Sobolev Transport: A Scalable Metric for Probability Measures with Graph Metrics

Tam Le[*] RIKEN AIP
Truyen Nguyen[*] The University of Akron
Dinh Phung Monash University
Viet Anh Nguyen VinAI Research

Abstract

Optimal transport (OT) is a popular measure to compare probability distributions. However, OT suffers a few drawbacks such as (i) a high complexity for computation, (ii) indefiniteness which limits its applicability to kernel machines. In this work, we consider probability measures supported on a graph metric space and propose a novel Sobolev transport metric. We show that the Sobolev transport metric yields a *closed-form* formula for fast computation and it is negative definite. We show that the space of probability measures endowed with this transport distance is isometric to a bounded convex set in a Euclidean space with a weighted ℓ_p distance. We further exploit the negative definiteness of the Sobolev transport to design positive-definite kernels, and evaluate their performances against other baselines in document classification with word embeddings and in topological data analysis.

1 INTRODUCTION

Optimal transport (OT) is a powerful tool to compare probability measures. OT is widely used in machine learning (Courty et al., 2017; Bunne et al., 2019; Nadjahi et al., 2019; Peyré and Cuturi, 2019; Kuhn et al., 2019; Titouan et al., 2019; Janati et al., 2020; Muzellec et al., 2020; Paty et al., 2020; Altschuler et al., 2021; Fatras et al., 2021; Klicpera et al., 2021; Le et al., 2021b; Mukherjee et al., 2021; Nguyen et al., 2021b; Scetbon et al., 2021; Si et al., 2021), statistics (Mena and Niles-Weed, 2019; Weed and Berthet, 2019; Blanchet et al., 2021), computer graphics and vision (Rabin et al., 2011; Solomon et al., 2015; Lavenant et al., 2018; Nguyen et al., 2021a). However, evaluating the OT incurs a high computational complexity in general (Peyré and Cuturi, 2019) which leads to several proposals in the recent literature to address this drawback of OT, e.g., approximate using entropic regularization (Cuturi, 2013), or exploit geometric structure of supports (Rabin et al., 2011; Le et al., 2019; Le and Nguyen, 2021). Among them, tree-Wasserstein (Evans and Matsen, 2012; Le et al., 2019) (TW) leverages the tree structure over supports to obtain a closed-form for fast computation. However, the requirement about tree structure for supports may be restricted in applications. In this work, we exploit the graph structure, which appears in several applications, and propose a scalable variant of OT to compare probability measures supported on a graph metric space.

Given any two distributions μ and ν supporting on nodes of a tree with nonnegative weights, it is known from Evans and Matsen (2012); Le et al. (2019) that the 1-Wasserstein distance $\mathcal{W}_1$ w.r.t. the tree distance (i.e., TW) admits a closed-form expression, which allows a fast computation (i.e., its complexity is linear to the number of edges in the tree). The key techniques in deriving this formula are to leverage the dual formulation of $\mathcal{W}_1$ and exploit the fact that there is a unique path between any two nodes on the tree. Due to a different nature of the dual formulation between $p = 1$ and $p > 1$, it is, unfortunately, unknown whether the closed-form expression still holds for the p-Wasserstein distance with ground tree metric when $p > 1$. It is also not known if the closed-form for $\mathcal{W}_1$ with ground tree metric can be extended to general graphs where there are multiple paths connecting two nodes (i.e., graph metric ground cost). The approaches proposed in Evans and Matsen (2012); Le et al. (2019); Le and Nguyen (2021) do not resolve these questions, either.

Related Work. Our proposed Sobolev transport is an instance of the integral probability metric (Müller, 1997) and closely related to $\mathcal{W}_1$ for probability mea-

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

[*]: Two authors contributed equally.

sures supported on a graph metric space. Similar to TW, the Sobolev transport exploits the structure of supports for a fast computation and has similar properties as the TW (e.g., both of them are negative definite which is the key to build positive definite kernels for applications with kernel machines). Moreover, Sobolev transport has more flexibility and degrees of freedom than TW since it requires a graph structure rather than tree structure over supports.

We further note that the Sobolev transport leverages a graph structure for probability measures supported on a *graph metric space*, rather than a *general graph* over supports. For example, the edge weight in the graph, corresponding with a graph metric space, is a cost to move from one node to the other node of that edge (i.e., the distance between two edge nodes), rather than an affinity between these edge nodes of the graph used in diffusion earth mover's distance (Tong et al., 2021).

Contributions. We propose a novel distance, named the Sobolev transport $\mathcal{S}_p$ of any order $p \geq 1$, to measure the distance between probability measures supported on a graph metric space. Moreover,

- we show that $\mathcal{S}_p$ (i) admits a fast closed-form computation and (ii) is negative definite. Consequently, we can derive positive-definite kernels using our proposed Sobolev transport distance $\mathcal{S}_p$, which can be applied for many kernel-dependent frameworks in machine learning.
- when $p = 1$ and with a tree structure, we draw a connection of our proposed Sobolev transport $\mathcal{S}_1$ to the 1-Wasserstein distance $\mathcal{W}_1$.
- we also prove that the space of probability measures with Sobolev transport metric $\mathcal{S}_p$ is isometric to a bounded convex set in a Euclidean space with a weighted ℓ_p distance.

In Section 2, we provide the setup of our problem. The Sobolev transport is formally introduced in Section 3, and we discuss its nice properties in Section 4. In Section 5, we illustrate empirically that the kernel machines using our proposed Sobolev transport distance perform favorably compared to other baselines in real-world applications. Proofs are placed in the supplementary (Section A). Furthermore, we have released code for our proposals.[1]

2 PRELIMINARIES

Let $\mathbb{G} = (V, E)$ be an undirected and connected graph with positive edge lengths $\{w_e\}_{e\in E}$. We consider a physical graph in the sense that V is a subset of the vector space $\mathbb{R}^n$ and each edge $e \in E$ is the standard line segment in $\mathbb{R}^n$ connecting the two end-points of e. The most important case for our applications is when w_e coincides with the Euclidean length of the edge e.

Henceforth, by mentioning the graph $\mathbb{G}$, we mean the set of *all* nodes V together with *all* points forming the edges E.[2] This general consideration allows us to work with a continuous setting to derive a closed-form formula for a newly proposed transport distance. Notice that we can canonically measure the weighted length for any path in $\mathbb{G}$ whose end-points might not be nodes in V. Indeed, for any two points x and y belonging to the same edge $e = \langle u, v\rangle$ connecting two nodes u and v, we can express $x = (1 - s)u + sv$ and $y = (1 - t)u + tv$ for some numbers $t, s \in [0, 1]$. Then, the length of the path connecting x and y along edge e (i.e., the line segment $\langle x, y\rangle$) is defined by $|t-s|w_e$. The length for an arbitrary path in $\mathbb{G}$ is defined similarly by breaking down into pieces and summing over their corresponding lengths.

We impose on $\mathbb{G}$ the following graph metric d: for every $x, y \in \mathbb{G}$, $d(x, y)$ equals to the length of the shortest path on $\mathbb{G}$ between x and y. Because the edges are undirected and the lengths $\{w_e\}_{e\in E}$ are positive, it is easy to show that d satisfies the non-negativity, the symmetry and the triangle inequality properties. Thus, d, by construction, is a metric.

Further, we assume that $\mathbb{G}$ satisfies the following uniqueness property of the shortest paths.

Assumption 2.1 (Unique-path root node)**.** *There exists a root node $z_0 \in V$ such that for every $x \in \mathbb{G}$, $d(x, z_0)$ is attained by a* unique *shortest path connecting x and z_0.*

Recall that a graph is geodetic if for every pair of nodes the shortest path between them is unique. Thus, geodetic graphs are special examples satisfying Assumption 2.1. An example of geodetic graph is given in Figure 1.

For $1 \leq p \leq \infty$ and for a nonnegative Borel measure λ on $\mathbb{G}$, let $L^p(\mathbb{G}, \lambda)$ denote the space of all Borel measurable functions $f : \mathbb{G} \to \mathbb{R}$ satisfying $\int_{\mathbb{G}} |f(y)|^p \lambda(\mathrm{d}y) < \infty$. Two functions $f_1, f_2 \in L^p(\mathbb{G}, \lambda)$ are considered to be the same if $f_1(x) = f_2(x)$ for λ almost every x in $\mathbb{G}$. Then, $L^p(\mathbb{G}, \lambda)$ is a normed space with the norm defined by

$$\|f\|_{L^p(\mathbb{G},\lambda)} \coloneqq \Big(\int_{\mathbb{G}} |f(y)|^p \lambda(\mathrm{d}y) \Big)^{\frac{1}{p}}.$$

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

[2] I.e., the collection of all points in $\mathbb{R}^n$ belongs to one of the edges.

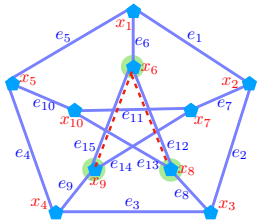


Figure 1: An illustration for a geodetic graph with 10 nodes $\{x_i\}_{i=1}^{10}$ and 15 edges $\{e_j\}_{j=1}^{15}$ where each edge weight equals to one, i.e., $w_{e_j} = 1, \forall j$. For any x_i, x_j, there is a unique shortest path between them, with a length 2. Let x_1 be the unique-path root node (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 $\Lambda(x_6) = \gamma(e_6) = \widetilde{\mathbb{G}}$.

Throughout the paper, we use $\langle x_1, x_2\rangle$ to denote the line segment in $\mathbb{R}^n$ connecting two points x_1, x_2, while (x_1, x_2) means the same line segment but without its two end-points. The symbol $[z_0, y]$ denotes the shortest path in $\mathbb{G}$ connecting z_0 and $y \in \mathbb{G}$. Under Assumption 2.1, $[z_0, y]$ is a unique path. Also, $\mathcal{P}(\mathbb{G})$ represents the set of all Borel probability measures on $\mathbb{G}$. The conjugate of a number $1 \le p \le \infty$ is denoted by p'. This is the number in $[1, \infty]$ satisfying $\frac{1}{p} + \frac{1}{p'} = 1$. In case $p = 1$, we have $p' = \infty$. Given $x \in \mathbb{G}$, define

$$\Lambda(x) \coloneqq \{y \in \mathbb{G} : x \in [z_0, y]\}. \tag{2.1}$$

Notice that $\Lambda(x)$ is always non-empty, and $x \in \Lambda(x)$. On the other hand, let γ_e denote the collection of all points $y \in \mathbb{G}$ such that the unique shortest path connecting y and z_0 contains the edge e. That is,

$$\gamma_e \coloneqq \{y \in \mathbb{G} : e \subset [z_0, y]\}. \tag{2.2}$$

In Figure 1, we show a computation for $\Lambda(x)$ and γ_e. Furthermore, we write $|E|$ and $|V|$ for the cardinality of sets E and V respectively. For a measure μ, let $\mathrm{supp}(\mu)$ denote the set of supports of μ.

3 SOBOLEV TRANSPORT DISTANCE

In this section, we define an instance of integral probability metrics between probability distributions on the graph. Our definition is inspired by the dual form of the 1-Wasserstein distance $\mathcal{W}_1$, and by Mroueh et al. (2018); Xu et al. (2021). Instead of using the Lipschitz constraint for the critic as in $\mathcal{W}_1$, we relax it by considering the constraint in a Sobolev space. We first propose a generalized version of the fundamental theorem of calculus, which defines the derivative of a function at any point $x \in \mathbb{G}$ dependent on the shortest path from the root node z_0 to x.

Definition 3.1 (Graph-based Sobolev space). *Let λ be a nonnegative Borel measure on $\mathbb{G}$, and let $1 \le p \le \infty$. A continuous function $f : \mathbb{G} \to \mathbb{R}$ is said to belong to the Sobolev space $W^{1,p}(\mathbb{G}, \lambda)$ if there exists a function $h \in L^p(\mathbb{G}, \lambda)$ satisfying*

$$f(x) - f(z_0) = \int_{[z_0, x]} h(y)\lambda(\mathrm{d}y) \quad \forall x \in \mathbb{G}. \tag{3.1}$$

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

The integral in Definition 3.1 is a line integral. We now formally define the Sobolev transport distance between two distributions supported on $\mathbb{G}$.

Definition 3.2 (Sobolev transport distance on graphs). *Let λ be a nonnegative Borel measure on $\mathbb{G}$. Let $1 \le p \le \infty$ and let p' be its conjugate, i.e., the number $p' \in [1, \infty]$ satisfying $\frac{1}{p} + \frac{1}{p'} = 1$. For $\mu, \nu \in \mathcal{P}(\mathbb{G})$, we define*

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

By definition, the quantity $\mathcal{S}_p(\mu, \nu)$ depends on the measure λ and on the choice of the unique-path root node z_0 via the graph derivative f'; however, we omit these dependencies when no confusion may arise. The role of λ will be displayed in Section 4 when we make a connection between our transport distance and the Wasserstein distance. Specifically, if $p = 1$ and $\lambda([z_0, x]) = d(z_0, x)$, then the constrain for f in Definition 3.2 is the same as $|f(x) - f(z_0)| \le d(z_0, x)$ for every $x \in \mathbb{G}$. Thus, the Sobolev transport distance $\mathcal{S}_1$ coincides with the 1-Wasserstein distance in this particular case. The next result asserts that $\mathcal{S}_p(\mu, \nu)$ is an integral probability metric on the graph $\mathbb{G}$.

Lemma 3.3 (Metrization). *For any $1 \le p \le \infty$, the Sobolev transport $\mathcal{S}_p$ is a metric on the space $\mathcal{P}(\mathbb{G})$.*

The next result gives a comparison between Sobolev transport distances with different exponent p.

Proposition 3.4 (Upper bound). *Assume that λ is a finite and nonnegative Borel measure on $\mathbb{G}$. Then, for any $1 \le p < q \le \infty$ with conjugates $1 \le q' < p' \le \infty$, we have*

$$\mathcal{S}_p(\mu, \nu) \le \lambda(\mathbb{G})^{\frac{1}{q'} - \frac{1}{p'}}\ \mathcal{S}_q(\mu, \nu).$$

Our proposed Sobolev transport distance $\mathcal{S}_p$ admits a closed-form formula as follows.

Proposition 3.5 (Closed-form formula). *Let λ be any nonnegative Borel measure on $\mathbb{G}$, and let $1 \le p \le \infty$.*

Then, we have

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

where $\Lambda(x)$ *is the subset of* $\mathbb{G}$ *defined by* (2.1).

Sketch of Proof of Proposition 3.5. By using representation (3.1) and employing Fubini's theorem to interchange the order of integration, we have for any function $f \in W^{1,p'}(\mathbb{G},\lambda)$ and any measure $\sigma \in \mathcal{P}(\mathbb{G})$ that

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

This together with the definition of distance $\mathcal{S}_p$ and by taking $g = f'$, we deduce that $\mathcal{S}_p(\mu,\nu)$ is the same as

$$\sup_{\|g\|_{L^{p'}(\mathbb{G},\lambda)}\le 1} \int_{\mathbb{G}} g(x)\big[\mu(\Lambda(x)) - \nu(\Lambda(x))\big]\,\lambda(\mathrm{d}x).$$

This last optimization problem admits a maximizer $g^*(x) = \frac{|r(x)|^{p-2}r(x)}{\|r\|^{p-1}_{L^p(\mathbb{G},\lambda)}}$ with $r(x) \coloneqq \mu(\Lambda(x)) - \nu(\Lambda(x))$, and the conclusion of the proposition follows.

In the particular case where the probability distributions μ and ν are supported only on nodes V, the expression in Proposition 3.5 can be rewritten more explicitly using the definition of γ_e in (2.2).

Corollary 3.6 (Discrete case)**.** *Assume that the measure* λ *has no atom, i.e.,* $\lambda(\{x\}) = 0$ *for every* $x \in \mathbb{G}$*. Then, if both measures* μ *and* ν *in* $\mathcal{P}(\mathbb{G})$ *are supported on* V*, we have*

$$\mathcal{S}_p(\mu,\nu)^p = \sum_{e\in E} \lambda(e)\,\big|\mu(\gamma_e) - \nu(\gamma_e)\big|^p. \tag{3.2}$$

Remark 3.7 (Two-step computational procedure)**.** *Our calculation of the Sobolev transport distance between* μ *and* ν *can be split into two separate steps. The first step is the preprocessing process involving only the graph structure and nothing about the probability distributions, and is done only once regardless how many pairs* (μ,ν) *that we have to measure. In this step by identifying shortest paths (e.g., Dijkstra algorithm), we calculate the set* γ_e *for each edge* $e \in E$*. In fact, any edge* e *with* $\gamma_e = \varnothing$ *does not contribute to the computation of the Sobolev transport. Therefore, we can remove such edge* e *for the summation over edges in* E *(in Equation* (3.2)*). In the second step, we just simply use the result in Step 1 and Corollary 3.6 to compute the Sobolev transport distance.*

Complexity. For preprocessing, the complexity of Dijkstra for shortest paths from the root node z_0 to all other supports (or vertices) is $\mathcal{O}(|E| + |V|\log|V|)$.

A key observation is that for any support point z of μ, i.e., $z \in \mathrm{supp}(\mu)$, its mass contributes to $\mu(\gamma_e)$ if and only if the edge e is a subset of the shortest path from the root node z_0 to z, i.e., $e \subset [z_0, z]$. Let $E_{\mu,\nu}$ be a subset of E, defined as:

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

then we can rewrite $\mathcal{S}_p(\mu,\nu)^p$ in (3.2) as

$$\mathcal{S}_p(\mu,\nu)^p = \sum_{e\in E_{\mu,\nu}} \lambda(e)\,\big|\mu(\gamma_e) - \nu(\gamma_e)\big|^p. \tag{3.3}$$

Therefore, the computation of Sobolev transport $\mathcal{S}_p(\mu,\nu)$ is linear to the number of edges in $E_{\mu,\nu}$.

4 PROPERTIES OF SOBOLEV TRANSPORT

This section shows a connection between our Sobolev transport distance and the Wasserstein distance when the measure λ is chosen as the length measure of the graph. We also demonstrate that the space of distributions $\mathcal{P}(V)$ is isometric to a bounded convex set in a Euclidean space. We then prove that for $1 \le p \le 2$, both $\mathcal{S}_p$ and its p-power $\mathcal{S}_p^p$ are negative definite which allows us to build positive definite kernels upon Sobolev transport. We also propose a slice variant for Sobolev transport.

4.1 A Connection to Wasserstein Distance

We will specifically construct a measure λ^* under which the distance $\mathcal{S}_1$ is the same as the 1-Wasserstein distance $\mathcal{W}_1$ w.r.t. the graph metric d.

Definition 4.1 (Length measure)**.** *Let* λ^* *be the unique Borel measure on* $\mathbb{G}$ *such that the restriction of* λ^* *on any edge is the length measure of that edge. That is,* λ^* *satisfies:*

i) For any edge e *connecting two nodes* u *and* v*, we have* $\lambda^*(\langle x,y\rangle) = (t-s)w_e$ *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*

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

The next lemma asserts that λ^* is closely connected to the graph metric d, and thus justifies the terminology of a length measure.

Lemma 4.2 (λ^* is the length measure on graph)**.** *Suppose that $\mathbb{G}$ has no short cuts, namely, any edge e is a shortest path connecting its two end-points. Then, λ^* is a length measure in the sense that*

$$\lambda^*([x,y]) = d(x,y)$$

for any shortest path $[x,y]$ connecting x and y. In particular, λ^* has no atom.

The measure λ^* is special as it is linked to the metric distance. For trees, $\mathcal{S}_1$ defined w.r.t. λ^* is the same as the Wasserstein distance with cost $d(x,y)$.

Corollary 4.3 (Tree topology)**.** *Suppose that the graph $\mathbb{G}$ is a tree and the distance $\mathcal{S}_1$ is defined w.r.t. the measure λ^*. Then, we have*

$$\mathcal{S}_1 \equiv \mathcal{W}_1,$$

where $\mathcal{W}_1$ is the Wasserstein distance[3] with cost d.

We do not know the exact relationship between $\mathcal{S}_p$ and the p-Wasserstein distance $\mathcal{W}_p$ when $p > 1$. However, the following result shows that $\mathcal{S}_p$ is always lower bounded by $\mathcal{W}_1$.

Lemma 4.4 (Bounds)**.** *Suppose the graph $\mathbb{G}$ is a tree and the distance $\mathcal{S}_p$ is defined w.r.t. the measure λ^*. Then, for any $1 \le p \le \infty$, we have*

$$\mathcal{W}_1(\mu,\nu) \le \lambda^*(\mathbb{G})^{\frac{1}{p'}} \mathcal{S}_p(\mu,\nu).$$

4.2 Isometry Between $\mathcal{P}(V)$ and a Bounded Convex Set in a Euclidean Space

Assume that $V = \{z_0 \equiv x_1, x_2, ..., x_n\}$. For a node x_i, let $N(x_i)$ denote the collection of all neighbor nodes of x_i, and let

$$N'(x_i) \coloneqq \Big\{ v \in N(x_i) :\ d(v,z_0) = d(x_i,z_0) + w_{\langle x_i, v\rangle} \Big\}.$$

Also, for $2 \le i \le n$, let $\hat{x}_i$ denote the unique node $x \in N(x_i)$ such that the shortest path $[z_0, x_i]$ passes through x, i.e., $x \in [z_0, x_i]$.

Let us now take a closer look at the feature map

$$\rho \in \mathcal{P}(V) \longmapsto \alpha \coloneqq (\rho(\gamma_e \cap V))_{e\in E} \in \mathbb{R}^m.$$

Observe that the representation vector $\alpha = (\alpha_e)_{e\in E}$ satisfies $\alpha_e \ge 0$, $\alpha_e = 0$ if $\gamma_e = \varnothing$ and

$$\sum_{e=\langle x_1, v\rangle : v\in N(x_1)} \alpha_e \le 1,$$

$$\sum_{e=\langle x_i, v\rangle : v\in N'(x_i)} \alpha_e \le \alpha_{\langle \hat{x}_i, x_i\rangle} \ \ \forall i = 2, ..., n.$$

[3]The definition of $\mathcal{W}_1$ is recalled in the supplementary.

Hereafter we use the convention that if $N'(x_i) = \varnothing$, then the corresponding summation is interpreted as zero. We note that $N'(x_i) = \varnothing$ happens precisely when there is no shortest path from other nodes to z_0 that passes through x_i (this, in particular, occurs for nodes in the "last level").

Let $\mathcal{K}$ denote the set of all vectors $\alpha \in \mathbb{R}^m$ having the above specified properties. Clearly, this is a bounded and convex set which is closed w.r.t. the Euclidean metric in $\mathbb{R}^m$. In the next proposition, we assume that the distance $\mathcal{S}_p$ is defined w.r.t. the measure λ^* defined in Section 4.1. This result shows that there is a one-to-one correspondence between $\mathcal{P}(V)$ and the set $\mathcal{K}$.

Proposition 4.5 ($\mathcal{P}(V)$ isometric to $\mathcal{K}$)**.** *The map*

$$\rho \in \mathcal{P}(V) \longmapsto \alpha \coloneqq (\rho(\gamma_e \cap V))_{e\in E} \in \mathcal{K} \tag{4.1}$$

is one-to-one and onto. In addition, for any $\alpha = (\alpha_e)_{e\in E} \in \mathcal{K}$, if we let

$$\begin{aligned} a^1 &\coloneqq 1 - \textstyle\sum_{e=\langle x_1, v\rangle : v\in N(x_1)} \alpha_e, \\ a^i &\coloneqq \alpha_{\langle \hat{x}_i, x_i\rangle} - \textstyle\sum_{e=\langle x_i, v\rangle : v\in N'(x_i)} \alpha_e \end{aligned} \tag{4.2}$$

for $i = 2, ..., n$, then $\rho \coloneqq \sum_{i=1}^n a^i \delta_{x_i} \in \mathcal{P}(V)$. Finally, the distance $\mathcal{S}_p$ on $\mathcal{P}(V)$ is the same as the weighted ℓ_p distance on $\mathcal{K}$, that is,

$$\mathcal{S}_p(\rho_1, \rho_2) = \Big(\sum_{e\in E} w_e \big|\alpha_e^1 - \alpha_e^2\big|^p \Big)^{\frac{1}{p}},$$

with $\alpha^i \coloneqq \alpha(\rho_i)$ for $i = 1, 2$.

The isometry is an useful properties of the Sobolev transport since any problem on the space of probability measures with Sobolev transport metric $\mathcal{S}_p$ can be recasted as a corresponding problem on a bounded convex set of vectors in a Euclidean space with ℓ_p metric.

4.3 Kernels for Sobolev Transport

Our next result about negative definiteness[4] is the key to build positive definite kernels upon Sobolev transport for kernel machines.

Proposition 4.6 (Negative definiteness)**.** *Suppose that the Sobolev transport distance $\mathcal{S}_p$ is defined w.r.t. the length measure λ^* on graph $\mathbb{G}$ for probability measures in $\mathcal{P}(V)$, then for $1 \le p \le 2$, $\mathcal{S}_p$ and $\mathcal{S}_p^p$ are negative definite.*

From Proposition 4.6 and following (Berg et al., 1984, Theorem 3.2.2, pp.74), given $t > 0$, $1 \le p \le 2$ and

[4]We follow the definition of negative-definiteness in (Berg et al., 1984, pp. 66–67). A review about kernels is placed in the supplementary (Section B).

$\mu, \nu \in \mathcal{P}(V)$, the kernels

$$k_{\mathcal{S}_p}(\mu,\nu) \coloneqq \exp(-t\mathcal{S}_p(\mu,\nu)),$$
$$k_{\mathcal{S}_p^p}(\mu,\nu) \coloneqq \exp(-t\mathcal{S}_p^p(\mu,\nu))$$

are positive definite.

4.4 Sliced Sobolev Transport Distance

As remark after Definition 3.2, our Sobolev transport distance depends on the choice of the root node z_0 satisfying Assumption 2.1. When there are multiple possible root nodes, each choice of z_0 imposes its own geometry on the graph, which characterizes differently the graph derivative f' of the function f. To alleviate the dependence in this case, and inspired by the slicing approach in optimal transport (Rabin et al., 2011; Le et al., 2019; Le and Nguyen, 2021) for practical applications, we propose the *sliced* Sobolev transport distance that fuses the Sobolev transport distance in Section 3. Towards this end, let $\mathcal{Z}_0 \subset V$ be a (sub)set of unique-path root nodes:

$$\mathcal{Z}_0 \coloneqq \{z_0 \in V : z_0 \text{ satisfies Assumption 2.1}\}.$$

The sliced Sobolev transport averages over a sampling distribution η on $\mathcal{Z}_0$, and is formally defined as follows.

Definition 4.7 (Sliced Sobolev transport)**.** *Let η be a probability measure on $\mathcal{Z}_0$. The sliced Sobolev transport is defined as*

$$\mathcal{S}_p^\eta(\mu,\nu) \coloneqq \int_{\mathcal{Z}_0} \mathcal{S}_p^{z_0}(\mu,\nu)\eta(\mathrm{d}z_0) = \sum_{z_0\in\mathcal{Z}_0} \eta(\{z_0\})\, \mathcal{S}_p^{z_0}(\mu,\nu),$$

where $\mathcal{S}_p^{z_0}$ is the Sobolev transport distance in Definition 3.2 that is specific to the choice of the unique-path root node z_0.

Because $\mathcal{S}_p^\eta$ is a convex combination of $\mathcal{S}_p^{z_0}$, we can readily verify that $\mathcal{S}_p^\eta$ is also a distance. The proof is relegated to the supplementary.

Proposition 4.8 (Metric)**.** *The sliced Sobolev transport $\mathcal{S}_p^\eta$ is a distance on $\mathcal{P}(\mathbb{G})$.*

5 NUMERICAL EXPERIMENTS

We evaluate the performance of our proposed Sobolev transport on two applications: (i) document classification with word embedding and (ii) topological data analysis (TDA).

Probability Measures Representation. We first describe probability measure representation for documents with word embedding in document classification and persistent diagrams for geometric structured data in TDA.

- **Documents with Word Embedding.** We consider each document as a probability measure where each word and its frequency in the document are regarded as a support and a corresponding weight in the probability measure respectively. We then follow the approach in Kusner et al. (2015); Le et al. (2019) to use *word2vec* word embedding (Mikolov et al., 2013) pretrained on Google News[5] for documents. The pretrained *word2vec* contains about 3 millions words/phrases. Consequently, each word in a document is mapped into a vector in $\mathbb{R}^{300}$. We also remove SMART stop words (Salton and Buckley, 1988) or words in documents which are not available in the pretrained *word2vec*.

- **Persistence Diagrams.** TDA provides a powerful toolkit to analyze complicated geometric structured data, e.g., object shape, material data, or linked twist maps (Adams et al., 2017; Le et al., 2019). TDA leverages algebraic topology methods (e.g., persistence homology) to extract robust topological features (e.g., connected components, rings, cavities) and yield a multiset of points in $\mathbb{R}^2$ which is also known as persistence diagram (PD). The two coordinates of a point in PD are corresponding to the birth and death time of a topological feature respectively. Therefore, each point in PD summarizes a life span of a particular topological feature. We regard each PD as an empirical measure where each 2-dimensional point in PD is considered as a support with a uniform weight in the empirical measure.

Note that supports in document classification are in a high-dimensional space (i.e., $\mathbb{R}^{300}$) while supports in TDA are in a low-dimensional space (i.e., $\mathbb{R}^2$). Therefore, these applications allow us to observe how the dimension of supports affects performances. We next describe various graphs with different sizes (i.e., given graph metrics which we assume in applications) considered in our experiments.

Graph Metric Construction. For simplicity, we use a random graph metric for supports of probability measures as follow:

We first apply a clustering method, e.g., the farthest-point clustering, to partition supports of probability measures into at most M clusters.[6] We assign V to be the set of centroids of these clusters. For edges, we consider two options: randomly choose (i) $M\log(M)$

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

[6] We set M for the number of clusters when running the clustering method. Depending on input data, we obtain at most M clusters.

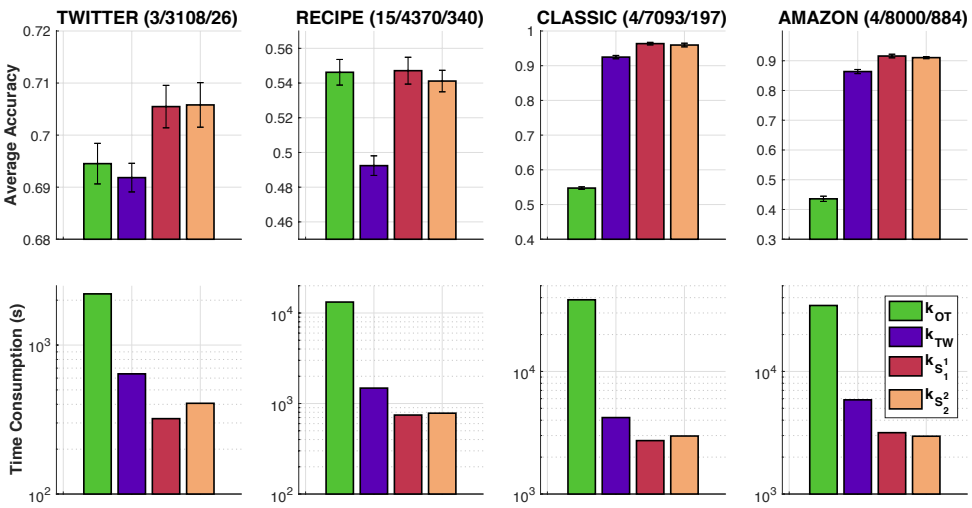


Figure 2: SVM results and time consumption for kernel matrices in document classification with graph $\mathbb{G}_{\text{Log}}$. For each dataset, the numbers in the parenthesis are the number of classes; the number of documents; and the maximum number of unique words for each document respectively.

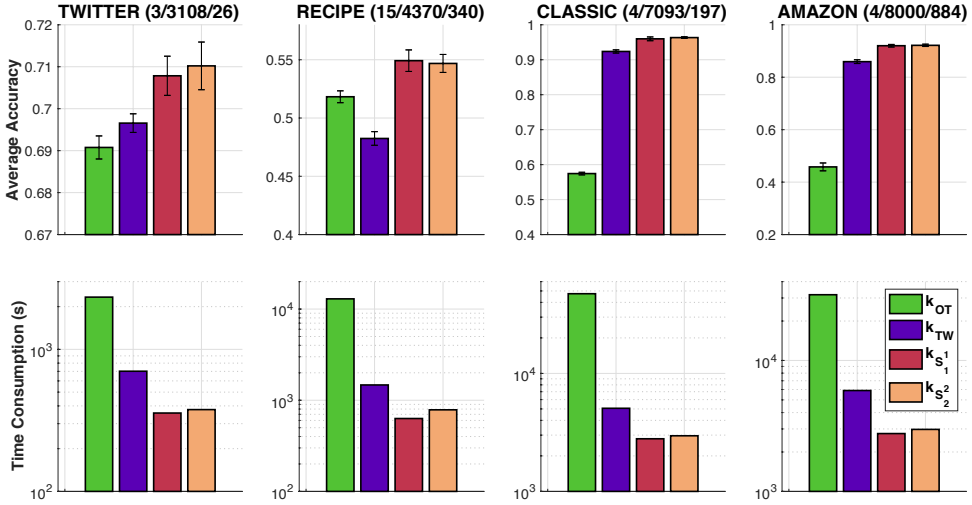


Figure 3: SVM results and time consumption for kernel matrices in document classification with graph $\mathbb{G}_{\text{Sqrt}}$.

edges or (ii) $M^{3/2}$ edges. For an edge e, its corresponding weight w_e is computed by the Euclidean distance between the two nodes of that edge e. Let $\tilde{E}$ be the set of those randomly sampled edges and n_c be the number of connected components in the graph $\tilde{\mathbb{G}}(V, \tilde{E})$, we then randomly add $(n_c - 1)$ more edges between these n_c connected components to construct a connected graph $\mathbb{G}$ from $\tilde{\mathbb{G}}$. Denote E_c as the set of these $(n_c - 1)$ added edges and $E = \tilde{E} \cup E_c$, then $\mathbb{G}(V, E)$ is the considered graph.

We next describe baseline methods and detailed setup for our experiments.

Baselines and Setup. We consider two typical baseline distances based on OT theory for probability measures supported on a graph metric space: (i) the optimal transport (OT) d_{OT} with a graph metric cost (i.e., an instance of min-cost flow problem via Beckman formulation (Peyré and Cuturi, 2019, Section 6.3)) and (ii) the tree-Wasserstein (Le et al., 2019) (TW) d_{TW} where the tree structure is randomly sampled from the graph $\mathbb{G}$. In all experiments, we consider the kernels $k_{\mathcal{S}_1}$ and $k_{\mathcal{S}_2^2}$ for the proposed Sobolev transport distances and baseline kernels $k_{\text{OT}}(\cdot, \cdot) \coloneqq \exp(-t d_{\text{OT}}(\cdot, \cdot))$ and $k_{\text{TW}}(\cdot, \cdot) \coloneqq \exp(-t d_{\text{TW}}(\cdot, \cdot))$ for the corresponding OT distance d_{OT} and TW distance d_{TW} respectively.

Following Le et al. (2019), we evaluate those kernels with support vector machine (SVM) for document classification with word embedding and some tasks in TDA, e.g., the orbit recognition and object shape classification. Note that $k_{\mathcal{S}_1}$, $k_{\mathcal{S}_2^2}$ and k_{TW} are positive

definite, but k_{OT} is empirically indefinite.[7] Similar as the approach in Le et al. (2019), we regularize for the Gram matrix of k_{OT} by adding a sufficiently large diagonal term. For multi-class classification, we employ 1-vs-1 strategy with Libsvm.[8]

For each dataset, we randomly split it into 70%/30% for training and test with 10 repeats. We typically choose hyper-parameters via cross validation. For kernel hyperparameter, we choose $1/t$ from $\{q_s, 2q_s, 5q_s\}$ with $s = 10, 20, \ldots, 90$ where q_s is the $s\%$ quantile of a subset of corresponding distances observed on a training set. For SVM regularization hyperparameter, we choose it from $\{0.01, 0.1, 1, 10, 100\}$. We also consider a various number of nodes $M = 10^2, 10^3, 10^4, 4 \times 10^4$ for $\mathbb{G}$. Reported time consumption for all methods includes their corresponding preprocessing, e.g., compute shortest paths for Sobolev transport and OT, or sample random tree structure from graph for TW.

5.1 Document Classification

We consider 4 document datasets: `TWITTER`, `RECIPE`, `CLASSIC` and `AMAZON`. The statistical characteristics of these datasets are summarized in Figure 2.

5.2 Topological Data Analysis (TDA)

For TDA, we consider the orbit recognition and the object shape classification.

5.2.1 Orbit Recognition

We consider the synthesized dataset as in Adams et al. (2017) for link twist map which are discrete dynamical systems to model flows in DNA microarrays (Hertzsch et al., 2007). There are 5 classes of orbits in the dataset. Following Le and Yamada (2018), for each class, we generated 1000 orbits where each orbit contains 1000 points. We consider the 1-dimensional topological features (i.e., connected components) for PD which are extracted with Vietoris-Rips complex filtration (Edelsbrunner and Harer, 2008). The statistical characteristics are summarized in Figure 4.

5.2.2 Object Shape Classification

We consider a subset of `MPEG7` dataset (Latecki et al., 2000) having 10 classes and each class has 20 samples as in Le and Yamada (2018). For simplicity, we follow the approach in Le and Yamada (2018) to extract 1-dimensional topological features (i.e., connected components) for PD with Vietoris-Rips complex filtration (Edelsbrunner and Harer, 2008).

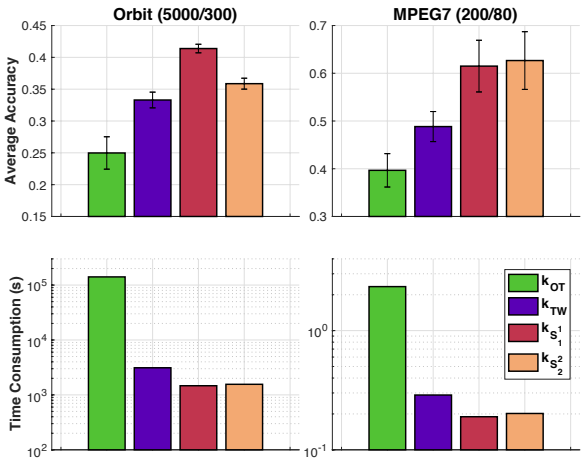


Figure 4: SVM results and time consumption for kernel matrices in TDA with graph $\mathbb{G}_{\mathrm{Log}}$. For each dataset, the numbers in the parenthesis are respectively the number of PD; and the maximum number of points in PD.

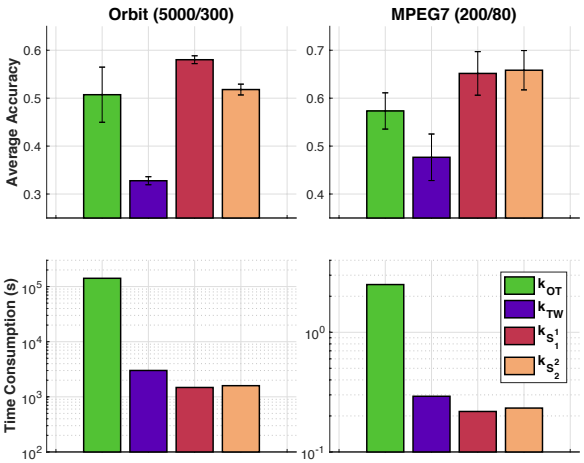


Figure 5: SVM results and time consumption for kernel matrices in TDA with graph $\mathbb{G}_{\mathrm{Sqrt}}$

5.3 SVM Results, Time Consumption and Discussions

We report results for graphs with $M = 10^4$ for all datasets except `MPEG7` where $M = 10^3$ due to its small size, and for both cases: (i) with $M \log(M)$ edges and (ii) with $M^{3/2}$ edges, and we denote those graphs as $\mathbb{G}_{\mathrm{Log}}$ and $\mathbb{G}_{\mathrm{Sqrt}}$ respectively.

In Figures 2 and 3, we illustrate the SVM results for document classification with word embedding with $\mathbb{G}_{\mathrm{Log}}$ and $\mathbb{G}_{\mathrm{Sqrt}}$, respectively. For TDA, we illustrate the results in Figures 4 and 5 for $\mathbb{G}_{\mathrm{Log}}, \mathbb{G}_{\mathrm{Sqrt}}$ respec-

[7] Generally, OT spaces are not Hilbertian (Peyré and Cuturi, 2019, Section 8.3). Additionally, we also empirically observe that the Gram matrix for k_{OT} has negative eigenvalues.

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

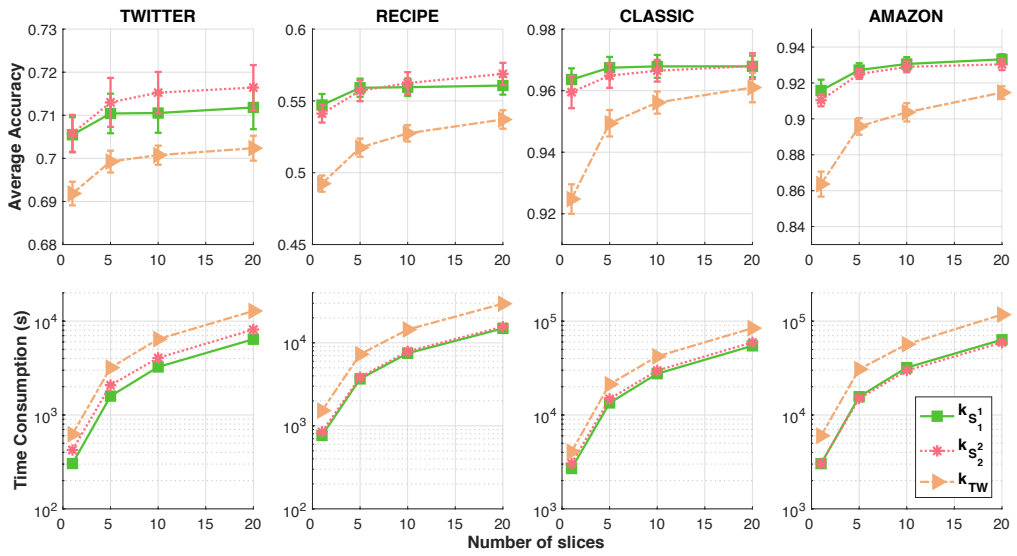


Figure 6: SVM results and time consumption for kernel matrices of slice variants for Sobolev transport and tree-Wasserstein in document classification with graph $\mathbb{G}_{\text{Log}}$.

tively. The performances of $k_{\mathcal{S}_1}$, $k_{\mathcal{S}_2^2}$ compare favorably with those of k_{OT} and k_{TW}. Moreover, the time consumption of the Gram matrices for $k_{\mathcal{S}_1}$, $k_{\mathcal{S}_2^2}$ is comparative with that of k_{TW} and is several-order faster than that of k_{OT}. Especially, in `Orbit` dataset, it took about more than 39 *hours* to compute the Gram matrix for k_{OT}, but only about 25 *minutes* for either $k_{\mathcal{S}_1}$ or $k_{\mathcal{S}_2^2}$. Recall that k_{OT} is indefinite, this infiniteness may affect the performances of k_{OT} in applications (in most of the experiments except the ones in `RECIPE` dataset with $\mathbb{G}_{\text{Log}}$ and in `Orbit` dataset with $\mathbb{G}_{\text{Sqrt}}$).

In Figure 6, we illustrate performances of slice variants for $k_{\mathcal{S}_1}$, $k_{\mathcal{S}_2^2}$ and k_{TW} for document classification with word embedding with $\mathbb{G}_{\text{Log}}$. When we use more slices, the performances are improved. However, its computation is also linearly increased.

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

6 CONCLUSION

In this paper, we have presented a scalable variant of optimal transport, namely the Sobolev transport, for probability measures supported on a graph (i.e., graph metric ground cost). By exploiting the graph-based Sobolev space structure, the proposed Sobolev transport distance admits a closed form solution for a fast computation. Moreover, the Sobolev transport is negative definite which allows to build positive definite kernels required in many kernel machine frameworks. We believe that exploiting local structures on supports such as tree or graph can improve the scalability for several optimal transport problems.

Acknowledgements

We thank anonymous reviewers and area chairs for their comments. TL acknowledges the support of JSPS KAKENHI Grant number 20K19873. The research of TN is supported in part by a grant from the Simons Foundation (#318995).

References

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

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

Berg, C., Christensen, J. P. R., and Ressel, P. (1984). *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Springer.

Blanchet, J., Murthy, K., and Nguyen, V. A. (2021). Statistical analysis of Wasserstein distributionally robust estimators. *INFORMS TutORials in Operations Research*.

Bunne, C., Alvarez-Melis, D., Krause, A., and Jegelka, S. (2019). Learning generative models across incomparable spaces. In *International Conference on Machine Learning (ICML)*, volume 97.

Courty, N., Flamary, R., Habrard, A., and Rako-

tomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 3730–3739.

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

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

Evans, S. and Matsen, F. (2012). The phylogenetic Kantorovich-Rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):569–592.

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

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

Janati, H., Muzellec, B., Peyré, G., and Cuturi, M. (2020). Entropic optimal transport between unbalanced Gaussian measures has a closed form. *Advances in Neural Information Processing Systems*, 33.

Klicpera, J., Lienen, M., and Günnemann, S. (2021). Scalable optimal transport in high dimensions for graph distances, embedding alignment, and more. In *International Conference on Machine Learning*, pages 5616–5627. PMLR.

Kuhn, D., Esfahani, P. M., Nguyen, V. A., and Shafieezadeh-Abadeh, S. (2019). Wasserstein distributionally robust optimization: Theory and applications in machine learning. *INFORMS TutORials in Operations Research*, pages 130–166.

Kusano, G., Fukumizu, K., and Hiraoka, Y. (2017). Kernel method for persistence diagrams via kernel embedding and weight factor. *The Journal of Machine Learning Research*, 18(1):6947–6987.

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

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

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

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

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

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

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

Le, T., Yamada, M., Fukumizu, K., and Cuturi, M. (2019). Tree-sliced variants of Wasserstein distances. In *Advances in Neural Information Processing Systems*.

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

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

Mroueh, Y., Li, C.-L., Sercu, T., Raj, A., , and Cheng, Y. (2018). Sobolev GAN. In *ICLR*, pages 5767–5777.

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

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

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

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

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

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

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

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

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

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

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

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

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

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. (2009). On integral probability metrics, ϕ-divergences and binary classification. *arXiv preprint arXiv:0901.2698*.

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

Tong, A. Y., Huguet, G., Natik, A., Macdonald, K., Kuchroo, M., Coifman, R., Wolf, G., and Krishnaswamy, S. (2021). Diffusion earth mover's distance and distribution embeddings. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10336–10346.

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

Xu, M., Zhou, Z., Lu, G., Tang, J., Zhang, W., and Yu, Y. (2021). Towards generalized implementation of Wasserstein distance in GANs. *arXiv preprint arXiv:2012.03420v2*.

Supplementary Material: Sobolev Transport: A Scalable Metric for Probability Measures with Graph Metrics

The supplementary is organized into two parts.

- In Section A, we provide the proofs for the theoretical results in the main manuscript.
- In Section B, we briefly review important aspects in our work, provide further experimental results and discussions about our proposed Sobolev transport.

We note that we have released code for our proposals at

`https://github.com/lttam/SobolevTransport`.

A PROOFS

A.1 Proofs and Results for Section 3

For Lemma 3.3.

Proof of Lemma 3.3. By taking $f = 0$ in Definition 3.2, we see that $\mathcal{S}_p(\mu, \nu) \geq 0$ for any (μ, ν), thus $\mathcal{S}_p$ is non-negative. Assume that $\mathcal{S}_p(\mu, \nu) = 0$. Then we must have

$$\int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(x)\nu(\mathrm{d}x) = 0 \tag{A.1}$$

for all $f \in W^{1,p'}(\mathbb{G}, \lambda)$ satisfying $\|f'\|_{L^{p'}(\mathbb{G},\lambda)} \leq 1$. Indeed, since otherwise there exists $\tilde{f} \in W^{1,p'}(\mathbb{G}, \lambda)$ with $\|\tilde{f}'\|_{L^{p'}(\mathbb{G},\lambda)} \leq 1$, and

$$\int_{\mathbb{G}} \tilde{f}(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} \tilde{f}(x)\nu(\mathrm{d}x) < 0.$$

Then by taking $f = -\tilde{f}$ in Definition 3.2, we see that $\mathcal{S}_p(\mu, \nu) > 0$ which contradicts the assumption $\mathcal{S}_p(\mu, \nu) = 0$. Thus (A.1) holds. It now follows from (A.1) that

$$\int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) = \int_{\mathbb{G}} f(x)\nu(\mathrm{d}x) \qquad \text{for every} \quad f \in W^{1,p'}(\mathbb{G}, \lambda),$$

giving $\mu = \nu$ as desired. To prove the symmetry of $\mathcal{S}_p(\mu, \nu)$, observe that if $f \in W^{1,p'}(\mathbb{G}, \lambda)$ with $\|f'\|_{L^{p'}(\mathbb{G},\lambda)} \leq 1$, then we also have $-f \in W^{1,p'}(\mathbb{G}, \lambda)$ with $\|-f'\|_{L^{p'}(\mathbb{G},\lambda)} = \|f'\|_{L^{p'}(\mathbb{G},\lambda)} \leq 1$. As a consequence, $\mathcal{S}_p(\mu, \nu) = \mathcal{S}_p(\nu, \mu)$. It remains to show that $\mathcal{S}_p$ satisfies the triangle inequality. For this, let $\mu, \nu, \sigma \in \mathcal{P}(\mathbb{G})$. Then for any function $f \in W^{1,p'}(\mathbb{G}, \lambda)$ satisfying $\|f'\|_{L^{p'}(\mathbb{G},\lambda)} \leq 1$, we have

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

This implies that $\mathcal{S}_p(\mu, \nu) \leq \mathcal{S}_p(\mu, \sigma) + \mathcal{S}_p(\sigma, \nu)$. We therefore conclude that $\mathcal{S}_p$ is a metric on the space $\mathcal{P}(\mathbb{G})$. □

For Proposition 3.4.

Proof of Proposition 3.4. Let $f \in W^{1,p'}(\mathbb{G}, \lambda)$ be such that $\|f'\|_{L^{p'}(\mathbb{G},\lambda)} \leq 1$. Since $q' < p'$ and $\lambda(\mathbb{G}) < +\infty$, it follows from Jensen's inequality that $f' \in L^{q'}(\mathbb{G}, \lambda)$. Now define

$$g(x) \coloneqq af(x) \text{ with } a \coloneqq \lambda(\mathbb{G})^{\frac{1}{p'} - \frac{1}{q'}}.$$

Then according to Definition 3.2 we have $g \in W^{1,q'}(\mathbb{G}, \lambda)$ with $g'(x) = af'(x)$. Hence by using Jensen's inequality we obtain

$$\Big(\frac{1}{\lambda(\mathbb{G})} \int_{\mathbb{G}} |f'(x)|^{q'} \lambda(\mathrm{d}x)\Big)^{\frac{p'}{q'}} \leq \frac{1}{\lambda(\mathbb{G})} \int_{\mathbb{G}} |f'(x)|^{p'} \lambda(\mathrm{d}x),$$

which yields

$$\begin{aligned} \|g'\|_{L^{q'}(\mathbb{G},\lambda)} = a\|f'\|_{L^{q'}(\mathbb{G},\lambda)} &\leq a\ \lambda(\mathbb{G})^{\frac{1}{q'} - \frac{1}{p'}} \|f'\|_{L^{p'}(\mathbb{G},\lambda)} \\ &= \|f'\|_{L^{p'}(\mathbb{G},\lambda)} \leq 1. \end{aligned}$$

Therefore,

$$a\Big[\int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(x)\nu(\mathrm{d}x)\Big] = \int_{\mathbb{G}} g(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} g(x)\nu(\mathrm{d}x) \leq \mathcal{S}_q(\mu, \nu).$$

Since this holds for any $f \in W^{1,p'}(\mathbb{G}, \lambda)$ satisfying$\|f'\|_{L^{p'}(\mathbb{G},\lambda)} \leq 1$, we conclude that

$$\mathcal{S}_p(\mu, \nu) \leq a^{-1}\, \mathcal{S}_q(\mu, \nu).$$

This completes the proof. □

For Proposition 3.5.

Proof of Proposition 3.5. For $f \in W^{1,p'}(\mathbb{G}, \lambda)$, we have by representation (3.1) that

$$f(x) = f(z_0) + \int_{[z_0,x]} f'(y)\lambda(\mathrm{d}y).$$

Let $\mathbf{1}_{[z_0,x]}(y)$ denote the indicator function of the shortest path $[z_0, x]$. That is, $\mathbf{1}_{[z_0,x]}(y)$ equals to 1 if $y \in [z_0, x]$ and equals to 0 otherwise. Then we obtain

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

By using Fubini's theorem to interchange the order of integration in the last expression, we further get

$$\begin{aligned} \int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) &= f(z_0)\mu(\mathbb{G}) + \int_{\mathbb{G}} \int_{\mathbb{G}} \mathbf{1}_{[z_0,x]}(y)\, f'(y)\mu(\mathrm{d}x)\lambda(\mathrm{d}y) \\ &= f(z_0)\mu(\mathbb{G}) + \int_{\mathbb{G}} \Big[\int_{\mathbb{G}} \mathbf{1}_{[z_0,x]}(y)\, \mu(\mathrm{d}x)\Big] f'(y)\lambda(\mathrm{d}y) \\ &= f(z_0)\mu(\mathbb{G}) + \int_{\mathbb{G}} f'(y)\mu(\Lambda(y))\, \lambda(\mathrm{d}y), \end{aligned}$$

where we have used the definition of $\Lambda(y)$ in (2.1) to obtain the last identity.

By exactly the same reason, we also have

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

Therefore, as $\mu(\mathbb{G}) = \nu(\mathbb{G})$ we infer from Definition 3.2 that

$$\mathcal{S}_p(\mu,\nu) = \sup_{f\in\mathbb{B}} \int_{\mathbb{G}} f'(x)\big[\mu(\Lambda(x)) - \nu(\Lambda(x))\big]\,\lambda(\mathrm{d}x),$$

where $\mathbb{B} \coloneqq \Big\{f \in W^{1,p'}(\mathbb{G},\lambda) :\, \|f'\|_{L^{p'}(\mathbb{G},\lambda)} \le 1\Big\}$.

Clearly, $\{f' :\, f\in\mathbb{B}\} \subset \{g \in L^{p'}(\mathbb{G},\lambda) :\, \|g\|_{L^{p'}(\mathbb{G},\lambda)} \le 1\}$. On the other hand, for any $g \in L^{p'}(\mathbb{G},\lambda)$ we have $g = f'$ with $f(x) \coloneqq \int_{[z_0,x]} g(y)\lambda(\mathrm{d}y) \in W^{1,p'}(\mathbb{G},\lambda)$. It follows that $\{f' :\, f\in\mathbb{B}\} = \{g \in L^{p'}(\mathbb{G},\lambda) :\, \|g\|_{L^{p'}(\mathbb{G},\lambda)} \le 1\}$, and hence we can rewrite $\mathcal{S}_p(\mu,\nu)$ as

$$\mathcal{S}_p(\mu,\nu) = \sup_{\|g\|_{L^{p'}(\mathbb{G},\lambda)}\le 1} \int_{\mathbb{G}} g(x)\big[\mu(\Lambda(x)) - \nu(\Lambda(x))\big]\,\lambda(\mathrm{d}x) = \Big[\int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p\,\lambda(\mathrm{d}x)\Big]^{\frac{1}{p}}, \tag{A.2}$$

which is the desired conclusion. Let us explain in details how to obtain the last identity in (A.2). Firstly, by Hölder's inequality we have

$$\int_{\mathbb{G}} g(x)\big[\mu(\Lambda(x)) - \nu(\Lambda(x))\big]\,\lambda(\mathrm{d}x) \le \Big[\int_{\mathbb{G}} |g(x)|^{p'}\,\lambda(\mathrm{d}x)\Big]^{\frac{1}{p'}} \Big[\int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p\,\lambda(\mathrm{d}x)\Big]^{\frac{1}{p}},$$

and so

$$\sup_{\|g\|_{L^{p'}(\mathbb{G},\lambda)}\le 1} \int_{\mathbb{G}} g(x)\big[\mu(\Lambda(x)) - \nu(\Lambda(x))\big]\,\lambda(\mathrm{d}x) \le \Big[\int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p\,\lambda(\mathrm{d}x)\Big]^{\frac{1}{p}}.$$

Secondly, by choosing

$$g^*(x) = \frac{|r(x)|^{p-2}r(x)}{\|r\|^{p-1}_{L^p(\mathbb{G},\lambda)}} \quad \text{with} \quad r(x) \coloneqq \mu(\Lambda(x)) - \nu(\Lambda(x))$$

we see that $\|g^*\|_{L^{p'}(\mathbb{G},\lambda)} = 1$ and $\int_{\mathbb{G}} g^*(x)\big[\mu(\Lambda(x)) - \nu(\Lambda(x))\big]\,\lambda(\mathrm{d}x) = \Big[\int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p\,\lambda(\mathrm{d}x)\Big]^{\frac{1}{p}}$. Thus we infer that last identity in (A.2) holds true, and the function g^* is a maximizer for the optimization problem in (A.2). □

For Corollary 3.6.

Proof of Corollary 3.6. We first recall that $\langle u,v\rangle$ denotes the line segment in $\mathbb{R}^n$ connecting two points u, v, while (u,v) means the same line segment but without its two end-points. Then from Proposition 3.5 and as λ has no atom, we get

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

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

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

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

$$\mathcal{S}_p(\mu,\nu)^p = \sum_{e=\langle u,v\rangle\in E} \int_{(u,v)} |\mu(\gamma_e) - \nu(\gamma_e)|^p\,\lambda(\mathrm{d}x) = \sum_{e\in E} \big|\mu(\gamma_e) - \nu(\gamma_e)\big|^p \lambda(e),$$

which leads to the postulated result. □

A.2 Proofs and Results for Section 4

For Lemma 4.2.

Proof of Lemma 4.2. Let $[x, y]$ be a shortest path connecting x and y. Assume that this path goes through nodes $v_1, ..., v_k$, then obviously $\langle x, v_1\rangle$, $\langle v_1, v_2\rangle$, ...,$\langle v_k, y\rangle$ are corresponding shortest paths w.r.t. its end-points. Therefore, it follows from Definition 4.1 for λ^* that

$$\begin{aligned}\lambda^*([x,y]) &= \lambda^*(\langle x, v_1\rangle) + \lambda^*(\langle v_1, v_2\rangle) + \cdots + \lambda^*(\langle v_k, y\rangle)\\ &= d(x, v_1) + d(v_1, v_2) + \cdots + d(v_k, y)\\ &= d(x, y),\end{aligned}$$

where the last identity is due to the assumption that $\mathbb{G}$ has no short cuts. □

For Corollary 4.3.

Proof of Corollary 4.3. This is a consequence of our Proposition 3.5 for $p = 1$ and the results obtained in (Evans and Matsen, 2012; Le et al., 2019). Indeed, when $\mathbb{G}$ is a tree with root z_0, it is shown in (Evans and Matsen, 2012, Equation (5)) and in the proof of Proposition 1 in (Le et al., 2019) that the 1-Wasserstein distance (see (B.1) for its definition) is given by

$$\mathcal{W}_1(\mu, \nu) = \int_{\mathbb{G}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|\, \lambda(\mathrm{d}x)$$

for any $\mu, \nu \in \mathcal{P}(\mathbb{G})$. By comparing this with our Proposition 3.5, we conclude that $\mathcal{S}_1(\mu, \nu) = \mathcal{W}_1(\mu, \nu)$. □

For Lemma 4.4.

Proof of Lemma 4.4. This is a direct consequence of Proposition 3.4 and Corollary 4.3. Indeed, we obtain from Proposition 3.4 that $\mathcal{S}_1(\mu, \nu) \leq \lambda^*(\mathbb{G})^{\frac{1}{p'}} \mathcal{S}_p(\mu, \nu)$ for any $1 \leq p \leq \infty$ (notice that the case $p = 1$ is trivial since $\frac{1}{p'} = 0$). Therefore, the conclusion follows as $\mathcal{S}_1(\mu, \nu) = \mathcal{W}_1(\mu, \nu)$ by Corollary 4.3. □

For Proposition 4.5.

Proof of Proposition 4.5. The last statement is just a consequence of Corollary 3.6. For the second statement, observe that the condition $\alpha \in \mathcal{K}$ ensures that $a^i \geq 0$ for all i. Also $\sum_{i=1}^n a^i = 1$ since by inspection it is easy to see that

$$\sum_{e=\langle x_1, v\rangle : v \in N(x_1)} \alpha_e + \sum_{i=2}^n \sum_{e=\langle x_i, v\rangle : v \in N'(x_i)} \alpha_e = \sum_{i=2}^n \alpha_{\langle \hat{x}_i, x_i\rangle}.$$

Therefore, $\rho \coloneqq \sum_{i=1}^n a^i \delta_{x_i}$ is a probability distribution on V. That is, $\rho \in \mathcal{P}(V)$.

The second statement also implies that the map (4.1) is onto. Indeed, for any given $\alpha = (\alpha_e)_{e \in E} \in \mathcal{K}$, let $\rho = \sum_{i=1}^n a^i \delta_{x_i} \in \mathcal{P}(V)$ be the corresponding measure given by the second statement in Proposition 4.5. Let $e \in E$ be arbitrary. Then either $\gamma_e = \varnothing$ or $\gamma_e \neq \varnothing$. In the first case, we obviously have $\rho(\gamma_e) = 0 = \alpha_e$. On the other hand, for the second case if we let x_i be the node on the edge e with the smaller distance to z_0, then $\gamma_e = \{x_i\} \cup \left(\cup_{e' = \langle x_i, v\rangle : v \in N'(x_i)} \gamma_{e'} \right)$ and this is the disjoint union. Thus,

$$\rho(\gamma_e) = \rho(\{x_i\}) + \sum_{e' = \langle x_i, v\rangle : v \in N'(x_i)} \rho(\gamma_{e'}) = a^i + \sum_{e = \langle x_i, v\rangle : v \in N'(x_i)} \alpha_{e'} = \alpha_e,$$

where the second equality is due to the induction process by repeating and tracing back to the base case $N'(x_i) = \varnothing$ to show that $\rho(\gamma_{e'}) = \alpha_{e'}$, and the last equality is by (4.2). Thus the map (4.1) is onto.

To show that the map (4.1) is one-to-one, assume that there exist $\rho_1,\ \rho_2 \in \mathcal{P}(V)$ such that $\rho_1(\gamma_e \cap V) = \rho_2(\gamma_e \cap V)$ for every $e \in E$. Let $\alpha \coloneqq (\rho_1(\gamma_e \cap V))_{e\in E} = (\rho_2(\gamma_e \cap V))_{e\in E}$, and define $\rho \coloneqq \sum_{i=1}^n a^i \delta_{x_i} \in \mathcal{P}(V)$ with a^i being given by (4.2). Since

$$\rho_1(\{x_1\}) = 1 - \sum_{e=\langle x_1,v\rangle : v\in N(x_1)} \rho_1(\gamma_e \cap V),$$
$$\rho_1(\{x_i\}) \coloneqq \rho_1(\gamma_{\langle \hat{x}_i, x_i\rangle}) - \sum_{e=\langle x_i,v\rangle : v\in N'(x_i)} \rho_1(\gamma_e \cap V) \quad \text{for } i = 2, ..., n,$$

we infer that $\rho_1(\{x_i\}) = a^i$ for all i. Due to the above choice of the distribution ρ, we therefore conclude that $\rho_1 = \rho$. By exactly the same reasoning, we also have $\rho_2 = \rho$. Thus $\rho_1 = \rho_2$, and hence the map (4.1) is one-to-one. So the first statement in Proposition 4.5 holds true, and the proof is complete. □

For Proposition 4.6

Proof of Proposition 4.6. Let ℓ_p be the distance on $\mathbb{R}^m$ defined by: for $x, z \in \mathbb{R}^m$, $\ell_p(x,z) = \|x - z\|_p = \left(\sum_{i=1}^m \left|x_{(i)} - z_{(i)}\right|^p\right)^{1/p}$ where $x_{(i)}$ is the i^{th} coordinate of x. We will first prove that for $1 \le p \le 2$, the ℓ_p distance and ℓ_p^p are negative definite.

For $a, b \in \mathbb{R}$, it is obvious that the function $(a,b) \mapsto (a-b)^2$ is negative definite. Consider $1 \le p \le 2$ and follow (Berg et al., 1984, Corollary 2.10, pp.78), the function $(a,b) \mapsto |a-b|^p$ is negative definite. It follows that ℓ_p^p is negative definite since it is a sum of negative definite functions. Using this and by applying (Berg et al., 1984, Corollary 2.10, pp.78) for the function ℓ_p^p, we also have that the function ℓ_p is negative definite.

We are now ready to prove the negative definiteness for $\mathcal{S}_p$ and $\mathcal{S}_p^p$. Let m be the number of edges in the graph $\mathbb{G}$. Due to Corollary 3.6, $\lambda^*(e)^{\frac{1}{p}}\mu(\gamma_e) = w_e^{\frac{1}{p}}\mu(\gamma_e)$ with $e \in E$ can be regarded as a feature map for probability measure μ onto $\mathbb{R}_+^m$. Therefore, $\mathcal{S}_p$ is equivalent to the ℓ_p distance between these feature maps (see also Proposition 4.5). Hence, $\mathcal{S}_p$ and $\mathcal{S}_p^p$ are negative definite for $1 \le p \le 2$. □

For Proposition 4.8.

Proof of Proposition 4.8. By Lemma 3.3 we know that $\mathcal{S}_p^{z_0}$ is a metric on $\mathcal{P}(\mathbb{G})$ for a given unique-root node z_0. On the other hand, according to Definition 4.7, $\mathcal{S}_p^\eta$ is a convex combination of the metric $\mathcal{S}_p^{z_0}$ with $z_0 \in \mathcal{Z}_0$[9]. Therefore, it follows immediately that $\mathcal{S}_p^\eta$ is also a metric. Indeed, the nonnegativity and symmetry are obvious. Also, if $\mathcal{S}_p^\eta(\mu,\nu) = 0$ then we have $\mathcal{S}_p^{z_0}(\mu,\nu) = 0$ for every point $z_0 \in \mathcal{Z}_0$ satisfying $\eta(\{z_0\}) > 0$. As $\sum_{z_0\in\mathcal{Z}_0} \eta(\{z_0\}) = 1$, there must exists a point $\tilde{z}_0 \in \mathcal{Z}_0$ such that $\eta(\{\tilde{z}_0\}) > 0$. Thus we obtain $\mathcal{S}_p^{\tilde{z}_0}(\mu,\nu) = 0$, and hence $\mu = \nu$ by Lemma 3.3. To check the triangle inequality, let $\mu, \nu, \sigma \in \mathcal{P}(\mathbb{G})$ be arbitrary. We then use Definition 4.7 and Lemma 3.3 to get

$$\begin{aligned}\mathcal{S}_p^\eta(\mu,\nu) = \sum_{z_0\in\mathcal{Z}_0} \eta(\{z_0\})\, \mathcal{S}_p^{z_0}(\mu,\nu) &\le \sum_{z_0\in\mathcal{Z}_0} \eta(\{z_0\}) \left[\mathcal{S}_p^{z_0}(\mu,\sigma) + \mathcal{S}_p^{z_0}(\sigma,\nu)\right] \\ &= \mathcal{S}_p^\eta(\mu,\sigma) + \mathcal{S}_p^\eta(\sigma,\nu).\end{aligned}$$

We thus conclude that $\mathcal{S}_p^\eta$ is a metric on $\mathcal{P}(\mathbb{G})$. □

B FUTHER RESULTS AND DISCUSSIONS

In this section, we give brief reviews about important aspects in our works, provide further experimental results and further discussions for our proposed Sobolev transport distance.

B.1 Brief Reviews

In this section, we briefly review about important aspects in our work and provide further experimental results.

[9] We assume that $\mathcal{Z}_0 \neq \varnothing$. This assumption is easily satisfied for general graph metric built from data points. See further discussion about the set $\mathcal{Z}_0$ (or the Assumption 2.1 in the main text) in §B.

For Kernels. We review some important definitions (e.g., positive/negative definite kernels (Berg et al., 1984)) and theorems (e.g., Theorem 3.2.2 in Berg et al. (1984)) about kernels used in our work.

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

$$\sum_{i,j} c_i c_j k(x_i, x_j) \geq 0, \qquad \forall c_i \in \mathbb{R}.$$

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

$$\sum_{i,j} c_i c_j k(x_i, x_j) \leq 0, \qquad \forall c_i \in \mathbb{R} \text{ s.t. } \sum_i c_i = 0.$$

- **Theorem 3.2.2 in (Berg et al., 1984, pp. 74) for Kernels.** If κ is a *negative definite* kernel, then $\forall t > 0$, kernel

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

  is positive definite.

For Persistence Diagrams and Definitions in Topological Data Analysis. We refer the reader to Kusano et al. (2017, §2) for a review about mathematical framework for persistence diagrams (e.g., persistence diagrams, filtrations, persistent homology).

For the Integral Probability Metric. Let $\mathfrak{F}$ be a class of real-valued bounded measurable functions on Ω; μ, ν be two Borel probability distributions on Ω, then the integral probability metric $\mathcal{I}$ associated with $\mathfrak{F}$ (Müller, 1997) is defined as follow:

$$\mathcal{I}_{\mathfrak{F}} \coloneqq \sup_{f \in \mathfrak{F}} \left| \int_\Omega f(x)\mu(dx) - \int_\Omega f(z)\nu(dz) \right|.$$

Some popular instances of the integral probability metrics are: (i) Dudley metric, (ii) Wasserstein metric, (iii) total variation metric, (iv) Kolmogorov metric, (v) maximum mean discrepancies, to name a few (Sriperumbudur et al., 2009; Müller, 1997).

For the 1-Wasserstein Distance. Let μ, ν be two Borel probability distributions on Ω, $R(\mu, \nu)$ be the set of probability distributions π on $\Omega \times \Omega$ such that $\pi(A \times \Omega) = \mu(A)$ and $\pi(\Omega \times B) = \nu(B)$ for all Borel sets A, B. The 1-Wasserstein distance $\mathcal{W}_1$ with a cost function c is defined as follow:

$$\mathcal{W}_1(\mu, \nu) = \inf \left\{ \int_{\Omega \times \Omega} c(x, z)\pi(dx, dz) \mid \pi \in R(\mu, \nu) \right\}. \tag{B.1}$$

Let $\mathcal{F}_c$ be the set of Lipschitz functions w.r.t. the cost function c, i.e. functions $f : \Omega \to \mathbb{R}$ such that $|f(x) - f(z)| \leq c(x, z), \forall x, z \in \Omega$. Then, the dual of (B.1) is:

$$\mathcal{W}_1(\mu, \nu) = \sup_{f \in \mathcal{F}_c} \left\{ \int_\Omega f(x)\mu(dx) - \int_\Omega f(z)\nu(dz) \right\}. \tag{B.2}$$

B.2 Further Discussions

About the Assumption 2.1. In our setting, the nodes in the graph are points in $\mathbb{R}^n$, edge weights are the distance (e.g., ℓ_2 distance) between two corresponding nodes (i.e., points in $\mathbb{R}^n$). Therefore, consider any two nodes in the graph, there may be several paths connecting one node to the other node, and with a high probability, lengths of those paths are different. Hence, it is almost surely that every node in the graph can be regarded as unique-path root node.

In case, we have some special graph, e.g., a grid of nodes. There is no unique-path root node for such graph. However, we can easily *adjust/approximate* such graph into a graph with unique-path root nodes by randomly perturbing each node of such graph in a ball (e.g., ℓ_2 ball) with a small radius.

About the Proposed Sobolev Transport Distance. In our setting, we assume that we know the graph metric space (i.e., the graph structure) which supports of probability measures are living. Giving such graph, we define our Sobolev transport for probability measures supported on that graph metric space.

In our experiments (in Section 5), we evaluate our proposed Sobolev transport on (i) various graph structures (e.g., $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$) (ii) with different graph sizes (e.g., the number of nodes in the graphs $M = 10^2, 10^3, 10^4, 4\times 10^4$. Performances of the Sobolev transport consistently compare favorably with those of the baseline approaches.

The question about learning the optimal graph metric structure from data for the Sobolev transport is left for future work.

A Further Note on Implementation for the Sobolev Transport. Following the closed-form solution of Sobolev transport for discrete probability measures supported on a graph metric space in Corollary 3.6 and Equation (3.3), we need to compute the mass of μ, ν on γ_e for each edge $e \in E_{\mu,\nu}$.

Recall that for any support z of a probability measure, it only contributes to the γ_e when e belongs to the shortest path in $\mathbb{G}$ from the unique-path root node z_0 to the considered support z. Therefore, we only need to run the Dijkstra algorithm for shortest paths one time for the source z_0 and the destination $(V \setminus \{z_0\})$.[10] Then, we can index for each support z in $\mathbb{G}$ for its contribution to each γ_e.[11]

Therefore, for a given probability measure μ, we only need to consider each support of μ one time to compute $\mu(\gamma_e)$ for all edge e in the graph $\mathbb{G}$ instead of a naive implementation where we need to consider all supports of μ for each γ_e in $\mathbb{G}$.

B.3 Further Experimental Results

In this section, we provide further experimental results.

Further Results for Document Classification with Word Embedding for Different Values of M (i.e., the Number of Nodes in the Graph).

- **For Graph $\mathbb{G}_{\text{Log}}$.** Similar to Figure 2 in the main text, we illustrate the SVM results and time consumption of kernel matrices for document classification with word-embedding for graph $\mathbb{G}_{\text{Log}}$ when $M = 10^3$ and $M = 10^2$ in Figures 7 and 8 respectively.

- **For Graph $\mathbb{G}_{\text{Sqrt}}$.** Similar to Figure 3 in the main text, we illustrate the SVM results and time consumption of kernel matrices for document classification with word-embedding for graph $\mathbb{G}_{\text{Sqrt}}$ when $M = 10^3$ and $M = 10^2$ in Figures 9 and 10 respectively.

Further Results for TDA for Different Values of M (i.e., the Number of Nodes in the Graph).

- **For Graph $\mathbb{G}_{\text{Log}}$.** Similar to Figure 4 in the main text, we illustrate the SVM results and time consumption for TDA for graph $\mathbb{G}_{\text{Log}}$ when $M = 10^3$ and $M = 10^2$ in Figure 11.

- **For Graph $\mathbb{G}_{\text{Sqrt}}$.** Similar to Figure 5 in the main text, we illustrate the SVM results and time consumption for TDA for graph $\mathbb{G}_{\text{Sqrt}}$ when $M = 10^3$ and $M = 10^2$ in Figure 12.

Further Results with Large Graph ($M = 40000$). We illustrate the SVM results and time consumption of kernel matrices for large graph with $M = 40000$ for both $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ in Figure 13.

[10] One can consider a set of all considered supports (exclude z_0) as the destination set for Dijkstra for a faster computation.

[11] We only need to compute this step one time (i.e., it can be considered as the preprocessing process involving only the graph structure and nothing about the probability distributions, and is done only once regardless how many pairs (μ, ν) that we have to measure. In this step by identifying shortest paths we calculate the set γ_e for each edge $e \in E$.), see our Remark 3.7.

Further Results for Slice Variants of Sobolev Transport and Tree-Wasserstein. Similar as Figure 6 in the main text, we illustrate further results for both document classification with word embedding and TDA for slice variants of Sobolev trransport and tree-Wasserstein: (i) for both $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{Sqrt}$, (ii) for different values of M (e.g., $10^2, 10^3, 10^4$).

- **For Document Classification with Word Embedding.**
  - **For Graph $\mathbb{G}_{\text{Log}}$.** We illustrate SVM results and time consumption of kernel matrices for sliced variants of Sobolev transport and tree-Wasserstein for document classification with word-embedding for graph $\mathbb{G}_{\text{Log}}$ when $M = 10^3$ and $M = 10^2$ in Figures 14 and 15 respectively.
  - **For Graph $\mathbb{G}_{\text{Sqrt}}$.** We illustrate SVM results and time consumption of kernel matrices for sliced variants of Sobolev transport and tree-Wasserstein for document classification with word-embedding for graph $\mathbb{G}_{\text{Log}}$ when $M = 10^4$, $M = 10^3$ and $M = 10^2$ in Figures 16, 17, and 18 respectively.
- **For TDA.**
  - **For Graph $\mathbb{G}_{\text{Log}}$.** We illustrate SVM results and time consumption of kernel matrices for sliced variants of Sobolev transport and tree-Wasserstein for TDA for graph $\mathbb{G}_{\text{Log}}$ when $M = 10^4$ for `Orbit` and $M = 10^3$ for `MPEG7` in Figure 19 (due to a small size of the dataset `MPEG7`); and with $M = 10^3$ and $M = 10^2$ for both datasets in Figure 20.
  - **For Graph $\mathbb{G}_{\text{Sqrt}}$.** We illustrate SVM results and time consumption of kernel matrices for sliced variants of Sobolev transport and tree-Wasserstein for TDA for graph $\mathbb{G}_{\text{Sqrt}}$ when $M = 10^4$ for `Orbit` and $M = 10^3$ for `MPEG7` in Figure 21 (due to a small size of the dataset `MPEG7`); and with $M = 10^3$ and $M = 10^2$ for both datasets in Figure 22.

Further Results with Large Graph ($M = 40000$) for sliced variants. We illustrate the SVM results and time consumption of kernel matrices for sliced variants of Sobolev transport and tree-Wasserstein for large graphs where the number of nodes is $M = 40000$ for both $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ in Figure 23.

Further Results for Tree-Wasserstein Kernel. We illustrate the SVM results for tree-Wasserstein kernel with the minimum spanning tree of the given graph, denote as $k_{\text{TW}^{\text{MST}}}$ for both graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ where the number of nodes is $M = 10000$ on document classification in Figure 24. The performances of $k_{\text{TW}^{\text{MST}}}$ improves those of k_{TW} (with random trees from a given graph).

Discussions. Through various tasks (e.g., document classification with work embedding and TDA), with various graph structure (e.g., $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$) with different graph sizes (e.g., the number of nodes in the graphs $M = 10^2, 10^3, 10^4, 4 \times 10^4$), the performances of the proposed Sobolev transport consistently compare favorably with those of other baselines. The Sobolev transport is several-order faster than the optimal transport with graph metric. Additionally, the Sobolev transport can leverage information from the graph which is more flexible and has more degree of freedom in applications than tree-Wasserstein (for tree structure). The question about learning the optimal graph structure from data is left for future work. We also think that local structures on supports such as graph structure in our work or tree structure in Le et al. (2019); Le and Nguyen (2021); Le et al. (2021a) play an important role to scale up problems in optimal transport, especially for large-scale applications.

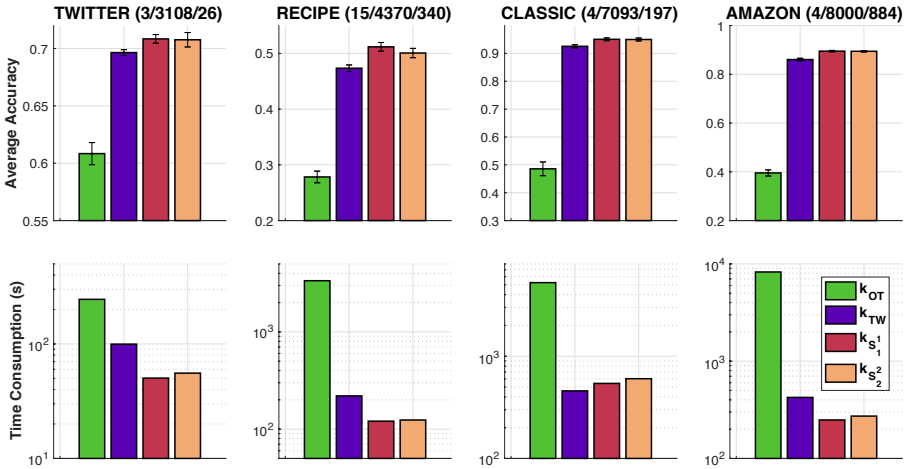


Figure 7: SVM results and time consumption for kernel matrices with $\mathbb{G}_{\text{Log}}$ where $M = 10^3$.

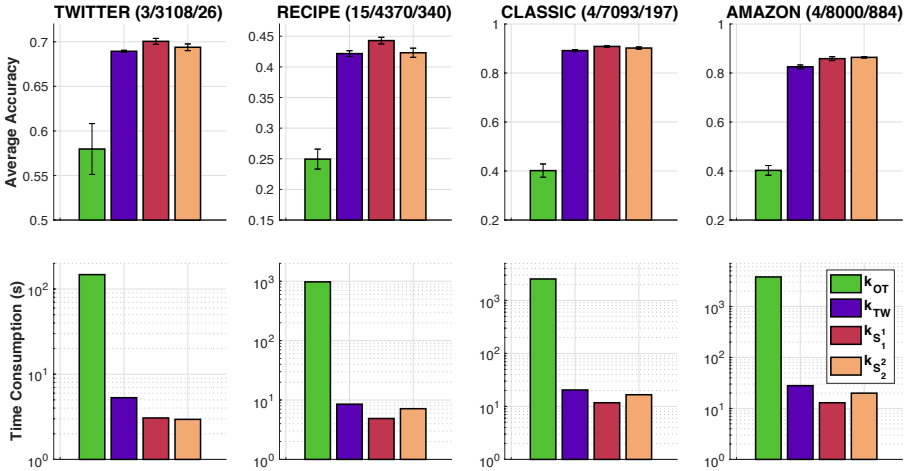


Figure 8: SVM results and time consumption for kernel matrices with $\mathbb{G}_{\text{Log}}$ where $M = 10^2$.

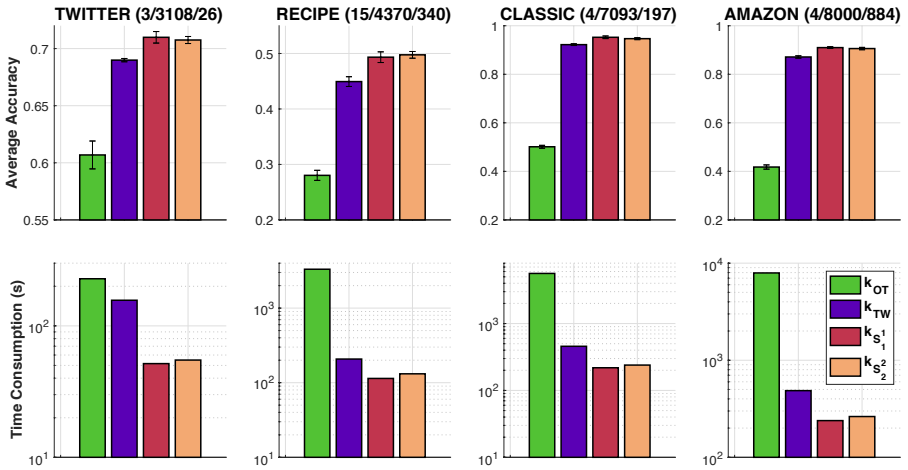


Figure 9: SVM results and time consumption for kernel matrices with $\mathbb{G}_{\text{Sqrt}}$ where $M = 10^3$.

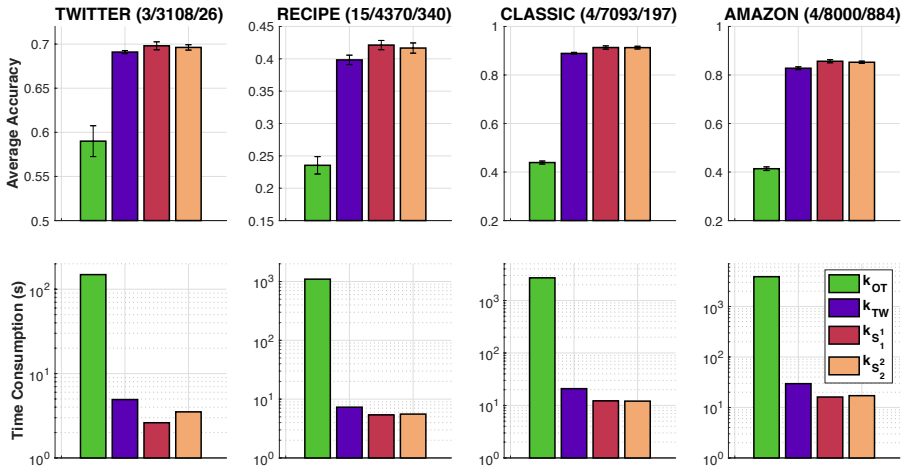
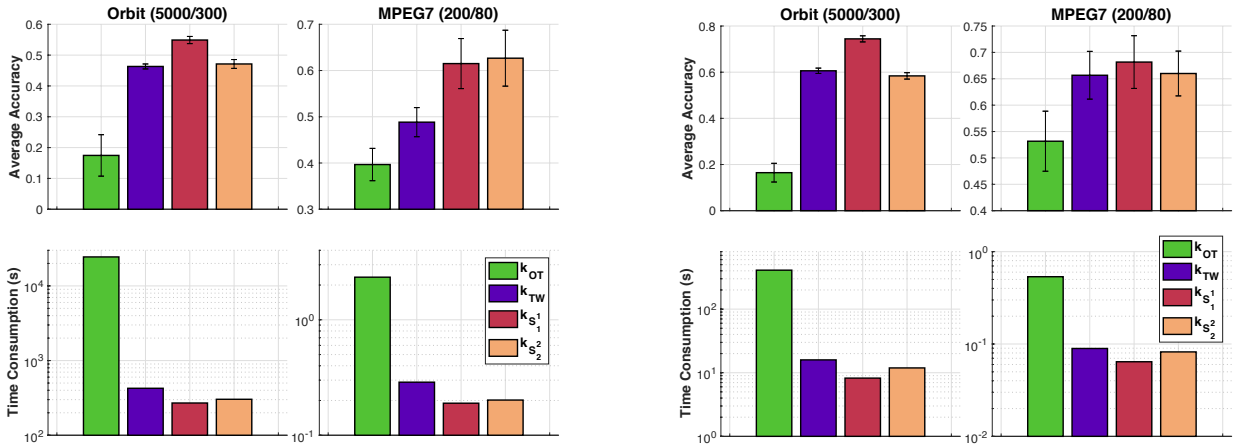
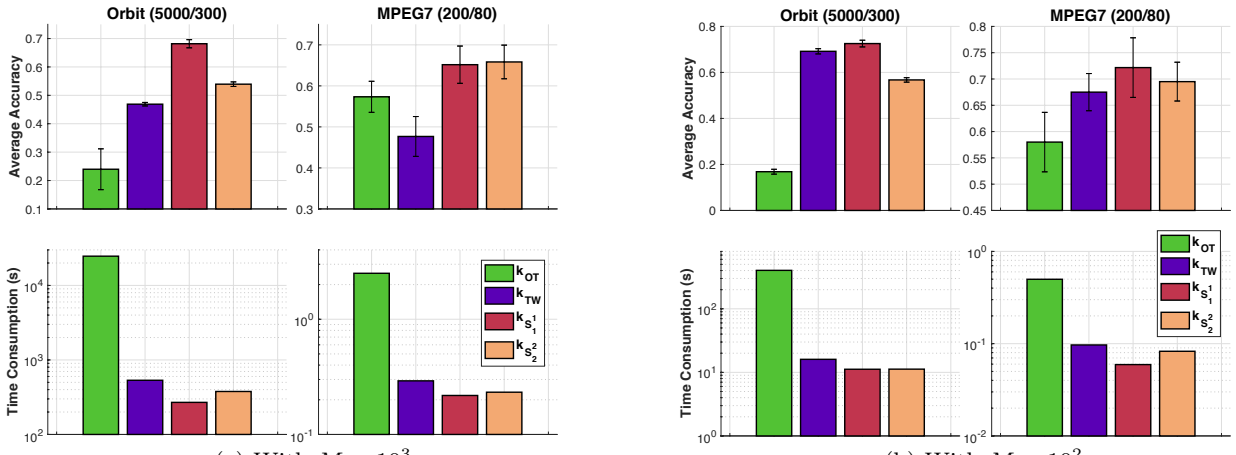


Figure 10: SVM results and time consumption for kernel matrices with $\mathbb{G}_{\text{Sqrt}}$ where $M = 10^2$.



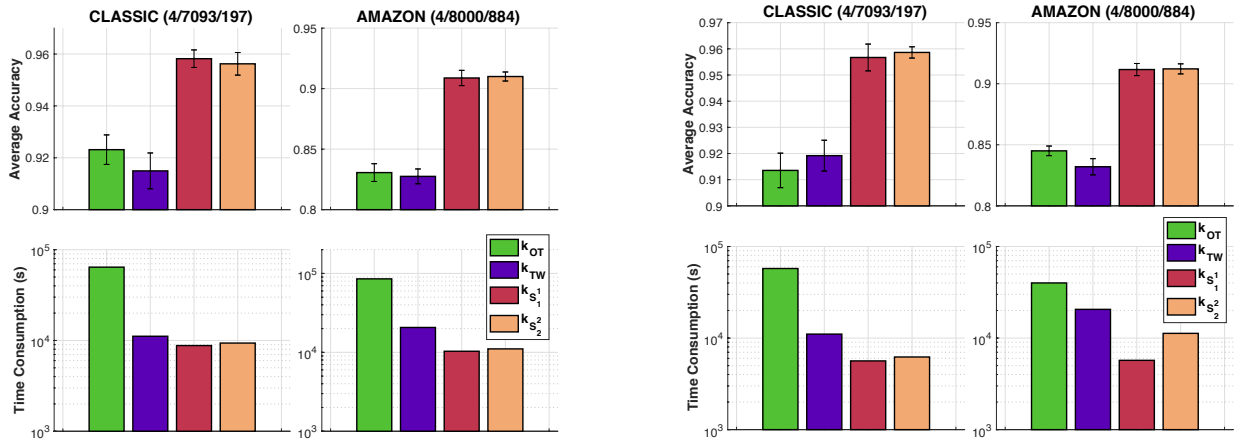
(a) With $M = 10^3$ (b) With $M = 10^2$.

Figure 11: SVM results and time consumption for kernel matrices with $\mathbb{G}_{\text{Log}}$.



(a) With $M = 10^3$ (b) With $M = 10^2$.

Figure 12: SVM results and time consumption for kernel matrices with $\mathbb{G}_{\text{Sqrt}}$.



(a) For graph $\mathbb{G}_{\text{Log}}$. (b) For graph $\mathbb{G}_{\text{Sqrt}}$.

Figure 13: SVM results and time consumption for kernel matrices with large graphs where $M = 40000$.

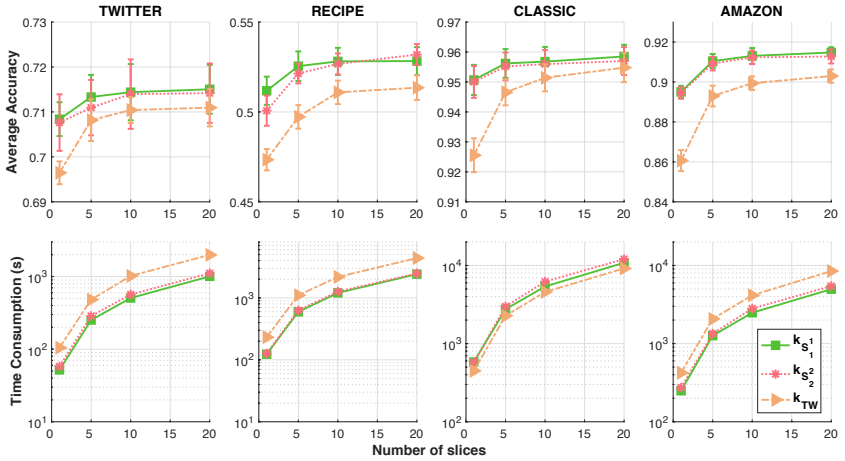


Figure 14: SVM results and time consumption for kernel matrices of slice variants with $\mathbb{G}_{\text{Log}}$ ($M = 10^3$).

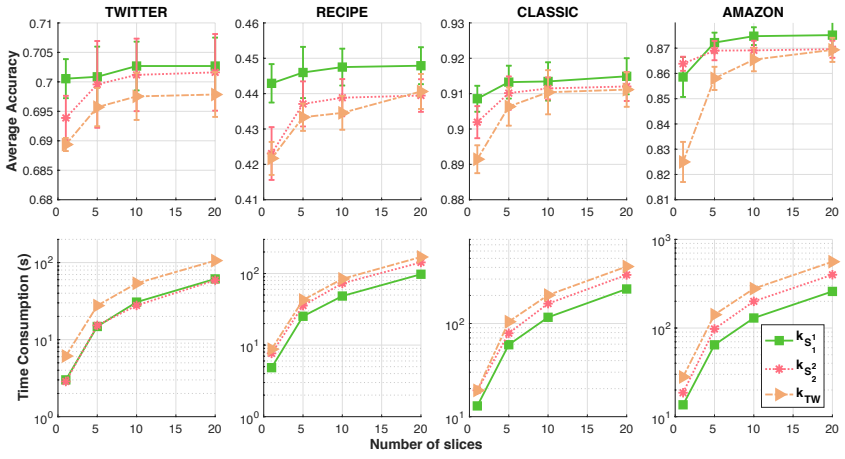


Figure 15: SVM results and time consumption for kernel matrices of slice variants with $\mathbb{G}_{\text{Log}}$ ($M = 10^2$).

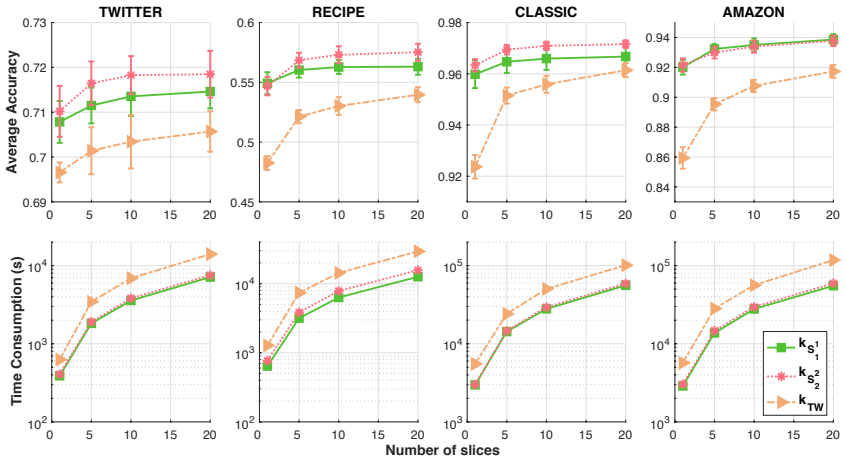


Figure 16: SVM results and time consumption for kernel matrices of slice variants with $\mathbb{G}_{\text{Sqrt}}$ ($M = 10^4$).

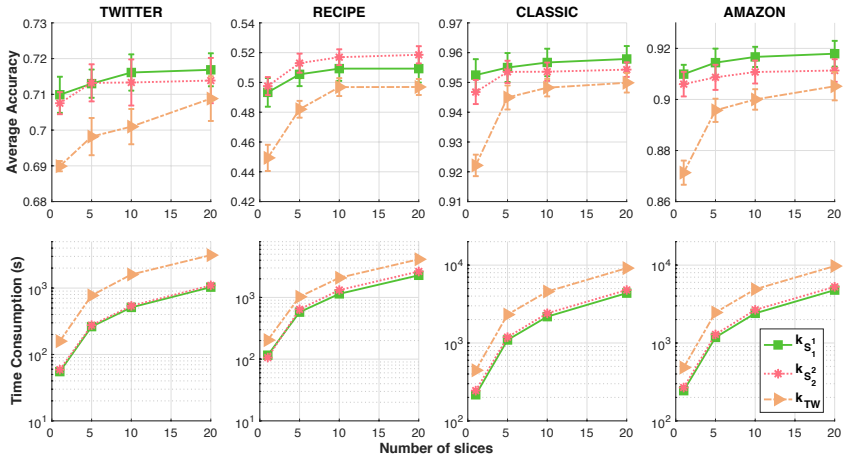


Figure 17: SVM results and time consumption for kernel matrices of slice variants with $\mathbb{G}_{\text{Sqrt}}$ ($M = 10^3$).

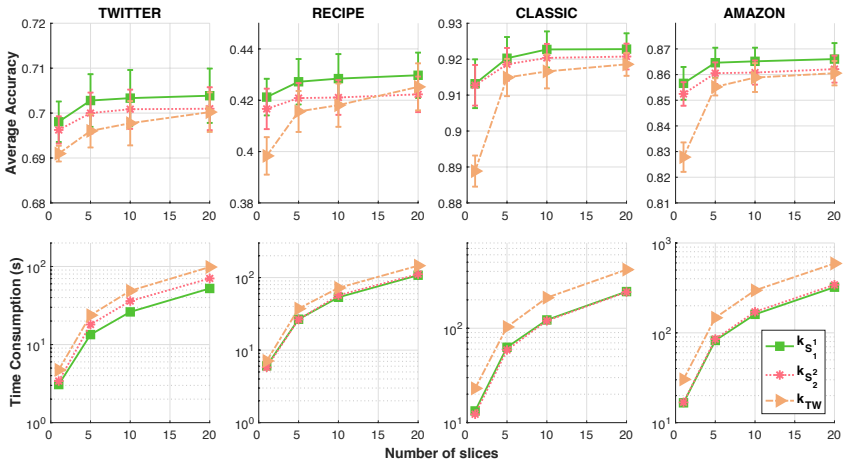


Figure 18: SVM results and time consumption for kernel matrices of slice variants with $\mathbb{G}_{\text{Sqrt}}$ ($M = 10^2$).

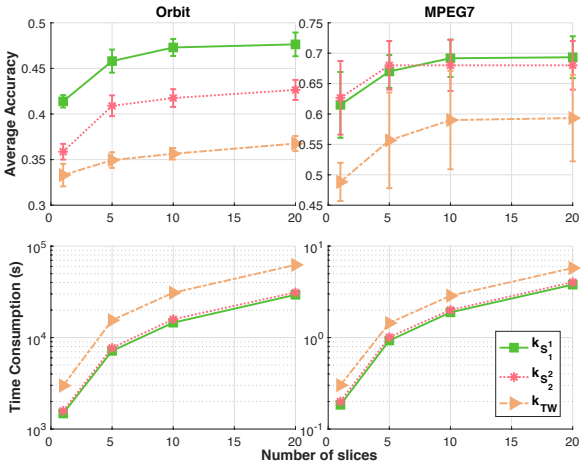
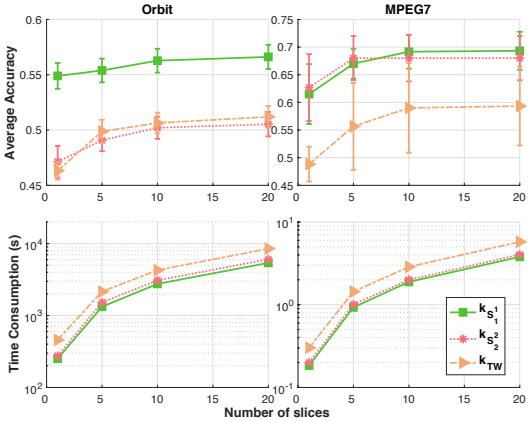
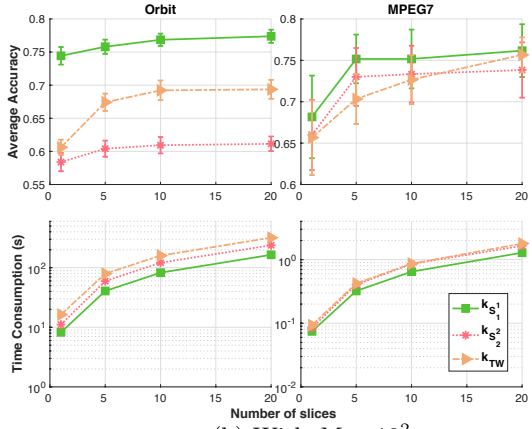


Figure 19: SVM results and time consumption for kernel matrices of slice variants with $\mathbb{G}_{\text{Log}}$ where $M = 10^4$ for `Orbit`, and $M = 10^3$ for `MPEG7`.



(a) With $M = 10^3$.



(b) With $M = 10^2$.

Figure 20: SVM results and time consumption for kernel matrices of slice variants with $\mathbb{G}_{\text{Log}}$ for TDA.

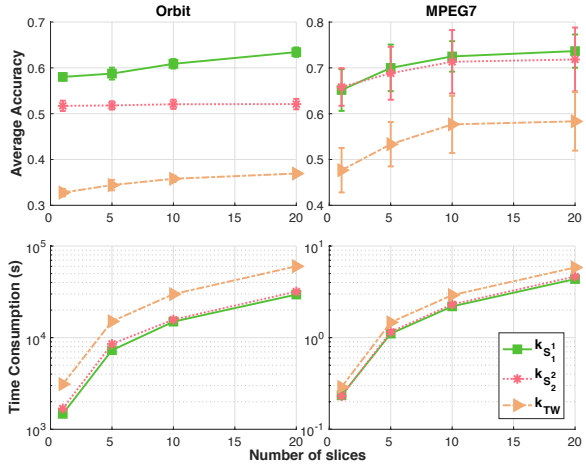
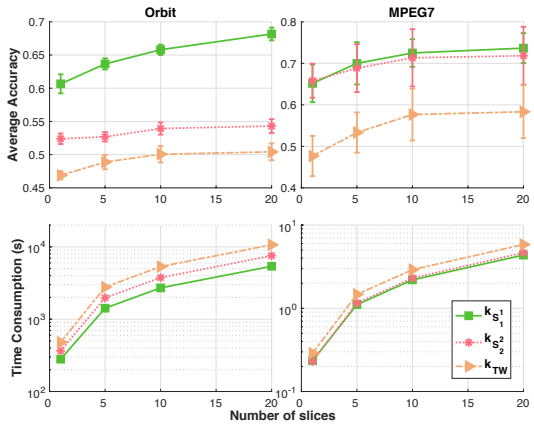
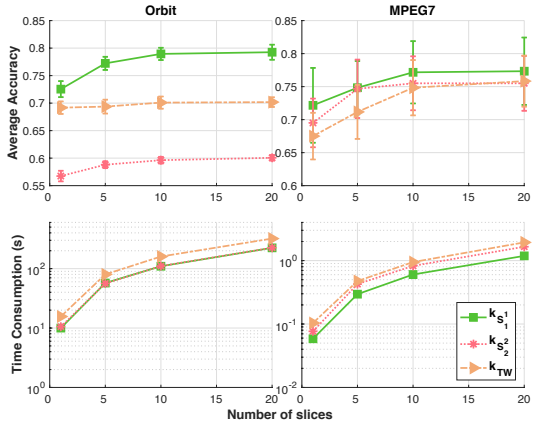


Figure 21: SVM results and time consumption for kernel matrices of slice variants with $\mathbb{G}_{\text{Sqrt}}$ where $M = 10^4$ for `Orbit`, and $M = 10^3$ for `MPEG7`.

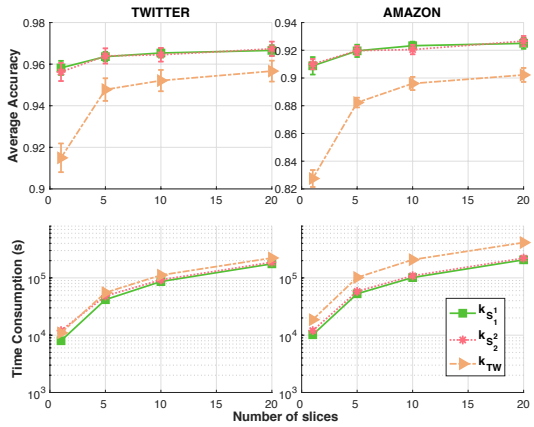


(a) With $M = 10^3$.

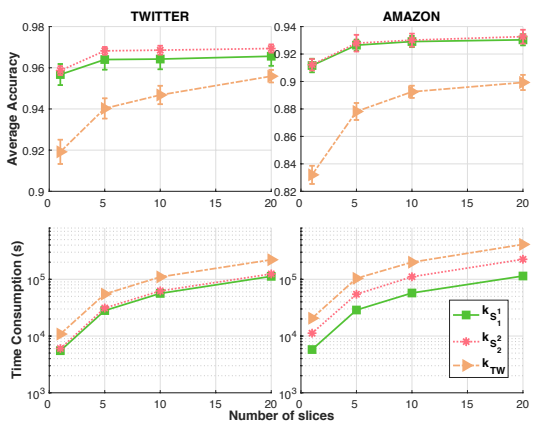


(b) With $M = 10^2$.

Figure 22: SVM results and time consumption for kernel matrices of slice variants with $\mathbb{G}_{\text{Sqrt}}$ for TDA.

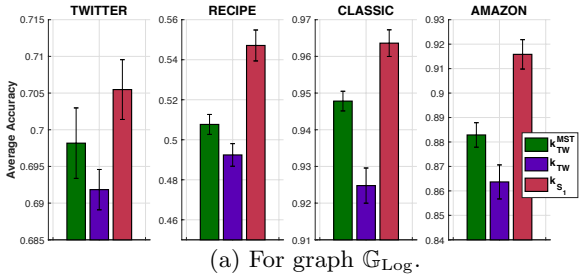


(a) For graph $\mathbb{G}_{\text{Log}}$.

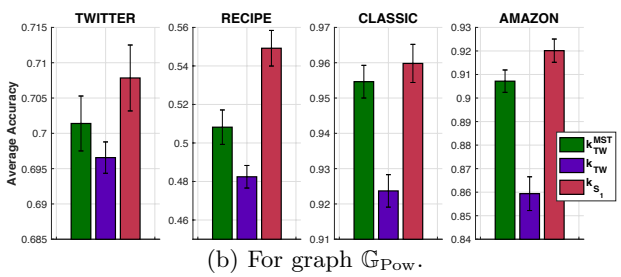


(b) For graph $\mathbb{G}_{\text{Pow}}$.

Figure 23: SVM results and time consumption for kernel matrices of slice variants with $\mathbb{G}_{\text{Sqrt}}$ for TDA with a large graph where the number of nodes is 40000.



(a) For graph $\mathbb{G}_{\text{Log}}$.



(b) For graph $\mathbb{G}_{\text{Pow}}$.

Figure 24: SVM results for document classification with $M = 10000$ graph nodes.