Generalized Sobolev Transport for Probability Measures on a Graph

Tam Le [* 1 2] **Truyen Nguyen** [* 3] **Kenji Fukumizu** [1]

Abstract

We study the optimal transport (OT) problem for measures supported on a graph metric space. Recently, Le et al. (2022) leverage the graph structure and propose a variant of OT, namely Sobolev transport (ST), which yields a closed-form expression for a fast computation. However, ST is essentially coupled with the L^p geometric structure within its definition which makes it nontrivial to utilize ST for other prior structures. In contrast, the classic OT has the flexibility to adapt to various geometric structures by modifying the underlying cost function. An important instance is the Orlicz-Wasserstein (OW) which moves beyond the L^p structure by leveraging the *Orlicz geometric structure*. Comparing to the usage of standard p-order Wasserstein, OW remarkably helps to advance certain machine learning approaches. Nevertheless, OW brings up a new challenge on its computation due to its two-level optimization formulation. In this work, we leverage a specific class of convex functions for Orlicz structure to propose the generalized Sobolev transport (GST). GST encompasses the ST as its special case, and can be utilized for prior structures beyond the L^p geometry. In connection with the OW, we show that one only needs to simply solve a univariate optimization problem to compute the GST, unlike the complex two-level optimization problem in OW. We empirically illustrate that GST is several-order faster than the OW. Moreover, we provide preliminary evidences on the advantages of GST for document classification and for several tasks in topological data analysis.

---

[*]Equal contribution [1]Department of Advanced Data Science, The Institute of Statistical Mathematics (ISM), Tokyo, Japan [2]RIKEN AIP, Tokyo, Japan [3]The University of Akron, Ohio, US. Correspondence to: Tam Le <tam@ism.ac.jp>.

1. Introduction

Optimal transport (OT) is a natural geometry to compare probability measures (Villani, 2008; Peyré & Cuturi, 2019). Intuitively, OT lifts the cost metric on support data points of input measures to the distance for the measures by minimizing the cost to transport a measure into the other. OT has been applied on many applications in various research fields, e.g., in machine learning (Nadjahi et al., 2019; Titouan et al., 2019; Janati et al., 2020; Mukherjee et al., 2021; Altschuler et al., 2021; Fatras et al., 2021; Scetbon et al., 2021; Le & Nguyen, 2021; Le et al., 2021a;b; Liu et al., 2021; Nguyen et al., 2021a;b; Takezawa et al., 2022; Fan et al., 2022; Bunne et al., 2022; 2023a; Bonet et al., 2023; Mahey et al., 2023; Nguyen et al., 2023a;b; Hua et al., 2023; Korotin et al., 2023; 2024; Le et al., 2024; Nguyen et al., 2024), statistics (Mena & Niles-Weed, 2019; Weed & Berthet, 2019; Liu et al., 2022; Nguyen et al., 2022; Wang et al., 2022; Nietert et al., 2022; 2023; Pham et al., 2024), and biology (Schiebinger et al., 2019; Tong et al., 2020; Bunne et al., 2023a;b; Korotin et al., 2024).

A main drawback of OT is that its computational complexity is high, i.e., super cubic with respect to the number of supports of input measures, which hinders its applications in large-scale applications. Recently, several approaches have been proposed in the literature to scale up the OT problem. Cuturi (2013) proposes to use entropic regularization for OT and leverages the Sinkhorn algorithm to reduce its complexity into quadratic. Scetbon et al. (2021) propose low-rank approach to further reduce the computation of Sinkhorn algorithm for entropic regularized OT. Another direction is to exploit the local structure on supports of input measures. Rabin et al. (2011) propose sliced-Wasserstein (SW) which projects supports of input measures into a randomly one-dimensional space and utilizes the closed-form computation of the univariate OT. However, by relying on the one-dimensional projection, SW limits its capacity to capture the topological structure of input measures, especially when measures are supported in a high-dimensional space. Le et al. (2019) propose to leverage tree structure which provides more flexibility and degrees of freedom, i.e., choosing a tree rather than a line as in SW, to alleviate the curse of dimensionality in SW. For practical applications, the tree structure might be a restricted requirement. Recently, Le et al. (2022) propose a variant of OT, namely

Sobolev transport (ST), for measures supported on a graph which admits a closed-form expression for fast computation.

Unlike the classic OT where one can easily utilize various prior geometric structures on supports of input measures, e.g., via the ground cost metric, it is nontrivial to use other structures rather than the L^p structure for the ST. Essentially, the definition of ST is based on the Kantorovich duality of 1-order Wasserstein, but ST considers the graph-based Sobolev constraint for the critic function. This makes the ST coupled with the L^p functional geometric structure within the graph-based Sobolev space.

One approach to move beyond the L^p geometry is to employ the Orlicz geometric structure, which is obtained by leveraging a specific class of convex functions. Indeed, Orlicz norm has been utilized to advance several machine learning approaches in the literature. In particular, Andoni et al. (2018) and Song et al. (2019) leverage the Orlicz norm as a loss for the classic linear regression. Orlicz norm loss regression not only generalizes both the least squared regression (i.e., squared ℓ_2 norm loss) and the least absolute deviation regression (i.e., ℓ_1 norm loss), but also provides a scale-invariant version of $\mathbb{M}$-estimator function loss (e.g., the Huber, Cauchy, Welsh, Tukey, Geman-McClure functions (Zhang, 1997, Table 1)).[1] Moreover, Andoni et al. (2018) illustrate that the Orlicz norm loss regression improves performances of linear regression with ℓ_1/ℓ_2 norms on Gaussian/sparse noise respectively. In addition, Deng et al. (2022) propose an efficient data structure for the Orlicz norm as its special case, which helps to scale up many machine learning problems including reinforcement learning, kernelized support vector machine, and clustering. Notably, in order to address the challenging polynomial growing nature of the underlying function class in the empirical process for random Fourier features on approximation of high-order kernel derivatives, Chamakh et al. (2020) leverage the Orlicz metric on the sample distribution to derive a finite-sample deviation bound for a general class of polynomial-growth functions. In particular, Chamakh et al. (2020) propose finite-sample uniform guarantee for random Fourier features which is almost surely convergence to approximate high-order derivatives for arbitrary kernel.

Furthermore, Orlicz metric has been recently explored in the context of OT. In particular, Lorenz & Mahler (2022) propose to leverage Orlicz norm as a regularization for OT problem with continuous probability measures. Especially, Orlicz metric is also utilized as a ground cost for OT, which is known as the Orlicz-Wasserstein (OW) (Sturm, 2011; Kell, 2017). Notably, Guha et al. (2023) and Altschuler & Chewi (2023) have shown that OW possesses unique

---

[1] The classic $\mathbb{M}$-estimation functional loss depends on the data scale, i.e., one may get different solutions if the data are rescaled (Andoni et al., 2018).

implicit essences, which may not exist in the classic OT with L^p structure cost, and help to advance certain results in machine learning. In particular, Guha et al. (2023) propose to consider OW metric for studying the Bayesian contraction convergence behavior of parameters arising from hierarchical Bayesian nonparametric models. OW metric helps to alleviate a number of raised concerns caused from the usage of the classic OT with Euclidean cost for quantifying the rates of parameter convergence within infinite Gaussian mixtures to significantly improve the contraction rate. Additionally, Altschuler & Chewi (2023) propose to also leverage OW as a metric shift for Rényi divergence to develop novel differential-privacy-inspired techniques to overcome longstanding challenges for proving fast convergence of hypocoercive differential equations. In spite of these achievements, OW brings up a difficult challenge on its computation. Precisely, OW is a two-level optimization problem: one level is for the transportation plan as in the classic OT problem, and another level is for an extra positive scalar within the Orlicz metric structure.

In this work, we focus on OT problem for probability measures supported on a graph metric space (Le et al., 2022). On one hand, ST efficiently exploits the graph structure to yield a closed-form for a fast computation. However, ST is essentially coupled with the L^p geometric structure, and it is therefore nontrivial to utilize ST with other prior structures. On the other hand, OW metric owns unique implicit essences to improve certain machine learning problems, but its optimization formulation is challenging for computation. Our goal is to alleviate these issues by adopting Orlicz geometry to generalize the ST for measures on a graph. This allows us to propose the generalized Sobolev transport (GST) which inherit the merits from both ST and OW.

Contribution. In summary, our contributions are three-fold:

- We leverage a certain class of convex functions corresponding to Orlicz geometric structure, and propose the GST for probability measures supported on a graph metric space. We show that GST can be computed by simply solving a univariate optimization problem.

- We demonstrate that ST is a special case of the proposed GST. Additionally, GST utilizes the Orlicz geometric structure in the same sense as the OW for OT problem. We further draw a connection between GST and OW.

- We empirically illustrate that GST is more computationally efficient than OW. We also show some preliminary evidences on the advantages of GST for document classification and for several tasks in topological data analysis (TDA).

Organization. We briefly review related notations used in

the development of our proposals in §2. In §3, we describe our proposed approach, the generalized Sobolev transport (GST). We show that GST is a metric, and then derive its properties and draw its connection to ST and OW in §4. In §5, we empirically illustrate the computational advantages of GST over OW, and show preliminary improvements of GST on document classification and TDA. In §6, we give concluding remarks. Additionally, we defer the proofs of key theoretical results and additional materials to the Appendices. We have also released code for our proposed approach.[2]

2. Preliminaries

In this section, we introduce the notations and give a brief review about graphs together with the Sobolev transport (ST) for measures on a graph, as well as the Orlicz geometric structure and the Orlicz-Wasserstein (OW).

2.1. Graph and functions on graph

We describe the graph setting for measures, and functions on graph.

Graph. We use the same graph setting as in (Le et al., 2022). Specifically, let V and E be respectively the sets of nodes and edges. We consider a connected, undirected, and physical[3] graph $\mathbb{G} = (V, E)$ with positive edge lengths $\{w_e\}_{e \in E}$. Following the convention in (Le et al., 2022) for continuous graph setting, $\mathbb{G}$ is regarded as the set of all nodes in V together with all points forming the edges in E. Also, $\mathbb{G}$ is equipped with the graph metric $d_{\mathbb{G}}(x, y)$ which equals to the length of the shortest path in $\mathbb{G}$ between x and y. 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 (Le et al., 2022).

Let $[x, z]$ denote the shortest path connecting x and z in $\mathbb{G}$. For $x \in \mathbb{G}$, edge $e \in E$, define the sets $\Lambda(x)$, γ_e as follow:

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

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.

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

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

2.2. Sobolev transport (ST)

For probability measures on a graph, Le et al. (2022) propose the ST which is a scalable variant of OT. More specifically, Le et al. (2022) leverage the graph structure and propose ST by building upon the dual form of the 1-order Wasserstein distance, but considering its Lipschitz constraint for the critic function within the graph-based Sobolev space.

In particular, let ω be a nonnegative Borel measure on $\mathbb{G}$. Given $1 \leq p \leq \infty$, and let p' be its conjugate, i.e., the number $p' \in [1, \infty]$ satisfying $\frac{1}{p} + \frac{1}{p'} = 1$. For $\mu, \nu \in \mathcal{P}(\mathbb{G})$, the p-order Sobolev transport (ST) (Le et al., 2022, Definition 3.2) is defined as

$$\mathcal{S}_p(\mu, \nu) := \begin{cases} \sup \left[\int_{\mathbb{G}} f(x) \mu(\mathrm{d}x) - \int_{\mathbb{G}} f(x) \nu(\mathrm{d}x) \right] \\ \text{s.t. } f \in W^{1,p}(\mathbb{G}, \omega), \ \|f'\|_{L^{p'}(\mathbb{G}, \omega)} \leq 1, \end{cases} \qquad (2)$$

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 $\mathbb{G}$.[4] Notably, the ST yields a closed-form expression for a fast computation (Le et al., 2022, Proposition 3.5).

Additionally, when graph $\mathbb{G}$ is a tree and ω is a length measure[5], the 1-order Sobolev transport coincides with the 1-order Wasserstein distance (Le et al., 2022, Corollary 4.3).

However, unlike standard OT where one can easily adjust the ground cost following prior structure for applications, it is nontrivial for such adaptation in the ST due to its graph structure and the generalized graph derivation within the graph-based Sobolev space.

2.3. Orlicz functional space and Orlicz-Wasserstein

We describe a family of convex functions which we leverage to generalize the ST, and briefly review the Orlicz functional geometric structure, as well as the OW which utilizes Orlicz norm as its ground cost.

A family of convex functions. We consider the collection of N-functions (Adams & Fournier, 2003, §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 $\lim_{t \to +\infty} \frac{\Phi(t)}{t} = +\infty$.

---

[2]The code repository is on https://github.com/lttam/Generalized-Sobolev-Transport

[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]See Appendix §B.1 for a review on the graph-based Sobolev space, the generalized graph derivative function, and the L^p functional space.

[5]See Definition B.3 in appendix for the length measure.

Examples. Some popular examples for N-functions are (i) $\Phi(t) = t^p$ with $1 < p < \infty$; (ii) $\Phi(t) = \exp(t) - t - 1$; (iii) $\Phi(t) = \exp(t^p) - 1$ with $1 < p < \infty$; and (iv) $\Phi(t) = (1+t)\log(1+t) - t$ (Adams & Fournier, 2003, §8.2).

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 being defined by

$$\|f\|_{L_\Phi} := \inf \left\{ t > 0 \mid \int_{\mathbb{G}} \Phi\left(\frac{|f(x)|}{t}\right) \omega(\mathrm{d}x) \le 1 \right\}. \quad (3)$$

The infimum in Equation (3) for $\|f\|_{L_\Phi}$ is attained (Adams & Fournier, 2003, §8.9).

Orlicz-Wasserstein (OW). Following Guha et al. (2023, Definition 3.2), the OW with the N-function Φ for measures $\mu, \nu \in \mathcal{P}(\mathbb{G})$ is defined as follows:

$$W_\Phi(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \inf \Big[t > 0 :$$
$$\int_{\mathbb{G} \times \mathbb{G}} \Phi\left(\frac{d_{\mathbb{G}}(x, z)}{t}\right) \mathrm{d}\pi(x, z) \le 1 \Big], \quad (4)$$

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

In this work, we propose to leverage the collection of N-functions to generalize the ST, which can adopt Orlicz geometric structure in the same sense as the OW. Therefore, the proposed generalized Sobolev transport can inherit advantages from both ST and OW.

3. Generalized Sobolev Transport (GST)

In this section, we examine a special family of convex functions which grows faster than linear. By leveraging it, we generalize the ST in Equation (2) by proposing the *graph-based Orlicz-Sobolev space* for the Lipschitz constraint on the critic function for measures on a graph.

Graph-based Orlicz-Sobolev space.

Definition 3.1 (Graph-based Orlicz-Sobolev space). *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*

$$f(x) - f(z_0) = \int_{[z_0, x]} h(y)\omega(\mathrm{d}y), \quad \forall x \in \mathbb{G}. \quad (5)$$

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. Henceforth, this generalized graph derivative of f is denoted f'.

We remark that the Orlicz-Sobolev space $WL_\Phi(\mathbb{G}, \omega)$ depends on the point z_0 and the choice of the N-function Φ.

For brevity, we however do not explicitly display these dependencies in its notation. Additionally, the formulation in Equation (5) can be considered as a generalized version of the fundamental theorem of calculus, which defines the generalized graph derivative for a continuous function f at any point $x \in \mathbb{G}$. Moreover, the graph-based Orlicz-Sobolev space may be regarded as a generalized version of the graph-based Sobolev space (analyzed rigorously in §4).

Unlike the OW which directly utilizes the given N-function within the Luxemburg norm of the Orlicz functional space (Equation (4)), in order to generalize the ST, we also require a notion of the complementary function of the given N-function.

Complementary function. For the given N-function Φ, its complementary function $\Psi : \mathbb{R}_+ \to \mathbb{R}_+$ (Adams & Fournier, 2003, §8.3) is the N-function, defined as follows

$$\Psi(t) = \sup\left[at - \Phi(a) \mid a \ge 0\right], \quad \text{for } t \ge 0. \quad (6)$$

Examples. Some popular complementary pairs of N-functions (Adams & Fournier, 2003, §8.3), (Rao & Ren, 1991, §2.2) are: (i) $\Phi(t) = \frac{t^p}{p}$ and $\Psi(t) = \frac{t^q}{q}$ where q is the conjugate of p, i.e., $\frac{1}{p} + \frac{1}{q} = 1$ and $1 < p < \infty$, and (ii) $\Phi(t) = \exp(t) - t - 1$ and $\Psi(t) = (1+t)\log(1+t) - t$. Additionally, for the N-function $\Phi(t) = \exp(t^p) - 1$ with $1 < p < \infty$, its complementary N-function admits an explicit expression, but not simple (Rao & Ren, 1991, §2.2).[6]

Inspired by the ST for measures on a graph (Le et al., 2022), we exploit the dual formulation of the 1-order Wasserstein distance to propose the *generalized Sobolev transport* (GST). More precisely, we replace the Lipschitz constraint for the critic function by a constraint involving the graph-based Orlicz-Sobolev space.

Definition 3.2 (Generalized Sobolev transport distance on graph). *Let Φ be an N-function and ω be a nonnegative Borel measure on $\mathbb{G}$. For $\mu, \nu \in \mathcal{P}(\mathbb{G})$, we define*

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

where Ψ is the complementary function of Φ (see (6)).

The Definition 3.2 implies that the GST for probability measures supported on a graph metric space is an instance of the integral probability metric (Müller, 1997).

Computation. We next show that one can compute $\mathcal{GS}_\Phi$ by simply solving a univariate optimization problem.

Theorem 3.3 (GST as univariate optimization problem). *The generalized Sobolev transport $\mathcal{GS}_\Phi(\mu, \nu)$ in Defini-*

---

[6]See §A.8 for rigorous details on this complementary function.

tion 3.2 can be computed as follows:

$$\mathcal{GS}_{\Phi}(\mu,\nu) = \inf_{k>0} \frac{1}{k}\left(1 + \int_{\mathbb{G}} \Phi(k\,|h(x)|)\omega(dx)\right), \quad (7)$$

where $h(x) := \mu(\Lambda(x)) - \nu(\Lambda(x))$ for all $x \in \mathbb{G}$.

The proof is placed in Appendix §A.1.

We next derive the discrete case for the GST in Equation (7).

Corollary 3.4 (Discrete case)**.** *Assume that $\omega(\{x\}) = 0$ for every $x \in \mathbb{G}$, and suppose that $\mu, \nu \in \mathcal{P}(\mathbb{G})$ are supported on nodes in V of graph $\mathbb{G}$.[7] Then, we have*

$$\mathcal{GS}_{\Phi}(\mu,\nu) = \inf_{k>0} \frac{1}{k}\Big[1 + \sum_{e\in E} w_e \Phi(k|\bar{h}(e)|)\Big], \quad (8)$$

where $\bar{h}(e) := \mu(\gamma_e) - \nu(\gamma_e)$ for every edge $e \in E$.

The proof is placed in Appendix §A.2.[8]

From Corollary 3.4, one only needs to simply solve the univariate optimization problem to compute the GST.

Remark 3.5 (GST for non-physical graph)**.** *Similar to the ST, we have assumed that $\mathbb{G}$ is a physical graph in §2.1. However, Corollary 3.4 implies that the GST $\mathcal{GS}_\Phi$ does not depend on the physical assumption when input measures are supported on graph nodes. In particular, it only depends on the graph structure (V, E) and edge weights w_e. Thus, we can apply the GST for non-physical graph $\mathbb{G}$.*

Remark 3.6 (Complementary pairs of N-functions for GST.)**.** *The Definition 3.2 for GST with the N-function Φ involves its complementary N-function Ψ. However, we can reformulate GST as a univariate optimization problem without involving the complementary function Ψ as in Equation (7). Note that in order to obtain the univariate optimization formulation, the complementary function Ψ is finite-valued, which is satisfied by any N-function Φ, i.e., growing faster than linear (see (6)).*

Preprocessing. For the computation of GST in Equation (8), observe that set γ_e (see Equation (1)) can be precomputed for all edge e in $\mathbb{G}$. This preprocessing step only involves the graph $\mathbb{G}$ itself, and is not related to input measures. Moreover, we only need to precompute it once, regardless the number of input pairs of probability measures to be measured by the GST. In particular, we recompute the shortest paths from the root node z_0 to all other input supports (or vertices) by Dijkstra algorithm with the complexity $\mathcal{O}(|E| + |V|\log|V|)$, where $|\cdot|$ is the cardinality of a set.

Remark 3.7 (Sparsity in Problem (8))**.** *Let $supp(\mu)$ be the set of supports of μ. Observe that for any support $x \in$*

---

[7]We discuss an extension for measures supported in $\mathbb{G}$ in §B.4.

[8]Rao & Ren (1991, Theorem 13) derived the necessary and sufficient conditions to obtain the infimum for problem (8).

$supp(\mu)$, its mass is accumulated into $\mu(\gamma_e)$ if and only if $e \subset [z_0, x]$. Let define set $E_{\mu,\nu} \subset E$ as

$$E_{\mu,\nu} := \{e \in E \mid \exists z \in (supp(\mu) \cup supp(\nu)), e \subset [z_0, z]\}.$$

Then in (8), we in fact only need to take the summation over all edges $e \in E_{\mu,\nu}$, i.e., removing all edges $e \in E \setminus E_{\mu,\nu}$.

4. Properties of the GST

In this section, we derive the metric property for the GST and establish a relationship for the GST with different N-functions. Additionally, we draw connections of the GST to the ST, the OW, and OT.

Theorem 4.1 (Metrization)**.** *The generalized Sobolev transport $\mathcal{GS}_\Phi(\mu,\nu)$ is a metric on the space $\mathcal{P}(\mathbb{G})$.*

The proof is placed in Appendix §A.3.

The GST is monotone with respect to the N-function Φ as shown in the next result. Consequently, it may enclose a stronger notion of metrics than the ST for comparing measures on a graph.

Proposition 4.2 (GST with different N-functions)**.** *For any two N-functions Φ_1, Φ_2 satisfying $\Phi_1(t) \le \Phi_2(t)$ for all $t \in \mathbb{R}_+$, we have*

$$\mathcal{GS}_{\Phi_1}(\mu,\nu) \le \mathcal{GS}_{\Phi_2}(\mu,\nu) \text{ for every } \mu,\nu \in \mathcal{P}(\mathbb{G}).$$

The proof is placed in Appendix §A.4.

Connection with Sobolev transport. We next show the connection between the graph-based Orlicz-Sobolev space and the graph-based Sobolev space (Le et al., 2022).

Proposition 4.3. *Let $1 < p < \infty$ and ω be a nonnegative Borel measure on graph $\mathbb{G}$. By taking $\Phi(t) := t^p$ for the N-function, then the associated graph-based Sobolev-Orlicz space is the same as the graph-based Sobolev space of order p, denoted as $W^{1,p}(\mathbb{G},\omega)$.[9] That is,*

$$WL_\Phi(\mathbb{G},\omega) = W^{1,p}(\mathbb{G},\omega).$$

The proof is placed in Appendix §A.5.

By leveraging the result in Proposition 4.3, we next derive the closed-form expression for the univariate optimization problem in (7) for the GST with certain N-functions. This allows us to draw its connection to the ST (Le et al., 2022).

Proposition 4.4 (Connection between GST and ST)**.** *Let $\Phi(t) := \frac{(p-1)^{p-1}}{p^p}t^p$ with $1 < p < \infty$, and let ω be a nonnegative Borel measure on graph $\mathbb{G}$. Then, the GST admits the following closed-form expression:*

$$\mathcal{GS}_\Phi(\mu,\nu) = \left(\int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p\,\omega(dx)\right)^{\frac{1}{p}}, \quad (9)$$

---

[9]See Definition B.1 for the graph-based Sobolev space.

for all measures $\mu, \nu \in \mathcal{P}(\mathbb{G})$. Consequently, we have

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

The proof is placed in Appendix §A.6.

Connection with Orlicz-Wasserstein and OT. In the special case $\Phi(t) = t^p$ with $p \geq 1$, the OW coincides with the p-order Wasserstein (Guha et al., 2023).[10] It is worth noting that OW in Equation (4) is a two-level optimization problem involving the transportation plan and the positive scalar t. Thus, it is more challenging to compute the OW than the classic OT (e.g., with squared Euclidean cost).

Remark 4.5. *Following Guha et al. (2023), the OW is equal to the 1-order Wasserstein if we take $\Phi(t) = t^p$ and then take the limit $p \to 1^+$. In the similar way and by taking the limit $p \to 1^+$, the closed-form GST in Equation (9) is equal to the 1-order ST.[11] With an additional assumption that graph $\mathbb{G}$ is a tree, it follows from Le et al. (2022) that the 1-order ST is in turn equal to the 1-order Wasserstein. Thus, the GST coincides with the OW when $\Phi(t) = t$ and the graph $\mathbb{G}$ is a tree.*

Due to Remark 4.5, the GST can be regarded as a variant of the OW. We summarize it in the following Proposition 4.6.

Proposition 4.6 (Connection between GST and OW). *When $\Phi(t) = t$ and the graph $\mathbb{G}$ is a tree, the generalized Sobolev transport and Orlicz-Wasserstein coincide, i.e.,*

$$\mathcal{GS}_\Phi(\mu, \nu) = W_\Phi(\mu, \nu),$$

for all measures $\mu, \nu \in \mathcal{P}(\mathbb{G})$.

Notice that our obtained results show that one only needs to solve a univariate optimization problem to compute the GST. On the other hand, it is much more challenging to solve the two-level optimization problem (4) for the OW. Notably, a recent realization is that solving exactly OT problems leads to overfitting (Peyré & Cuturi, 2019, §8.4). Therefore, it would be self-defeating to give excessive efforts to optimize the OT problems since it would lead to overfitting within the computation of the OT problems themselves. Consequently, GST can be regarded as a regularization approach for OW.

5. Experiments

In this section, we illustrate that the computation of OW for general N-function Φ is very expensive, and the GST is several-order faster than OW. Then, we evaluate the GST

---

[10]The p-order Wasserstein is reviewed in §B.5. Although $\Phi(t) = t$ is not an N-function due to its linear growth, it can be regarded as the limit $p \to 1^+$ for the function $\Phi(t) = t^p$. In the similar way, Andoni et al. (2018) leverages the Huber function with linear growth for $\Phi(t)$ for Orlicz loss regression in applications.

[11]Note that $\lim_{p \to 1^+} \frac{(p-1)^{p-1}}{p^p} t^p = t$, see §A.7 for the proof.

and show preliminary evidences on its advantages for document classification and for several tasks in TDA.
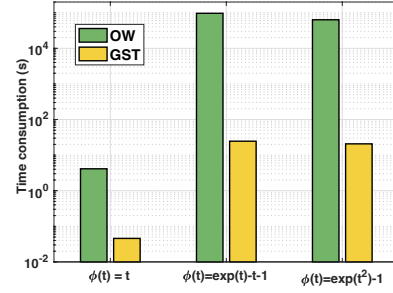

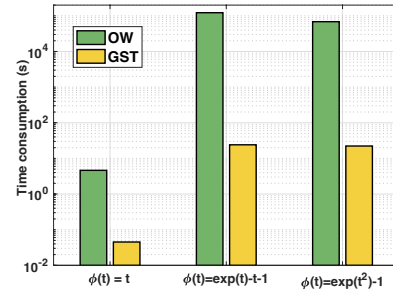*Figure 1.* Time consumption for GST and OW on $\mathbb{G}_{\text{Log}}$.

Figure 2. Time consumption for GST and OW on $\mathbb{G}_{\text{Sqrt}}$.

Document classification. We consider 4 traditional document datasets: TWITTER, RECIPE, CLASSIC, AMAZON. The characteristic properties of these datasets are summarized in Figure 3. Following Le et al. (2022), we use word embedding, e.g., word2vec (Mikolov et al., 2013) pretrained on Google News for documents, to map words into vectors in $\mathbb{R}^{300}$. Additionally, we remove SMART stop words (Salton & Buckley, 1988), and words in documents which are not in the pretrained word2vec. We then regard each document as a probability measure by considering each word in the document as its support in $\mathbb{R}^{300}$, and using word frequency as the mass on that corresponding word.

TDA. We consider two tasks: orbit recognition on the synthesis Orbit dataset (Adams et al., 2017), and object shape classification on MPEG7 dataset (Latecki et al., 2000) as in Le et al. (2022). The characteristic properties of these datasets are summarized in Figure 5. We utilize the persistence diagrams (PD) to represent objects of interest for these tasks. Particularly, PD is a multiset of points in $\mathbb{R}^2$, where each point summarizes a life span (i.e., birth and death time) of a particular topological feature (e.g., connected component, ring, or cavity), extracted by algebraic topology methods (e.g., persistence homology) (Edelsbrunner & Harer, 2008). We consider each PD as a probability measure of the 2-dimensional topological feature data points with a uniform mass.
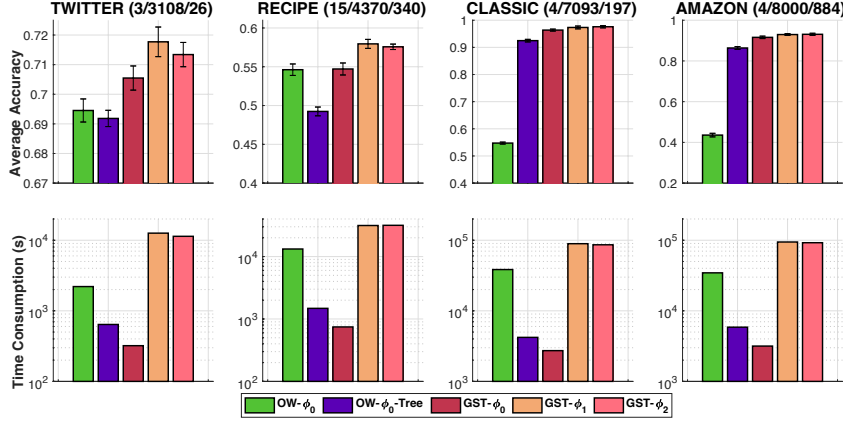
Figure 3. Document classification on graph $\mathbb{G}_{\text{Log}}$. For each dataset, the numbers in the parenthesis are respectively the number of classes; the number of documents; and the maximum number of unique words for each document.
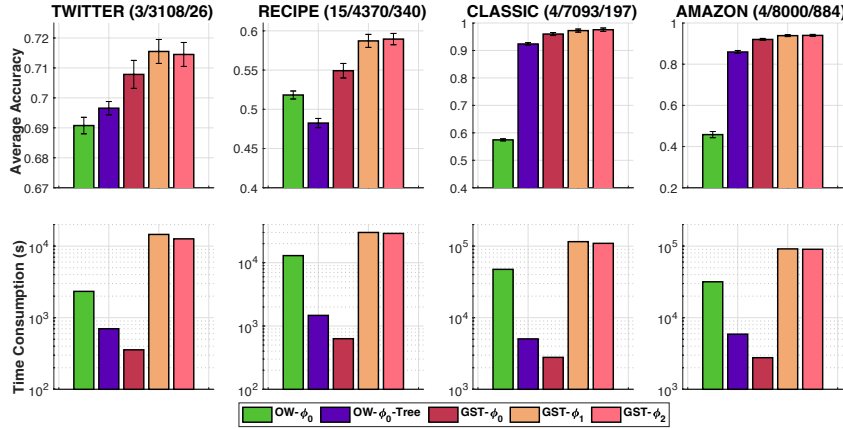

Figure 4. Document classification on graph $\mathbb{G}_{\text{Sqrt}}$.

Graph. We employ the graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ (Le et al., 2022, §5) for our simulations, which empirically satisfy the assumptions in §2.[12] Additionally, we consider $M = 10^2, 10^3, 10^4$ as the number of nodes for these graphs.

N-function. We consider popular N-functions Φ for the GST and the OW: $\Phi_1(t) = \exp(t) - t - 1$, and $\Phi_2(t) = \exp(t^2) - 1$. We also examine the limit case: $\Phi_0(t) = t$.

Optimization algorithm. For general N-functions, we use Guha et al. (2023, Algorithm 1) to compute OW, and a second-order method, e.g., fmincon Trust Region Reflective solver in MATLAB, for solving the *univariate* optimization problem to compute the GST.

Classification. We use support vector machine (SVM) for both document classification and the tasks in TDA. We employ kernel $\exp(-\bar{t}\bar{d}(\cdot, \cdot))$, where $\bar{d}$ is a distance (e.g., GST, OW) for probability measures supported on a graph, and $\bar{t} > 0$. Following Cuturi (2013), we regularize for the Gram

Table 1. The number of pairs on datasets for SVM.

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

matrices by adding a sufficiently large diagonal term for SVM with indefinite kernels. We use 1-vs-1 strategy to carry out SVM for multi-class classification, e.g., with Libsvm.[13] For each dataset, we randomly split it into $70\%/30\%$ for training and test with 10 repeats. We typically choose hyperparameters via cross validation. For kernel hyperparameter, we choose $1/\bar{t}$ from $\{q_s, 2q_s, 5q_s\}$ with $s = 10, 20, \dots, 90$ where q_s is the $s\%$ quantile of a subset of distances observed on a training set. For SVM regularization hyperparameter,

---

[12]See Appendix §B.4 for a review and further discussions.

[13]https://www.csie.ntu.edu.tw/~cjlin/libsvm/

we choose it from $\{0.01, 0.1, 1, 10\}$. For the root node z_0 in graph $\mathbb{G}$, we choose it from a random 10-root-node subset of V in $\mathbb{G}$. For document classification and TDA, reported time consumption includes preprocessing procedures, e.g., computing shortest paths for GST and OW.

We also summarize the number of pairs which we need to compute distances in training and test with kernel SVM for each run in Table 1, e.g., more than 29 million pairs of measures on AMAZON dataset.

5.1. Computation

We compare the time consumption of GST and OW with popular N-functions Φ_1, Φ_2, and the limit case Φ_0.[14]

Setup. We randomly sample 10^4 pairs of measures from AMAZON dataset, and consider graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ with the number of nodes $M = 10^3$.

Results and discussions. We illustrate the time consumption on $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ in Figures 1 and 2 respectively. GST is several-order faster than OW. More concretely, GST is $100\times, 5000\times, 3000\times$ faster than OW for Φ_0, Φ_1, Φ_2 respectively. Especially, for Φ_1 and Φ_2, GST takes less than 25 *seconds*, but OW takes at least 17 *hours*, and up to 33 *hours*.

Notice that the OW with Φ_0 is equal to the OT with graph metric $d_{\mathbb{G}}$ as its ground cost. Additionally, GST with Φ_0 yields a closed-form expression for a fast computation, following the Proposition 4.4 and Remark 4.5. Therefore, the computation of GST and OW with Φ_0 is more efficient than with Φ_1, Φ_2.

5.2. Document classification

Set up. We evaluate GST with Φ_0, Φ_1, Φ_2 as in §5.1 (denoted as GST-Φ_i for $i = 0, 1, 2$). For OW, we only consider Φ_0 (denoted as OW-Φ_0), but exclude Φ_1, Φ_2 due to their heavy computations (illustrated and discussed in §5.1). Following Le et al. (2022) and Proposition 4.6, we also consider a special case for OW-Φ_0, where we randomly sample a tree from the given graph $\mathbb{G}$ (denoted as OW-Φ_0-Tree). Notice that OW-Φ_0-Tree admits a closed-form expression for a fast computation (Le et al., 2019). We report results on graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ with the number of nodes $M = 10^4$. Further empirical results with different values for M are placed in Appendix §B.5.

Results and discussions. We illustrate SVM results and the time consumption of kernel matrices on graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ in Figures 3 and 4 respectively. The performances of GST with all Φ functions compare favorably to those of OW. Additionally, the computation of GST-Φ_0 is several-order

---

[14] Notice that GST-Φ_0 is equal to 1-order Sobolev transport (see Remark 4.5), and OW-Φ_0 is equal to the OT on a graph (i.e., OT with graph metric ground cost).

faster than OW-Φ_0. We also reemphasize that it is prohibitively expensive to evaluate OW with Φ_1, Φ_2 functions on many pairs of measures (See Table 1 and §5.1). GST-Φ_1 and GST-Φ_2 improve performances of GST-Φ_0, but their computational time is several-order higher (i.e., GST-Φ_0 has a closed-form expression for a fast computation). Thus, it may imply that Orlicz geometric structure in GST may be useful for document classification. Recall that OW-Φ_0-Tree uses a partial information of $\mathbb{G}$, while OW-Φ_0 uses information of the whole $\mathbb{G}$. On the other hand, kernel for OW-Φ_0-Tree is positive definite, while kernel for OW-Φ_0 is indefinite. Performances of OW-Φ_0-Tree are worse than those of OW-Φ_0 in TWITTER, RECIPE, but better than those of OW-Φ_0 in CLASSIC, AMAZON, which agrees with observations in Le et al. (2022).

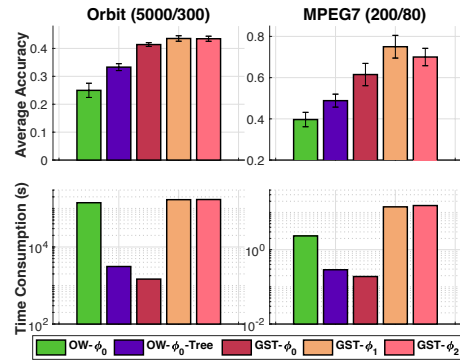*Figure 5.* TDA on graph $\mathbb{G}_{\text{Log}}$. For each dataset, the numbers in the parenthesis are respectively the number of PD; and the maximum number of points in PD.
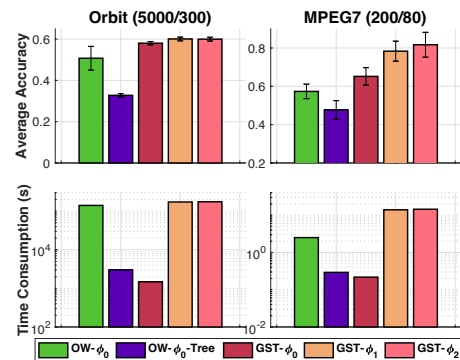

Figure 6. TDA on graph $\mathbb{G}_{\text{Sqrt}}$.

5.3. Topological Data Analysis

Set up. We evaluate the same distances as in document classification, i.e., GST-Φ_i for $i = 0, 1, 2$; OW-Φ_0; and OW-Φ_0-Tree. We report results on graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ with $M = 10^4$ for Orbit dataset, and $M = 10^3$ for MPEG7 dataset due to its small size. We also include further empirical results with different values for M in Appendix §B.5.

Results and discussions. We illustrate SVM results and the time consumption of kernel matrices on graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ in Figures 5 and 6 respectively. We have similar observations as for document classification. GST is several-order faster than OW (for the same Φ function), and improve performances of OW. Orlicz geometry structure is also helpful for TDA. Although only using a partial information of $\mathbb{G}$, positive definiteness may play an important role for OW-Φ_0-Tree to improve performances of OW-Φ_0 in TDA, also observed in Le et al. (2022).

6. Conclusion

We propose the generalized Sobolev transport (GST) for probability measures supported on a graph metric space. This is achieved by leveraging a special family of convex functions (i.e., the set of N-functions) to extend the Sobolev transport (ST), which is a scalable variant of OT on a graph. This novel approach enables us to adopt Orlicz geometric structure for GST. Consequently, GST can inherit advantages from both ST and the Orlicz-Wasserstein (OW). An important feature is that unlike OW which involves a complex two-level optimization problem, one can simply solve a univariate optimization problem for the GST computation. Thus, GST is a computationally efficient variant for the computationally-demanding OW, which paves a way to use OW in applications. For future work, it is interesting to extend the GST for measures possibly having different total mass (i.e., unbalanced setting), which is essential for applications having noisy supports or outliers in input measures (Frogner et al., 2015; Balaji et al., 2020; Mukherjee et al., 2021; Fatras et al., 2021).

Acknowledgements

Impact Statement

The paper presents a generalized version for the Sobolev transport (ST) which is a scalable variant of optimal transport for probability measures on a graph. By leveraging a special family of convex functions, we can adopt Orlicz geometric structure for the generalized Sobolev transport (GST). Furthermore, GST can be regarded as a computationally efficient variant of the computationally-demanding Orlicz-Wasserstein (OW), which possesses implicit essences to advance certain results in machine learning. Therefore, the proposed GST can inherit the merits from both ST and OW. To our knowledge, there are no foresee potential societal consequences of our work.

References

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.

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

Altschuler, J. M. and Chewi, S. Faster high-accuracy log-concave sampling via algorithmic warm starts. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 2169–2176, 2023.

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

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.

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.

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

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

Bunne, C., Hsieh, Y.-P., Cuturi, M., and Krause, A. The Schrödinger bridge between Gaussian measures has a closed form. In *International Conference on Artificial Intelligence and Statistics*, pp. 5802–5833. PMLR, 2023a.

Bunne, C., Stark, S. G., Gut, G., Del Castillo, J. S., Levesque, M., Lehmann, K.-V., Pelkmans, L., Krause, A., and Rätsch, G. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023b.

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

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

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.

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

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

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

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.

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.

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.

Hua, X., Nguyen, T., Le, T., Blanchet, J., and Nguyen, V. A. Dynamic flows on curved space generated by labeled data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 3803–3811, 2023.

Huber, P. J. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

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

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

Korotin, A., Selikhanovych, D., and Burnaev, E. Neural optimal transport. In *The Eleventh International Conference on Learning Representations*, 2023.

Korotin, A., Gushchin, N., and Burnaev, E. Light Schröodinger bridge. In *The Eleventh International Conference on Learning Representations*, 2024.

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.

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

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.

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*, pp. 3934–3942. PMLR, 2021a.

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

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. PMLR, 2022.

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. PMLR, 2023.

Le, T., Nguyen, T., and Fukumizu, K. Optimal transport for measures with noisy tree metric. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024.

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

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

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

Mahey, G., Chapel, L., Gasso, G., Bonet, C., and Courty, N. Fast optimal transport through sliced Wasserstein generalized geodesics. 2023.

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

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

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.

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

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

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

Nguyen, H., Le, K., Nguyen, Q. M., Pham, T., Bui, H., and Ho, N. On robust optimal transport: Computational complexity and barycenter computation. *Advances in Neural Information Processing Systems*, 34:21947–21959, 2021a.

Nguyen, K., Ren, T., and Ho, N. Markovian sliced Wasserstein distances: Beyond independent projections. *Advances in Neural Information Processing Systems*, 2023a.

Nguyen, K., Zhang, S., Le, T., and Ho, N. Sliced Wasserstein with random-path projecting directions. In *International Conference on Machine Learning (ICML)*, 2024.

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*, 2023b.

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

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

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

Nietert, S., Goldfeld, Z., and Shafiee, S. Outlier-robust Wasserstein DRO. *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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.

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

Pham, T., Shimizu, S., Hino, H., and Le, T. Scalable counterfactual distribution estimation in multivariate causal models. In *Conference on Causal Learning and Reasoning (CLeaR)*, 2024.

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.

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

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

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

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

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.

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

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

Tam, E. and Dunson, D. Multiscale graph comparison via the embedded laplacian discrepancy. *arXiv:2201.12064*, 2022.

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

Tong, A., Huang, J., Wolf, G., Van Dijk, D., and Krishnaswamy, S. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International conference on machine learning*, pp. 9526–9536. PMLR, 2020.

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

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

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

Zhang, Z. Parameter estimation techniques: A tutorial with application to conic fitting. *Image and vision Computing*, 15(1):59–76, 1997.

Supplement to "Generalized Sobolev Transport for Probability Measures on a Graph"

In §A of this appendix, we give the detailed proofs for our theoretical results. Further results and discussions are given in §B.

A. Detailed Proofs

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

A.1. Proof for Theorem 3.3

Proof. Let us consider a critic function $f \in WL_\Psi(\mathbb{G}, \omega)$. Then by Definition 3.1, we have

$$f(x) = f(z_0) + \int_{[z_0, x]} f'(y) \omega(\mathrm{d}y) \quad \text{for} \quad x \in \mathbb{G}. \tag{10}$$

For convenience, let $\mathbf{1}_{[z_0, x]}(y)$ denote the indicator function of the shortest path $[z_0, x]$, i.e.,

$$\mathbf{1}_{[z_0, x]}(y) = \begin{cases} 1 & \text{if } y \in [z_0, x] \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Using (10)–(11) and due to $\mu(\mathbb{G}) = 1$, we can rewrite $\int_\mathbb{G} f(x) \mu(\mathrm{d}x)$ as

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

$$= f(z_0) + \int_\mathbb{G} \int_\mathbb{G} \mathbf{1}_{[z_0, x]}(y) \, f'(y) \omega(\mathrm{d}y) \mu(\mathrm{d}x).$$

It then follows by applying Fubini's theorem to interchange the order of integration in the above last integral that

$$\int_\mathbb{G} f(x) \mu(\mathrm{d}x) = f(z_0) + \int_\mathbb{G} \int_\mathbb{G} \mathbf{1}_{[z_0, x]}(y) \, f'(y) \mu(\mathrm{d}x) \omega(\mathrm{d}y)$$

$$= f(z_0) + \int_\mathbb{G} \left[\int_\mathbb{G} \mathbf{1}_{[z_0, x]}(y) \, \mu(\mathrm{d}x) \right] f'(y) \omega(\mathrm{d}y).$$

Owing to the definition of $\Lambda(y)$ in (1), the above expression is equivalent to

$$\int_\mathbb{G} f(x) \mu(\mathrm{d}x) = f(z_0) + \int_\mathbb{G} f'(y) \mu(\Lambda(y)) \, \omega(\mathrm{d}y).$$

By exactly the same arguments, we also have

$$\int_\mathbb{G} f(x) \nu(\mathrm{d}x) = f(z_0) + \int_\mathbb{G} f'(y) \nu(\Lambda(y)) \, \omega(\mathrm{d}y).$$

As a consequence, the generalized Sobolev transport in Definition 3.2 can be rewritten as

$$\mathcal{GS}_\Phi(\mu, \nu) = \sup_{f \in \mathbb{F}} \left| \int_\mathbb{G} f'(x) \big[\mu(\Lambda(x)) - \nu(\Lambda(x)) \big] \omega(\mathrm{d}x) \right|, \tag{12}$$

where $\mathbb{F} := \left\{ f \in WL_\Psi(\mathbb{G}, \omega) : \|f'\|_{L_\Psi} \leq 1 \right\}$ with $\| \cdot \|_{L_\Psi}$ being the Luxemburg norm defined in (3).

Next, on one hand it is clear that $\{f' : f \in \mathbb{F}\} \subset \{g \in L_\Psi(\mathbb{G}, \omega) : \|g\|_{L_\Psi} \leq 1\}$. On the other hand, for any $g \in L_\Psi(\mathbb{G}, \omega)$ we have $g = f'$ with $f(x) := \int_{[z_0, x]} g(y) \omega(\mathrm{d}y) \in WL_\Psi(\mathbb{G}, \omega)$. Therefore, we conclude that $\{f' : f \in \mathbb{F}\} = \{g \in L_\Psi(\mathbb{G}, \omega) : \|g\|_{L_\Psi} \leq 1\}$. As a consequence, we can rewrite identity (12) as

$$\mathcal{GS}_\Phi(\mu, \nu) = \sup_{g \in L_\Psi(\mathbb{G}, \omega) : \|g\|_{L_\Psi} \leq 1} \left| \int_\mathbb{G} g(x) h(x) \, \omega(\mathrm{d}x) \right|, \tag{13}$$

where $h(x) := \mu(\Lambda(x)) - \nu(\Lambda(x))$ for all $x \in \mathbb{G}$.

Recall from Definition 3.2 that $\Psi : \mathbb{R}_+ \to \mathbb{R}_+$ is the complementary function of Φ. Therefore, we obtain from (13) and (Rao & Ren, 1991, Proposition 10, pp.81) that

$$\mathcal{GS}_\Phi(\mu, \nu) = \|h\|_\Phi \tag{14}$$

with $\|h\|_\Phi$ being the Orlicz norm define by (see (Rao & Ren, 1991, Definition 2, pp.58))

$$\|h\|_\Phi := \sup \left\{ \int_{\mathbb{G}} |h(x)g(x)| \omega(\mathrm{d}x) : \int_{\mathbb{G}} \Psi(|g(x)|) \omega(\mathrm{d}x) \le 1 \right\}. \tag{15}$$

By applying (Rao & Ren, 1991, Theorem 13, pp.69), we have

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

This together with (14) yields

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

This completes the proof of the theorem. ∎

A.2. Proof for Corollary 3.4

Proof. By Theorem 3.3, we have

$$\mathcal{GS}_\Phi(\mu, \nu) = \inf_{k>0} \frac{1}{k} \left(1 + \int_{\mathbb{G}} \Phi(k\,|h(x)|) \omega(\mathrm{d}x) \right) \tag{16}$$

with $h(x) := \mu(\Lambda(x)) - \nu(\Lambda(x))$. We next compute the integral in (16).

Give two data points u, v in $\mathbb{R}^n$, let $\langle u, v \rangle$ be the line segment in $\mathbb{R}^n$ connecting the two points u, v, and denote (u, v) as the same line segment but without its two end-points.

Using the assumption $\omega(\{x\}) = 0$ for every $x \in \mathbb{G}$, we have

$$\int_{\mathbb{G}} \Phi(k\,|h(x)|) \omega(\mathrm{d}x) = \sum_{e=\langle u,v \rangle \in E} \int_{(u,v)} \Phi(k\,|h(x)|) \omega(\mathrm{d}x).$$

As the input measures μ and ν are assumed to be supported on nodes V of graph $\mathbb{G}$, we also have $h(x) = \mu(\Lambda(x)) - \nu(\Lambda(x)) = \mu(\Lambda(x) \setminus (u,v)) - \nu(\Lambda(x) \setminus (u,v))$ for every edge $e = \langle u, v \rangle \in E$. Therefore, the above identity can be further rewritten as

$$\int_{\mathbb{G}} \Phi(k\,|h(x)|) \omega(\mathrm{d}x) = \sum_{e=\langle u,v \rangle \in E} \int_{(u,v)} \Phi(k\,|\mu(\Lambda(x) \setminus (u,v)) - \nu(\Lambda(x) \setminus (u,v))|) \omega(\mathrm{d}x). \tag{17}$$

Consider edge e between two nodes $u, v \in V$ of graph $\mathbb{G}$, i.e., $e = \langle u, v \rangle$. For $x \in (u,v)$, 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). It follows that $\Lambda(x) \setminus (u,v) = \gamma_e$, and thus we deduce from (17) that

$$\int_{\mathbb{G}} \Phi(k\,|h(x)|) \omega(\mathrm{d}x) = \sum_{e=\langle u,v \rangle \in E} \int_{(u,v)} \Phi(k\,|\mu(\gamma_e) - \nu(\gamma_e)|) \omega(\mathrm{d}x)$$

$$= \sum_{e=\langle u,v \rangle \in E} \Phi(k\,|\mu(\gamma_e) - \nu(\gamma_e)|) \int_{(u,v)} \omega(\mathrm{d}x)$$

$$= \sum_{e \in E} w_e\, \Phi(k\,|\mu(\gamma_e) - \nu(\gamma_e)|).$$

14

This together with (16) yields

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

Thus, the proof is complete. ∎

A.3. Proof for Theorem 4.1

Proof. We will prove that the generalized Sobolev transport $\mathcal{GS}_\Phi(\mu, \nu)$ satisfies: (i) nonnegativity, (ii) indiscernibility, (iii) symmetry, and (iv) triangle inequality.

(i) Nonnegativity. By choosing $f = 0$ in Definition 3.2, we see that $\mathcal{GS}_\Phi(\mu, \nu) \geq 0$ for every (μ, ν) in $\mathcal{P}(\mathbb{G}) \times \mathcal{P}(\mathbb{G})$. Therefore, the generalized Sobolev transport is nonnegative.

(ii) Indiscernibility. Assume that $\mathcal{GS}_\Phi(\mu, \nu) = 0$, then we have

$$\int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(x)\nu(\mathrm{d}x) = 0, \tag{18}$$

for all $f \in WL_\Psi(\mathbb{G}, \omega)$ satisfying the constraint $\|f'\|_{L_\Psi} \leq 1$.

Now let $g \in WL_\Psi(\mathbb{G}, \omega)$ be any nonconstant function. Then $c := \|g'\|_{L_\Psi} > 0$. Then by taking $f := \frac{g}{c}$, we have $f \in WL_\Psi(\mathbb{G}, \omega)$ with $\|f'\|_{L_\Psi} = \|\frac{g'}{c}\|_{L_\Psi} = \frac{1}{c}\|g'\|_{L_\Psi} = 1$. Hence, it follows from (18) that

$$\int_{\mathbb{G}} \frac{g(x)}{c}\mu(\mathrm{d}x) - \int_{\mathbb{G}} \frac{g(x)}{c}\nu(\mathrm{d}x) = 0,$$

which implies that

$$\int_{\mathbb{G}} g(x)\mu(\mathrm{d}x) = \int_{\mathbb{G}} g(x)\nu(\mathrm{d}x). \tag{19}$$

Thus, we have shown that (19) holds true for every nonconstant function $g \in WL_\Psi(\mathbb{G}, \omega)$. But (19) is also obviously true for any constant function g. So, we in fact obtain

$$\int_{\mathbb{G}} g(x)\mu(\mathrm{d}x) = \int_{\mathbb{G}} g(x)\nu(\mathrm{d}x) \qquad \text{for every} \quad g \in WL_\Psi(\mathbb{G}, \omega),$$

which gives $\mu = \nu$ as desired.

(iii) Symmetry. This property is obvious from Definition 3.2 as the value $\mathcal{GS}_\Phi(\mu, \nu)$ is unchanged when the role of μ and ν is interchanged. That is, $\mathcal{GS}_\Phi(\mu, \nu) = \mathcal{GS}_\Phi(\nu, \mu)$.

(iv) Triangle inequality. Let μ, ν, σ be probability measures in $\mathcal{P}(\mathbb{G})$. Then for any function $f \in WL_\Psi(\mathbb{G}, \omega)$ satisfying $\|f'\|_{L_\Psi} \leq 1$, we have

$$\left| \int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(x)\nu(\mathrm{d}x) \right| = \left| \left[\int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(x)\sigma(\mathrm{d}x) \right] + \left[\int_{\mathbb{G}} f(x)\sigma(\mathrm{d}x) - \int_{\mathbb{G}} f(x)\nu(\mathrm{d}x) \right] \right|$$

$$\leq \left| \int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(x)\sigma(\mathrm{d}x) \right| + \left| \int_{\mathbb{G}} f(x)\sigma(\mathrm{d}x) - \int_{\mathbb{G}} f(x)\nu(\mathrm{d}x) \right|$$

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

By taking the supremum over f, this implies that $\mathcal{GS}_\Phi(\mu, \nu) \leq \mathcal{GS}_\Phi(\mu, \sigma) + \mathcal{GS}_\Phi(\sigma, \nu)$.

Due to the above properties, we conclude that the generalized Sobolev transport $\mathcal{GS}_\Phi(\mu, \nu)$ is a metric on the space $\mathcal{P}(\mathbb{G})$ of probability measures on graph $\mathbb{G}$.

∎

A.4. Proof for Proposition 4.2

Proof. Let Ψ_1 and Ψ_2 be respectively the complementary functions of Φ_1 and Φ_2 according to definition (6).

Let $t \in \mathbb{R}_+$ be arbitrary. Since $\Phi_1 \leq \Phi_2$ we have:

$$
\begin{aligned}
at - \Phi_1(a) &\geq at - \Phi_2(a) \text{ for every } a \in \mathbb{R}_+, \\
\Rightarrow \sup_{a \geq 0} (at - \Phi_1(a)) &\geq \sup_{a \geq 0} (at - \Phi_2(a)).
\end{aligned}
$$

This implies that

$$\Psi_1(t) \geq \Psi_2(t), \text{ for all } t \in \mathbb{R}_+.$$

It follows that $L_{\Psi_1}(\mathbb{G}, \omega) \subset L_{\Psi_2}(\mathbb{G}, \omega)$ and $WL_{\Psi_1}(\mathbb{G}, \omega) \subset WL_{\Psi_2}(\mathbb{G}, \omega)$. Moreover, for any fixed Orlicz function f' and any number $t > 0$, we have

$$\int_{\mathbb{G}} \Psi_1 \left(\frac{|f'(x)|}{t} \right) \omega(\mathrm{d}x) \geq \int_{\mathbb{G}} \Psi_2 \left(\frac{|f'(x)|}{t} \right) \omega(\mathrm{d}x).$$

Consequently, we obtain

$$\left\{ t > 0 \mid \int_{\mathbb{G}} \Psi_1 \left(\frac{|f'(x)|}{t} \right) \omega(\mathrm{d}x) \leq 1 \right\} \subset \left\{ t > 0 \mid \int_{\mathbb{G}} \Psi_2 \left(\frac{|f'(x)|}{t} \right) \omega(\mathrm{d}x) \leq 1 \right\}.$$

Since the infimum of a set is smaller than or equal to the infimum of its subset, we deduce that

$$\|f'\|_{L_{\Psi_1}} \geq \|f'\|_{L_{\Psi_2}}.$$

It follows from this and $WL_{\Psi_1}(\mathbb{G}, \omega) \subset WL_{\Psi_2}(\mathbb{G}, \omega)$ that

$$\left\{ f \mid f \in WL_{\Psi_1}(\mathbb{G}, \omega), \|f'\|_{L_{\Psi_1}(\mathbb{G}, \omega)} \leq 1 \right\} \subset \left\{ f \mid f \in WL_{\Psi_2}(\mathbb{G}, \omega), \|f'\|_{L_{\Psi_2}(\mathbb{G}, \omega)} \leq 1 \right\}.$$

Since the supremum of a set is larger than or equal to the supremum of its subset, we conclude that

$$\mathcal{GS}_{\Phi_1}(\mu, \nu) \leq \mathcal{GS}_{\Phi_2}(\mu, \nu)$$

for any input measures μ, ν in $\mathcal{P}(\mathbb{G})$. Therefore, the proof is completed.

We further note that one can directly leverage the result in Theorem 3.3, where generalized Sobolev transport can be computed by solving univariate optimization problem, to simplify the proof. ∎

A.5. Proof for Proposition 4.3

Proof. For the N-function $\Phi(t) = t^p$ with $1 < p < \infty$, we have

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

where $L^p(\mathbb{G}, \omega)$ is the standard L^p functional space which is reviewed in §B.1 (appendix).

Thus, following Definition 3.1 and the definition of graph-based Sobolev transport (Le et al., 2022, Definition 3.1) (a review is given in Definition B.1 (appendix)), we have

$$WL_\Phi(\mathbb{G}, \omega) = W^{1,p}(\mathbb{G}, \omega).$$

Thus, the proof is complete.

∎

A.6. Proof for Proposition 4.4

Proof. Let $h(x) := \mu(\Lambda(x)) - \nu(\Lambda(x))$ for $x \in \mathbb{G}$. Then since $\Phi(t) = \frac{(p-1)^{p-1}}{p^p} t^p$ with $1 < p < \infty$, we obtain from Theorem 3.3 that

$$
\begin{aligned}
\mathcal{GS}_\Phi(\mu, \nu) &= \inf_{k>0} \frac{1}{k} \left(1 + \int_\mathbb{G} \frac{(p-1)^{p-1}}{p^p} k^p \, |h(x)|^p \, \omega(\mathrm{d}x) \right) \\
&= \inf_{k>0} F(k),
\end{aligned}
\tag{20}
$$

where $F(k) := \frac{1}{k} + \frac{(p-1)^{p-1}}{p^p} k^{p-1} \int_\mathbb{G} |h(x)|^p \, \omega(\mathrm{d}x)$ for $k > 0$.

Let us consider the following two possibilities:

Case 1: $\int_\mathbb{G} |h(x)|^p \, \omega(\mathrm{d}x) = 0$. In that case, we have from (20) that

$$
\mathcal{GS}_\Phi(\mu, \nu) = \inf_{k>0} \frac{1}{k} = 0 = \left(\int_\mathbb{G} |h(x)|^p \, \omega(\mathrm{d}x) \right)^{\frac{1}{p}} = \mathcal{S}_p(\mu, \nu).
$$

Case 2: $\int_\mathbb{G} |h(x)|^p \, \omega(\mathrm{d}x) \neq 0$. Then $\lim_{k \to 0^+} F(k) = \lim_{k \to +\infty} F(k) = +\infty$. Therefore, it follows from (20) that

$$
\mathcal{GS}_\Phi(\mu, \nu) = F(k_0)
\tag{21}
$$

for some finite number $k_0 \in (0, +\infty)$ satisfying $F'(k_0) = 0$. As $F'(k) = -\frac{1}{k^2} + \left(\frac{p-1}{p} \right)^p k^{p-2} \int_\mathbb{G} |h(x)|^p$, we can solve the equation $F'(k_0) = 0$ for k_0 to obtain

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

Plugging this value of k_0 into Equation (21) and using the formula for $F(k)$, we get

$$
\begin{aligned}
\mathcal{GS}_\Phi(\mu, \nu) &= \frac{1}{k_0} \left(1 + \frac{(p-1)^{p-1}}{p^p} k_0^p \int_\mathbb{G} |h(x)|^p \right) \\
&= \frac{p-1}{p} \left(\int_\mathbb{G} |h(x)|^p \, \omega(\mathrm{d}x) \right)^{\frac{1}{p}} \left(1 + \frac{(p-1)^{p-1}}{p^p} \frac{1}{\frac{(p-1)^p}{p^p} \left(\int_\mathbb{G} |h(x)|^p \, \omega(\mathrm{d}x) \right)} \int_\mathbb{G} |h(x)|^p \, \omega(\mathrm{d}x) \right) \\
&= \left(\int_\mathbb{G} |h(x)|^p \, \omega(\mathrm{d}x) \right)^{\frac{1}{p}} \\
&= \mathcal{S}_p(\mu, \nu).
\end{aligned}
$$

Thus, we have shown that $\mathcal{GS}_\Phi(\mu, \nu) = \mathcal{S}_p(\mu, \nu)$ in both cases, and hence the proof is complete. ∎

Remark A.1. *For the N-function $\Phi(t) = t^q$ where $1 < q < \infty$, its complementary function is $\Psi(t) = \frac{(p-1)^{p-1}}{p^p} t^p$ where p is the conjugate of q, i.e., $\frac{1}{q} + \frac{1}{p} = 1$.*

A.7. Proof for $\lim_{p \to 1^+} \frac{(p-1)^{p-1}}{p^p} t^p = t$ in Remark 4.5

We will prove that $\lim_{p \to 1^+} \frac{(p-1)^{p-1}}{p^p} t^p = t$ which is used in Remark 4.5. For this and since $\lim_{p \to 1^+} \frac{t^p}{p^p} = t$, it is enough to show that

$$
\lim_{p \to 1^+} (p-1)^{p-1} = 1.
$$

By taking logarithm, this in turn is equivalent to proving that

$$
\lim_{p \to 1^+} \ln (p-1)^{p-1} = 0.
\tag{22}
$$

Proof of (22).

$$\lim_{p \to 1^+} \ln(p-1)^{p-1} = \lim_{p \to 1^+} \frac{\ln(p-1)}{1/(p-1)} = \lim_{p \to 1^+} \frac{1/(p-1)}{-1/(p-1)^2},$$

where we apply the L'Hopital's rule for the last equality. It follows that $\lim_{p \to 1^+} \ln(p-1)^{p-1} = \lim_{p \to 1^+} -(p-1) = 0$ as desired. Hence, the proof is complete. ∎

A.8. Complementary function for $\Phi(t) = \exp(t^p) - 1$ with $1 < p < \infty$

For the given N-function $\Phi(t) = \exp(t^p) - 1$ with $1 < p < \infty$, its complementary function Ψ is computed as follow

$$\Psi(t) = \sup_{a \geq 0} (at - \Phi(a)). \tag{23}$$

Let us fix $t \geq 0$ and denote the objective function in Equation (23) as $F(a)$. Then $F(a) = at - \exp(a^p) + 1$, and we have

$$F'(a) = t - pa^{p-1} \exp(a^p),$$

$$F''(a) = -p(p-1)a^{p-2} \exp(a^p) - p^2 a^{2(p-1)} \exp(a^p) = -\left[p(p-1)a^{p-2} + p^2 a^{2(p-1)} \right] \exp(a^p).$$

Since $1 < p < \infty$, we obtain $F''(a) < 0$ for all $a > 0$. Thus, $F'(a)$ is a strictly decreasing function on $(0, +\infty)$. Additionally, we have

$$F'(0) = t \geq 0,$$

$$\lim_{a \to +\infty} F'(a) = -\infty.$$

Therefore, $F'(a) = 0$ has a unique root, which is denoted by a^*. Then, the complementary function Ψ is

$$\Psi(t) = \sup_{a \geq 0} F(a) = F(a^*) = a^* t - \exp((a^*)^p) - 1.$$

B. Further Results and Discussions

In this section, we give further results and discussions.

Notations. Let $\langle x, y \rangle$ be the line segment in $\mathbb{R}^n$ connecting two points x, y, and denote (x, y) as the same line segment but without its two end-points.

B.1. On Sobolev transport

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.

Functions $f_1, f_2 \in L^p(\mathbb{G}, \omega)$ are considered to be the same if $f_1(x) = f_2(x)$ for ω-a.e. $x \in \mathbb{G}$.

Then, $L^p(\mathbb{G}, \omega)$ is a normed space with the norm defined by

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

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

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

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

Definition B.1 (Graph-based Sobolev space (Le et al., 2022))**.** *Let ω be a nonnegative Borel measure on $\mathbb{G}$, and let $1 \leq p \leq \infty$. A continuous function $f : \mathbb{G} \to \mathbb{R}$ is 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*

$$f(x) - f(z_0) = \int_{[z_0, x]} h(y) \omega(\mathrm{d}y) \quad \forall x \in \mathbb{G}. \tag{24}$$

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

Proposition B.2 (Closed-form expression of ST (Le et al., 2022))**.** *Let ω be any nonnegative Borel measure on $\mathbb{G}$, and let $1 \le p \le \infty$. Then, we have*

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

where $\Lambda(x)$ is the subset of $\mathbb{G}$ defined by Equation (1)*.*

Definition B.3 (Length measure (Le et al., 2022))**.** *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 $s \le t$. Here, $\langle x, y\rangle$ is the line segment in e connecting x and y.*

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

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

Lemma B.4 (ω^* is the length measure on graph (Le et al., 2022))**.** *Suppose that $\mathbb{G}$ has no short cuts, namely, any edge e is a shortest path connecting its two end-points. Then, ω^* is a length measure in the sense that*

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

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

B.2. On Wasserstein distance

We review the definition of the p-Wasserstein distance with graph metric cost for measures on graph $\mathbb{G}$.

Definition B.5. *Let $1 \le p < \infty$, suppose that μ and ν are two nonnegative Borel measures on $\mathbb{G}$ satisfying $\mu(\mathbb{G}) = \nu(\mathbb{G}) = 1$. Then, the p-order Wasserstein distance between μ and ν is defined as follows:*

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

where

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

where γ_1, γ_2 are the first and second marginals of γ respectively.

B.3. On generalized Sobolev transport

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

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

Equivalence (Adams & Fournier, 2003, §8.17) (Musielak, 2006, §13.11). The Luxemburg norm is equivalent to the Orlicz norm

$$\|f\|_{L_\Phi} \le \|f\|_\Phi \le 2 \|f\|_{L_\Phi}, \tag{25}$$

where we recall that the Luxemburg norm $\|\cdot\|_{L_\Phi}$ is defined in Equation (3), and the Orlicz norm $\|\cdot\|_\Phi$ is defined in Equation (15) (see (Rao & Ren, 1991, Definition 2, pp.58)).

Generalized Hölder inequality. Let Φ, Ψ be a pair of complementary N-functions, then Hölder inequality w.r.t. Luxemburg norm (Adams & Fournier, 2003, §8.11) is as follows:

$$\left| \int_{\mathbb{G}} f(x)g(x)\omega(dx) \right| \le 2 \|f\|_{L_\Phi} \|g\|_{L_\Psi}. \tag{26}$$

Additionally, Hölder inequality w.r.t. Luxemburg norm and Orlicz norm (Musielak, 2006, §13.13) is as follows:

$$\left| \int_{\mathbb{G}} f(x)g(x)\omega(dx) \right| \le \|f\|_{L_\Phi} \|g\|_\Psi. \tag{27}$$

B.4. Further discussions

For completeness, we recall important discussions on the underlying graph in Le et al. (2022) for ST, since they are also applied and/or adapted for the GST.

OT problems for measures on a graph. In this work, our considered problem is to compute OT distance between two input probability measures supported on the *same* graph. We distinguish this line of research to the OT problem between *two (different) input graphs*, considered in Petric Maretic et al. (2019); Dong & Sawin (2020); Tam & Dunson (2022), where its goal is to compute the distance between two input graphs by using OT approach.

Algorithms for the univariate optimization to compute GST. For empirical simulations, we leverage the `fmincon` in MATLAB with the *trust-region-reflective* algorithm since it is easy to compute the gradient and Hessian for the objective function for the univariate optimization problem.

Path length for points in $\mathbb{G}$ **(Le et al., 2023).** We can canonically measure a path length connecting any two points $x, y \in \mathbb{G}$ where x, y are not necessary to be nodes in V. Indeed, for $x, y \in \mathbb{R}^n$ belonging to the same edge $e = \langle u, v \rangle$ which connects two nodes $u, v \in V$, we have

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

for some numbers $t, s \in [0, 1]$. Therefore, the length of the path connecting x, y along the edge e (i.e., the line segment $\langle x, y \rangle$) is defined by $|t - s| w_e$.

Thus, the length for an arbitrary path in $\mathbb{G}$ can be similarly defined by breaking down into pieces over edges and summing over their corresponding lengths (Le et al., 2022).

Extension to measures supported on $\mathbb{G}$**.** Similar to ST (Le et al., 2022), the discrete case of GST in Equation (8) can be extended for measures with finite supports on $\mathbb{G}$ (i.e., measures which may have supports on edges) by using the same strategy to measure a path length for points in $\mathbb{G}$ (discussed in the above paragraph). More precisely, we break down edges containing supports into pieces and sum over their corresponding values instead of the sum over edges as in Equation (8).

About the assumption of uniqueness property of the shortest paths on $\mathbb{G}$**.** As discussed in (Le et al., 2022) for the ST, note that $w_e \in \mathbb{R}$ 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 (with a high probability, lengths of paths connecting any two nodes in graph $\mathbb{G}$ are different).

Additionally, for some special graph, e.g., a grid of nodes, there is *no* unique-path root node for such graph. However, by perturbing each node (and/or perturbing lengths of edges in case $\mathbb{G}$ is a non-physical graph) with a small deviation, we can obtain a graph satisfying the unique-path root node assumption.

About the generalized Sobolev transport (GST). Similar to the ST (Le et al., 2022), we assume that the graph metric space (i.e., the graph structure) is given, and leave the question to learn an optimal graph metric structure from data for future work.

About graphs $\mathbb{G}_{\mathbf{Log}}$ **and** $\mathbb{G}_{\mathbf{Sqrt}}$ **(Le et al., 2022).** We use a clustering method, e.g., the farthest-point clustering, to partition supports of measures into at most M clusters.[15] Then, let V be the set of centroids of these clusters, i.e., graph vertices. For edges, in graph $\mathbb{G}_{\mathbf{Log}}$, we randomly choose $M \log(M)$ edges; and $M^{3/2}$ edges for graph $\mathbb{G}_{\mathbf{Sqrt}}$. We further denote the set of those randomly sampled edges as $\tilde{E}$.

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

Huber function and its corresponding normalized function (Andoni et al., 2018). Consider the Huber function (Huber,

---

[15]M is the input number of clusters used for the clustering method. Therefore, the clustering result has at most M clusters, depending on input data.
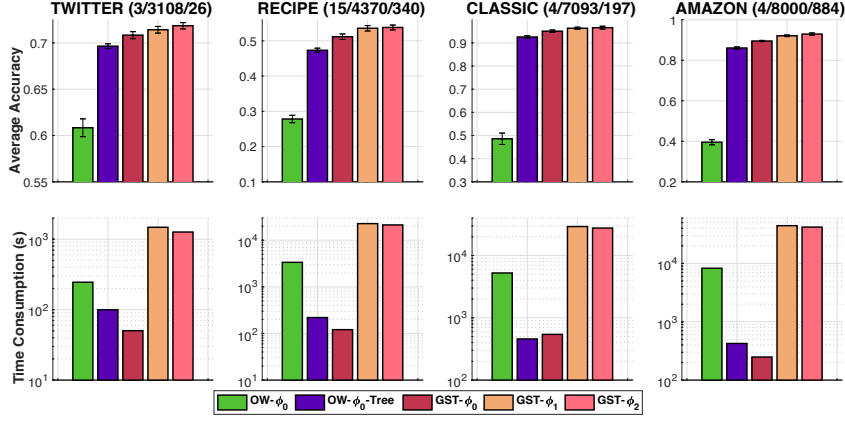
Figure 7. Document classification on graph $\mathbb{G}_{\text{Log}}$ ($M = 10^3$).

1992) on the domain $[0, +\infty)$, defined as,

$$f_H(t) = \begin{cases} t^2/2 & \text{if } 0 \le t \le \delta \\ \delta(t - \delta/2) & \text{if } t > \delta, \end{cases} \tag{28}$$

where $\delta > 0$ is a constant. Then, the normalized Huber function (Andoni et al., 2018) on the domain $[0, +\infty)$ is defined as

$$\Phi_H(t) = \begin{cases} f_H(f_H^{-1}(1)t) & \text{if } 0 \le t \le 1 \\ \Phi'_{H_-}(1)t - (\Phi'_{H_-}(1) - 1) & \text{if } t > 1, \end{cases} \tag{29}$$

where Φ'_{H_-} is the left derivative of Φ_H.

Hyperparamter validation. For validation, we further randomly split *the training set* into $70\%/30\%$ for validation-training and validation with 10 repeats to choose hyper-paramters in our simulations.

The number of pairs in training and test for kernel SVM. Let N_{tr}, N_{te} be the number of measures used for training and test respectively. For the kernel SVM training, the number of pairs which we compute the distances is $(N_{tr} - 1) \times \frac{N_{tr}}{2}$. For the test phase, the number of pairs which we compute the distances is $N_{tr} \times N_{te}$. Therefore, for 1 repeat, the number of pairs which we compute the distances for both training and test is totally $N_{tr} \times (\frac{N_{tr}-1}{2} + N_{te})$.

B.5. Further empirical results

Document classification. We further illustrate document classification on graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ for:

- $M = 10^3$ in Figures 7 and 8 respectively.

- $M = 10^2$ in Figures 9 and 10 respectively.

TDA. We also further illustrate TDA on graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ for:

- $M = 10^3$ in Figures 11 and 12 respectively.
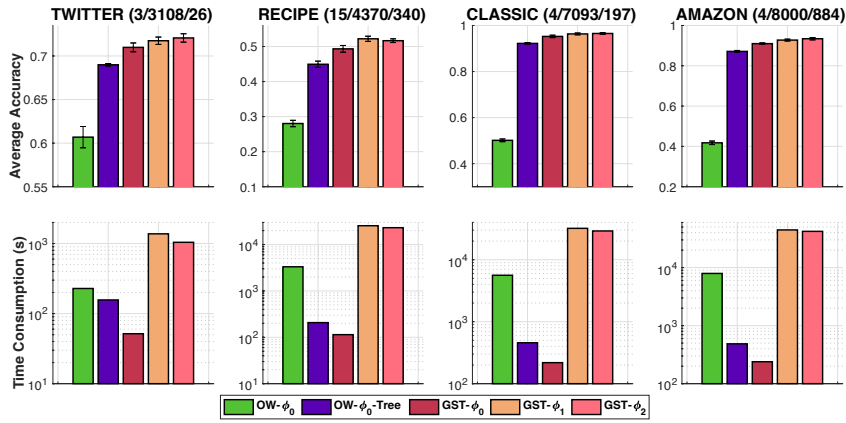
- $M = 10^2$ in Figures 13 and 14 respectively.

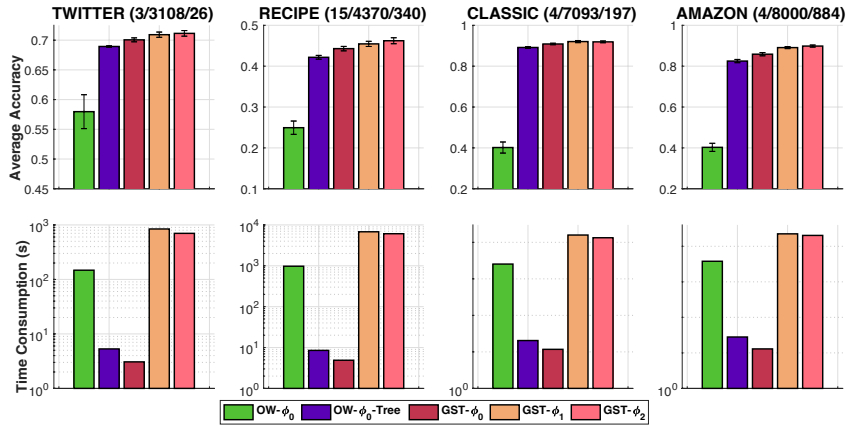*Figure 8.* Document classification on graph $\mathbb{G}_{\mathrm{Sqrt}}$ ($M = 10^3$).

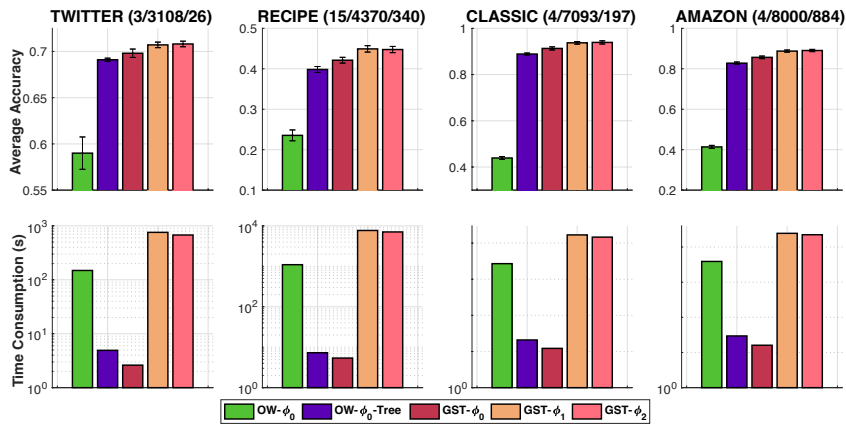*Figure 9.* Document classification on graph $\mathbb{G}_{\mathrm{Log}}$ ($M = 10^2$).

Figure 10. Document classification on graph $\mathbb{G}_{\mathrm{Sqrt}}$ ($M = 10^2$).
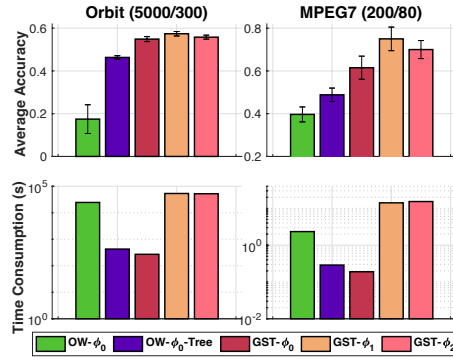
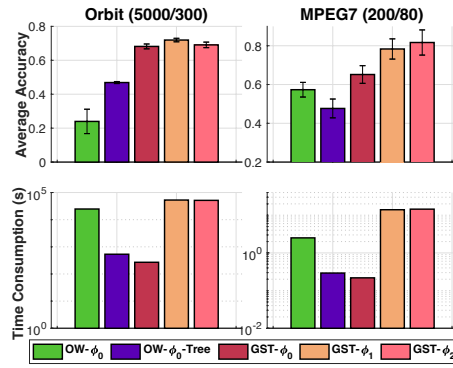*Figure 11.* Topological data analysis on graph $\mathbb{G}_{\text{Log}}$ ($M = 10^3$).

Figure 12. Topological data analysis on graph $\mathbb{G}_{\text{Sqrt}}$ ($M = 10^3$).
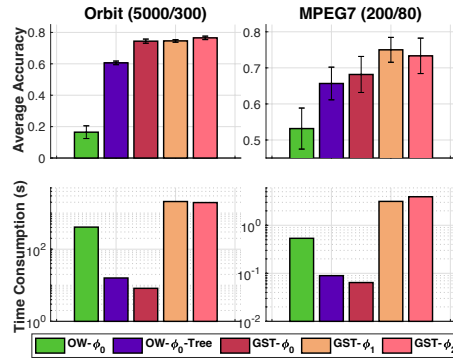

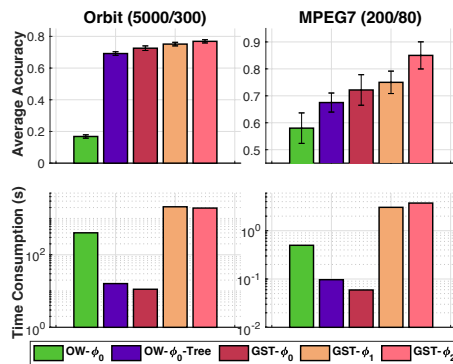*Figure 13.* Topological data analysis on graph $\mathbb{G}_{\text{Log}}$ ($M = 10^2$).

Figure 14. Topological data analysis on graph $\mathbb{G}_{\text{Sqrt}}$ ($M = 10^2$).