionally, for $\Phi(t) = t$, from Proposition 5.6, we have

$$\mathcal{OE}_{\Phi}(\mu,\nu) = \mathrm{KT}(\mu,\nu).$$

With additional assumptions that $\lambda \geq 0$ and the nonnegative weight functions w_1, w_2 are b-Lipschitz w.r.t. $d_{\mathbb{G}}$, then by applying [40, Corollary 3.2], we have

$$\mathcal{OE}_{\Phi}(\mu,\nu) = \mathrm{KT}(\mu,\nu) = \mathrm{ET}_{\lambda}(\mu,\nu).$$

Consequently, for $\alpha = 0$, and the length measure ω on $\mathbb{G}$, then following [40, Proposition 5.2], we have

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) \geq \mathcal{OE}_{\Phi}(\mu,\nu) + \frac{b\lambda}{2}(\mu(\mathbb{G}) + \nu(\mathbb{G})).$$

The proof is completed. ∎

A.2.14 Proof for Proposition A.1

Proof. From Proposition 5.5, we have

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) = \Theta|\mu(\mathbb{G}) - \nu(\mathbb{G})| + b\sum_{e\in E} w_e\,|\mu(\gamma_e) - \nu(\gamma_e)| \tag{44}$$

$$= \mathcal{US}_{1,\alpha}(\mu,\nu). \tag{45}$$

For the case when $\mathbb{G}$ is a tree, then following [40, Proposition 5.3 i)], we further have

$$\mathcal{US}_{1,\alpha}(\mu,\nu) = d_{\alpha}(\mu,\nu). \tag{46}$$

Thus, from Equations (44) and (46), we have

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) = d_{\alpha}(\mu,\nu).$$

The proof is completed. ∎

A.2.15 Proof for Proposition A.2

Proof. From Equation (44) in the proof of Proposition A.1, we have

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) = \mathcal{US}_{1,\alpha}(\mu,\nu). \tag{47}$$

Additionally, when $\mathbb{G}$ is a tree, and with an additional assumption that $\mu(\mathbb{G}) = \nu(\mathbb{G})$, by applying [40, Proposition 5.3 ii)], and notice that $p = 1$ and $b = 1$, we obtain

$$\mathcal{US}_{1,\alpha}(\mu,\nu) = \mathcal{W}_{d_{\mathbb{G}}}(\mu,\nu), \tag{48}$$

where recall that $\mathcal{W}_{d_{\mathbb{G}}}$ is the standard optimal transport with graph metric ground cost $d_{\mathbb{G}}$.

Hence, from Equations (47) and (48), we get

$$\mathcal{OS}_{\Phi,\alpha}(\mu,\nu) = \mathcal{W}_{d_{\mathbb{G}}}(\mu,\nu).$$

The proof is completed.

■

A.2.16 Proof for Proposition A.3

Proof. Following [19, Theorem 2.2.1] and definition of conditional entropy [19, Equation 2.10], for any $\tilde{\gamma} \in \Pi(\hat{\mu},\hat{\nu})$, we have

$$\bar{\mathcal{H}}(\tilde{\gamma}) \geq \frac{1}{2}(\bar{\mathcal{H}}(\hat{\mu}) + \bar{\mathcal{H}}(\hat{\nu})) \tag{49}$$

$$\bar{\mathcal{H}}(\tilde{\gamma}) + 1 \geq \frac{1}{2}(\bar{\mathcal{H}}(\hat{\mu}) + \bar{\mathcal{H}}(\hat{\nu})) + 1 \tag{50}$$

$$H(\tilde{\gamma}) \geq \frac{1}{2}(H(\hat{\mu}) + H(\hat{\nu})), \tag{51}$$

where we recall that $\bar{\mathcal{H}}$ and H are defined in Equation (26) and in Proposition 3.4 respectively.

Therefore, as in the proof for Proposition 3.5 in §A.2.4, from Equation (28), we have

$$\mathcal{T} \geq \Phi\left(\frac{1}{t}\int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \hat{c}(x,y)\mathrm{d}\tilde{\gamma}(x,y)\right) - \varepsilon\left(H(\hat{\mu}) + H(\hat{\nu}) - 1\right) \tag{52}$$

$$\geq \Phi\left(\frac{1}{t}\left[\int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \hat{c}(x,y)\mathrm{d}\tilde{\gamma}(x,y) - \varepsilon H(\tilde{\gamma}) + \frac{\varepsilon}{2}\left(H(\hat{\mu}) + H(\hat{\nu})\right)\right]\right) - \varepsilon\left(H(\hat{\mu}) + H(\hat{\nu}) - 1\right) \tag{53}$$

where we assume that the entropic regularized input of N-function Φ is nonnegative in the second row (Equation (53)), i.e.,

$$\int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \hat{c}(x,y)\mathrm{d}\tilde{\gamma}(x,y) - \varepsilon H(\tilde{\gamma}) + \frac{\varepsilon}{2}\left(H(\hat{\mu}) + H(\hat{\nu})\right) \geq 0,$$

for any $\tilde{\gamma} \in \Pi(\hat{\mu},\hat{\nu})$.

Taking the infimum of $\tilde{\gamma}$ in $\Pi(\hat{\mu},\hat{\nu})$, we obtain

$$\mathcal{A}_\varepsilon\left(t;\hat{\mu},\hat{\nu}\right) \geq \Phi\left(\frac{1}{t}\left[\mathcal{W}_\varepsilon(\hat{\mu},\hat{\nu}) + \frac{\varepsilon}{2}\left(H(\hat{\mu}) + H(\hat{\nu})\right)\right]\right) - \varepsilon\left(H(\hat{\mu}) + H(\hat{\nu}) - 1\right) \tag{54}$$

Therefore, by choosing $t = \frac{\mathcal{W}_\varepsilon(\hat{\mu},\hat{\nu}) + \frac{\varepsilon}{2}(H(\hat{\mu})+H(\hat{\nu}))}{\Phi^{-1}(1+\varepsilon[H(\hat{\mu})+H(\hat{\nu})-1])}$, then we have

$$\mathcal{A}_\varepsilon\left(\frac{\mathcal{W}_\varepsilon(\hat{\mu},\hat{\nu}) + \frac{\varepsilon}{2}\left(H(\hat{\mu}) + H(\hat{\nu})\right)}{\Phi^{-1}(1+\varepsilon\left[H(\hat{\mu}) + H(\hat{\nu}) - 1\right])}; \hat{\mu},\hat{\nu}\right) \geq 1.$$

The proof is completed. ■

A.2.17 Proof for Proposition A.4

Proof. We use the same reason as in the proof for Proposition 5.6. From Equation (11), we have

$$\mathcal{OE}_{\Phi,\varepsilon}(\mu,\nu) := (\mu(\mathbb{G}) + \nu(\mathbb{G}))\,(\mathcal{W}_{\Phi,\varepsilon}(\hat{\mu},\hat{\nu}) - b\lambda).$$

For $\Phi(t) = t$, we further have

$$\mathcal{W}_{\Phi,\varepsilon}(\hat{\mu},\hat{\nu}) = \inf_{\tilde{\gamma}\in\Pi(\hat{\mu},\hat{\nu})} \inf\left[t > 0 : \int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \left(\frac{\hat{c}(x,y)}{t}\right) \mathrm{d}\tilde{\gamma}(x,y) - \varepsilon H(\tilde{\gamma}) \leq 1\right]$$

Then, let $\tilde{\gamma}_\varepsilon^*$ is the optimal solution for the entropic regularized OT

$$\mathcal{W}_{\hat{c},\varepsilon}(\hat{\mu},\hat{\nu}) = \inf_{\tilde{\gamma}\in\Pi(\hat{\mu},\hat{\nu})} \left[\int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \hat{c}(x,y)\tilde{\gamma}(\mathrm{d}x,\mathrm{d}y) - \varepsilon H(\tilde{\gamma})\right].$$

Thus, for the infimum $(t^*, \tilde{\gamma}_\varepsilon^*)$, we have

$$\int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \left(\frac{\hat{c}(x,y)}{t^*}\right) \mathrm{d}\tilde{\gamma}_\varepsilon^*(x,y) = 1.$$

Therefore, we obtain

$$t^* = \int_{\hat{\mathbb{G}}\times\hat{\mathbb{G}}} \hat{c}(x,y)\mathrm{d}\tilde{\gamma}_\varepsilon^*(x,y) = \mathcal{W}_{\hat{c},\varepsilon}(\hat{\mu},\hat{\nu}).$$

Hence, we have

$$\mathcal{OE}_{\Phi,\varepsilon}(\mu,\nu) = (\mu(\mathbb{G}) + \nu(\mathbb{G}))\,(\mathcal{W}_{\hat{c},\varepsilon}(\hat{\mu},\hat{\nu}) - b\lambda).$$

The proof is completed. ■

B Brief Reviews, Further Discussions and Empirical Results

In this section, we give brief reviews on important related notions to our proposed approaches. We next give further discussions on several aspects, and provide further empirical results.

B.1 Brief Reviews

We provide brief reviews on important related notions to our proposed approaches.

B.1.1 Sobolev Transport (ST)

We briefly review main notions for Sobolev transport (ST) [39] for probability measures on a graph.

***L^p* functional space.** For a nonnegative Borel measure ω on $\mathbb{G}$, denote $L^p(\mathbb{G},\omega)$ as the space of all Borel measurable functions $f : \mathbb{G} \to \mathbb{R}$ such that $\int_{\mathbb{G}} |f(y)|^p \omega(\mathrm{d}y) < \infty$. For $p = \infty$, we instead assume that f is bounded ω-a.e. Then, $L^p(\mathbb{G},\omega)$ is a normed space with the norm defined by

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

and for $p = \infty$,

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

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. $x \in \mathbb{G}$.

Graph-based Sobolev space [39]. 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 said to belong to the Sobolev space $W^{1,p}(\mathbb{G},\omega)$ if there exists a function $h \in L^p(\mathbb{G},\omega)$ satisfying

$$f(x) - f(z_0) = \int_{[z_0,x]} h(y)\omega(\mathrm{d}y), \quad \forall x \in \mathbb{G}. \tag{55}$$

Such function h is unique in $L^p(\mathbb{G},\omega)$ and is called the generalized graph derivative of f w.r.t. the measure ω. The generalized graph derivative of $f \in W^{1,p}(\mathbb{G},\omega)$ is denoted $f' \in L^p(\mathbb{G},\omega)$.

Sobolev transport [39]. Let ω be a nonnegative Borel measure on $\mathbb{G}$. Given $1 \le p \le \infty$, and let p' be its conjugate, i.e., the number $p' \in [1, \infty]$ satisfying $\frac{1}{p} + \frac{1}{p'} = 1$. For probability measures μ, ν supported on graph $\mathbb{G}$, the p-order Sobolev transport (ST) [39, Definition 3.2] is defined as

$$\mathcal{S}_p(\mu, \nu) \coloneqq \begin{cases} \sup \left[\int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(x)\nu(\mathrm{d}x) \right] \\ \text{s.t. } f \in W^{1,p'}(\mathbb{G}, \omega),\ \|f'\|_{L^{p'}(\mathbb{G},\omega)} \le 1, \end{cases} \tag{56}$$

where we write f' for the generalized graph derivative of f, $W^{1,p'}(\mathbb{G}, \omega)$ for the graph-based Sobolev space on $\mathbb{G}$, and $L^{p'}(\mathbb{G}, \omega)$ for the L^p functional space on $\mathbb{G}$.

B.1.2 Length measure

We briefly review the length measure on graph $\mathbb{G}$ in [39].

Definition B.1 (Length measure [39])**.** 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 $s \le t$, where recall that $\langle x, y \rangle$ denotes the line segment in e connecting x and y.

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

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

Lemma B.2 (ω^* is the length measure on graph [39])**.** *Suppose that $\mathbb{G}$ has no short cuts, i.e., 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, y. Particularly, ω^* has no atom in the sense that $\omega^*(\{x\}) = 0$ for every $x \in \mathbb{G}$.

B.1.3 Orlicz functions

We describe a brief review on Orlicz functions as summarized in [41] for completeness. Please see [2, 60], for in-depth studies on Orlicz functions.

A family of convex functions. We consider the collection of N-functions [2, §8.2] which are special convex functions on $\mathbb{R}_+$. Hereafter, a strictly increasing and convex function $\Phi : [0, \infty) \to [0, \infty)$ is called an N-function if $\lim_{t \to 0} \frac{\Phi(t)}{t} = 0$ and $\lim_{t \to +\infty} \frac{\Phi(t)}{t} = +\infty$.

Examples of N-functions. Some popular examples for N-functions [2, §8.2] are

1. $\Phi(t) = t^p$ with $1 < p < \infty$.
2. $\Phi(t) = \exp(t) - t - 1$.
3. $\Phi(t) = \exp(t^p) - 1$ with $1 < p < \infty$.
4. $\Phi(t) = (1 + t)\log(1 + t) - t$.

For Luxemburg norm. For Luxemburg norm (see Equation (2)) for Orlicz functional space, the infimum in its definition is attained [2, §8.9].

Complementary function. For N-function Φ, its complementary function $\Psi : \mathbb{R}_+ \to \mathbb{R}_+$ [2, §8.3] is the N-function, defined as follows

$$\Psi(t) = \sup \left[at - \Phi(a) \mid a \ge 0 \right], \quad \text{for } t \ge 0. \tag{57}$$

Examples of complementary pairs of N-functions. Some popular complementary pairs of N-functions [2, §8.3], [60, §2.2] are as follows:

1. $\Phi(t) = \frac{t^p}{p}$ and $\Psi(t) = \frac{t^q}{q}$ where q is the conjugate of p, i.e., $\frac{1}{p} + \frac{1}{q} = 1$ and $1 < p < \infty$.
2. $\Phi(t) = \exp(t) - t - 1$ and $\Psi(t) = (1+t)\log(1+t) - t$.
3. For the N-function $\Phi(t) = \exp(t^p) - 1$ with $1 < p < \infty$, its complementary N-function yields an explicit for, but not simple [60, §2.2], see [41, §A.8] for the detailed derivation of the complementary N-function.

Young inequality. Let Φ, Ψ be a pair of complementary N-functions, then we have

$$st \leq \Psi(s) + \Phi(t).$$

Orlicz norm. Besides the Luxemburg norm, the Orlicz norm [60, §3.3, Definition 2] is also a popular norm for $L_\Phi(\mathbb{G}, \omega)$, defined as

$$\|f\|_\Phi := \sup\Big\{ \int_{\mathbb{G}} |f(x)g(x)|\omega(\mathrm{d}x) \;\Big|\; \int_{\mathbb{G}} \Psi(|g(x)|)\omega(\mathrm{d}x) \leq 1 \Big\}, \tag{58}$$

where Ψ is the complementary N-function of Φ.

Computation for Orlicz norm. By applying [60, §3.3, Theorem 13], we can rewrite the Orlicz norm as follows:

$$\|f\|_\Phi = \inf_{k>0} \frac{1}{k} \left(1 + \int_{\mathbb{G}} \Phi(k\,|f(x)|)\omega(\mathrm{d}x) \right).$$

Therefore, one can use any second-order method, e.g., `fmincon` solver in MATLAB (with trust region reflective algorithm), for solving the *univariate* optimization problem.

Equivalence [2, §8.17] [50, §13.11]. The Luxemburg norm is equivalent to the Orlicz norm. In fact, we have

$$\|f\|_{L_\Phi} \leq \|f\|_\Phi \leq 2\,\|f\|_{L_\Phi}. \tag{59}$$

Connection between L^p and L_Φ functional spaces. When the convex function $\Phi(t) = t^p$, for $1 < p < \infty$, we have

$$L^p(\mathbb{G}, \omega) = L_\Phi(\mathbb{G}, \omega).$$

Generalized Hölder inequality. Let Φ, Ψ be a pair of complementary N-functions, then generalized Hölder inequality w.r.t. Luxemburg norm [2, §8.11] is as follows:

$$\left| \int_{\mathbb{G}} f(x)g(x)\omega(dx) \right| \leq 2\,\|f\|_{L_\Phi}\,\|g\|_{L_\Psi}. \tag{60}$$

Additionally, we have the generalized Hölder inequality w.r.t. Luxemburg norm and Orlicz norm [50, §13.13] is as follows:

$$\left| \int_{\mathbb{G}} f(x)g(x)\omega(dx) \right| \leq \|f\|_{L_\Phi}\,\|g\|_{\Psi}. \tag{61}$$

B.1.4 Wasserstein Distance and Orlicz Wasserstein (OW)

We briefly review p-Wasserstein distance with graph metric cost, and the Orlicz Wasserstein (OW) for probability measures on a graph.

Wasserstein distance with graph metric cost. Given $1 \leq p < \infty$, and probability measures μ and ν supported on graph $\mathbb{G}$, then the p-order Wasserstein distance is defined as follows:

$$\mathcal{W}_p(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{G}\times\mathbb{G}} d_{\mathbb{G}}(x,y)^p \gamma(\mathrm{d}x, \mathrm{d}y) \right)^{\frac{1}{p}},$$

where $\Pi(\mu,\nu) := \Big\{ \gamma \in \mathcal{P}(\mathbb{G}\times\mathbb{G}) :\ \gamma_1 = \mu,\ \gamma_2 = \nu \Big\}$, and γ_1, γ_2 are the first and second marginals of γ respectively.

Orlicz Wasserstein (OW). Following Guha et al. [28, Definition 3.2], the OW with the N-function Φ for probability measures μ, ν supported on graph $\mathbb{G}$ is defined as follows:

$$W_\Phi(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \inf\Big[t>0 : \int_{\mathbb{G}\times\mathbb{G}} \Phi\left(\frac{d_{\mathbb{G}}(x,z)}{t}\right) \mathrm{d}\pi(x,z) \le 1\Big], \tag{62}$$

where recall that $\Pi(\mu,\nu)$ is the set of all couplings between μ and ν.

B.1.5 Generalized Sobolev transport (GST)

We briefly review main results on generalized Sobolev transport (GST) [41] for probability measures on a graph.

Graph-based Orlicz-Sobolev space [41]. Let Φ be an N-function and ω be a nonnegative Borel measure on graph $\mathbb{G}$. A continuous function $f : \mathbb{G} \to \mathbb{R}$ is said to belong to the graph-based Orlicz-Sobolev space $WL_\Phi(\mathbb{G},\omega)$ if there exists a function $h \in L_\Phi(\mathbb{G},\omega)$ satisfying

$$f(x) - f(z_0) = \int_{[z_0,x]} h(y)\omega(\mathrm{d}y), \quad \forall x \in \mathbb{G}. \tag{63}$$

Such function h is unique in $L_\Phi(\mathbb{G},\omega)$ and is called the generalized graph derivative of f w.r.t. the measure ω. This generalized graph derivative of f is denoted as f'.

Generalized Sobolev transport (GST) [41]. Let Φ be an N-function and ω be a nonnegative Borel measure on $\mathbb{G}$. For probability measures μ, ν on a graph $\mathbb{G}$, the generalized Sobolev transport (GST) is defined as follows:

$$\mathcal{GS}_\Phi(\mu,\nu) := \left\{ \begin{array}{ll} \sup & \left| \int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(x)\nu(\mathrm{d}x) \right| \\ \text{s.t.} & f \in WL_\Psi(\mathbb{G},\omega),\ \|f'\|_{L_\Psi} \le 1, \end{array} \right.$$

where Ψ is the complementary function of Φ (see (57)).

B.1.6 Unbalanced Sobolev transport (UST)

We give a brief review on main results of unbalanced Sobolev transport (UST) [40] for measures on a graph, possibly having different total masses.

The regularized set $\mathbb{U}^\alpha_{p'}$ for critic function [40]. For $1 \le p \le \infty$ and $0 \le \alpha \le \frac{1}{2}[b\lambda + w_1(z_0) + w_2(z_0)]$, let $\mathbb{U}^\alpha_{p'}$ be the collection of all functions $f \in W^{1,p'}(\mathbb{G},\omega)$ satisfying

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

and

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

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

$$f(x) = s + \int_{[z_0,x]} h(y)\omega(\mathrm{d}y) \tag{64}$$

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

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

Unbalanced Sobolev transport (UST) [40]. Let ω be a nonnegative Borel measure on graph $\mathbb{G}$. Given $1 \le p \le \infty$ and $0 \le \alpha \le \frac{1}{2}[b\lambda + w_1(z_0) + w_2(z_0)]$. For unbalanced measures $\mu, \nu \in \mathcal{P}(\mathbb{G})$, the unbalanced Sobolev transport (UST) is defined as follows

$$\mathcal{US}_{p,\alpha}(\mu,\nu) := \sup_{f\in\mathbb{U}^\alpha_{p'}} \left[\int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(x)\nu(\mathrm{d}x) \right].$$

For simplicity, we also use $\mathcal{US}_p$ for the p-order UST when the context for α is clear.

Equal measures on a graph [40].

Lemma B.3 (Lemma A.9 in [40]). *For unbalanced measures $\mu, \nu \in \mathcal{P}(\mathbb{G})$, then $\mu = \nu$ if and only if $\mu(\Lambda(x)) = \nu(\Lambda(x))$ for every x in $\mathbb{G}$.*

B.1.7 Regularized EPT and Distance d_α

We briefly review main results for the regularized EPT and distance d_α in [36] for probability measures on tree $\mathcal{T}$.

Regularized set of critic functions [36]. Let $\mathbb{L}_\alpha$ be a collection of all functions f of the form

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

where r is the tree root, and s is a constant in the interval $\Big[- w_2(r) - \frac{b\lambda}{2} + \alpha, w_1(r) + \frac{b\lambda}{2} - \alpha \Big]$ and with $\|g\|_{L^\infty(\mathcal{T})} \leq b$.

Regularized EPT [36]. For unbalanced measures μ, ν supported on tree $\mathcal{T}$, the regularized EPT is defined as follows:

$$\widetilde{\mathrm{ET}}^\alpha_\lambda(\mu, \nu) := \sup \left\{ \int_{\mathcal{T}} f(\mu - \nu) : \ f \in \mathbb{L}_\alpha \right\} - \frac{b\lambda}{2} \big[\mu(\mathcal{T}) + \nu(\mathcal{T}) \big]. \tag{65}$$

Following [36, Proposition 3.8], we have

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

with $i := 1$ if $\mu(\mathcal{T}) \geq \nu(\mathcal{T})$ and $i := 2$ if $\mu(\mathcal{T}) < \nu(\mathcal{T})$.

Distance d_α [36]. We briefly review the definition of distance d_α in [36]

$$d_\alpha(\mu, \nu) := \widetilde{\mathrm{ET}}^\alpha_\lambda(\mu, \nu) + \frac{b\lambda}{2} \big[\mu(\mathcal{T}) + \nu(\mathcal{T}) \big]. \tag{66}$$

Following [36, Proposition 3.10], the distance d_α is a metric.

B.2 Further Discussions

We give further discussions and details for various aspects in our work. For completeness, we recall important discussions on the graph in [39] (for Sobolev transport for probability measures), since these discussions and results are also applied and/or easily adapted for our proposed approaches.

Further details for the computation of Orlicz-EPT and OST. We describe further details for the computation for Orlicz-EPT and OST.

• **For Orlicz-EPT.** Following the theoretical ground derived for the computation of the entropic regularized Orlicz-EPT, i.e., the objective function is monotone non-increasing (Proposition 3.4), and the lower and upper limits for the objective functions (Proposition 3.5), one can compute the entropic regularized Orlicz-EPT by a binary search algorithmic approach. For completeness, we straightforwardly describe the pseudo-code for it in Algorithm 1. Additionally, notice that we can leverage Proposition A.3 in §A.1.2 to alternatively set the initial value for t_ℓ in Algorithm 1 (line 3).

• **For OST.** For popular N-function, it is easy to derive its gradient and Hessian for the objective function of the univariate optimization problem. Therefore, in our experiments, we leverage the `fmincon` MATLAB solver with the *trust-region-reflective* algorithm to solve the univariate optimization problem for OST computation.

Algorithm 1 Compute entropic regularized Orlicz-EPT $\mathcal{OE}_{\Phi,\varepsilon}$

Input: Input measures μ, ν, function Φ, graph $\mathbb{G}$, parameters b, λ, ε, and stopping threshold $\bar{\varepsilon}$
Output: entropic regularized Orlicz-EPT $\mathcal{OE}_{\Phi,\varepsilon}(\mu, \nu)$

Construct $\hat{\mathbb{G}} = \mathbb{G} \cup \{\hat{s}\}$ and corresponding nonnegative cost function $\hat{c}$ (§3)
Construct corresponding probability measures $\hat{\mu} = \frac{\mu + \nu(\mathbb{G})\delta_{\hat{s}}}{\mu(\mathbb{G})+\nu(\mathbb{G})}$ and $\hat{\nu} = \frac{\nu + \mu(\mathbb{G})\delta_{\hat{s}}}{\mu(\mathbb{G})+\nu(\mathbb{G})}$.
Set $t_r = \frac{L_{\hat{\mu},\hat{\nu}}}{\Phi^{-1}(1+\varepsilon)}$ and $t_\ell = \frac{\mathcal{W}_{\hat{c}}(\hat{\mu},\hat{\nu})}{\Phi^{-1}(1+\varepsilon[H(\hat{\mu})+H(\hat{\nu})-1])}$
while $t_r - t_\ell > \bar{\varepsilon}$ **do**
    Set $t_m = \frac{t_\ell + t_r}{2}$
    Compute $f_m = \mathcal{A}_\varepsilon(t_m; \hat{\mu}, \hat{\nu})$
    if $f_m \leq 1$ **then**
        Set $t_r = t_m$
        if $f_m == 1$ **then**
            Break
    else
        Set $t_\ell = t_m$
Return $\mathcal{OE}_{\Phi,\varepsilon}(\mu, \nu) = (\mu(\mathbb{G}) + \nu(\mathbb{G}))\,(t_r - b\lambda)$

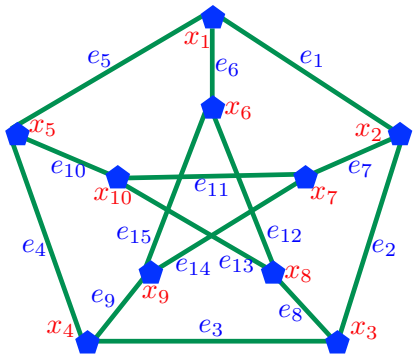

Figure 4: A geodetic graph illustration. The set of nodes V has 10 nodes, i.e., $V = \{x_1, x_2, \ldots, x_{10}\}$. The set of edges E has 15 edges, i.e., $E = \{e_1, e_2, \ldots, e_{15}\}$ where each edge weight/length is set to one, i.e., $w_{e_j} = 1$, for $1 \leq j \leq 15$. For any x_i, x_j, there is a unique shortest path between them, with a length 2. Let x_1 be the unique-path root node (i.e., $z_0 = x_1$) and $\widetilde{\mathbb{G}}$ be a subgraph containing 3 nodes $\{x_6, x_8, x_9\}$ and 2 edges $\{e_{12}, e_{15}\}$, then we have $\Lambda(x_6) = \gamma(e_6) = \widetilde{\mathbb{G}}$.

Further discussion on Orlicz-EPT and OST. Following Proposition 5.2, we can view OST as an extension of GST [41] for handling unbalanced measures, and Orlicz-EPT as an extension of OW [67] for unbalanced measures. Notably, when the input measures have equal mass (i.e., $\mu(\mathbb{G}) = \nu(\mathbb{G})$), and $b = 1$, Orlicz-EPT reduces to OW. Furthermore, Orlicz-EPT shares the same computational complexity as OW, as shown in Equation (8). It is worth noting that GST is a scalable variant of OW for balanced measures, in the same sense, OST can be seen as a scalable variant of Orlicz-EPT.

As a result, Orlicz-EPT is applicable for all applications of OW and extends OW to handle unbalanced input measures, whereas OST may not reserve all the properties of OW.

Like OW, Orlicz-EPT involves a two-level optimization problem, limiting its applicability to small-scale domains. In contrast, OST, similar to GST, is scalable and can be applied to large-scale domains.

Illustrations for notations on a graph. We illustrate notions on a graph in our work in Figure 4.

Geodetic graph. Recall that for a graph, for every pairs of nodes, the shortest path between them is unique, then it is called geodetic graph [39]. Therefore, geodetic graphs are special examples

satisfying the uniqueness property of the shortest paths (§2). We give an illustration of geodetic graph in Figure 4.

Path length. As discussed in [39], we can compute a path length connecting any two points $x, y \in \mathbb{G}$, where the points x, y are not necessary to be nodes in V (but even for any point on an edge as well). Indeed, for the points $x, y \in \mathbb{R}^n$ belonging to the same edge $e = \langle u, v\rangle$, connecting two nodes $u, v \in V$. We have

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

for some scalars $t, s \in [0, 1]$. Thus, the length of the path connecting x, y along the edge e (i.e., the line segment $\langle x, y\rangle$) can be computed by $|t - s|w_e$.

Consequently, the length for an arbitrary path in $\mathbb{G}$ can be computed by breaking down into pieces over edges, and then summing over their corresponding lengths [39].

Extension to measures supported on $\mathbb{G}$. For the discrete case, similar to ST [39], OST (17) is extendable for measures with finite supports on $\mathbb{G}$, i.e., measures which may have supports on edges, by following the strategy to compute a path length for points in $\mathbb{G}$. Precisely, we break down edges containing supports into pieces, and then sum over their corresponding values, instead of the sum over edges as in (17) for OST.

About the uniqueness property of the shortest paths on $\mathbb{G}$. As discussed in [39], note that edge length is a real nonnegative scale, $w_e \in \mathbb{R}_+$ for any edge $e \in E$ in $\mathbb{G}$., it is almost surely that every node in V can be regarded as unique-path root node since with a high probability, lengths of paths connecting any two nodes in graph $\mathbb{G}$ are different.

Moreover, for some special graph, e.g., a grid of nodes, there is *no* unique-path root node for such graphs. However, by perturbing each node (or also perturbing edge lengths in case $\mathbb{G}$ is a non-physical graph) with a small deviation, the perturbed graph will satisfy the unique-path root node assumption.

For continuous case, when measures are extended to support on $\mathbb{G}$, and input measures are fully supported, then for some finite special nodes where there are multiple shortest paths, we fix one of the shortest paths to those points, and treat it as the chosen shortest path for those special points. However, for practical applications, the number of supports are finite, which bypasses the mentioned issue.

For the given graph setting. As in [39], we assume that the graph metric space is given. The question to learn an optimal graph metric structure from data is not considered in this work and leave for future investigation.

Measures on a graph. In this work, we consider OT problem for *two input unbalanced measures* supported on the *same* graph. See [40] for the same problem setting.

The proposed approaches, Orlicz-EPT and OST, are for *input unbalanced measures*, i.e., to compute the distance between two unbalanced measures, on the *same* graph. We distinguish our considered problem to the following related problems:

• **Compute distances/discrepancies between two (different) input graphs.** For examples, [55, 22] compute distance/discrepancy between *two (different) input graphs*. They are essentially different to our considered problem which computes distance between *two input probability measures* supported on the *same* graph. In particular, Le et al. [38] consider a variant of Gromov-Wasserstein problem for *two input probability measures*, but possibly supported on *different* tree metric spaces.

• **Compute kernels between two (different) input graphs.** Graph kernels are kernel functions between two input (different) graphs to assess their similarity. See [10, 34, 53] for comprehensive reviews on graph kernels. Essentially, it is different to our proposed approaches to compute distances between *two input unbalanced measures* on the *same* graph.

For (variational) OT problems on a graph. OT problems for measures supported on a graph metric space have been explored in previous studies [39–41]. Additionally, Le et al. [43] have recently studied a variational OT problem, e.g., Sobolev IPM where the critic function is constrained within

a unit ball of Sobolev norm involving both the critic function and its gradient, for graph-based measures. Notably, graph metrics extend tree metrics, which are utilized in scalable optimal transport methods like tree-sliced Wasserstein (TSW) [37, 70, 68, 71, 69]. TSW, in turn, generalizes the popular sliced-Wasserstein (SW) approach [59, 9, 51]. The graph structure offers greater flexibility and degrees of freedom compared to the tree structure in TSW and the line structure in SW. For a more in-depth discussion on the motivation for OT on a graph, please refer to [39].

Persistence diagrams for TDA. Persistence diagrams (PD) are multisets of data points in $\mathbb{R}^2$, containing the birth and death time respectively of topological features (e.g., connected component, ring, or cavity), extracted by algebraic topology methods (e.g., persistence homology) [23].

Graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ [39]. For completeness, we review the construction for graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ in [39]. We utilize the farthest-point clustering method to cluster supports of measures into at most M clusters. Then, let the vertex set V be the collection of centroids of these clusters, i.e., graph vertices. For edges, we randomly select $(M \log M)$ edges, and $M^{3/2}$ edges for graphs $\mathbb{G}_{\text{Log}}$, and $\mathbb{G}_{\text{Sqrt}}$ respectively. Let $\tilde{E}$ be the set of those randomly sampled edges. For each edge e, its edge length/weight w_e is computed by Euclidean distance between the two corresponding end points (i.e., corresponding nodes of edge e). Let n_c be the number of connected components in $\tilde{\mathbb{G}}(V, \tilde{E})$. We randomly add $(n_c - 1)$ more edges between these n_c connected components to form 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 constructed graph.

Further discussion on graphs in experiments. For all datasets, except `MPEG7` dataset, $\mathbb{G}_{\text{Sqrt}}$ consists of 10K nodes and 1 million edges, while $\mathbb{G}_{\text{Log}}$ comprises 10K nodes and 100K edges. Due to the smaller size of the `MPEG7` dataset, we constructed $\mathbb{G}_{\text{Sqrt}}$ with 1K nodes and 32K edges, and $\mathbb{G}_{\text{Log}}$ with 1K nodes and 7K edges.

Datasets and Computational Devices. For the datasets in our experiments, one can contact the authors of Sobolev transport [39] to access to them. Additionally, all of our experiments are run on commodity hardware.

In our experiments, we demonstrate the effectiveness of our approach in document classification on four real-world datasets and topological data analysis (TDA), including orbit recognition for linked twist maps, i.e., a discrete dynamical system modeling flows in DNA microarrays [31], and object shape recognition in `MPEG7`. These evaluations on document classification and TDA are often used for tasks involving comparing measures on a graph, see [39, 41]. We believe that such experimental coverage is rich and diverse enough.

Hyperparameter validation. We use the same validation as in [40]. Precisely, we further randomly split *the training set* into $70\%/30\%$ for validation-training and validation with 10 repeats to choose hyper-parameters for the experiments.

Further discussion on hyperparameters. The performance of OST/Orlicz-EPT typically depends on the choice of the N-function Φ, similar to how kernel functions impact performance in kernel-dependent machine learning frameworks. In our experiments with N-functions Φ_1 and Φ_2 (§7), we observed slightly different performances. Similar findings have been reported for GST [41].

Determining the optimal N-function Φ for OST/Orlicz-EPT in a given task is an open problem that warrants further investigation. We leave it for future work. As an interim solution, cross-validation can be used to select Φ from a set of candidate functions.

Regarding the regularization weight b, we used $b = 1$ in our experiments based on the results in Propositions 5.2 and 5.3, as well as for simplicity. This choice of b is supported by theoretical results on EPT on a graph [40, Lemma A.6, Remark 4.8] and has been used in previous experiments for EPT [36, 40].

For α, from the result in Proposition 5.7 and for simplicity, we use $\alpha = 0$ for experiments. Such value for α is also supported by theoretical results of EPT on a graph [40, Lemma 4.4, Proposition 5.2] and used in experiments for EPT [36, 40]. Additionally, recall that $\mathcal{I}_\alpha$ is the largest interval when $\alpha = 0$.

Table 1: The number of pairs for SVM.

Datasets	#pairs
TWITTER	4394432
RECIPE	8687560
CLASSIC	22890777
AMAZON	29117200
Orbit	11373250
MPEG7	18130

The number of pairs for kernel SVM [41]. Denote N_{tr}, N_{te} for the number of measures used for training and test respectively. For the kernel SVM training, the number of pairs for computing the distances is $(N_{tr}-1)\times\frac{N_{tr}}{2}$. For the test, the number of pairs for computing the distances is $N_{tr}\times N_{te}$. Thus, for each run, the number of pairs for computing the distances for both training and test is totally $N_{tr}\times(\frac{N_{tr}-1}{2}+N_{te})$. In Table 1, we summarize the number of pairs which we need to evaluate distances/discrepancies for SVM in each run to illustrate the experimental scale, e.g., more than 29M pairs for AMAZON.

Debiased approaches for (Sinkhorn-based) UOT [63, 65]. We review the discussion and evaluation in [40] for the debiased approaches for (Sinkhorn-based) UOT in [63, 65], with the same setup as in our experiments. As noted in the main text, the debiased version for Sinkhorn-based approach for UOT [26, 63] which may be helpful for applications, and the debiased version is empirically indefinite. Both the UOT and its debiased version have the same computational complexity.

Empirical results for the debiased version [63, 65] are given in [40, Figures 41–44]. As in [40], the debiased version improve performances of UOT in some datasets, especially for datasets in TDA tasks (Orbit and MPEG7). For document datasets, performances of the debiased version and UOT are comparative, i.e., the role of debias property is not clear for advantages in applications.

Broader impacts. In this work, we propose novel approaches to extend OW/GST for unbalanced measures on a graph. The proposed Orlicz-EPT is directly derived from standard OT, bypassing challenges raised from unbalanced measures, as in OW. Additionally, we formulate a novel regularization approach, resulting in the proposed OST, which is efficient in computation. Therefore, our proposals pave the ways to use OT approach endowed with Orlicz geometric structure for applications with unbalanced measures, which is common in real-world scenarios. To our knowledge, there is no foresee negative social impacts for our research.

B.3 Further Empirical Results

Further empirical results on graph $\mathbb{G}_{\text{Log}}$. We provide corresponding results as in §7 for graph $\mathbb{G}_{\text{Log}}$.

- In Figure 5, we compare the time consumption of OST and Orlicz-EPT with Φ_0, Φ_1, Φ_2 on graph $\mathbb{G}_{\text{Log}}$.
- In Figure 6, we illustrate the SVM results and time consumptions of kernels on document classification with graph $\mathbb{G}_{\text{Log}}$.
- In Figure 7, we illustrate the SVM results and time consumptions of kernels on TDA with graph $\mathbb{G}_{\text{Log}}$.

Further empirical results for different hyperparameters b and α. We carry out more additional experiments for different b, α. The results for average accuracy, and the standard deviation (in the parentheses) are as follows:

• For document classification.

- We present further SVM results for document classification on TWITTER dataset with graph $\mathbb{G}_{\text{Sqrt}}$ and different hyperparameters b and α in Tables 2 and 3 respectively.
- We present further SVM results for document classification on TWITTER dataset with graph $\mathbb{G}_{\text{Log}}$ and different hyperparameters b and α in Tables 4 and 5 respectively.

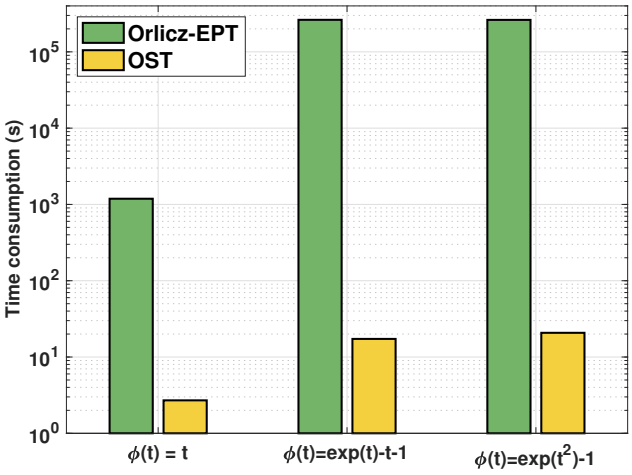

Figure 5: Time consumption on graph $\mathbb{G}_{\text{Log}}$.

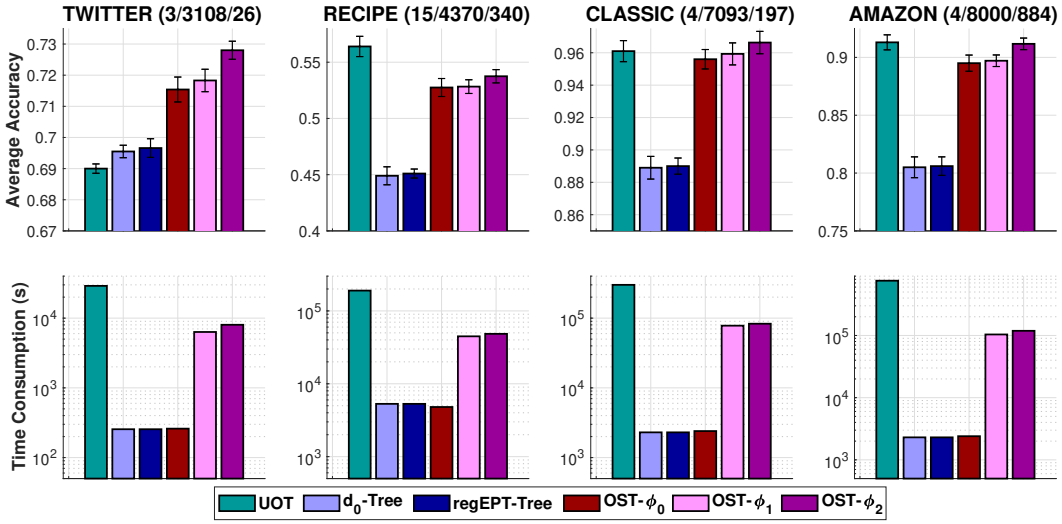

Figure 6: Document classification on graph $\mathbb{G}_{\text{Log}}$.

• **For TDA.**

- We present further SVM results for TDA on `MPEG7` dataset with graph $\mathbb{G}_{\text{Sqrt}}$ and different hyperparameters b and α in Tables 6 and 7 respectively.
- We present further SVM results for TDA on `MPEG7` dataset with graph $\mathbb{G}_{\text{Log}}$ and different hyperparameters b and α in Tables 8 and 9 respectively.

We observe that turning these hyperparameters b and α, e.g., via cross-validation, may help to improve the performances further.

Table 2: SVM results for document classification on `TWITTER` dataset with graph $\mathbb{G}_{\text{Sqrt}}$ and different hyperparameter b.

	Φ_0	Φ_1	Φ_2
$b = 0.5$	71.66 ± 0.63	72.03 ± 0.56	72.12 ± 0.57
$b = 2$	71.54 ± 0.49	72.15 ± 0.42	72.28 ± 0.34

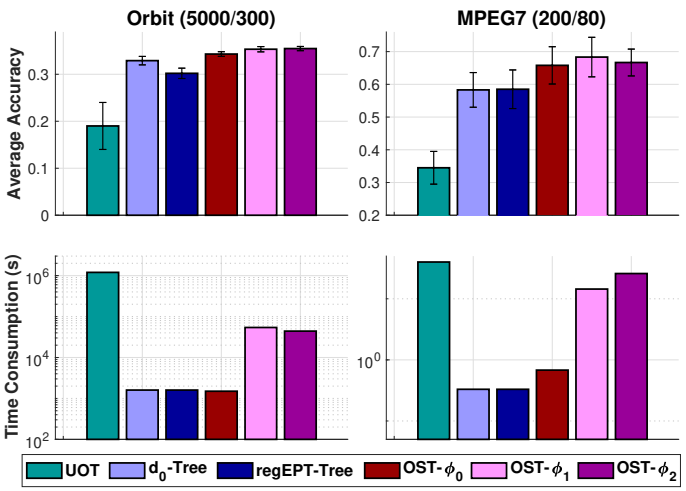

Figure 7: TDA on graph $\mathbb{G}_{\text{Log}}$.

Table 3: SVM results for document classification on TWITTER dataset with graph $\mathbb{G}_{\text{Sqrt}}$ and different hyperparameter α.

	Φ_0	Φ_1	Φ_2
$\alpha = 0.1$	71.48 ± 0.43	72.34 ± 0.55	72.64 ± 0.50
$\alpha = 0.2$	71.64 ± 0.50	72.12 ± 0.36	72.27 ± 0.30

Table 4: SVM results for document classification on TWITTER dataset with graph $\mathbb{G}_{\text{Log}}$ and different hyperparameter b.

	Φ_0	Φ_1	Φ_2
$b = 0.5$	70.97 ± 0.29	71.74 ± 0.33	71.78 ± 0.41
$b = 2$	71.58 ± 0.41	71.92 ± 0.16	71.96 ± 0.22

Table 5: SVM results for document classification on TWITTER dataset with graph $\mathbb{G}_{\text{Log}}$ and different hyperparameter α.

	Φ_0	Φ_1	Φ_2
$\alpha = 0.1$	71.57 ± 0.23	71.87 ± 0.36	72.04 ± 0.28
$\alpha = 0.2$	71.59 ± 0.35	71.74 ± 0.22	71.95 ± 0.19

Table 6: SVM results for TDA on MPEG7 dataset with graph $\mathbb{G}_{\text{Sqrt}}$ and different hyperparameter b.

	Φ_0	Φ_1	Φ_2
$b = 0.5$	66.67 ± 3.45	71.67 ± 5.40	68.83 ± 3.09
$b = 2$	65.00 ± 3.58	69.13 ± 3.87	67.33 ± 3.55

Table 7: SVM results for TDA on MPEG7 dataset with graph $\mathbb{G}_{\text{Sqrt}}$ and different hyperparameter α.

	Φ_0	Φ_1	Φ_2
$\alpha = 0.1$	66.43 ± 4.81	70.47 ± 4.51	68.50 ± 4.54
$\alpha = 0.2$	66.82 ± 4.01	68.43 ± 4.63	69.17 ± 3.60

Table 8: SVM results for TDA on MPEG7 dataset with graph $\mathbb{G}_{\text{Log}}$ and different hyperparameter b.

	Φ_0	Φ_1	Φ_2
$b = 0.5$	65.87 ± 5.68	67.00 ± 5.60	68.67 ± 5.61
$b = 2$	65.93 ± 5.11	68.91 ± 6.00	66.17 ± 4.60

Table 9: SVM results for TDA on MPEG7 dataset with graph $\mathbb{G}_{\text{Log}}$ and different hyperparameter α.

	Φ_0	Φ_1	Φ_2
$\alpha = 0.1$	65.94 ± 4.19	67.83 ± 5.83	66.46 ± 4.58
$\alpha = 0.2$	65.84 ± 5.45	69.61 ± 5.56	69.86 ± 5.22

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The proposed Orlicz-EPT is presented in §3, where we revisit the EPT problem, leverage Caffarelli & McCann [12]'s insights to reformulate EPT as a corresponding standard OT. By carefully calibrating the ground cost of the corresponding standard OT, we guarantee that its ground cost is nonnegative, which is essential to develop Orlicz-EPT, a variant of EPT with Orlicz geometric structure, since the N-function is only defined on the nonnegative domain. Additionally, we provide theoretical background to solve Orlicz-EPT by a binary search approach.

   Additionally, we present the proposed Orlicz-Sobolev transport (OST) in §4. By leveraging the dual EPT and the underlying graph structure, we derive a novel regularization for the critic function, to develop the proposed OST. We prove that OST can be computed by simply solving a univariate optimization problem.

   Furthermore, in §5, we derive geometric structures for OST, and show its connections to other transport distances. In §7, we empirically illustrate that OST is several-order faster than Orlicz-EPT. We also show evidences on the advantages of OST for unbalanced measures on a graph for document classification and TDA.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We consider OT problem for unbalanced measures supported on a graph metric space.

   For Orlicz-EPT, we show that it is directly derived from a standard OT, with a guarantee on the nonnegativity for its ground cost via calibration, similar to OW approach (derived from corresponding standard OT with nonnegative ground cost [67]). Therefore, we can bypass all challenges raised from unbalanced measures. We develop theoretical backgrounds to solve it by a binary search approach. However, Orlicz-EPT is still a two-level optimization problem, it has a high computational cost, illustrated in §7.

   For OST, it is efficient in computation by simply solving a univariate optimization problem (§4). Although OST is a regularized approach, we theoretically show that OST generalize GST for unbalanced measures on a graph (Proposition 5.2).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Detailed proofs of theoretical results are placed in Appendix §A.2.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: we provide detailed setup for our experiments in §7. We discuss further details in Appendix B.2. We also submit the code together with our submission.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often

one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Details of the experiments are given in §7. We use public datasets for our experiments. Code is submitted together with the submission. We also discuss further details in Appendix B.2.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please see §7, Appendix B.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Please see reported empirical results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see Appendix §B.2, where we describe that all experiments are carried on commodity hardware.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research in our submission is to advance the machine learning field. To our knowledge, there is no foresee harmful consequences.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Please see Appendix §B.2, where we describe that our research aims to advance to the machine learning field. The proposed OST has an efficient computation, which may help to reduce the computational cost. To our knowledge, there is no foresee negative social impacts for our research.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our research poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Please see §7, where we give proper credit/citation to the original owners.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.