Orlicz–Sobolev with Musielak: An Efficient Regularization Approach for Graph-based IPM

Tam Le[*]
The Institute of Statistical Mathematics
& Tohoku University
tam@ism.ac.jp

Truyen Nguyen[*]
The University of Akron
truyennguyen3@gmail.com

Hideitsu Hino[†]
The Institute of Statistical Mathematics
hino@ism.ac.jp

Kenji Fukumizu[†]
The Institute of Statistical Mathematics
fukumizu@ism.ac.jp

Abstract

We study the Sobolev IPM problem for probability measures supported on a graph metric space, where critic function is constrained to lie within the unit ball defined by Sobolev norm. Sobolev IPM is intrinsically coupled with L^p geometric structure within its definition, limiting its ability to incorporate other prior geometry beyond the L^p paradigm. Conversely, the classic optimal transport is flexible, easy to adapt to various geometric structures by simply changing its ground cost. An important example is Orlicz-Wasserstein (OW) which utilizes *Orlicz* geometric structure to generalize the L^p within the standard p-order Wasserstein, and remarkably play a vital role to advance machine learning methodologies. Inspired by recent advantages of OW, in this work, we leverage a specific class of convex functions for Orlicz geometry to mitigate such limitation for Sobolev IPM, and propose the generalized Sobolev IPM (GSI). Our GSI approach encompasses Sobolev IPM as a special case while accommodating diverse geometric priors beyond L^p. It however brings up significant computational hurdles that compound those already notoriously inherent in Sobolev IPM. To address these challenges, we theoretically establish a novel connection between *Orlicz-Sobolev* norm and *Musielak* norm which facilitates a novel efficient regularization for GSI. By further exploiting the underlying graph structure, we show that the regularized GSI reduces to a simple univariate optimization problem, achieving notably computational efficiency, enabling its usage in practical applications. We empirically illustrate that the regularized GSI is several-order faster than the popular OW in computation, and its performances compare favorably to other transport baselines for comparing graph-based measures in document classification and topological data analysis.

1 Introduction

Probability measures serve as canonical mathematical representations for diverse objects of interest across various research fields, e.g., documents in natural language processing [26, 67], persistence diagrams in topological data analysis (TDA) [17, 33], point clouds in computer vision and graphics [22, 65, 44, 13]. To compare such measures, integral probability metrics (IPM) offer a versatile and principled class of metric functions [40, 53]. Conceptually, IPM operate by determining an optimal critic function that achieves maximal discrimination between probability measures. This

[*]: equal contribution, [†]: co-last author. Correspondence to: Tam Le <tam@ism.ac.jp>.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

mathematical elegance and versatility has facilitated the widespread adoption of IPM throughout statistics and machine learning [52, 18, 46, 36, 62, 63, 42, 24, 33, 60].

In this work, we study the Sobolev IPM problem for measures supported on a graph metric space, where critic function is constrained within the unit ball induced by Sobolev norm [2, 35, 37, 50, 33, 60]. Sobolev IPM has proven fundamental to numerous theoretical analyses, including convergence rates in density estimation and approximation theory for deep architectures [35, 37, 50]. Additionally, by exploiting the underlying graph structure, Le et al. [33] pioneered a tractable algorithmic approach for the computation of Sobolev IPM, overcoming long-standing limitations in practical use. Moreover, Tran et al. [60] leveraged a tree system [56–58, 61], and propose injective Radon transforms to further extend the utilization of Sobolev IPM for applications with probability measures in the popular Euclidean spaces and hyperspheres, including gradient flows, generative modeling with diffusion models, and topic modeling. However, Sobolev IPM is intrinsically coupled with L^p geometric structure, thereby limiting its ability to incorporate other prior geometry beyond the L^p [33, 60]. Furthermore, we note that one should distinguish Sobolev IPM problem from Sobolev transport [30] and its generalized approach (i.e., generalized Sobolev transport (GST) [32]) which only constraint on gradient of critic function, while the constraints of Sobolev IPM problem involve both critic function and its gradient within the Sobolev norm.

Unlike Sobolev IPM, the classic optimal transport (OT) has a flexibility to accommodate prior structures by simply modifying the underlying cost function. In a seminal work, Sturm [54] utilized Orlicz geometric structure to propose Orlicz-Wasserstein (OW) for moving beyond the L^p structure in OT. Comparing to the standard p-order Wasserstein, OW [54, 23, 19, 3] has demonstrated remarkable effectiveness in advancing machine learning methodologies. Notably, Altschuler & Chewi [3] leverage OW to facilitate the development of differential-privacy-inspired methodologies that address long-lasting convergence challenges in hypocoercive differential equations. Moreover, Guha et al. [19] demonstrate that OW substantially enhances Bayesian contraction rates, effectively circumventing limitations inherent to traditional OT with Euclidean ground cost. Furthermore, Orlicz structure has found broad applicability across diverse machine learning paradigms [32, 34, 4, 51, 15, 9, 38]. For comprehensive studies of Orlicz geometry, see [2, 48].

Inspired by advantages of OW, and to move beyond the L^p paradigm for Sobolev IPM, we advocate to employ a specific class of convex functions for Orlicz geometric structure, and propose the generalized Sobolev IPM (GSI). Our approach includes Sobolev IPM as a special case while enabling diverse prior geometric structures beyond L^p for graph-based IPM. Analogous to the historically computational challenges notoriously inherent in Sobolev IPM, the GSI approach poses significant computational obstacles, which are beyond the recent advancements developed for OW [32] and Sobolev IPM [33]. To address these challenges, we theoretically establish a novel connection between *Orlicz-Sobolev* norm and *Musielak* norm which in turn motivates a novel regularization scheme for the GSI problem.[2] Exploiting the underlying graph structure, we further show that the regularized GSI reduces to a simple univariate optimization problem, yielding substantial computational efficiency and enabling practical deployment at scale. Therefore, our approach not only helps to accommodate geometric structures beyond L^p regime for Sobolev IPM, but also overcomes the arisen computational barriers of GSI, and paves the way for its applications, even at scale, in practice.

Contribution. Our contributions are three-fold as follows:

- We leverage a certain class of convex functions, enabling diverse geometric structures beyond L^p paradigm for Sobolev IPM for the first time, to our knowledge. Additionally, we propose a novel regularization approach for the resulting GSI problem that yields an efficient computation by simply solving a univariate optimization problem.
- We prove that the regularized GSI is a metric, and it is equivalent to the original GSI. Moreover, we theoretically establish its connections to Sobolev IPM, and other transport distances.
- Empirically, we illustrate that the regularized GSI is several-order faster than the popular OW, and its performances compare favorably with other transport baselines for comparing graph-based measures under the same settings in document classification and TDA.

[2] As a by-product of the theoretical finding of such connection, one can draw a connection between GST and the proposed GSI problem, as well as its regularization approach, detailed in §4.

Organization. In §2, we review related backgrounds and notations. We describe the GSI and its novel regularization in §3. In §4, we provide theoretical properties for the regularized GSI. Then, in §5, we empirically show that the regularized GSI is fast in computation, and its performances compare favorably to other transport baselines for comparing graph-based measures under the same settings in document classification and TDA, following by concluding remarks in §6. Supplementary materials including further theoretical and empirical results, detailed discussions, reviews on related notions, and full proofs are available in the Appendix. We also release code for the proposed approach.[3]

2 Backgrounds and Notations

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

Graph. We follow the problem setting for probability measures on a given graph, as in [30, 32, 33]. We consider a connected, undirected, finite, and physical graph $\mathbb{G}$ with sets of nodes and edges V, E respectively, and positive edge lengths $\{w_e\}_{e\in E}$. For continuous graph setting, $\mathbb{G}$ is regarded as the set of all nodes in V and all points forming the edges in E. Additionally, we equip $\mathbb{G}$ with graph metric $d_{\mathbb{G}}(x, z)$ which equals to the length of the shortest path $[x, z]$ between $x, z \in \mathbb{G}$, and assume that there exists a root node $z_0 \in V$ such that for any considered point $x \in \mathbb{G}$, then $[z_0, x]$ is unique. For $x \in \mathbb{G}, e \in E$, define the sets $\Lambda(x)$ and γ_e as follows:

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

Let $\mathcal{P}(\mathbb{G})$ (resp. $\mathcal{P}(\mathbb{G} \times \mathbb{G})$) be the set of all nonnegative Borel measures on $\mathbb{G}$ (resp. $\mathbb{G} \times \mathbb{G}$) with a finite mass. 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. Similar notation is used for continuous functions on $\mathbb{G} \times \mathbb{G}$.

A family of convex functions. We consider the set of N-functions [2, §8.2], where a strictly increasing and convex function $\Phi : [0, \infty) \to [0, \infty)$ is called an N-function if $\lim_{t\to 0} \frac{\Phi(t)}{t} = 0$ and $\lim_{t\to+\infty} \frac{\Phi(t)}{t} = +\infty$.

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

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

For $\Phi(t) = t^p$ with $1 < p < \infty$, then $\|f\|_{L_\Phi}$ is equal to the L^p norm $\|f\|_{L^p}$ [2]. For positive weight function $\hat{w}$ on $\mathbb{G}$, consider the weighted $L^{\hat{w}}_\Phi(\mathbb{G}, \lambda)$ as the linear hull of the set of all Borel measurable functions $f : \mathbb{G} \to \mathbb{R}$ satisfying $\int_{\mathbb{G}} \hat{w}(x)\Phi(|f(x)|)\lambda(\mathrm{d}x) < \infty$. Then, $L^{\hat{w}}_\Phi(\mathbb{G}, \lambda)$ is a normed space with *Musielak* norm [41, §10.2][4], defined as

$$\|f\|_{L^{\hat{w}}_\Phi} := \inf\left\{t > 0 \mid \int_{\mathbb{G}} \hat{w}(x)\, \Phi\left(\frac{|f(x)|}{t}\right) \lambda(\mathrm{d}x) \le 1\right\}. \tag{3}$$

Sobolev IPM. For an exponent $1 \le p \le \infty$ and its conjugate p', i.e., $p' \in [1, \infty]$ satisfying $\frac{1}{p} + \frac{1}{p'} = 1$ and if $p = 1$, then $p' = \infty$, let $W^{1,p}_0(\mathbb{G}, \lambda)$ be the subspace consisting of all functions f in the graph-based Sobolev space $W^{1,p}(\mathbb{G}, \lambda)$ [30, Definition 3.1] satisfying $f(z_0) = 0$, $\|f\|_{W^{1,p}}$ is the Sobolev norm [2, §3.1], defined as

$$\|f\|_{W^{1,p}} := \left(\|f\|^p_{L^p} + \|f'\|^p_{L^p}\right)^{\frac{1}{p}}. \tag{4}$$

Let $\mathcal{B}_p := \left\{f \in W^{1,p}_0(\mathbb{G}, \lambda), \|f\|_{W^{1,p}} \le 1\right\}$, then Sobolev IPM between $\mu, \nu \in \mathcal{P}(\mathbb{G})$ is defined as

$$\mathcal{S}_p(\mu, \nu) := \sup_{f\in\mathcal{B}_p} \left|\int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(y)\nu(\mathrm{d}y)\right|. \tag{5}$$

[3] The code repository is on `https://github.com/lttam/GenSobolevIPM-Musielak`.

[4] The weighted $L^{\hat{w}}_\Phi(\mathbb{G}, \lambda)$ is a specific instance of the *Musielak-Orlicz* space, where the generalized N-function $\bar{\Phi}(x, t) = \hat{w}(x)\Phi(t)$ for all $t > 0$ and $x \in \mathbb{G}$, and also see [20, Definition 3.2.1] for Musielak norm.

From Equation (5), Sobolev IPM is intrinsically coupled with the L^p geometric structure in its definition. Consequently, unlike optimal transport, where one can easily adapt to various prior geometric structures by simply tailoring its cost, it is nontrivial for Sobolev IPM to accommodate other structures beyond the L^p regime. Inspired by the seminal work of Sturm [54], and advantages of Orlicz-Wasserstein, we will leverage the set of convex N-functions for Orlicz geometry for Sobolev IPM to go beyond the coupled L^p structures, and propose the generalized Sobolev IPM in §3.[5]

3 Generalized Sobolev IPM and the Regularization Approach

Sobolev IPM provides a powerful, yet rigid framework, essentially coupled with L^p within its definition. Thus, it is nontrivial to utilize Sobolev IPM with other structures beyond L^p, which is in stark contrast to the flexibility of OT for its adaptivity to diverse geometric structures by simply modifying the underlying cost function. In this section, we leverage convex N-functions to generalize Sobolev IPM beyond its coupled L^p structure. We first introduce the *graph-based Orlicz-Sobolev* space [32] and its *Orlicz-Sobolev* norm [48, §9.3], [2, §3.1, §8.30]. Based on these components, we then describe the definition of the *generalized Sobolev IPM* (GSI) for graph-based measures.

Definition 3.1 (Graph-based Orlicz-Sobolev space [32])**.** Let Φ be an N-function and λ be a nonnegative Borel measure on $\mathbb{G}$. A continuous function $f : \mathbb{G} \to \mathbb{R}$ is said to belong to the graph-based Orlicz-Sobolev space $WL^1_\Phi(\mathbb{G}, \lambda)$ if there exists a function $h \in L_\Phi(\mathbb{G}, \lambda)$ satisfying

$$f(x) - f(z_0) = \int_{[z_0, x]} h(y)\lambda(\mathrm{d}y), \quad \forall x \in \mathbb{G}. \tag{6}$$

Such function h is unique in $L_\Phi(\mathbb{G}, \lambda)$ and is called the generalized graph derivative of f w.r.t. λ, and this generalized graph derivative of f is denoted f'.

Orlicz-Sobolev norm. $WL^1_\Phi(\mathbb{G}, \lambda)$ is a normed space with Orlicz-Sobolev norm [48, 2], defined as

$$\|f\|_{WL^1_\Phi} := \|f\|_{L_\Phi} + \|f'\|_{L_\Phi}. \tag{7}$$

Let $WL^1_{\Phi,0}(\mathbb{G}, \lambda)$ be the subspace consisting of all functions f in $WL^1_\Phi(\mathbb{G}, \lambda)$ satisfying $f(z_0) = 0$. For N-function $\Phi(t) = t^p$ with $1 < p < \infty$, following [32, Proposition 4.3], we have $WL^1_\Phi(\mathbb{G}, \lambda) = W^{1,p}(\mathbb{G}, \lambda)$ where $W^{1,p}$ is the graph-based Sobolev space.[6]

Generalized Sobolev IPM (GSI). Similar to Sobolev IPM, the GSI is an instance of IPM where its critic function belongs to the graph-based Orlicz-Sobolev space, and is constrained within the unit ball of that space. More concretely, given a nonnegative Borel measure λ on $\mathbb{G}$, a pair of complementary N-functions Φ, Ψ,[7] the GSI problem between probability measures $\mu, \nu \in \mathcal{P}(\mathbb{G})$ is defined as

$$\mathcal{GS}_\Phi(\mu, \nu) := \sup_{f \in \mathcal{B}_\Psi} \left| \int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(y)\nu(\mathrm{d}y) \right|, \tag{8}$$

where $\mathcal{B}_\Psi := \left\{ f \in WL^1_{\Psi,0}(\mathbb{G}, \lambda), \|f\|_{WL^1_\Psi} \le 1 \right\}$ is the unit ball defined by Orlicz-Sobolev norm.[8] For N-function $\Psi(t) = t^p$ with $1 < p < \infty$, the GSI turns into Sobolev IPM. Furthermore, notice that the quantity inside the absolute signs is unchanged if f is replaced by $f - f(z_0)$. Thus, we can assume without loss of generality that $f(z_0) = 0$. This is the motivation for the introduction of the Orlicz-Sobolev space $WL^1_{\Psi,0}(\mathbb{G}, \lambda)$, which shares the same sense to approaches for Sobolev GAN [39], and for Sobolev IPM [33].

Analogous to the historically long-standing computational challenges notoriously inherent in Sobolev IPM, the GSI poses significant obstacles in computation for practical use, which are beyond advancements developed for OW and Sobolev IPM in [32, 33].

[5] To ease the readers, in §D, we further give reviews covering related notions and approaches in the literature.

[6] See a review in §D.3.

[7] See a review in §D.6.

[8] GSI may be regarded as *Orlicz-Sobolev IPM*.

The regularization approach for the GSI problem. To address these aforementioned challenges, we theoretically establish a novel relation between the *Orlicz-Sobolev* norm and *Musielak*, ground-based for our novel efficient regularization scheme approach for GSI.

Theorem 3.2 (Equivalence)**.** *For the length measure λ on $\mathbb{G}$, and function $f \in WL^1_{\Phi,0}(\mathbb{G}, \lambda)$, then*

$$\frac{1}{2}\left\|f'\right\|_{L^{\hat{w}}_{\Phi}} \leq \|f\|_{WL^1_{\Phi}} \leq (1+\lambda(\mathbb{G}))\left\|f'\right\|_{L^{\hat{w}}_{\Phi}}, \tag{9}$$

where the weight function $\hat{w}$ is defined as

$$\hat{w}(x) := 1 + \frac{\lambda(\Lambda(x))}{\lambda(\mathbb{G})}, \quad \forall x \in \mathbb{G}. \tag{10}$$

Theorem 3.2 implies that the *Orlicz-Sobolev* norm of a critic function $f \in WL^1_{\Phi,0}(\mathbb{G}, \lambda)$ is equivalent to the *Musielak* norm of its gradient f', where weight function $\hat{w}$ is given explicitly in Equation (10).

Remark 3.3. The novel theoretical finding result in Theorem 3.2 plays the key role for the development of our regularization scheme approach. More precisely, we relax the constraint of critic function f within the unit ball induced by Orlicz-Sobolev norm by the Musielak norm constraint of its gradient f' with a specified weight function $\hat{w}$ (Equation (10)).

Discussions. In [33], Le et al. established an equivalence between the *Sobolev* norm of a critic function and *weighted* L^p norm of its gradient for their weight function $\hat{w}_{\hat{\mathcal{S}}}(x) := 1 + \lambda(\Lambda(x))$ for all $x \in \mathbb{G}$ by relying on closed-form expressions of (weighted) L^p-norm. Such developments are *not* applicable for general convex N-functions, beyond L^p regime, in *Orlicz-Sobolev* norm (Equations (7), (2)) and *Musielak* norm (Equation (3)). Indeed, in this work, we employ different weight function $\hat{w}$ (Equation (10)), formed from our novel theoretical findings,[9] to establish an equivalence between the *Orlicz-Sobolev* norm of a critic function and *Musielak* norm of its gradient for the theoretical finding result in Theorem 3.2, which is novel, and in fact nontrivial. Unlike the (weighted) L^p geometric structures exploited for Sobolev IPM in [33], there are no closed-form expressions for *Orlicz-Sobolev* norm and *Musielak* norm for general convex N-functions in GSI. In fact, to establish the novel equivalence in Theorem 3.2, we have developed several novel auxiliary theoretical results (Appendix A.1), beyond developments in [33, 32] for Sobolev IPM and OW. Moreover, we emphasize that Theorem 3.2 holds for our weight function $\hat{w}$ (Equation (10)), and such result may not still hold for any given general weight functions.

Generalized Sobolev IPM with Musielak regularization (GSI-M). Hereafter, we employ the weight function $\hat{w}$ in Theorem 3.2, and define $\mathcal{B}^{\hat{w}}_{\Psi} := \left\{ f \in WL^1_{\Psi,0}(\mathbb{G}, \lambda), \left\|f'\right\|_{L^{\hat{w}}_{\Psi}} \leq 1 \right\}$. Based on the equivalent relation given by Theorem 3.2, we propose to regularize GSI (Equation (8)) by relaxing the constraint on the critic function f in the graph-based Orlicz-Sobolev space $WL^1_{\Psi,0}$. More precisely, instead of critic f belonging to the unit ball $\mathcal{B}_{\Psi}$ of the *Orlicz-Sobolev* space, we propose to constraint critic f within the unit ball $\mathcal{B}^{\hat{w}}_{\Psi}$ of the *Musielak* norm of its gradient f'. We now formally define the *generalized Sobolev IPM with Musielak regularization* (GSI-M).[10]

Definition 3.4 (Generalized Sobolev IPM with Musielak regularization)**.** Let λ be a nonnegative Borel measure on $\mathbb{G}$ and a pair of complementary N-functions Φ, Ψ. Then, for probability measures $\mu, \nu \in \mathcal{P}(\mathbb{G})$, the generalized Sobolev IPM with Musielak regularization (GSI-M) is defined as

$$\widehat{\mathcal{GS}}_{\Phi}(\mu, \nu) := \sup_{f \in \mathcal{B}^{\hat{w}}_{\Psi}} \left| \int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(y)\nu(\mathrm{d}y) \right|. \tag{11}$$

Computation. We next show that one can compute the regularized GSI $\widehat{\mathcal{GS}}_{\Phi}$ by simply solving a univariate optimization problem, paving ways for its practical use in applications.

Theorem 3.5 (GSI-M as univariate optimization problem)**.** *The generalized Sobolev IPM with Musielak regularization $\widehat{\mathcal{GS}}_{\Phi}(\mu, \nu)$ can be computed as follows:*

$$\widehat{\mathcal{GS}}_{\Phi}(\mu, \nu) = \inf_{k>0} \frac{1}{k} \left\{ 1 + \int_{\mathbb{G}} \hat{w}(x)\, \Phi\left(\frac{k}{\hat{w}(x)} \left| \mu(\Lambda(x)) - \nu(\Lambda(x)) \right| \right) \lambda(dx) \right\}. \tag{12}$$

[9] See Theorem A.3 in Appendix A.1, and its important role for the proof of Theorem 3.2 in Appendix E.1.

[10] Unlike a traditional convention for regularization by adding constraints into a problem, we relax the constraint with Orlicz-Sobolev norm by its corresponding Musielak norm, following a similar sense in [33] for Sobolev IPM for simplicity.

Notice that in Equation (12), both the subgraph $\Lambda(x)$ (Equation (1)) and the weight function $\hat{w}(x)$ (Equation (10)) depend on input point x under the integral over $\mathbb{G}$. For practical applications, we next derive an explicit formula for the integral over graph $\mathbb{G}$ in Equation (12) when the input probability measures are supported on nodes V of graph $\mathbb{G}$. This gives an efficient method for computing the GSI-M $\widehat{\mathcal{GS}}_\Phi$. Note that to achieve this result, we use the length measure on graph $\mathbb{G}$ [30] for the nonnegative Borel measure λ, i.e., we have $\lambda([x,z]) = d_{\mathbb{G}}(x,z), \forall x,z \in \mathbb{G}$. We summarize the result in the following theorem.

Theorem 3.6 (Discrete case)**.** *Let λ be the length measure on $\mathbb{G}$, and Φ be an N-function. Suppose that $\mu,\nu \in \mathcal{P}(\mathbb{G})$ are supported on nodes V of graph $\mathbb{G}$.*[11] *Then we have*

$$\widehat{\mathcal{GS}}_\Phi(\mu,\nu) = \inf_{k>0} \frac{1}{k}\left\{1 + \sum_{e\in E}\left[\int_0^1 \bar{w}_t(e)\,\Phi\left(\frac{k|\bar{h}(e)|}{\bar{w}_t(e)}\right) w_e \mathrm{d}t\right]\right\}, \tag{13}$$

where $\bar{h}(e) := \mu(\gamma_e) - \nu(\gamma_e)$, and $\bar{w}_t(e) := \frac{w_e}{\lambda(\mathbb{G})}t + 1 + \frac{\lambda(\gamma_e)}{\lambda(\mathbb{G})}$ for all edge $e \in E$ and $t \in [0,1]$.

From Theorem 3.6, by exploiting the graph structure for the integral over graph $\mathbb{G}$ in Equation (12), one only needs to simply solve the univariate optimization problem to compute the regularized GSI $\widehat{\mathcal{GS}}_\Phi$.[12] We further note that for each edge e, given a specific popular N-function, then the integral w.r.t. scalar t has an explicit form for efficient computation, see Appendix §A.3 for details.[13]

Special case with closed-form expression. We illustrate that for a specific N-function, the regularized GSI $\widehat{\mathcal{GS}}_\Phi$ can yield a closed-form expression for fast computation.

Proposition 3.7 (Closed-form discrete case)**.** *Suppose that $\mu,\nu \in \mathcal{P}(\mathbb{G})$ are supported on nodes V of graph $\mathbb{G}$. For N-function $\Phi(t) = \frac{(p-1)^{p-1}}{p^p}t^p$ with $1 < p < \infty$, length measure λ on $\mathbb{G}$, then*

$$\widehat{\mathcal{GS}}_\Phi(\mu,\nu) = \left(\sum_{e\in E}\beta_e\,|\mu(\gamma_e) - \nu(\gamma_e)|^p\right)^{\frac{1}{p}}, \tag{14}$$

where for each edge $e \in E$ of graph $\mathbb{G}$, the scalar number β_e is given by

$$\beta_e := \begin{cases} \lambda(\mathbb{G})\log\left(1 + \frac{w_e}{\lambda(\mathbb{G})+\lambda(\gamma_e)}\right) & \text{if } p = 2, \\ \frac{(\lambda(\mathbb{G})+\lambda(\gamma_e)+w_e)^{2-p} - (\lambda(\mathbb{G})+\lambda(\gamma_e))^{2-p}}{(2-p)\lambda(\mathbb{G})^{1-p}} & \text{otherwise.} \end{cases} \tag{15}$$

The proofs for these theoretical results in §3 are respectively placed in §E.1 – §E.4.

Related work. We study Sobolev IPM problem for *two probability measures* supported on the *same* graph, as in [30, 32, 33, 11]. We distinguish the considered problem with the research lines on computing either kernels [7, 49, 12] or distances/discrepancies [45, 66, 16, 8, 5] between *two input graphs* which are possibly *different*. See Appendix §C for a further discussion.

• **Relation with regularized Sobolev IPM [33].** In [33], Le et al. pioneered a tractable computational approach for Sobolev IPM, addressing historically long-lasting limitations in practical use. More concretely, by exploiting closed-form expressions of L^p norms, Le et al. [33] developed an efficient regularization for Sobolev IPM by relying on the connections between L^p-norm of a critic function and weighted L^p-norm of its gradient for the specific weight function $\hat{w}_{\hat{\mathcal{S}}}$. Therefore, such developments in [33] are neither applicable nor directly extendable for the GSI problem (8), involving *Luxemburg* norm (Equation (2)) without closed-form expressions for general convex N-functions, for accommodating geometric structures beyond L^p.

• **Relation with generalized Sobolev transport (GST) [32].** In [32], Le et al. developed the generalized Sobolev transport (GST), which is a scalable variant of OW [54] for graph-based measures. More concretely, Le et al. [32] adopted Orlicz geometry for *Sobolev transport* (ST) [30],

[11] An extension for measures supported in graph $\mathbb{G}$ is discussed in Appendix §C.

[12] Rao & Ren [48, Theorem 13] derived the necessary and sufficient conditions to obtain the infimum for problem (13).

[13] See §B.1 for further implementation notes for γ_e in GSI-M including practical preprocessing procedure, exploiting its sparsity to improve computational complexity, and further discussion on its computational complexity.

which relaxed the Lipschitz constraint for critic functions as in 1-order Wasserstein by considering such constraint in a *Sobolev* space. It is worth noting that similar to Sobolev transport, GST only constraints on gradient of critic function. Therefore, the GST problem is essentially different to the considered Sobolev IPM problem where its constraint involves both critic function and its gradient. Consequently, although GST also utilizes Orlicz geometry, such developments in [32] are neither suitable nor directly relevant for the GSI problem (8). In the next section, by leveraging our novel finding results in Theorem 3.2, we derive a by-product result on the relation between GST and the GSI problem as well as the proposed regularization approach.

4 Theoretical Properties of the Regularized GSI

In this section, we derive the metric property for the GSI with Musielak regularization (GSI-M) and establish a relationship for the GSI-M with different N-functions. Additionally, we draw connections of the regularized GSI to the original GSI, and other related IPM and transport distances including the regularized Sobolev IPM [33], generalized Sobolev transport (GST) [32] and OW [54].[14]

Theorem 4.1 (Metrization)**.** *The generalized Sobolev IPM with Musielak regularization $\widehat{\mathcal{GS}}_\Phi$ is a metric on the space $\mathcal{P}(\mathbb{G})$ of probability measures on graph $\mathbb{G}$.*

The GSI-M 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 Sobolev IPM for comparing graph-based measures.

Proposition 4.2 (GSI-M with different N-functions)**.** *For any two N-functions Φ_1, Φ_2 satisfying $\Phi_1(t) \leq \Phi_2(t)$ for all $t \in \mathbb{R}_+$, and $\mu, \nu \in \mathcal{P}(\mathbb{G})$, then we have*

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

Connections to other IPM and transport distances. We then theoretically establish relations between the regularized GSI (11) and other related IPM and transport approaches. Consequently, we can theoretically position the regularized GSI within the research picture of the literature.

• **Original GSI.** We show that the regularized GSI (11) is equivalent to the original GSI (8).

Theorem 4.3 (Connection between GSI-M and original GSI)**.** *For a nonnegative Borel measure λ on $\mathbb{G}$, N-function Φ, $\mu, \nu \in \mathcal{P}(\mathbb{G})$, then*

$$\frac{1}{2}\,\mathcal{GS}_\Phi(\mu, \nu) \leq \widehat{\mathcal{GS}}_\Phi(\mu, \nu) \leq (1 + \lambda(\mathbb{G}))\,\mathcal{GS}_\Phi(\mu, \nu), \tag{16}$$

• **Regularized Sobolev IPM [33].** Let $\hat{\mathcal{S}}_p$ be the regularized p-order Sobolev IPM.

Proposition 4.4 (Connection between GSI-M and regularized Sobolev IPM)**.** *Let $\Phi(t) := \frac{(p-1)^{p-1}}{p^p} t^p$ with $1 < p < \infty$, $\hat{c}_1 := \max(1, \lambda(\mathbb{G})^{-1})^{\frac{1-p}{p}}$, and $\hat{c}_2 := \min(1, \lambda(\mathbb{G})^{-1})^{\frac{1-p}{p}}$. Then, for a nonnegative Borel measure λ on graph $\mathbb{G}$, and measures $\mu, \nu \in \mathcal{P}(\mathbb{G})$, we have*

$$\hat{c}_1\,\hat{\mathcal{S}}_p(\mu, \nu) \leq \widehat{\mathcal{GS}}_\Phi(\mu, \nu) \leq \hat{c}_2\,\hat{\mathcal{S}}_p(\mu, \nu). \tag{17}$$

Proposition 4.4 illustrates that for a certain given N-function, the regularized GSI $\widehat{\mathcal{GS}}$ (11) is provably equivalent to the regularized Sobolev IPM $\hat{\mathcal{S}}_p$ [33]. Additionally, $\widehat{\mathcal{GS}}$ can be considered as a variational generalization for $\hat{\mathcal{S}}_p$ to incorporate more general geometric priors, beyond L^p paradigm.

• **Generalized Sobolev transport (GST) [32].** Denote $\mathcal{GST}_\Phi$ for the GST with N-function Φ.

Proposition 4.5 (Connection between GSI-M and generalized Sobolev transport)**.** *For a nonnegative Borel measure λ on $\mathbb{G}$, then for all measures $\mu, \nu \in \mathcal{P}(\mathbb{G})$, we have*

$$\frac{1}{2}\,\mathcal{GST}_\Phi(\mu, \nu) \leq \widehat{\mathcal{GS}}_\Phi(\mu, \nu) \leq \mathcal{GST}_\Phi(\mu, \nu). \tag{18}$$

Propositions 4.5 illustrates that GSI-M is provably equivalent to GST for the same convex N-functions. Similar to GST, GSI-M constraints on gradient of a critic function, thereby GSI-M may be regarded

[14] See §D for a review on these IPM and transport distances.

as a weighted variant of GST. However, we emphasize that the weight function $\hat{w}$ (Equation (10)) plays the key role to establish the equivalence in Theorem 3.2 between *Orlicz-Sobolev* norm of a critic function in original GSI problem (8) and *Musielak* norm of its gradient in the regularized GSI problem (11). Therefore, GSI-M with the weight function $\hat{w}$ provides an efficient regularization approach for the GSI problem, theoretically ground-based by novel results in Theorem 3.2.

• **Orlicz-Wasserstein (OW) [54].** Let $\mathcal{OW}$ denote the Orlicz-Wasserstein.

Proposition 4.6 (Connection between GSI-M and Orlicz-Wasserstein)**.** *Consider the limit case $\Phi(t) := t$, and graph $\mathbb{G}$ is a tree.*[15] *For a nonnegative Borel measure λ on $\mathbb{G}$, $\mu, \nu \in \mathcal{P}(\mathbb{G})$, then*

$$\frac{1}{2}\,\mathcal{OW}(\mu,\nu) \leq \widehat{\mathcal{GS}}_{\Phi}(\mu,\nu) \leq \mathcal{OW}(\mu,\nu). \tag{19}$$

Propositions 4.6 shows that GSI-M $\widehat{\mathcal{GS}}$ is provably equivalent to OW when graph $\mathbb{G}$ is a tree, and $\Phi(t) = t$. Both GSI-M and OW are capable to employ Orlicz structures with general convex N-functions. However, OW is challenging for computation, hinders its practical applications while GSI-M is computationally efficient, by simply solving a univariate optimization problem.

The proofs for the theoretical results in §4 are respectively placed in §E.5 – §E.10. See Appendix §A.2 for further theoretical results of the regularized GSI.

5 Experiments

In this section, we illustrate that GSI-M is fast for computation, which is several-order faster than the popular OW. Following the problem setups in [33], we evaluate GSI-M for probability measures on a given graph. We show that performances of GSI-M compare favorably to other transport distances to compare graph-based measures under the same settings for document classification and TDA. Note that there are different approaches for document classification and TDA. However, our goals are to compare different transport distances for graph-based probability measures under the same settings as considered in [33].

Datasets. For document classification, we consider 4 real-world datasets: `TWITTER`, `RECIPE`, `CLASSIC`, `AMAZON` where their properties are given in Figure 2. For TDA, we consider 2 tasks: (i) orbit recognition on the synthesized `Orbit` dataset [1] for linked twist map, a discrete dynamical system modeling flow, which is used to model flows in DNA microarrays [21], and (ii) object shape image classification on a 10-class subset of `MPEG7` dataset [27]. We summarize the dataset properties in Figure 3. The objects of interests for TDA are represented by persistence diagrams (PD) [17].

Following [33, §5], we regard a document as a distribution over its words, and PD as a distribution over its topological features. We employ the graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ for words and topological features in document classification and TDA respectively.[16] Thus, document and PD are represented as probability measures supported on these given graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$.

Set up. We employ kernelized support vector machine (SVM) for both document classification and TDA. We use kernel $\exp(-\tilde{t}\tilde{d})$ with hyperparameter $\tilde{t} > 0$ and $\tilde{d}$ is the considered IPM/transport distances for graph-based measures. Following [14], we regularize the Gram matrices of indefinite kernels by adding a sufficiently large diagonal term, and use 1-vs-1 strategy for multi-class classification with SVM. We randomly split each dataset into $70\%/30\%$ for training and test with 10 repeats. We typically utilize cross validation for hyper-parameters. For SVM regularization hyperparameter, we choose it from $\{0.01, 0.1, 1, 10\}$. For the kernel hyperparameter $\tilde{t}$, we choose it from $\{1/q_s, 1/(2q_s), 1/(5q_s)\}$ where q_s is the $s\%$ quantile of a random subset of distances observed on a training set and $s = 10, 20, \ldots, 90$. For the root node z_0 in graph $\mathbb{G}$, we choose it from a random 10-root-node subset of V in $\mathbb{G}$. Note that we include preprocessing procedures, e.g., computing shortest paths, into reported time consumptions. Additionally, we use a second-order solver (i.e., *fmincon* with Trust Region Reflective solver) for solving the *univariate* optimization problem for GSI-M. Following [19, 3, 32, 4, 51, 9, 2, 48, 10, 64], we evaluate two most popular N-functions widely used in the literature: $\Phi_1(t) = \exp(t) - t - 1$, and $\Phi_2(t) = \exp(t^2) - 1$. We also examine

[15] 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 $\Phi(t) = t^p$.
[16] See a review in Appendix §C.

the limit case: $\Phi_0(t) = t$ as in [19, 32]. To illustrate the scale, e.g., in AMAZON, we need to evaluate distances for more than 29 million pairs of probability measures for each SVM run.[17]

5.1 Computation

We compare the time consumption of GSI-M and OW with Φ_0, Φ_1, Φ_2 on random 10K pairs of measures in AMAZON with graph $\mathbb{G}_{\text{Log}}$ having 1K nodes and about 7K edges. Figure 1 illustrates that the computation of GSI-M is several-order faster than OW. More concretely, GSI-M is $90\times, 5100\times, 2900\times$ faster than OW for Φ_0, Φ_1, Φ_2 respectively. For N-functions Φ_1 and Φ_2, GSI-M takes less than 22 *seconds* while OW needs at least 17 *hours*, and up to 27 *hours* for the computation. We remark that following Proposition 3.7, by taking the limit $p \to 1^+$, GSI-M with Φ_0 has closed-form expression for fast computation. Additionally, OW with Φ_0 is equal to the OT with graph metric ground cost [19, 32]. Therefore, the computations of GSI-M and OW with Φ_0 are more efficient than with Φ_1, Φ_2 as expectation.

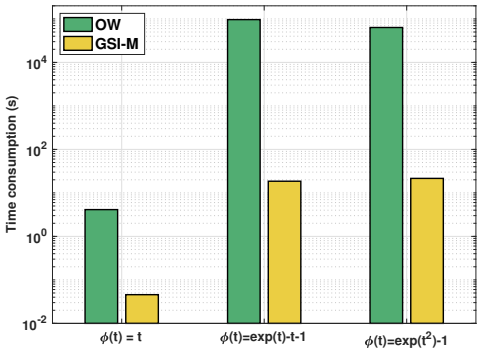


Figure 1: Time consumption on $\mathbb{G}_{\text{Log}}$.

5.2 Document classification

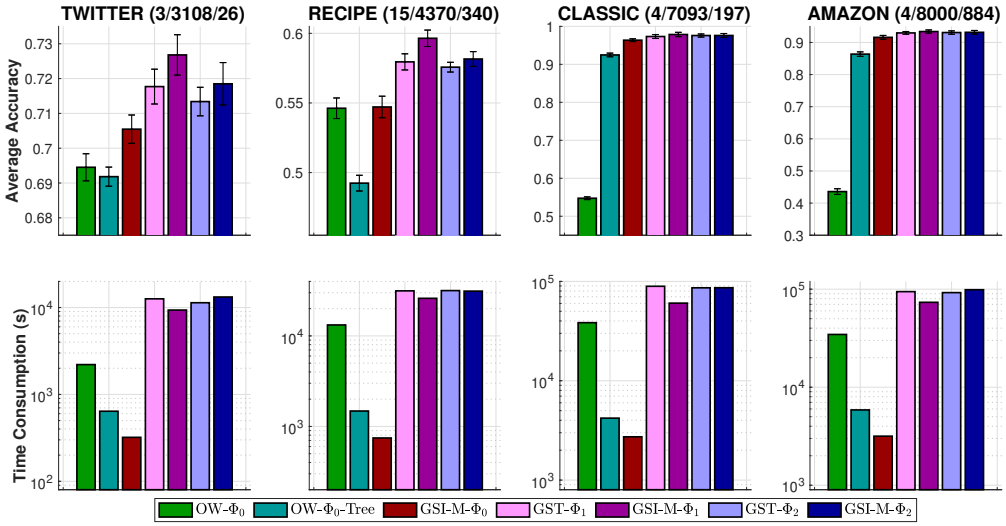


Figure 2: SVM results and time consumption for kernel matrices with graph $\mathbb{G}_{\text{Log}}$. For each dataset, the numbers in the parenthesis are the number of classes; the number of documents; and the maximum number of unique words for each document respectively.

Set up. We evaluate GSI-M with Φ_0, Φ_1, Φ_2, denoted as GSI-M-Φ_i for $i = 0, 1, 2$ respectively. For OW, we only use Φ_0, denoted as OW-Φ_0, but exclude Φ_1, Φ_2 due to their prohibited computations (Figure 1). We include GST [32], which is a scalable alternative for OW in applications, for Φ_0, Φ_1, Φ_2, denoted as GST-Φ_i for $i = 0, 1, 2$ respectively.[18] Additionally, we also carry out a special case of OW-Φ_0 where it is computed on a random tree extracted from the given graph $\mathbb{G}$, denoted as OW-Φ_0-Tree. We remark that OW-Φ_0-Tree is equal to tree-Wasserstein [29]. We use graph $\mathbb{G}_{\text{Log}}$ with 10K nodes and about 100K edges; and graph $\mathbb{G}_{\text{Sqrt}}$ with 10K nodes, about 1M edges [33, §5].

Results. Figure 2 illustrates that the performances of GSI-M with all Φ functions compare favorably to those baselines. GSI-M-Φ_1 and GSI-M-Φ_2 improve performances of GSI-M-Φ_0, especially in

[17] See further details in §B.

[18] Note that for the limit Φ_0, GSI-M is equal to GST, i.e., GSI-M-Φ_0 = GST-Φ_0.

RECIPE, but their computational time is slower since GSI-M-Φ_0 has a closed-form expression (§5.1). Therefore, it implies that *Orlicz geometric structure* in GSI-M may be helpful for graph-based document classification, similar to observations in [32]. Additionally, performances of GSI-M compare favorably to those of GST with the same N-functions Φ_1, Φ_2, especially in TWITTER and RECIPE. Performances of OW-Φ_0-Tree are better in CLASSIC and AMAZON, but worse in TWITTER and RECIPE than those of OW-Φ_0. Although OW-Φ_0-Tree only leverages a partial information of $\mathbb{G}$, it forms positive definite kernels while kernels of OW-Φ_0 are indefinite.

5.3 Topological Data Analysis

Set up. We use the same set up as document classification. Except in MPEG7, due to its smaller size, we use $\mathbb{G}_{\text{Log}}$ with 1K nodes and about 7K edges for $\mathbb{G}_{\text{Log}}$; and $\mathbb{G}_{\text{Sqrt}}$ with 1K nodes and about 32K edges as in [33].

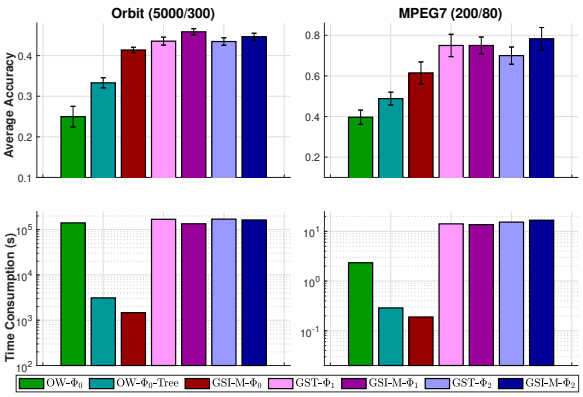


Figure 3: SVM results and time consumption for kernel matrices with graph $\mathbb{G}_{\text{Log}}$. For each dataset, the numbers in the parenthesis are respectively the number of PD; and the maximum number of points in PD.

Results. Figure 3 shows that we obtain similar observations as for document classification. For the same Φ function, GSI-M is several-order faster OW. Performances of GSI-M-Φ_1 GSI-M-Φ_2 compare favorably to those of GSI-M-Φ_0, especially in MPEG7, but they are also slower. Therefore, similar to observations in [32], *Orlicz geometric structure* in GSI-M may be also useful for graph-based TDA. Additionally, performances of GSI-M compare favorably to those of GST with the same N-functions Φ_1, Φ_2, especially in MPEG7 for N-function Φ_2. Although OW-Φ_0-Tree use a partial information of $\mathbb{G}$, the positive definiteness of its corresponding kernel may help to improve performances of OW-Φ_0.

For further empirical results for both document classification and TDA, e.g., ablation studies on different graph structures (e.g., number of vertices, number of edges), different N-functions, different root nodes, effects of minor perturbation on edge weights of a given graph, empirical gaps between GSI-M and GST/OW, we refer the readers in Appendix B for rigorous details due to the limited space.

6 Conclusion

In this work, we leverage convex N-functions to adopt Orlicz geometric structure for Sobolev IPM to go beyond L^p. Moreover, we propose a novel regularization for GSI, which is efficient for computation by simply solving a univariate optimization problem. For future works, it is interesting to consider whether it is possible to derive efficient algorithmic approaches for the original Sobolev IPM and its generalization (without regularization), and go beyond the Sobolev geometric structures, e.g., Besov IPM, for applications. Additionally, automatically learning N-functions for GSI beyond the traditional cross-validation will be also an impactful direction for further investigations in depth. Moreover, in this work, we investigate GSI for measures on a given graph. Therefore, constructing the best graph for a given task is also an interesting open problem for research on graph-based IPM. Also, see Appendix §C for further discussions and limitations.

Acknowledgments and Disclosure of Funding

We thank the area chairs and anonymous reviewers for their comments. KF has been supported in part by Grant-in-Aid for Transformative Research Areas (A) JP22H05106 and JSPS KAKENHI JP26K02990. HH acknowledges the support of JSPS KAKENHI JP25H01494, 26K02989, and 26K23861. TL gratefully acknowledges the support of JST BOOST Grant number JPMJBY25F8.

References

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

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

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

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

[5] Bai, Y., Tran, H., Du, H., Liu, X., and Kolouri, S. Fused partial gromov-wasserstein for structured objects. *arXiv preprint arXiv:2502.09934*, 2025.

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

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

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

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

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

[11] Choromanski, K., Long, D., Parashar, A., and Saha, D. Near-linear time generalized sinkhorn algorithms for bounded genus graphs. *arXiv preprint arXiv:2605.09782*, 2026.

[12] Choromanski, K. M. Taming graph kernels with random features. In *International Conference on Machine Learning*, pp. 5964–5977, 2023.

[13] Choromanski, K. M., Sehanobish, A., Lin, H., Zhao, Y., Berger, E., Parshakova, T., Pan, A., Watkins, D., Zhang, T., Likhosherstov, V., et al. Efficient graph field integrators meet point clouds. In *International Conference on Machine Learning*, pp. 5978–6004, 2023.

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

[15] 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.

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

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

[18] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[19] 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.

[20] Harjulehto, P. and Hästö, P. *Orlicz Spaces and Generalized Orlicz Spaces*. Springer, 2019.

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

[22] Hua, B.-S., Tran, M.-K., and Yeung, S.-K. Pointwise convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 984–993, 2018.

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

[24] Kolouri, S., Ketz, N. A., Soltoggio, A., and Pilly, P. K. Sliced Cramer synaptic consolidation for preserving deeply learned representations. In *International Conference on Learning Representations*, 2020.

[25] Krasnoselskii, M. A. and Rutickii, Y. B. *Convex functions and Orlicz spaces*. P. Noordhoff Ltd., Groningen, 1961.

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

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

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

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

[34] Le, T., Nguyen, T., Hino, H., and Fukumizu, K. An efficient Orlicz-Sobolev approach for transporting unbalanced measures on a graph. *Advances in Neural Information Processing Systems*, 38:81586–81627, 2026.

[35] Liang, T. How well can generative adversarial networks learn densities: A nonparametric view. *arXiv preprint*, 2017.

[36] Liang, T. Estimating certain integral probability metric (IPM) is as hard as estimating under the IPM. *arXiv preprint*, 2019.

[37] Liang, T. How well generative adversarial networks learn distributions. *Journal of Machine Learning Research*, 22(228):1–41, 2021.

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

[39] Mroueh, Y., Li, C.-L., Sercu, T., Raj, A., , and Cheng, Y. Sobolev GAN. In *International Conference on Learning Representations*, pp. 5767–5777, 2018.

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

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

[42] Nadjahi, K., Durmus, A., Chizat, L., Kolouri, S., Shahrampour, S., and Simsekli, U. Statistical and topological properties of sliced probability divergences. *Advances in Neural Information Processing Systems*, 33:20802–20812, 2020.

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

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

[45] 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.

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

[47] 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.

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

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

[50] Singh, S., Uppal, A., Li, B., Li, C.-L., Zaheer, M., and Póczos, B. Nonparametric density estimation under adversarial losses. *Advances in Neural Information Processing Systems*, 31, 2018.

[51] 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.

[52] Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. On integral probability metrics, ϕ-divergences and binary classification. *arXiv preprint*, 2009.

[53] Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6: 1550–1599, 2012.

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

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

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

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

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

[59] Tran, V.-H., Tran, T., Chu, T., Le, T., and Nguyen, T. Tree-sliced entropy partial transport. *Advances in Neural Information Processing Systems*, 38:16503–16560, 2026.

[60] Tran, V.-H., Tran, T., Chu, T., Pham, D.-T., Tran, T.-K., Le, T., and Nguyen, T. M. Tree-sliced Sobolev IPM. In *The Fourteenth International Conference on Learning Representations*, 2026.

[61] Tran, V.-H., Tran, T. Q., Chu, T., Tran, T.-K., Pham, D.-T., Le, T., and Nguyen, T. Revisiting tree-sliced Wasserstein distance through the lens of the Fermat–Weber problem. In *International Conference on Learning Representations*, volume 2026, pp. 62831–62860, 2026.

[62] Uppal, A., Singh, S., and Póczos, B. Nonparametric density estimation & convergence rates for GANs under Besov IPM losses. *Advances in neural information processing systems*, 32, 2019.

[63] Uppal, A., Singh, S., and Póczos, B. Robust density estimation under Besov IPM losses. *Advances in Neural Information Processing Systems*, 33:5345–5355, 2020.

[64] Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[65] Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., and Solomon, J. M. Dynamic graph CNN for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 38(5):1–12, 2019.

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

[67] Yurochkin, M., Claici, S., Chien, E., Mirzazadeh, F., and Solomon, J. M. Hierarchical optimal transport for document representation. *Advances in neural information processing systems*, 32, 2019.

[68] Zhang, M., Lin, J. A., Choromanski, K., Weller, A., Turner, R. E., and Reid, I. Graph random features for scalable gaussian processes. In *International Conference on Learning Representations*, volume 2026, pp. 45157–45180, 2026.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Please see §1 for the main claims of the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Please see the discussion in §6 and further discussion in Appendix §C.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: Please see theoretical results in §3 and §4; and further theoretical results in Appendix §A. We provide full proofs in Appendix §E.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide code in the supplementary. Our experiments are described in §5, and further details are provided in Appendix §B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

Answer: [Yes]

Justification: We provide code in the supplementary. We use available data for the experiments in §5.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Our experiments are described in §5, and further details are provided in Appendix §B.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Please see empirical results presented in §5, and further results in Appendix §B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Please see §5, and further details in Appendix §B and §C.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: This paper presents work whose goal is to advance the field of Machine Learning.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Please see Appendix §C.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: There are no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We cite original papers for codes used in our work.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: We only use LLM for editing.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Supplement to "Orlicz–Sobolev with Musielak: An Efficient Regularization Approach for Graph-based IPM"

In this appendix, we provide further theoretical results in §A, further experiment experimental details and empirical results in §B, and further discussions in §C. Then, we provide reviews for related notions and approaches in the literature in §D. Detailed proofs for theoretical results in the main manuscript and additional theoretical results in §A are placed in §E.

A Further Theoretical Results

A.1 Auxiliary Results

In this section, we describe auxiliary theoretical finding results, which play important roles to derive our theoretical results presented in the main text.

Lemma A.1. *Let Φ be an N-function and $f \in WL^1_{\Phi,0}(\mathbb{G},\lambda)$. Then for any nonnegative weight function $\hat{w}$ and $t>0$, we have*

$$\int_{\mathbb{G}} \hat{w}(x)\,\Phi\left(\frac{1}{t}\,|f(x)|\right)\lambda(\mathrm{d}x) \le \frac{1}{\lambda(\mathbb{G})}\int_{\mathbb{G}} \Phi\left(\frac{\lambda(\mathbb{G})}{t}\,|f'(y)|\right)\bar{\lambda}_{\hat{w}}(\Lambda(y))\lambda(\mathrm{d}y), \tag{20}$$

where we write $\bar{\lambda}_{\hat{w}}$ for measure λ weighted by function $\hat{w}$, i.e., $\int \bar{\lambda}_{\hat{w}}(\mathrm{d}x) := \int \hat{w}(x)\lambda(\mathrm{d}x)$.

The proof is placed in Appendix §E.11.

Lemma A.2. *Let w_0 be any positive weight function such that $w_0(x) \ge \lambda(\Lambda(x))$ for all $x \in \mathbb{G}$. Then if function $f \in WL^1_{\Phi,0}(\mathbb{G},\lambda)$ and scalar $t>0$ satisfying $\int_{\mathbb{G}} w_0(x)\,\Phi\left(\frac{|f'(x)|}{t}\right)\lambda(\mathrm{d}x) \le \lambda(\mathbb{G})$, we must have*

$$t \ge \frac{\|f\|_{L_\Phi}}{\lambda(\mathbb{G})}. \tag{21}$$

The proof is placed in Appendix §E.12.

Theorem A.3. *Let w_0 be any positive weight function satisfying $w_0(x) \ge \lambda(\Lambda(x))$ for all $x \in \mathbb{G}$. For every function $f \in WL^1_{\Phi,0}(\mathbb{G},\lambda)$, then*

$$\|f\|_{L_\Phi} \le \lambda(\mathbb{G})\,\|f'\|_{L_\Phi^{w_0/\lambda(\mathbb{G})}}. \tag{22}$$

The proof is placed in Appendix §E.13.

Lemma A.4. *Let Φ be an N-function. For any two positive weight functions w_1, w_2 such that $w_1(x) \ge w_2(x)$ for all $x \in \mathbb{G}$, and any Borel measurable function f on $\mathbb{G}$, then*

$$\|f\|_{L_\Phi^{w_1}} \ge \|f\|_{L_\Phi^{w_2}}. \tag{23}$$

The proof is placed in Appendix §E.14.

Lemma A.5. *For N-function Φ, scalar $t>0$, and scalar $k \ge 1$, then*

$$\Phi(kt) \ge k\Phi(t). \tag{24}$$

The proof is placed in Appendix §E.15.

Lemma A.6. *For N-function Φ, positive weight function $\hat{w}$ on $\mathbb{G}$, $k>0$, then we have*

$$\|kf\|_{L_\Phi^{\hat{w}}} = k\,\|f\|_{L_\Phi^{\hat{w}}}. \tag{25}$$

The proof is placed in Appendix §E.16.

A.2 Further Theoretical Results

In this section, we provide further theoretical results, complemented to our results in the main text.

Special case with closed-form expression. For specific N-function, the generalized Sobolev IPM can yield a closed-form expression for computation.

Proposition A.7 (Closed-form expression)**.** *For N-function $\Phi(t) = \frac{(p-1)^{p-1}}{p^p} t^p$ with $1 < p < \infty$, the generalized Sobolev IPM with Musielak regularization has a closed-form expression as follows:*

$$\widehat{\mathcal{GS}}_\Phi(\mu, \nu) = \left[\int_{\mathbb{G}} \hat{w}(x)^{1-p} \, |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p \, \lambda(\mathrm{d}x) \right]^{\frac{1}{p}}. \tag{26}$$

The proof is placed in Appendix §E.17.

Further connections between the regularized GSI with other IPM/transport distances. We establish further connections between the regularized GSI with original Sobolev IPM, Sobolev transport [30], and optimal transport.

• **Original Sobolev IPM.** Denote $\mathcal{S}_p$ for the original p-order Sobolev IPM.

Proposition A.8 (Connection between GSI-M and original Sobolev IPM)**.** *Let $\Phi(t) := \frac{(p-1)^{p-1}}{p^p} t^p$ with $1 < p < \infty$. For a nonnegative Borel measure λ on $\mathbb{G}$, and measures $\mu, \nu \in \mathcal{P}(\mathbb{G})$, then*

$$c_1 \, \mathcal{S}_p(\mu, \nu) \leq \widehat{\mathcal{GS}}_\Phi(\mu, \nu) \leq c_2 \, \mathcal{S}_p(\mu, \nu), \tag{27}$$

where $c_1 := \left[\frac{\min(1, \lambda(\mathbb{G})^{p-1}) \max(1, \lambda(\mathbb{G})^{-1})^{1-p}}{1+\lambda(\mathbb{G})^p} \right]^{\frac{1}{p}}$, $c_2 := \left[\min(1, \lambda(\mathbb{G})^{-1})^{1-p} \max(1, \lambda(\mathbb{G})^{p-1}) \right]^{\frac{1}{p}}$.

The proof is placed in Appendix §E.18.

We note that the original GSI (8) includes Sobolev IPM as a special case. As discussed in §3, for N-function $\Psi(t) = t^p$ with $1 < p < \infty$, the GSI turns into Sobolev IPM. Proposition A.8 shows a connection between the regularized GSI (11) with the original Sobolev IPM.

• **Sobolev transport.** Denote $\mathcal{ST}_p$ for the p-order Sobolev transport.

Proposition A.9 (Connection between GSI-M and Sobolev transport)**.** *For a nonnegative Borel measure λ on $\mathbb{G}$, N-function $\Phi(t) := \frac{(p-1)^{p-1}}{p^p} t^p$ with $1 < p < \infty$, and $\mu, \nu \in \mathcal{P}(\mathbb{G})$, then we have*

$$\frac{1}{2} \, \mathcal{ST}_p(\mu, \nu) \leq \widehat{\mathcal{GS}}_\Phi(\mu, \nu) \leq \mathcal{ST}_p(\mu, \nu). \tag{28}$$

The proof is placed in Appendix §E.19.

• **Optimal transport.** Denote $\mathcal{W}_1$ for the 1-order Wasserstein.

Proposition A.10 (Connection between GSI-M and optimal transport)**.** *Under the same assumptions in Proposition 4.6, then for all measures $\mu, \nu \in \mathcal{P}(\mathbb{G})$, we have*

$$\frac{1}{2} \, \mathcal{W}_1(\mu, \nu) \leq \widehat{\mathcal{GS}}_\Phi(\mu, \nu) \leq \mathcal{W}_1(\mu, \nu). \tag{29}$$

The proof is placed in Appendix §E.20.

A.3 Discrete Case for Generalized Sobolev IPM with Musielak Regularization with Popular N-Functions

In this section, we provide rigorous details for discrete case for generalized Sobolev IPM with Musielak regularization with popular N-functions, which are used for implementations.

Following Theorem 3.6, the discrete case of the regularized generalized Sobolev IPM $\widehat{\mathcal{GS}}_\Phi$ is

$$\widehat{\mathcal{GS}}_\Phi(\mu, \nu) = \inf_{k>0} \frac{1}{k} \left(1 + \sum_{e \in E} \left[\int_0^1 \bar{w}_t(e) \, \Phi\left(\frac{k|\bar{h}(e)|}{\bar{w}_t(e)} \right) w_e \mathrm{d}t \right] \right), \tag{30}$$

where $\bar{h}(e) := \mu(\gamma_e) - \nu(\gamma_e)$ for every edge $e \in E$, and $\bar{w}_t(e) := \frac{w_e}{\lambda(\mathbb{G})}t + 1 + \frac{\lambda(\gamma_e)}{\lambda(\mathbb{G})}$ for all edge $e \in E$ and $t \in [0, 1]$.

We consider popular practical 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: $\Phi_0(t) = t$.

For each edge $e \in E$, we want to compute

$$\mathcal{A}_\Phi(e) := \int_0^1 \left[\frac{w_e}{\lambda(\mathbb{G})}t + 1 + \frac{\lambda(\gamma_e)}{\lambda(\mathbb{G})}\right] \Phi\left(\frac{k|\bar{h}(e)|}{\frac{w_e}{\lambda(\mathbb{G})}t + 1 + \frac{\lambda(\gamma_e)}{\lambda(\mathbb{G})}}\right) w_e \mathrm{d}t \tag{31}$$

For simplicity, let denote the exponential integral function for $z > 0$ as follow:

$$Ei(z) = \int_{-\infty}^{z} \frac{\exp(t)}{t}\mathrm{d}t, \tag{32}$$

and notice that

$$\frac{\mathrm{d}}{\mathrm{d}z}Ei(z) = \frac{\exp(z)}{z}, \quad z \neq 0, \tag{33}$$

and $Ei(z) = -E_1(-z)$, where $E_1(z) = \int_z^\infty \frac{\exp(-t)}{t}\mathrm{d}t$.

For the limit case: $\Phi_0(t) = t$. We have

$$\mathcal{A}_{\Phi_0}(e) := \int_0^1 k|\bar{h}(e)|w_e \mathrm{d}t = k|\bar{h}(e)|w_e. \tag{34}$$

For N-function: $\Phi_1(t) = \exp(t) - t - 1$. For $a > 0, b > 0$, we consider

$$A_1 := \beta \int_0^1 (at + b)\, \Phi_1\left(\frac{\alpha}{at + b}\right) \mathrm{d}t. \tag{35}$$

See Appendix §E.21 for the detailed computation of A_1 where its results are summarized as follow:

$$A_1 = \frac{\beta}{2a}\Bigg[- \alpha^2 \left[Ei\left(\frac{\alpha}{a+b}\right) - Ei\left(\frac{\alpha}{b}\right)\right] + \alpha \left[(a+b)\exp\left(\frac{\alpha}{a+b}\right) - b\exp\left(\frac{\alpha}{b}\right)\right] + \left[(a+b)^2 \exp\left(\frac{\alpha}{a+b}\right) - b^2 \exp\left(\frac{\alpha}{b}\right)\right] - 2\alpha a - \left[(a+b)^2 - b^2\right]\Bigg]. \tag{36}$$

Therefore, we obtain

$$\mathcal{A}_{\Phi_1} = A_1, \tag{37}$$

where $\beta = w_e, a = \frac{w_e}{\lambda(\mathbb{G})}, b = 1 + \frac{\lambda(\gamma_e)}{\lambda(\mathbb{G})}, \alpha = k|\bar{h}(e)|$.

Additionally, the first-order derivative of A_1 w.r.t. k is as follows:

$$\frac{\mathrm{d}}{\mathrm{d}k}A_1 = \frac{\beta|\bar{h}(e)|}{a}\left\{-\alpha\left[Ei\left(\frac{\alpha}{a+b}\right) - Ei\left(\frac{\alpha}{b}\right)\right] + (a+b)\exp\left(\frac{\alpha}{a+b}\right) - b\exp\left(\frac{\alpha}{b}\right) - a\right\}. \tag{38}$$

Moreover, its second-order derivative of A_1 w.r.t. k is as follows:

$$\frac{\mathrm{d}^2}{\mathrm{d}k^2}A_1 = -\frac{\beta|\bar{h}(e)|^2}{a}\left[Ei\left(\frac{\alpha}{a+b}\right) - Ei\left(\frac{\alpha}{b}\right)\right]. \tag{39}$$

For N-function: $\Phi_2(t) = \exp(t^2) - 1$. For $a > 0, b > 0$, we consider

$$A_2 := \beta \int_0^1 (at + b)\, \Phi_2\left(\frac{\alpha}{at + b}\right) \mathrm{d}t. \tag{40}$$

See Appendix §E.22 for the detailed computation of A_2 where its result is summarized as follow:

$$A_2 = \frac{\beta}{2a}\Bigg[-\alpha^2\left[Ei\left(\frac{\alpha^2}{(a+b)^2}\right) - Ei\left(\frac{\alpha^2}{b^2}\right)\right] + \left[(a+b)^2\exp\left(\frac{\alpha^2}{(a+b)^2}\right) - b^2\exp\left(\frac{\alpha^2}{b^2}\right)\right] - \left[(a+b)^2 - b^2\right]\Bigg]. \tag{41}$$

Therefore, we get

$$\mathcal{A}_{\Phi_2} = A_2, \tag{42}$$

where $\beta = w_e, a = \frac{w_e}{\lambda(\mathbb{G})}, b = 1 + \frac{\lambda(\gamma_e)}{\lambda(\mathbb{G})}, \alpha = k|\bar{h}(e)|$.

Additionally, the first-order derivative of A_2 w.r.t. k is as follows:

$$\frac{\mathrm{d}A_2}{\mathrm{d}k} = -\frac{\beta\alpha|\bar{h}(e)|}{a}\left[Ei\left(\frac{\alpha^2}{(a+b)^2}\right) - Ei\left(\frac{\alpha^2}{b^2}\right)\right]. \tag{43}$$

Moreover, its second-order derivative of A_2 w.r.t. k is as follows:

$$\frac{\mathrm{d}^2}{\mathrm{d}k^2}A_2 = -\frac{\beta|\bar{h}(e)|^2}{a}\left[Ei\left(\frac{\alpha^2}{(a+b)^2}\right) - Ei\left(\frac{\alpha^2}{b^2}\right) + 2\exp(\alpha^2/(a+b)^2) - 2\exp(\alpha^2/b^2)\right]. \tag{44}$$

B Further Experimental Details and Empirical Results

B.1 Further Experimental Details

We summarize the number of pairs of probability measures for each dataset which is required to evaluate for kernelized SVM in Table 1.

Implementation notes for γ_e in GSI-M.

• **Preprocessing procedure for γ_e.** Following [33], we precompute γ_e for all edge e in the given graph $\mathbb{G}$. It only needs once for such preprocessing procedure since it does not depend on input measures, but only the graph structure itself. More concretely, we apply the Dijkstra algorithm to recompute the shortest paths from root node z_0 to all other input supports (or vertices) with complexity $\mathcal{O}(|E| + |V|\log|V|)$ where we write $|\cdot|$ for the set cardinality. Then, we can evaluate γ_e for each edge e in E.

• **Sparsity of γ_e.** As observed in [33], for any support of input measure μ, its mass is contributed to $\mu(\gamma_e)$ if and only if $e \subseteq [z_0, x]$ [30]. Therefore, let $\mathrm{supp}(\mu)$ be the set of supports of measure μ, and define $E_{\mu,\nu} \subseteq E$ as follows

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

Additionally, note that $\Phi(0) = 0$ for all N-function Φ. Then, in fact, we can remove all edges $e \in E \setminus E_{\mu,\nu}$ in the summation in Equation (13) for the univariate optimization problem for computing GSI-M.

Table 1: The number of pairs of probability measures on datasets for kernelized SVM.

Datasets	The number of pairs
TWITTER	4394432
RECIPE	8687560
CLASSIC	22890777
AMAZON	29117200
Orbit	11373250
MPEG7	18130

Notes on IPM and transport baselines. We carry out the popular OW as baselines for our experiments. Since OW is notoriously computational heavy in practical applications, we employ generalized Sobolev transport (GST) [32] which is a scalable alternative of OW in applications.

As discussed in §5, we emphasize that for the limit case Φ_0, OW-Φ_0 is equal to the standard optimal transport where its ground cost is the underlying graph metric in $\mathbb{G}$. Additionally, we also have GSI-M-Φ_0 is equal to GST-Φ_0, which is further equal to the 1-order Sobolev transport [30].

Especially, we note that following Proposition 3.7, by taking the limit $p \to 1^+$, GSI-M-Φ_0 has closed-form expression for fast computation. Furthermore, Φ_0 is not a convex N-function due to its linear growth. By leveraging convex N-functions, one can adopt Orlicz geometric structure for Sobolev IPM, beyond L^p paradigm.

Further discussion on computational complexity. Following Theorem 3.6, we simply need to solve the univariate optimization problem for the computation of GSI-M. In practice, we use a second-order solver (i.e., *fmincon* with Trust Region Reflective solver) in our experiments.

As in the implementation notes discussed in this section, we apply a preprocessing for γ_e with computational complexity $\mathcal{O}(|E| + |V| \log |V|)$. It only needs once for such preprocessing procedure since it does not depend on input measures, but only the graph structure itself.

Furthermore, the computational complexity for each iteration is linear to the number of edges, i.e., $\mathcal{O}(|E|)$. Moreover, we exploit the sparsity of γ_e to reduce the computational complexity into $\mathcal{O}(|E_{\mu,\nu}|)$ since we can remove all edges $e \in E \setminus E_{\mu,\nu}$ in the summation in Equation (13) for the computation of GSI-M between μ and ν. Therefore, let T be the number of iterations used in the *fmincon* solver, then the computational complexity of GSI-M between measures μ and ν is $\mathcal{O}(T|E_{\mu,\nu}|)$. Empirically, we observe that the solver converges after about 40–50 iterations.

B.2 Further Empirical Results

We provide further results, corresponding to the empirical results in §5 for graph $\mathbb{G}_{\text{Sqrt}}$.

Computational comparison. We illustrate corresponding results for computational comparison on graph $\mathbb{G}_{\text{Sqrt}}$ with 1K nodes and 32K edges in Figure 4. The computation of GSI-M is also several-order faster than OW, and comparable to GST. More concretely, $\widehat{\mathcal{GS}}$ is $100\times, 6800\times, 2900\times$ faster than OW for Φ_0, Φ_1, Φ_2 respectively. For N-functions Φ_1 and Φ_2, GSI-M takes less than 23 *seconds* while OW needs at least 19 *hours*, and up to 34 *hours* for the computation.

Additionally, we also include generalized Sobolev transport (GST) [32], which is a scalable alternative of OW in practical applications, in the computational comparison, and illustrated in Figures 5 and 6 for graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ respectively. We observe that the computations of the regularized GSI (GSI-M) are comparable to GST for all Φ functions.

Document classification. Figure 7 illustrates corresponding results for document classification on graph $\mathbb{G}_{\text{Sqrt}}$ with 10K nodes and about 1M edges.

TDA. Figure 8 shows corresponding results for TDA on graph $\mathbb{G}_{\text{Sqrt}}$ with 10K nodes and about 1M edges.

Ablation study on different graphs and different N-functions. Besides the graphs $\mathbb{G}_{\text{Log}}$(10K, 100K) and $\mathbb{G}_{\text{Sqrt}}$(10K, 1M) in our experiments where the numbers in the parenthesis, we run additional experiments on various graphs with different connectivities, and different number of nodes: G_{Log1K}(1K, 7K), G_{Sqrt1K}(1K, 32K), G_{Log100}(100, 500), and G_{Sqrt100}(100, 1K). Furthermore, following Proposition 3.7 with $p = 2$, we consider an additional N-function $\Phi_3(t) = \frac{1}{4}t^2$, besides the two popular Φ functions Φ_1, Φ_2 in the literature, used in our experiments. We illustrate results for ablation study on different graphs and different N-functions for document classification on `TWITTER` dataset in Table 2 and for TDA on `MPEG7` dataset in Table 3. We observe that for the same given graphs, performances of GSI-M-Φ_i ($i = 1, 2, 3$), consistently improve performances of GSI-M-Φ_0.

Ablation study on different root nodes. In our experiments, we use cross validation to choose a root node from a random 10-root-node subset of V in the given graph $\mathbb{G}$. We further carry out

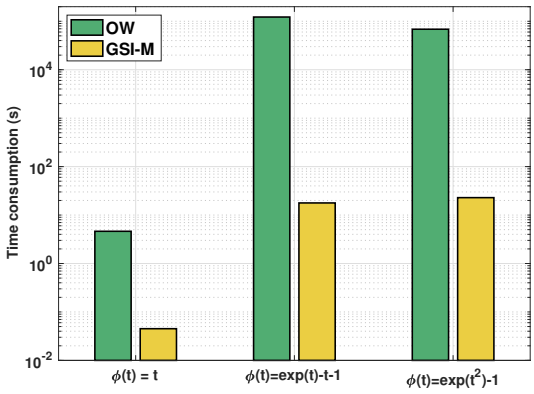


Figure 4: Time consumption for GSI-M and OW on $\mathbb{G}_{\text{Sqrt}}$ with 1K nodes and about 32K edges.

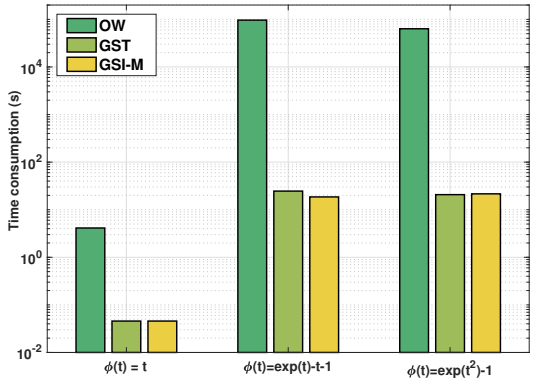


Figure 5: Time consumption for GSI-M, GST and OW on $\mathbb{G}_{\text{Log}}$ with 1K nodes and about 7K edges.

additional experiments where the root node is randomly selected, and repeat it 50 times. Tables 4 and 5 illustrate ablation studies on different root nodes for document classification on TWITTER dataset, and for TDA on MPEG7 dataset respectively. We observe that the root node may affect performances of GSI-M. The cross-validation can help to improve performances of GSI-M in our experiments.

For given graphs with a minor perturbation of edge weights. We further consider a minor perturbation of edge weights on given graphs by adding random deviations in the range $(0, 0.1)$.

- In Tables 6 and 7, we illustrate results for document classification on TWITTER dataset with graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ with a minor perturbation of their edge weights respectively.
- In Tables 8 and 9, we illustrate results for TDA on MPEG7 dataset with graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ with a minor perturbation of their edge weights respectively.

We observe that the perturbation on edge weights slightly affects performances of GSI-M.

Table 2: Ablation study on different graphs and different N-functions for document classification on TWITTER dataset.

	Φ_0	Φ_1	Φ_2	Φ_3
$\mathbb{G}_{\text{Log}}$(10K, 100K)	70.55 ± 0.41	72.68 ± 0.58	71.85 ± 0.61	71.78 ± 0.61
$\mathbb{G}_{\text{Sqrt}}$(10K, 1M)	70.78 ± 0.47	72.45 ± 0.46	72.15 ± 0.46	71.86 ± 0.50
$\mathbb{G}_{\text{Log}}$(1K, 7K)	70.84 ± 0.38	72.52 ± 0.36	72.35 ± 0.34	72.03 ± 0.40
$\mathbb{G}_{\text{Sqrt}}$(1K, 32K)	70.99 ± 0.50	72.25 ± 0.42	72.57 ± 0.48	72.05 ± 0.44
$\mathbb{G}_{\text{Log}}$(100, 500)	70.05 ± 0.33	71.92 ± 0.44	71.36 ± 0.42	71.22 ± 0.26
$\mathbb{G}_{\text{Sqrt}}$(100, 1K)	69.81 ± 0.45	71.71 ± 0.30	71.29 ± 0.43	70.96 ± 0.43

Table 3: Ablation study on different graphs and different N-functions for TDA on MPEG7 dataset.

	Φ_0	Φ_1	Φ_2	Φ_3
$\mathbb{G}_{\text{Log}}$(1K, 7K)	61.50 ± 5.41	75.00 ± 4.16	78.33 ± 5.49	72.67 ± 4.66
$\mathbb{G}_{\text{Sqrt}}$(1K, 32K)	65.17 ± 4.54	81.67 ± 6.25	81.67 ± 4.78	79.33 ± 6.58
$\mathbb{G}_{\text{Log}}$(100, 500)	68.17 ± 5.00	76.67 ± 4.29	75.67 ± 5.44	74.17 ± 5.24
$\mathbb{G}_{\text{Sqrt}}$(100, 1K)	72.17 ± 5.67	78.33 ± 6.15	77.67 ± 6.35	76.33 ± 5.33

Table 4: Ablation study on different root nodes for document classification on TWITTER dataset.

	Φ_1	Φ_2
$\mathbb{G}_{\text{Log}}$	72.42 ± 0.41	71.53 ± 0.44
$\mathbb{G}_{\text{Sqrt}}$	72.31 ± 0.28	72.04 ± 0.25

Table 5: Ablation study on different root nodes for TDA on MPEG7 dataset.

	Φ_1	Φ_2
$\mathbb{G}_{\text{Log}}$	74.50 ± 1.39	77.23 ± 2.02
$\mathbb{G}_{\text{Sqrt}}$	79.92 ± 2.65	80.03 ± 2.61

Table 6: Ablation study on graph $\mathbb{G}_{\text{Log}}$ with a minor perturbation of its edge weights for document classification on TWITTER dataset.

	Φ_1	Φ_2
No perturbation	72.68 ± 0.58	71.85 ± 0.61
With perturbation	72.14 ± 0.46	71.42 ± 0.39

Table 7: Ablation study on graph $\mathbb{G}_{\text{Sqrt}}$ with a minor perturbation of its edge weights for document classification on TWITTER dataset.

	Φ_1	Φ_2
No perturbation	72.45 ± 0.46	72.15 ± 0.46
With perturbation	71.95 ± 0.44	71.87 ± 0.48

Table 8: Ablation study on graph $\mathbb{G}_{\text{Log}}$ with a minor perturbation of its edge weights for TDA on MPEG7 dataset.

	Φ_1	Φ_2
No perturbation	75.00 ± 4.16	78.33 ± 5.49
With perturbation	73.33 ± 4.67	76.17 ± 3.93

Table 9: Ablation study on graph $\mathbb{G}_{\text{Sqrt}}$ with a minor perturbation of its edge weights for TDA on MPEG7 dataset.

	Φ_1	Φ_2
No perturbation	81.67 ± 6.25	81.67 ± 4.78
With perturbation	81.33 ± 6.31	81.17 ± 6.06

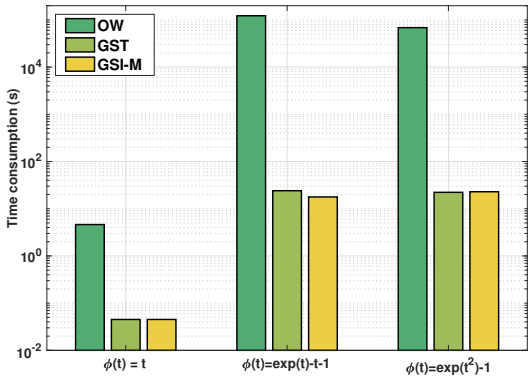


Figure 6: Time consumption for GSI-M, GST and OW on $\mathbb{G}_{\text{Sqrt}}$ with 1K nodes and about 32K edges.

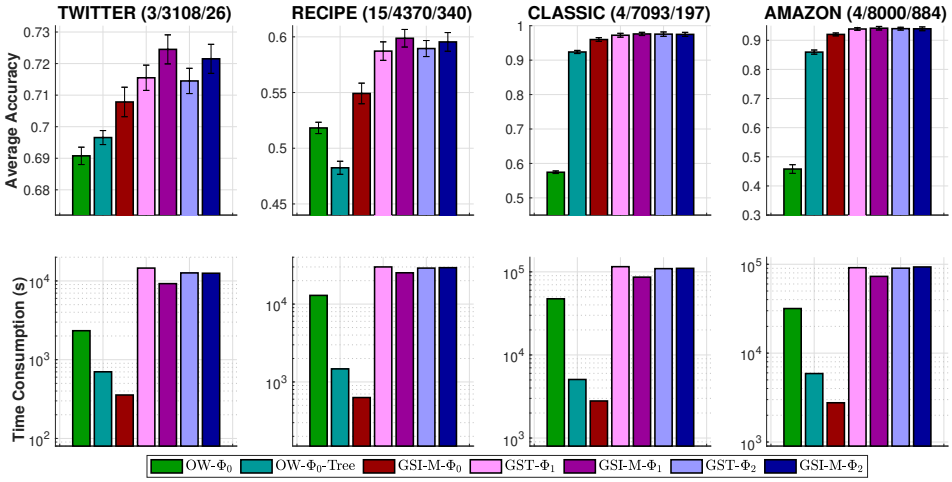


Figure 7: SVM results and time consumption for kernel matrices with graph $\mathbb{G}_{\text{Sqrt}}$.

Empirical gaps between GSI-M and GST/OW. To our knowledge, for general non-tree graph structures, it is unknown whether there is any connection between GSI-M and OW or not. Their relation in Proposition 4.6 is only held when the given graph is a tree. For the connection between GSI-M and GST, their relation in Proposition 4.5 is held for any given general non-tree graph structures.

Additionally, we clarify that the goal of GSI-M is to provide an efficient algorithmic approach which yields an equivalent metric to the GSI problem. We agree that GSI-M leverages Orlicz geometry for the *Sobolev IPM* problem similar to GST and OW which leverage Orlicz geometry for *Sobolev transport* and *optimal transport* respectively. Moreover, we also clarify that the goal of GSI-M is not to approximate either GST or OW.

Furthermore, with the setting in §5.1, we compute the gap between GSI-M and GST, and the gap between GSI-M and OW, illustrated in Tables 10 and 11 respectively. We observe that the empirical gap between GSI-M and GST is relatively small, while the empirical gap between GSI-M and OW is relatively large.

Table 10: The empirical gap between GSI-M and GST.

	Φ_1	Φ_2
$\mathbb{G}_{\text{Log}}$	0.019 ± 0.011	0.026 ± 0.015
$\mathbb{G}_{\text{Sqrt}}$	0.004 ± 0.002	0.005 ± 0.002

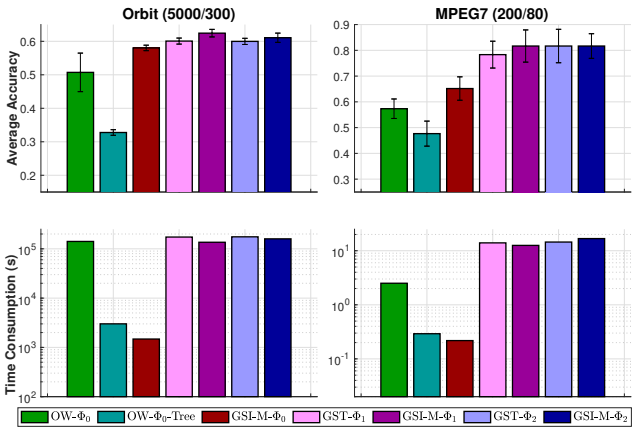


Figure 8: SVM results and time consumption for kernel matrices with graph $\mathbb{G}_{\text{Sqrt}}$.

Table 11: The empirical gap between GSI-M and OW.

	Φ_1	Φ_2
$\mathbb{G}_{\text{Log}}$	293.06 ± 19.43	403.47 ± 26.75
$\mathbb{G}_{\text{Sqrt}}$	165.45 ± 28.88	227.79 ± 39.75

C Further Discussions

For completeness, we recall important discussions on the underlying graph in [30, 32, 33], since they are also applied and/or adapted to the proposed generalized Sobolev IPM with Musielak regularization.

Measures on a graph [33]. We reemphasize that in this work we consider the Sobolev IPM problem between *two input probability measures* supported on the *same* graph, which is also explored in [33]. Such measures supported on a graph metric space are also considered in OT problem, explored in [30, 32]. Our work generalizes Sobolev IPM, and we also derive a novel regularization for the generalized Sobolev IPM, which admits an efficient algorithmic approach (i.e., simply solving a univariate optimization problem for its computation).

The generalized Sobolev IPM with Musielak regularization (GSI-M) is for *input probability measures*, i.e., to compute distance between two probability measures, on the *same* graph. We further distinguish the considered problem to the following related problems:

• **Compute distance between two (different) input graphs.** Petric Maretic et al. [45], Dong & Sawin [16] compute OT problem between *two input graphs*, where their goals are to compute distance between such two input graphs. Therefore, they are essentially different to our considered problem which computes distance between *two input probability measures* supported on the *same* graph.

• **Graph kernels between two (different) input graphs.** Graph kernels are functions between *two input graphs* to measure their similarity. See Borgwardt et al. [7] for a comprehensive review on graph kernels. Obviously, this research direction is different to our considered kernels for SVM which are built upon the GSI-M used for measuring similarity between *two input probability measures* on the *same* graph.

Applications with graph-based measures. One of motivations of considering the graph-based measures in applications is from the manifold perspective for supports of measures. For such applications, one of popular approaches is to use graph structure as a representation for data manifolds (e.g., k-nearest-neighbor graphs). Additionally, some applications may have graph-based structure within themselves, e.g., one example is the popular transportation problem with a given map (regarded as a graph-structure over locations). A concrete example is to transport minerals from a distribution of mines to a distribution of factories, where locations of mines and factories are physically connected in a graph-based map, and the transportation (e.g., by vehicles) follows such a real physical map.

Additionally, please see [11] for further application settings of optimal transport for measures supported on a graph, and also Gaussian process on a given graph [68], or graph field integrator [13]. Therefore, leveraging graph structure for machine learning is a promising research, not only for optimal transport and IPM, but also several other machine learning problems.

Path length for points in graph $\mathbb{G}$ [30]. 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 of graph $\mathbb{G}$.

Consider the edge $e = \langle u, v\rangle$ connecting two nodes $u, v \in V$, for $x, y \in \mathbb{R}^n$ and $x, y \in e$, we have

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

for some scalars $t, s \in [0, 1]$. Therefore, the length of the path $[x, y]$ along edge e (i.e., the line segment $\langle x, y\rangle$) is equal to $|t-s|w_e$. As a result, 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 [30].

Extension to measures supported on $\mathbb{G}$. Similar to the regularized Sobolev IPM [33], the discrete case of the GSI-M in Equation (13) can be easily extended for measures with finite supports on $\mathbb{G}$ (i.e., supports of the input measures may not be nodes in V, but possibly points on edges in E) by using the same strategy to measure a path length for support data points in graph $\mathbb{G}$. More precisely, we break down edges containing supports into pieces and sum over their corresponding values instead of the sum over edges.

The assumption of uniqueness property of the shortest paths on $\mathbb{G}$. As discussed in [30, 32, 33], note that $w_e \in \mathbb{R}_+$ for any edge $e \in E$ in graph $\mathbb{G}$., it is almost surely that every node in V can be regarded as unique-path root node since with a high probability, lengths of paths connecting any two nodes in graph $\mathbb{G}$ are different.

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 if $\mathbb{G}$ is a non-physical graph, with a small deviation, we can obtain a graph satisfying the unique-path root node assumption.

Besides that, for input probability measures with full supports in graph $\mathbb{G}$, or at least full supports in any cycle in graph $\mathbb{G}$, then it exists a special support data point where there are multiple shortest paths from the root node to it. In this case, we simply choose one fixed shortest path among them for this support data point (or we can add a virtual edge from the root node to this support data point where the edge length is deducted by a small deviation). In many practical applications (e.g., document classification and TDA in our experiments), one can neglect this special case since input probability measures have a finite number of supports.

Recent advantages of Orlicz geometric structure. In this work, we propose generalized Sobolev IPM (GSI) which leverages Orlicz geometric structure for Sobolev IPM for going beyond the L^p paradigm, which is mainly motivated from recent advances and breakthroughs of Orlicz-Wasserstein (OW) [54].

As discussed in §1, the two breakthroughs results on OW [3, 19] illustrate that OW possesses unique implicit essences, which may not exist in the classic optimal transport (OT) with L^p geometric structure cost.

- Altschuler & Chewi [3] leverage OW to facilitate the development of differential privacy-inspired methodologies that address long-standing convergence challenges in hypocoercive differential equations.
- Guha et al. [19] demonstrate that OW substantially enhances Bayesian contraction rates, effectively circumventing limitations inherent to traditional OT with Euclidean ground cost.

Additionally, we also summarize some recent advantages of Orlicz geometric structure as follows:

- Andoni et al. [4] and Song et al. [51] used Orlicz norm as a loss for regression to generalize both the least-square (squared L^2-norm loss) and least absolute deviation (L^1-norm loss) regressions, and also provide a scale-invariant version of M-estimator function loss (e.g., the Huber, Cauchy, Welsh, Tukey, Geman-McClure functions)

- Andoni et al. [4] also illustrated that regression with Orlicz norm loss improves performances of linear regression with L^1/L^2 norms on Gaussian/sparse noise respectively.
- Deng et al. [15] leveraged Orlicz structure to scale up many machine learning problems including reinforcement learning, kernelized support vector machine, and clustering.
- Chamakh et al. [9] leveraged Orlicz structure to propose finite-sample uniform guarantee for random Fourier features which is almost surely convergence to approximate high-order derivatives for arbitrary kernels.
- Lorenz & Mahler [38] studied Orlicz norm as a regularization for OT problem with continuous probability measures.
- Le et al. [32] illustrated that Orlicz geometric structure is useful for the generalize Sobolev transport in applications, including document classification, orbit recognition, image classification. These applications are also included for our experimental setups (§5).

Advantages of Orlicz geometric structures in applications are a promising on-going research directions. For comprehensive studies of Orlicz geometry, see [2, 48].

The generalized Sobolev IPM with Musielak regularization (GSI-M). Similar to regularized Sobolev IPM [33], we assume that the graph metric space is given. The question of adaptively learning an optimal graph metric structure from given data is left for future work for further investigation.

Further computational discussion on Sobolev IPM, generalized Sobolev IPM, and regularization approaches. Additionally, we clarify that for the original Sobolev IPM, and original generalized Sobolev IPM (GSI), to our knowledge, we have not been aware of any algorithmic approaches to compute them efficiently yet (i.e., without regularization).

The regularized Sobolev IPM has a closed-form expression for computation [33, Theorem 3.5]. Consequently, there is no need for any optimization in the computation of the regularized Sobolev IPM.

Our proposed GSI-M is an efficient approach which yields an equivalent metric to GSI only. We agree that it is interesting to consider whether it is possible to derive efficient algorithmic approaches for the original Sobolev IPM and original GSI in future work as described in §6.

Following Theorem 13, we simply need to solve the univariate optimization problem for the computation of GSI-M. In practice, we use a second-order solver (i.e., *fmincon* with Trust Region Reflective solver) in our experiments

The graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$ [30, 33]. For an efficient and fast computation, we apply the farthest-point clustering method to cluster supports of measures into at most M clusters.[19] Then, let the set of vertices V be the set of centroids of these clusters, i.e., graph vertices. For edges, in graph $\mathbb{G}_{\text{Log}}$, we randomly choose $(M \log M)$ edges; and $M^{3/2}$ edges for graph $\mathbb{G}_{\text{Sqrt}}$. We further denote the set of those randomly sampled edges as $\tilde{E}$.

For each edge e, its corresponding edge length (i.e., edge 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.

• For document classification, following [33], we use word2vec to map words into vectors in $\mathbb{R}^{300}$, and apply the above procedures for the word embeddings to construct graphs $\mathbb{G}_{\text{Log}}$ and $\mathbb{G}_{\text{Sqrt}}$. Therefore, we regard each document as a probability distribution of words supported on these given graphs.

• For TDA, we use the above procedure for topological features, which are 2-dimensional data points summarizing its life span. Thus, we regard each PD as a probability measure of topological features supported on these given graphs.

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

Datasets. For the datasets in our experiments (i.e., `TWITTER, RECIPE, CLASSIC, AMAZON` for document datasets, and `Orbit, MPEG7` for TDA), one can contact the authors of Sobolev transport [30] to access to them.

Computational devices. We run all of our experiments on commodity hardware.

Hyperparamter validation. We use the same strategy as in [33]. For validation, we further randomly split *the training set* into $70\%/30\%$ for validation-training and validation with 10 repeats to choose hyper-parameters in experiments.

Further discussion on hyperparameters. The performance of the generalized Sobolev IPM with Musielak regularization (GSI-M) basically depends on the choice of the N-function Φ, which is much similar to how kernel functions impact performance in kernel-dependent machine learning framework. In our experiments with N-functions Φ_1 and Φ_2, performances with N-function Φ_1 is slightly more favorable than those with N-function Φ_2, except in `MPEG7` dataset with graph $\mathbb{G}_{\text{Sqrt}}$.

Determining the optimal N-function Φ for the generalized Sobolev IPM with Musielak regularization (GSI-M) in a given task is an open problem that warrants further investigation. We leave it for future work. As an interim solution, cross-validation can be used to select N-function Φ from a set of candidate functions.

The number of pairs in training and test for kernelized SVM [33]. Let N_{tr}, N_{te} be the number of measures for training and test respectively. For the kernelized SVM training, the number of pairs which we need to evaluate distances is $(N_{tr}-1)\times\frac{N_{tr}}{2}$. On the test phase, the number of pairs which we need to evaluate distances is $N_{tr}\times N_{te}$. Therefore, for each run, the number of pairs which we require to evaluate distances for both training and test is totally $N_{tr}\times(\frac{N_{tr}-1}{2}+N_{te})$. See Table 1 for the number of pairs we need to evaluate distances for kernelized SVM in experiments.

Graph structure construction. In our experiments, we consider the given graphs, which are constructed from supports of all measures in the training set. Additionally, one may consider the tree system for the generalized Sobolev IPM (GSI) problem which may by-pass a bias of graph structure construction from finite data points due to the unboundedness of the tree system, i.e., with an infinite number of nodes.

Limitations and discussions. We study Sobolev IPM for probability measures supported on a *given* graph metric, as in [30, 32, 33, 11]. The approach of graph-based probability measures extends those involving the tree-based measures and 1d-projected measures, since graph structure provides more flexibility and degrees of freedom compared to the tree structure and 1d-line structure (also see [30]). Some examples using tree-structure and 1d-line structure include tree-sliced Wasserstein [29, 55–58, 61], tree-sliced entropy partial transport (EPT) [28, 59], tree-sliced Sobolev IPM [60], and the popular sliced-Wasserstein [47, 6, 43]. As in discussed in §1, [60] leveraged tree system [55–58, 61] and proposed efficient Radon transforms to extend finding results for graph-based measures (e.g., regularized Sobolev IPM [33]) to applications with probability measures on the popular Euclidean spaces and hyperspheres.

Therefore, the tree system framework provides a mechanism to bridge finding results for graph-based measures and applications with probability measures on the popular Euclidean spaces and hyperspheres. Since we focus on Sobolev IPM for graph-based measures in this work, we leave this application research direction as future works for in-depth investigations.

D Brief Reviews

In this section, to ease the readers, we give a brief review for related notions which we have used in the development of our proposal. For completeness, we also summarize some notions reviewed in [32, 33].

D.1 Graph Illustration

We illustrate graph notions, reviewed in §2, in Figure 9.

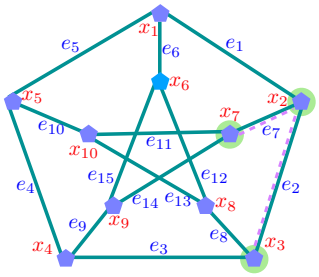


Figure 9: A geodetic graph with 10 nodes $\{x_1, x_2, \ldots, x_{10}\}$ and 15 edges $\{e_1, e_2, \ldots, e_{15}\}$, and each edge length equals to 1, i.e., $w_{e_j} = 1, \forall j$. For any x_i, x_j, there is a unique shortest path between them, with a length 2. Therefore, it satisfies the uniqueness property of the shortest paths. Let x_1 be the unique-path root node (i.e., $z_0 = x_1$), and subgraph $\widetilde{\mathbb{G}}$ containing 3 nodes $\{x_2, x_3, x_7\}$ and 2 edges $\{e_2, e_7\}$, then we have $\Lambda(x_2) = \gamma(e_1) = \widetilde{\mathbb{G}}$.

D.2 A Review on Functional Spaces

We describe a short review on the L^p space, the weighted L^p space.

L^p space. For a nonnegative Borel measure λ on $\mathbb{G}$, let $L^p(\mathbb{G}, \lambda)$ be the space of all Borel measurable functions $f : \mathbb{G} \to \mathbb{R}$ s.t. $\int_{\mathbb{G}} |f(y)|^p \lambda(\mathrm{d}y) < \infty$. For $p = \infty$, we instead assume that f is bounded λ-a.e. Then, $L^p(\mathbb{G}, \lambda)$ is a normed space with the norm being defined as follows:

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

On the other hand, for $p = \infty$, then we have

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

Additionally, 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-a.e. $x \in \mathbb{G}$.

$L^p_{\hat{w}}$ space. For a nonnegative Borel measure λ on $\mathbb{G}$, and a positive weight function $\hat{w}$ on G, let $L^p_{\hat{w}}(\mathbb{G}, \lambda)$ be the space of all Borel measurable functions $f : \mathbb{G} \to \mathbb{R}$ s.t. $\int_{\mathbb{G}} \hat{w}(x)|f(x)|^p \lambda(\mathrm{d}x) < \infty$. For $p = \infty$, we instead assume that f is bounded $\hat{w}\lambda$-a.e. Then, $L^p_{\hat{w}}(\mathbb{G}, \lambda)$ is a normed space with the norm being defined as follows:

$$\|f\|_{L^p_{\hat{w}}} := \left(\int_{\mathbb{G}} \hat{w}(x)|f(x)|^p \lambda(\mathrm{d}x) \right)^{\frac{1}{p}} \quad \text{for } 1 \leq p < \infty.$$

For the case $p = \infty$, as $\hat{w}(x) > 0$ for every $x \in \mathbb{G}$, we have

$$\begin{aligned} \|f\|_{L^\infty_{\hat{w}}} &:= \inf \{t \in \mathbb{R} : \ |f(x)| \leq t \text{ for } (\hat{w}\lambda)\text{-a.e. } x \in \mathbb{G}\} \\ &= \inf \{t \in \mathbb{R} : \ |f(x)| \leq t \text{ for } \lambda\text{-a.e. } x \in \mathbb{G}\} \\ &= \|f\|_{L^\infty}. \end{aligned}$$

D.3 A Review on Sobolev Transport

In this section, we provide a brief review on the Sobolev transport (ST) [30] for graph-based measures, and the length measure on a graph.

Definition D.1 (Graph-based Sobolev space [30])**.** Let λ 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 $W^{1,p}(\mathbb{G}, \lambda)$ if there exists a function $h \in L^p(\mathbb{G}, \lambda)$ satisfying

$$f(x) - f(z_0) = \int_{[z_0, x]} h(y) \, \lambda(\mathrm{d}y) \quad \forall x \in \mathbb{G}. \tag{45}$$

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. The graph derivative of $f \in W^{1,p}(\mathbb{G}, \lambda)$ is denoted as $f' \in L^p(\mathbb{G}, \lambda)$.

Sobolev transport (ST) [30]. For probability measures $\mu, \nu \in \mathcal{P}(\mathbb{G})$, and $1 \leq p \leq \infty$, the p-order Sobolev transport (ST) [30, Definition 3.2] is defined as

$$\mathcal{ST}_p(\mu,\nu) \coloneqq \begin{cases} \sup \left[\int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(x)\nu(\mathrm{d}x) \right] \\ \text{s.t. } f \in W^{1,p'}(\mathbb{G},\lambda),\ \|f'\|_{L^{p'}(\mathbb{G},\lambda)} \leq 1, \end{cases} \tag{46}$$

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

ST is a scalable variant of optimal transport (OT) for probability measures supported on a graph (i.e., not OT itself) since it relaxes the Lipschitz constriant for the critic function as in 1-order Wasserstein by considering this Lipschitz constraint in a Sobolev space. Additionally, Le et al. [30, Corollary 4.3] showed that 1-order Sobolev transport is identical to 1-order Wasserstein *when the graph is a tree*. Furthermore, Le et al. [30, §4.1] admitted that the exact relationship between p-order Sobolev transport and p-order Wasserstein when $p > 1$ is *unknown*.

Furthermore, for input measures possibly having different total mass, ST can be extended into its unbalanced variant, e.g., see [31] for rigorous details of the unbalanced Sobolev transport.

Proposition D.2 (Closed-form expression of Sobolev transport [30])**.** *Let λ be any nonnegative Borel measure on $\mathbb{G}$, and let $1 \leq p < \infty$. Then, we have*

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

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

Definition D.3 (Length measure [30])**.** Let λ^* be the unique Borel measure on $\mathbb{G}$ s.t. 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 $s \leq t$. Here, recall that $\langle x, y\rangle$ is the line segment in e connecting x and y.

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

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

Lemma D.4 (λ^* is the length measure on graph [30])**.** *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*

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

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

D.4 A Review on Sobolev IPM and Its Scalable Regularized Approach

We give a brief review on the Sobolev norm [2], Sobolev IPM, and its scalable regularized approach [33] for graph-based measures.

Sobolev norm. $W^{1,p}(\mathbb{G},\lambda)$ is a normed space (reviewed in §D.3), with the Sobolev norm [2, §3.1] being defined as

$$\|f\|_{W^{1,p}} \coloneqq \left(\|f\|_{L^p}^p + \|f'\|_{L^p}^p \right)^{\frac{1}{p}}. \tag{47}$$

Additionally, let $W_0^{1,p}(\mathbb{G},\lambda)$ be the subspace consisting of all functions f in $W^{1,p}(\mathbb{G},\lambda)$ satisfying $f(z_0) = 0$. Denote $\mathcal{B}_p \coloneqq \left\{ f \in W_0^{1,p}(\mathbb{G},\lambda) : \|f\|_{W^{1,p}} \leq 1 \right\}$ as the unit ball in the Sobolev space.

Sobolev IPM. Sobolev IPM for graph-based measures is an instance of the IPM where its critic function belongs to the graph-based Sobolev space, and is constrained within the unit ball of that

space. More concretely, given a nonnegative Borel measure λ on $\mathbb{G}$, an exponent $1 \leq p \leq \infty$ and its conjugate p', the Sobolev IPM between measures $\mu, \nu \in \mathcal{P}(\mathbb{G})$ is defined as follows:

$$\mathcal{S}_p(\mu, \nu) := \sup_{f \in \mathcal{B}_{p'}} \left| \int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(y)\nu(\mathrm{d}y) \right|. \tag{48}$$

Note that one should distinguish Sobolev IPM problem from Sobolev transport [30], generalized Sobolev transport (GST) [32], and Wasserstein problems. In fact, for Sobolev IPM, the critic function is constrained within a unit ball of Sobolev norm, which involves both critic function and its gradient while Sobolev transport, GST and Wasserstein only constraint on gradient of critic function (i.e., Lipschitz constraint).

Scalable regularized Sobolev IPM [33]. For weight function $\hat{w}_{\hat{\mathcal{S}}}(x) = 1 + \lambda(\Lambda(x))$, for all $x \in \mathbb{G}$. Let $\mathcal{B}(p', \hat{w}_{\hat{\mathcal{S}}})$ be defined as follows:

$$\mathcal{B}(p', \hat{w}_{\hat{\mathcal{S}}}) := \left\{ f \in W_0^{1,p'}(\mathbb{G}, \lambda) : \|f'\|_{L^{p'}_{\hat{w}_{\hat{\mathcal{S}}}}} \leq 1 \right\}. \tag{49}$$

Then, the regularized Sobolev IPM is defined as

Definition D.5 (Regularized Sobolev IPM on graph [33])**.** Let λ be a nonnegative Borel measure on $\mathbb{G}$ and $1 \leq p \leq \infty$. Then for any given probability measures $\mu, \nu \in \mathcal{P}(\mathbb{G})$, the regularized Sobolev IPM is defined as

$$\hat{\mathcal{S}}_p(\mu, \nu) := \sup_{f \in \mathcal{B}(p', \hat{w}_{\hat{\mathcal{S}}})} \left| \int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(y)\nu(\mathrm{d}y) \right|. \tag{50}$$

Note that it is *unknown* whether the regularization approach in Le et al. [33] with weight function $\hat{w}_{\hat{\mathcal{S}}}$ can be extended for general cases beyond the L^p geometric structure, e.g., Orlicz geometric structure.

D.5 A Review on IPM and Wasserstein Distance

We provide a short review on IPM and Wasserstein distance for probability measures.

IPM. Integral probability metrics (IPM) for probability measures μ, ν are defined as follows:

$$\gamma_{\mathcal{F}}(\mu, \nu) = \sup_{f \in \mathcal{F}} \left| \int_{\lambda} f(x)\mu(\mathrm{d}x) - \int_{\lambda} f(y)\nu(\mathrm{d}y) \right|. \tag{51}$$

Special case: 1-Wasserstein distance (dual formulation). The 1-Wasserstein distance is a special case of IPM. In particular, for $\mathcal{F} = \mathcal{F}_W := \{f : |f(x) - f(y)| \leq d_{\mathbb{G}}(x, y)\}$ where $d_{\mathbb{G}}$ is the graph metric on graph $\mathbb{G}$, then IPM is equal to the 1-Wasserstein distance with ground cost $d_{\mathbb{G}}$

$$\mathcal{W}(\mu, \nu) = \sup_{f \in \mathcal{F}_W} \left| \int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(y)\nu(\mathrm{d}y) \right|. \tag{52}$$

p-Wasserstein distance (primal formulation). Let $1 \leq p < \infty$, for probability measures μ and ν on $\mathbb{G}$, then, the p-Wasserstein distance is defined as follows:

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

where $\Pi(\mu, \nu) := \left\{ \pi \in \mathcal{P}(\mathbb{G} \times \mathbb{G}) : \pi_1 = \mu, \pi_2 = \nu \right\}$; π_1, π_2 are the first and second marginals of π respectively.

We would like to remark that to our knowledge, there is *no* closed-form expression for optimal transport (OT) problem for probability measures on a graph in general. However, when the graph is a tree, OT admits closed-form expression, a.k.a., tree-Wasserstein [29].

D.6 Orlicz functions

We provide a brief review on Orlicz functions as summarized in [32] for completeness. For comprehensive studies on Orlicz functions, see [2, 48].

Popular examples of N-functions. Some popular examples for N-functions [2, §8.2] in the literature are as follows:

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

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

$$\Psi(t) = \sup\left[at - \Phi(a) \mid a \geq 0\right], \quad \text{for } t \geq 0. \tag{53}$$

Popular examples of complementary pairs of N-functions. Some popular complementary pairs of N-functions [2, §8.3], [48, §2.2] are as follows:

1. $\Phi(t) = \frac{t^p}{p}$ and $\Psi(t) = \frac{t^q}{q}$ where q is the conjugate of p, i.e., $\frac{1}{p} + \frac{1}{q} = 1$ and $1 < p < \infty$.
2. $\Phi(t) = \exp(t) - t - 1$ and $\Psi(t) = (1+t)\log(1+t) - t$.
3. For the N-function $\Phi(t) = \exp(t^p) - 1$ with $1 < p < \infty$, its complementary N-function yields an explicit for, but not simple [48, §2.2], see [32, §A.8] for the details.

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

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

Orlicz norm. Together with the Luxemburg norm, the Orlicz norm [48, §3.3, Definition 2] is a popular norm for $L_\Phi(\mathbb{G}, \lambda)$ in the literature, defined as

$$\|f\|_\Phi := \sup\Big\{ \int_{\mathbb{G}} |f(x)g(x)|\lambda(\mathrm{d}x) \mid \int_{\mathbb{G}} \Psi(|g(x)|)\lambda(\mathrm{d}x) \leq 1 \Big\}, \tag{54}$$

where Ψ is the complementary N-function of Φ.

Computation for Orlicz norm. Following [48, §3.3, Theorem 13], the Orlicz norm can be recasted as follows:

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

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

Equivalence [2, §8.17] [41, §13.11]. The Luxemburg norm is equivalent to the Orlicz norm. In fact, we have

$$\|f\|_{L_\Phi} \leq \|f\|_\Phi \leq 2\,\|f\|_{L_\Phi}. \tag{55}$$

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

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

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

$$\left|\int_{\mathbb{G}} f(x)g(x)\lambda(dx)\right| \leq 2\,\|f\|_{L_\Phi}\,\|g\|_{L_\Psi}. \tag{56}$$

Additionally, we have the generalized Hölder inequality w.r.t. Luxemburg norm and Orlicz norm [41, §13.13] is as follows:

$$\left|\int_{\mathbb{G}} f(x)g(x)\lambda(dx)\right| \leq \|f\|_{L_\Phi}\,\|g\|_{\Psi}. \tag{57}$$

D.7 Generalized Sobolev transport (GST)

We briefly review generalized Sobolev transport (GST) [32] for graph-based measures.

Generalized Sobolev transport (GST) [32]. Let Φ be an N-function and λ be a nonnegative Borel measure on $\mathbb{G}$. For probability measures μ, ν on a graph $\mathbb{G}$, the generalized Sobolev transport (GST) is defined as follows:

$$\mathcal{GS}_{\Phi}(\mu,\nu) \coloneqq \left\{ \begin{array}{ll} \sup & \left| \int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(x)\nu(\mathrm{d}x) \right| \\ \text{s.t.} & f \in WL^1_{\Psi}(\mathbb{G},\lambda),\ \|f'\|_{L_\Psi} \leq 1, \end{array} \right.$$

where Ψ is the complementary function of Φ (see Equation (53)).

Note that GST is a scalable variant of Orlicz-Wasserstein (OW) for graph-based probability measures (i.e., not OW itself). Moreover, Le et al. [32, Proposition 4.6] showed that GST is equal to OW *when the graph is a tree, and* $\Phi(t) = t$ *(i.e., the limite case of* N*-function).*

D.8 Orlicz-Wasserstein (OW)

We briefly review Orlicz-Wasserstein (OW) [54] for graph-based measures. Following Guha et al. [19, Definition 3.2], the OW with the N-function Φ for measures $\mu, \nu \in \mathcal{P}(\mathbb{G})$ is defined as follows:

$$W_{\Phi}(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \inf \left[t > 0 : \int_{\mathbb{G}\times\mathbb{G}} \Phi\left(\frac{d_{\mathbb{G}}(x,z)}{t}\right) \mathrm{d}\pi(x,z) \leq 1 \right], \tag{58}$$

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

E Proofs

In this section, we give detailed proofs for the theoretical results in the main manuscript and additional results in Appendix §A.

E.1 Proof for Theorem 3.2

Proof. We first derive the lower bound as follows:

$$\begin{aligned} \|f\|_{WL^1_{\Phi}} &= \|f\|_{L_\Phi} + \|f'\|_{L_\Phi} \\ &\geq \|f'\|_{L_\Phi}. \end{aligned} \tag{59}$$

Additionally, for any scalar $k \geq 1$, we have

$$\begin{aligned} \int_{\mathbb{G}} \Phi\left(k\frac{|f'(x)|}{t}\right) \lambda(\mathrm{d}x) &\geq \int_{\mathbb{G}} k\,\Phi\left(\frac{|f'(x)|}{t}\right) \lambda(\mathrm{d}x) && (60) \\ &\geq \int_{\mathbb{G}} \frac{k}{2}\left(1 + \frac{\lambda(\Lambda(x))}{\lambda(\mathbb{G})}\right) \Phi\left(\frac{|f'(x)|}{t}\right) \lambda(\mathrm{d}x) && (61) \\ &= \int_{\mathbb{G}} \frac{k}{2}\hat{w}(x)\,\Phi\left(\frac{|f'(x)|}{t}\right) \lambda(\mathrm{d}x). \end{aligned}$$

The first inequality in (60) is due to Lemma A.5, and the second inequality in (61) comes from the property of the length measure λ, i.e., $\lambda(\Lambda(x)) \leq \lambda(\mathbb{G})$ for all $x \in \mathbb{G}$.

Therefore, we obtain

$$\left\{ t > 0 \mid \int_{\mathbb{G}} \Phi\left(k\frac{|f'(x)|}{t}\right) \lambda(\mathrm{d}x) \leq 1 \right\} \subseteq \left\{ t > 0 \mid \int_{\mathbb{G}} \frac{k}{2}\hat{w}(x)\,\Phi\left(\frac{|f'(x)|}{t}\right) \lambda(\mathrm{d}x) \leq 1 \right\}.$$

Notice that the infimum of a set is smaller than or equal to the infimum of its subset. Consequently, we obtain

$$k\,\|f'\|_{L_\Phi} \geq \|f'\|_{L_\Phi^{k\hat{w}/2}}.$$

Therefore, we have shown that

$$\|f'\|_{L_\Phi} \geq \frac{1}{k} \|f'\|_{L_\Phi^{k\hat{w}/2}}, \quad \text{for every } k \geq 1. \tag{62}$$

By choosing $k = 2$ in (62) and combining with the inequality in (59), we obtain the lower bound as follows:

$$\|f\|_{WL_\Phi^1} \geq \frac{1}{2} \|f'\|_{L_\Phi^{\hat{w}}}.$$

Next, let weight function $w_1(x) := \frac{\lambda(\Lambda(x))}{\lambda(\mathbb{G})}$ for all $x \in \mathbb{G}$, we then derive the upper bound as follows:

$$\begin{aligned}
\|f\|_{WL_\Phi^1} &= \|f\|_{L_\Phi} + \|f'\|_{L_\Phi} \\
&\leq \lambda(\mathbb{G}) \|f'\|_{L_\Phi^{w_1}} + \|f'\|_{L_\Phi} &(63)\\
&\leq \lambda(\mathbb{G}) \|f'\|_{L_\Phi^{\hat{w}}} + \|f'\|_{L_\Phi^{\hat{w}}} &(64)\\
&= (1 + \lambda(\mathbb{G})) \|f'\|_{L_\Phi^{\hat{w}}},
\end{aligned}$$

where the first inequality in (63) is obtained by using Theorem A.3 with $w_0(x) := \lambda(\Lambda(x))$, and the second inequality (64) is obtained by using Lemma A.4.

Hence, the proof is complete. ■

E.2 Proof for Theorem 3.5

Proof. Consider a critic function $f \in WL^1_{\Psi,0}(\mathbb{G}, \lambda)$. Then by Definition 3.1, we have

$$f(x) = f(z_0) + \int_{[z_0,x]} f'(y)\lambda(\mathrm{d}y), \quad \forall x \in \mathbb{G}. \tag{65}$$

Using (65), leveraging the indicator function of the shortest path $[z_0, x]$ (denoted as $\mathbf{1}_{[z_0,x]}$), and notice that $\mu(\mathbb{G}) = 1$, we get

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

Then, applying Fubini's theorem to interchange the order of integration in the above last integral, we obtain

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

Using the definition of $\Lambda(y)$ in Equation (1), we can rewrite it as

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

By exactly the same arguments, we also have

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

Consequently, the regularized generalized Sobolev IPM in Equation (11) can be reformulated as

$$\widehat{\mathcal{GS}}_\Phi(\mu, \nu) = \sup_{f \in \mathcal{B}_\Psi^{\hat{w}}} \left| \int_{\mathbb{G}} f'(x) \big[\mu(\Lambda(x)) - \nu(\Lambda(x))\big]\, \lambda(\mathrm{d}x) \right|, \tag{66}$$

where we recall that $\mathcal{B}_{\Psi}^{\hat{w}} = \left\{ f \in WL^1_{\Psi,0}(\mathbb{G},\lambda) : \|f'\|_{L_{\Psi}^{\hat{w}}} \leq 1 \right\}$ (see Equation (49)).

Observe that, we have on one hand

$$\{f' : f \in \mathcal{B}_{\Psi}^{\hat{w}}\} \subset \{g \in L_{\Psi}(\mathbb{G},\lambda) : \|g\|_{L_{\Psi}^{\hat{w}}} \leq 1\}.$$

On the other hand, for any $g \in L_{\Psi}(\mathbb{G},\lambda)$, we have $g = f'$ with $f(x) \coloneqq \int_{[z_0,x]} g(y)\lambda(\mathrm{d}y) \in WL^1_{\Psi,0}(\mathbb{G},\lambda)$.

Therefore, we conclude that

$$\{f' : f \in \mathcal{B}_{\Psi}^{\hat{w}}\} = \{g \in L_{\Psi}(\mathbb{G},\lambda) : \|g\|_{L_{\Psi}^{\hat{w}}} \leq 1\}. \tag{67}$$

Consequently, if we let $\hat{f}(x) := \frac{\mu(\Lambda(x))-\nu(\Lambda(x))}{\hat{w}(x)}$ for $x \in \mathbb{G}$, then Equation (66) can be recasted as

$$\widehat{\mathcal{GS}}_{\Phi}(\mu,\nu) = \sup_{g \in L_{\Psi}(\mathbb{G},\lambda):\, \|g\|_{L_{\Psi}^{\hat{w}}} \leq 1} \left| \int_{\mathbb{G}} \hat{w}(x)\hat{f}(x)g(x)\lambda(\mathrm{d}x) \right| \tag{68}$$

Additionally, Φ, Ψ are a pair of complementary N-functions. Therefore, we obtain from (68) and [41, §13.20][20] that

$$\widehat{\mathcal{GS}}_{\Phi}(\mu,\nu) = \|\hat{f}\|_{\Phi,\hat{w}}, \tag{69}$$

where $\|\hat{f}\|_{\Phi,\hat{w}}$ is the Musileak-Orlicz norm with weight function $\hat{w}$, define by (see [41, §13.11][21])

$$\|\hat{f}\|_{\Phi,\hat{w}} := \left\{ \int_{\mathbb{G}} \hat{w}(x)\hat{f}(x)g(x)\lambda(\mathrm{d}x) : \int_{\mathbb{G}} \hat{w}(x)\Psi(|g(x)|)\lambda(\mathrm{d}x) \leq 1 \right\}. \tag{70}$$

By applying [25, Theorem 10.5, §10.8][22], we have

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

This together with (69) yields

$$\begin{aligned}
\widehat{\mathcal{GS}}_{\Phi}(\mu,\nu) \quad &= \inf_{k>0} \frac{1}{k} \left(1 + \int_{\mathbb{G}} \hat{w}(x)\, \Phi\Big(k \left| \hat{f}(x) \right| \Big)\, \lambda(\mathrm{d}x) \right) \\
&= \inf_{k>0} \frac{1}{k} \left(1 + \int_{\mathbb{G}} \hat{w}(x)\, \Phi\left(k \left| \frac{\mu(\Lambda(x)) - \nu(\Lambda(x))}{\hat{w}(x)} \right| \right) \lambda(\mathrm{d}x) \right) \\
&= \inf_{k>0} \frac{1}{k} \left(1 + \int_{\mathbb{G}} \hat{w}(x)\, \Phi\left(\frac{k}{\hat{w}(x)} \left| \mu(\Lambda(x)) - \nu(\Lambda(x)) \right| \right) \lambda(\mathrm{d}x) \right).
\end{aligned}$$

This completes the proof of the theorem.

■

E.3 Proof for Theorem 3.6

Proof. We consider the length measure on graph $\mathbb{G}$ for λ. Thus, we have $\lambda(\{x\}) = 0$ for all $x \in \mathbb{G}$.

Consequently, we have

$$\widehat{\mathcal{GS}}_{\Phi}(\mu,\nu) = \inf_{k>0} \frac{1}{k} \left(1 + \sum_{e=\langle u,v\rangle \in E} \int_{(u,v)} \hat{w}(x)\, \Phi\left(\frac{k}{\hat{w}(x)} \left| \mu(\Lambda(x)) - \nu(\Lambda(x)) \right| \right) \lambda(\mathrm{d}x) \right). \tag{71}$$

Additionally, we consider input probability measures μ, ν supported on nodes in V of graph $\mathbb{G}$. Thus, for all edge $e = \langle u, v\rangle \in E$, and any point $x \in (u, v)$, we have

$$\mu(\Lambda(x)) - \nu(\Lambda(x)) = \mu(\Lambda(x) \setminus (u,v)) - \nu(\Lambda(x) \setminus (u,v)). \tag{72}$$

[20] Also see [48, Proposition 10, §3.4].

[21] Also see [48, Definition 2, §3.3], [20, Definition 3.4.5].

[22] See also [48, Theorem 13, §3.3].

Moreover, let us consider edge $e = \langle u, v\rangle \in E$. Then for any $x \in (u, v)$, we have $y \in \mathbb{G} \setminus (u, v)$ belongs to $\Lambda(x)$ if and only if $y \in \gamma_e$ where we recall that $\Lambda(x)$ and γ_e are defined in Equation (1). Thus, we have

$$\Lambda(x) \setminus (u, v) = \gamma_e, \qquad \forall x \in (u, v). \tag{73}$$

Using Equations (72) and (73), we can rewrite Equation (71) as

$$\widehat{\mathcal{GS}}_\Phi(\mu, \nu) = \inf_{k>0} \frac{1}{k} \left(1 + \sum_{e=\langle u,v\rangle \in E} \int_{(u,v)} \hat{w}(x)\, \Phi\!\left(\frac{k}{\hat{w}(x)} \left|\mu(\gamma_e) - \nu(\gamma_e)\right| \right) \lambda(\mathrm{d}x) \right). \tag{74}$$

We next want to compute the integral in (74) for each edge $\langle u, v\rangle \in E$.

For this, recall that $\hat{w}(x) = 1 + \frac{\lambda(\Lambda(x))}{\lambda(\mathbb{G})}, \forall x \in \mathbb{G}$ (see Equation (10)). Without loss of generality, assume that $d_{\mathbb{G}}(z_0, u) \le d_{\mathbb{G}}(z_0, v)$, i.e., among two nodes u, v of the edge e, node v is farther away from the root node z_0 than node u.

Notice that for $x \in (u, v)$, we can write $x = v + t(u - v)$ for $t \in (0, 1)$. With this change of variable, we have

$$\begin{aligned}
&\int_{(u,v)} \left[1 + \frac{\lambda(\Lambda(x))}{\lambda(\mathbb{G})} \right] \Phi\left(\frac{k}{1 + \frac{\lambda(\Lambda(x))}{\lambda(\mathbb{G})}} \left|\mu(\gamma_e) - \nu(\gamma_e)\right| \right) \lambda(\mathrm{d}x) \\
&\qquad = \int_0^1 \left[1 + \frac{\lambda(\Lambda(v + t(u - v)))}{\lambda(\mathbb{G})} \right] \Phi\left(\frac{k}{1 + \frac{\lambda(\Lambda(v+t(u-v)))}{\lambda(\mathbb{G})}} \left|\mu(\gamma_e) - \nu(\gamma_e)\right| \right) w_e \mathrm{d}t.
\end{aligned}$$

Moreover, we have

$$\lambda(\Lambda(v + t(u - v))) = \lambda(\Lambda(v)) + \lambda([v, v + t(u - v)]) = \lambda(\gamma_e) + w_e t.$$

Therefore,

$$\begin{aligned}
&\int_{(u,v)} \left[1 + \frac{\lambda(\Lambda(x))}{\lambda(\mathbb{G})} \right] \Phi\left(\frac{k}{1 + \frac{\lambda(\Lambda(x))}{\lambda(\mathbb{G})}} \left|\mu(\gamma_e) - \nu(\gamma_e)\right| \right) \lambda(\mathrm{d}x) \\
&\qquad = \int_0^1 \left[1 + \frac{\lambda(\gamma_e) + w_e t}{\lambda(\mathbb{G})} \right] \Phi\left(\frac{k}{1 + \frac{\lambda(\gamma_e)+w_e t}{\lambda(\mathbb{G})}} \left|\mu(\gamma_e) - \nu(\gamma_e)\right| \right) w_e \mathrm{d}t \\
&\qquad = \int_0^1 \left[\frac{w_e}{\lambda(\mathbb{G})} t + 1 + \frac{\lambda(\gamma_e)}{\lambda(\mathbb{G})} \right] \Phi\left(\frac{k \left|\mu(\gamma_e) - \nu(\gamma_e)\right|}{\frac{w_e}{\lambda(\mathbb{G})} t + 1 + \frac{\lambda(\gamma_e)}{\lambda(\mathbb{G})}} \right) w_e \mathrm{d}t
\end{aligned}$$

Hence, the proof is complete.

■

E.4 Proof for Proposition 3.7

Proof. For the length measure λ on graph $\mathbb{G}$, we have $\lambda(\{x\}) = 0$ for all $x \in \mathbb{G}$. Consequently, for N-function $\Phi(t) = \frac{(p-1)^{p-1}}{p^p} t^p$ with $1 < p < \infty$, from the closed-form expression of the generalized Sobolev IPM with Musielak regularization in Proposition A.7, we obtain

$$\widehat{\mathcal{GS}}_\Phi(\mu, \nu)^p = \sum_{e=\langle u,v\rangle \in E} \int_{(u,v)} \hat{w}(x)^{1-p} \left|\mu(\Lambda(x)) - \nu(\Lambda(x))\right|^p \lambda(\mathrm{d}x). \tag{75}$$

Additionally, for input measures μ, ν supported on nodes in V of graph $\mathbb{G}$, then for all edge $e = \langle u, v\rangle \in E$, and any point $x \in (u, v)$, we have

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

Therefore, from Equation (75), we obtain

$$\widehat{\mathcal{GS}}_{\Phi}(\mu,\nu)^p = \sum_{e=\langle u,v\rangle\in E} \int_{(u,v)} \hat{w}(x)^{1-p} \left|\mu(\Lambda(x)\setminus(u,v)) - \nu(\Lambda(x)\setminus(u,v))\right|^p \lambda(\mathrm{d}x). \tag{76}$$

Moreover, consider edge $e = \langle u, v\rangle \in E$, for any $x \in (u, v)$, then we have $y \in \mathbb{G} \setminus (u, v)$ belongs to $\Lambda(x)$ if and only if $y \in \gamma_e$.[23] Thus, we have

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

Thus, from Equation (76), we obtain

$$\widehat{\mathcal{GS}}_{\Phi}(\mu,\nu)^p = \sum_{e=\langle u,v\rangle\in E} |\mu(\gamma_e) - \nu(\gamma_e)|^p \int_{(u,v)} \hat{w}(x)^{1-p}\lambda(\mathrm{d}x). \tag{77}$$

From Equation (3), for any x in $\mathbb{G}$, we have

$$\hat{w}(x) = 1 + \frac{\lambda(\Lambda(x))}{\lambda(\mathbb{G})}.$$

Without loss of generality, for any edge $e = \langle u, v\rangle \in E$, assume that $d_{\mathbb{G}}(z_0, u) \le d_{\mathbb{G}}(z_0, v)$, i.e., among two nodes u, v of the edge e, node v is farther away from the root node z_0 than node u.

Observe that for $x \in (u, v)$, then $x = v + t(u - v)$ for $t \in (0, 1)$. Using this change of variable, we obtain

$$\int_{(u,v)} \left[1 + \frac{\lambda(\Lambda(x))}{\lambda(\mathbb{G})}\right]^{1-p} \lambda(\mathrm{d}x) = \int_0^1 \left[1 + \lambda(\mathbb{G})^{-1}\lambda(\Lambda(v + t(u - v)))\right]^{1-p} w_e \mathrm{d}t$$

Additionally, notice that

$$\lambda(\Lambda(v + t(u - v))) = \lambda(\Lambda(v)) + \lambda([v, v + t(u - v)]) = \lambda(\gamma_e) + w_e t.$$

Thus, we obtain

$$\int_{(u,v)} \left[1 + \frac{\lambda(\Lambda(x))}{\lambda(\mathbb{G})}\right]^{1-p} \lambda(\mathrm{d}x) = \int_0^1 \left[1 + \lambda(\mathbb{G})^{-1}\lambda(\gamma_e) + \lambda(\mathbb{G})^{-1} w_e t\right]^{1-p} w_e \mathrm{d}t.$$

Furthermore, the last integral can be computed easily depending on the case $p = 2$ or $p \neq 2$. Consequently, we obtain

$$\int_{(u,v)} \left[1 + \frac{\lambda(\Lambda(x))}{\lambda(\mathbb{G})}\right]^{1-p} \lambda(\mathrm{d}x) = \begin{cases} \lambda(\mathbb{G}) \log\left(1 + \frac{w_e}{\lambda(\mathbb{G})+\lambda(\gamma_e)}\right) & \text{if } p = 2, \\ \frac{(\lambda(\mathbb{G})+\lambda(\gamma_e)+w_e)^{2-p} - (\lambda(\mathbb{G})+\lambda(\gamma_e))^{2-p}}{(2-p)\lambda(\mathbb{G})^{1-p}} & \text{otherwise.} \end{cases}$$

Thus, we have $\int_{(u,v)} \left[1 + \frac{\lambda(\Lambda(x))}{\lambda(\mathbb{G})}\right]^{1-p} \lambda(\mathrm{d}x) = \beta_e$ (see Equation (15)).

Consequently, from Equation (77), we obtain

$$\widehat{\mathcal{GS}}_{\Phi}(\mu,\nu) = \left(\sum_{e\in E} \beta_e |\mu(\gamma_e) - \nu(\gamma_e)|^p\right)^{\frac{1}{p}}. \tag{78}$$

Hence, the proof is completed.

■

[23] See Equation (1) for the definitions of $\Lambda(x)$ and γ_e.

E.5 Proof for Lemma 4.1

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

(i) Nonnegativity. By choosing $f = 0$ in Definition 3.4, we have that $\widehat{\mathcal{GS}}_\Phi(\mu, \nu) \geq 0$ for every pair of probability measures (μ, ν) in $\mathcal{P}(\mathbb{G}) \times \mathcal{P}(\mathbb{G})$. Therefore, the regularized generalized Sobolev IPM $\widehat{\mathcal{GS}}_\Phi$ is nonnegative.

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

$$\int_{\mathbb{G}} f(x)\mu(\mathrm{d}x) - \int_{\mathbb{G}} f(x)\nu(\mathrm{d}x) = 0, \tag{79}$$

for all $f \in WL^1_{\Psi,0}(\mathbb{G}, \lambda)$ satisfying the constraint $\|f'\|_{L^{\hat{w}}_\Psi} \leq 1$.

Let $g \in WL^1_{\Psi,0}(\mathbb{G}, \lambda)$ be any nonconstant function. Then $\alpha := \|g'\|_{L_\Psi} > 0$. Then by choosing $f := \frac{g}{\alpha}$, we have $f \in WL^1_{\Psi,0}(\mathbb{G}, \lambda)$ with $\|f'\|_{L_\Psi} = \|\frac{g'}{\alpha}\|_{L_\Psi} = \frac{1}{\alpha}\|g'\|_{L_\Psi} = 1$ where we use Lemma A.6 for the second equation. Hence, it follows from (79) that

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

which implies that

$$\int_{\mathbb{G}} g(x)\mu(\mathrm{d}x) = \int_{\mathbb{G}} g(x)\nu(\mathrm{d}x). \tag{80}$$

Thus, we have shown that (80) holds true for every nonconstant function $g \in WL^1_{\Psi,0}(\mathbb{G}, \lambda)$. Additionally, Equation (80) is also obviously true for any constant function g. Therefore, we obtain

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

for every $g \in WL^1_{\Psi,0}(\mathbb{G}, \lambda)$, which gives $\mu = \nu$ as desired.

(iii) Symmetry. This property is obvious from Definition 3.4 as the value $\widehat{\mathcal{GS}}_\Phi(\mu, \nu)$ is unchanged when the role of μ and ν is interchanged, i.e.,

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

(iv) Triangle inequality. Let μ, ν, σ be probability measures in $\mathcal{P}(\mathbb{G})$, then for any function $f \in WL^1_{\Psi,0}(\mathbb{G}, \lambda)$ satisfying the constraint $\|f'\|_{L^{\hat{w}}_\Psi} \leq 1$, we have

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

By taking the supremum over f, this implies that

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

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

■

E.6 Proof for Proposition 4.2

Proof. Let Ψ_1 and Ψ_2 be respectively the complementary functions of Φ_1 and Φ_2, consider an arbitrary positive scale $t \in \mathbb{R}_+$, and notice that $\Phi_1 \leq \Phi_2$, then we have:

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

This implies that

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

It follows that $L^{\hat{w}}_{\Psi_1}(\mathbb{G}, \lambda) \subset L^{\hat{w}}_{\Psi_2}(\mathbb{G}, \lambda)$ and $WL^1_{\Psi_1,0}(\mathbb{G}, \lambda) \subset WL^1_{\Psi_2,0}(\mathbb{G}, \lambda)$. Moreover, for any fixed Orlicz function f' and any number $t > 0$, we have

$$\int_{\mathbb{G}} \hat{w}(x)\, \Psi_1\!\left(\frac{|f'(x)|}{t}\right) \lambda(\mathrm{d}x) \geq \int_{\mathbb{G}} \hat{w}(x)\, \Psi_2\!\left(\frac{|f'(x)|}{t}\right) \lambda(\mathrm{d}x).$$

Consequently, we obtain

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

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

$$\|f'\|_{L^{\hat{w}}_{\Psi_1}} \geq \|f'\|_{L^{\hat{w}}_{\Psi_2}}.$$

It follows from this and $WL^1_{\Psi_1,0}(\mathbb{G}, \lambda) \subset WL^1_{\Psi_2,0}(\mathbb{G}, \lambda)$ that

$$\left\{ f \mid f \in WL^1_{\Psi_1,0}(\mathbb{G}, \lambda),\ \|f'\|_{L^{\hat{w}}_{\Psi_1}} \leq 1 \right\} \subset \left\{ f \mid f \in WL^1_{\Psi_2,0}(\mathbb{G}, \lambda),\ \|f'\|_{L^{\hat{w}}_{\Psi_2}} \leq 1 \right\}.$$

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

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

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

■

E.7 Proof for Proposition 4.3

Proof. Given a positive scalar $c > 0$, a pair of complementary N-functions Φ, Ψ, and let $\mathcal{B}^{\hat{w}}_{\Psi,c} := \left\{ f \in WL^1_{\Psi,0}(\mathbb{G}, \lambda) : \|f'\|_{L^{\hat{w}}_{\Psi}} \leq \frac{1}{c} \right\}$. We define the IPM distance w.r.t. $\mathcal{B}^{\hat{w}}_{\Psi,c}$ as follows

$$\widetilde{\mathcal{GS}}_{\Phi,c}(\mu, \nu) := \sup_{f \in \mathcal{B}^{\hat{w}}_{\Psi,c}} \left| \int_{\mathbb{G}} f(x) \mu(\mathrm{d}x) - \int_{\mathbb{G}} f(y) \nu(\mathrm{d}y) \right|. \tag{81}$$

By exploiting the graph structure for the IPM objective function and applying a similar reasoning as in the proof of identity (66) in §E.2, we can rewrite Equation (81) as

$$\widetilde{\mathcal{GS}}_{\Phi,c}(\mu, \nu) = \sup_{f \in \mathcal{B}^{\hat{w}}_{\Psi,c}} \left| \int_{\mathbb{G}} f'(x) \big[\mu(\Lambda(x)) - \nu(\Lambda(x))\big]\, \lambda(\mathrm{d}x) \right|. \tag{82}$$

Additionally, by using a similar reasoning as in the proof of (67) in §E.2, we have

$$\left\{ f' :\ f \in \mathcal{B}^{\hat{w}}_{\Psi,c} \right\} = \left\{ g \in L_{\Psi}(\mathbb{G}, \lambda) :\ \|g\|_{L^{\hat{w}}_{\Psi}} \leq \frac{1}{c} \right\}. \tag{83}$$

Let $\tilde{f}(x) := \frac{\mu(\Lambda(x))-\nu(\Lambda(x))}{c\,\hat{w}(x)}$ for $x \in \mathbb{G}$. Then by using (83), Equation (82) can be recasted as

$$\begin{aligned}
\widetilde{\mathcal{GS}}_{\Phi,c}(\mu,\nu) &= \sup_{g\in L_\Psi(\mathbb{G},\lambda):\,\|g\|_{L^{\hat{w}}_\Psi}\le\frac{1}{c}} \left|\int_{\mathbb{G}} \hat{w}(x)\tilde{f}(x)[c\,g(x)]\lambda(\mathrm{d}x)\right| \\
&= \sup_{\tilde{g}\in L_\Psi(\mathbb{G},\lambda):\,\|\tilde{g}\|_{L^{\hat{w}}_\Psi}\le 1} \left|\int_{\mathbb{G}} \hat{w}(x)\tilde{f}(x)\tilde{g}(x)\lambda(\mathrm{d}x)\right| \\
&= \left\|\tilde{f}\right\|_{\Phi,\hat{w}} \qquad (84)\\
&= \inf_{k>0}\frac{1}{k}\left(1+\int_{\mathbb{G}}\hat{w}(x)\,\Phi\Big(k\left|\tilde{f}(x)\right|\Big)\,\lambda(\mathrm{d}x)\right) \qquad (85)\\
&= \inf_{k>0}\frac{1}{k}\left(1+\int_{\mathbb{G}}\hat{w}(x)\,\Phi\left(k\left|\frac{\mu(\Lambda(x))-\nu(\Lambda(x))}{c\,\hat{w}(x)}\right|\right)\lambda(\mathrm{d}x)\right) \\
&= \frac{1}{c}\inf_{(k/c)>0}\frac{1}{(k/c)}\left(1+\int_{\mathbb{G}}\hat{w}(x)\,\Phi\left(\frac{(k/c)}{\hat{w}(x)}\,|\mu(\Lambda(x))-\nu(\Lambda(x))|\right)\lambda(\mathrm{d}x)\right) \\
&= \frac{1}{c}\,\widehat{\mathcal{GS}}_\Phi(\mu,\nu), \qquad (86)
\end{aligned}$$

where the third equality in (84) following from [41, §13.20] and Equation (70), the fourth equality in (85) following from [25, Theorem 10.5, §10.8], and the last equality in (86) following from Theorem 3.5.

Additionally, notice that from Theorem 3.2, we have

$$c_1\,\|f'\|_{L^{\hat{w}}_\Psi} \le \|f\|_{WL^1_\Psi} \le c_2\,\|f'\|_{L^{\hat{w}}_\Psi},$$

where $c_1 = 1/2$, and $c_2 = 1+\lambda(\mathbb{G})$.

This implies that

$$\mathcal{B}^{\hat{w}}_{\Psi,c_1} \supseteq \mathcal{B}_\Psi \supseteq \mathcal{B}^{\hat{w}}_{\Psi,c_2}.$$

Therefore, for probability measures $\mu,\nu \in \mathcal{P}(\mathbb{G})$, we have

$$\widetilde{\mathcal{GS}}_{\Phi,c_1}(\mu,\nu) \ge \mathcal{GS}_\Phi(\mu,\nu) \ge \widetilde{\mathcal{GS}}_{\Phi,c_2}(\mu,\nu). \qquad (87)$$

It follows from Equations (86) and (87) that

$$\frac{1}{c_1}\,\widehat{\mathcal{GS}}_\Phi(\mu,\nu) \ge \mathcal{GS}_\Phi(\mu,\nu) \ge \frac{1}{c_2}\,\widehat{\mathcal{GS}}_\Phi(\mu,\nu).$$

Consequently, we obtain

$$c_1\,\mathcal{GS}_\Phi(\mu,\nu) \le \widehat{\mathcal{GS}}_\Phi(\mu,\nu) \le c_2\,\mathcal{GS}_\Phi(\mu,\nu).$$

Hence, we have

$$\frac{1}{2}\,\mathcal{GS}_\Phi(\mu,\nu) \le \widehat{\mathcal{GS}}_\Phi(\mu,\nu) \le (1+\lambda(\mathbb{G}))\,\mathcal{GS}_\Phi(\mu,\nu).$$

The proof is complete.

■

E.8 Proof for Proposition 4.4

Proof. Following Equation (68) in the proof of Theorem 3.5 in §E.2, we have

$$\widehat{\mathcal{GS}}_\Phi(\mu,\nu) = \sup_{g\in L_\Psi(\mathbb{G},\lambda):\,\|g\|_{L^{\hat{w}}_\Psi}\le 1}\left|\int_{\mathbb{G}}\hat{w}(x)\hat{f}(x)g(x)\lambda(\mathrm{d}x)\right|, \qquad (88)$$

where $\hat{f}(x) = \frac{\mu(\Lambda(x))-\nu(\Lambda(x))}{\hat{w}(x)}$ for $x \in \mathbb{G}$.

Additionally, following [32, Remark A.1], for $\Phi(t) = \frac{(p-1)^{p-1}}{p^p} t^p$ with $1 < p < \infty$, we have its complementary function $\Psi(t) = t^q$ where q is the conjugate of p, i.e., $1/q + 1/p = 1$.

Consequently, we have $L_\Psi = L^q$, and $L_\Psi^{\hat{w}} = L_{\hat{w}}^q$, where $L_{\hat{w}}^q$ is the weighted L^q space with weight function $\hat{w}$. Therefore, we can rewrite Equation (88) as follows:

$$\begin{aligned}
\widehat{\mathcal{GS}}_\Phi(\mu, \nu) &= \sup_{g \in L^q(\mathbb{G}, \lambda):\, \|g\|_{L_{\hat{w}}^q} \le 1} \left| \int_{\mathbb{G}} \hat{w}(x) \hat{f}(x) g(x) \lambda(\mathrm{d}x) \right| \\
&= \left\| \hat{f} \right\|_{L_{\hat{w}}^p} \\
&= \left[\int_{\mathbb{G}} \hat{w}(x) \left| \hat{f}(x) \right|^p \lambda(\mathrm{d}x) \right]^{\frac{1}{p}} \\
&= \left[\int_{\mathbb{G}} \hat{w}(x)^{1-p} \left| \mu(\Lambda(x)) - \nu(\Lambda(x)) \right|^p \lambda(\mathrm{d}x) \right]^{\frac{1}{p}}. \qquad (89)
\end{aligned}$$

Moreover, following [33, Theorem 3.4], the closed-form of the regularized p-order Sobolev IPM is as follows:

$$\hat{\mathcal{S}}_p(\mu, \nu) = \left[\int_{\mathbb{G}} \hat{w}_{\hat{\mathcal{S}}}(x)^{1-p} \left| \mu(\Lambda(x)) - \nu(\Lambda(x)) \right|^p \lambda(\mathrm{d}x) \right]^{\frac{1}{p}}, \qquad (90)$$

where the weight function $\hat{w}_{\hat{\mathcal{S}}}(x) = 1 + \lambda(\Lambda(x))$ for all $x \in \mathbb{G}$ [33, Equation (5)].

Recall that $\hat{w}(x) = 1 + \frac{\lambda(\Lambda(x))}{\lambda(\mathbb{G})}$ for all $x \in \mathbb{G}$. For any $x \in \mathbb{G}$, we have

$$\min(1, \lambda(\mathbb{G})^{-1})(1 + \lambda(\Lambda(x))) \le 1 + \frac{\lambda(\Lambda(x))}{\lambda(\mathbb{G})} \le \max(1, \lambda(\mathbb{G})^{-1})(1 + \lambda(\Lambda(x))).$$

Then, for $1 < p < \infty$, we obtain

$$\min(1, \lambda(\mathbb{G})^{-1})^{1-p} \hat{w}_{\mathcal{S}}(x)^{1-p} \ge \hat{w}(x)^{1-p} \ge \max(1, \lambda(\mathbb{G})^{-1})^{1-p} \hat{w}_{\mathcal{S}}(x). \qquad (91)$$

Therefore, from Equations (89), (90), (91), we get

$$\hat{c}_1 \, \hat{\mathcal{S}}_p(\mu, \nu) \le \widehat{\mathcal{GS}}_\Phi(\mu, \nu) \le \hat{c}_2 \, \hat{\mathcal{S}}_p(\mu, \nu),$$

where $\hat{c}_1 = \max(1, \lambda(\mathbb{G})^{-1})^{\frac{1-p}{p}}$ and $\hat{c}_2 = \min(1, \lambda(\mathbb{G})^{-1})^{\frac{1-p}{p}}$.

Hence, the proof is complete. ■

E.9 Proof for Proposition 4.5

Proof. From Equation (10), for any $x \in \mathbb{G}$, we have

$$1 \le \hat{w}(x) \le 2.$$

Then, for any $t > 0$, we have

$$\int_{\mathbb{G}} \Psi\left(\frac{|f'(x)|}{t} \right) \lambda(\mathrm{d}x) \le \int_{\mathbb{G}} \hat{w}(x) \, \Psi\left(\frac{|f'(x)|}{t} \right) \lambda(\mathrm{d}x) \le 2 \int_{\mathbb{G}} \Psi\left(\frac{|f'(x)|}{t} \right) \lambda(\mathrm{d}x).$$

Consequently, we obtain

$$\begin{aligned}
&\left\{ t > 0 \mid \int_{\mathbb{G}} \Psi\left(\frac{|f'(x)|}{t} \right) \lambda(\mathrm{d}x) \le 1 \right\} \qquad (92) \\
&\qquad \supseteq \left\{ t > 0 \mid \int_{\mathbb{G}} \hat{w}(x) \, \Psi\left(\frac{|f'(x)|}{t} \right) \lambda(\mathrm{d}x) \le 1 \right\} \\
&\qquad\qquad \supseteq \left\{ t > 0 \mid 2 \int_{\mathbb{G}} \Psi\left(\frac{|f'(x)|}{t} \right) \lambda(\mathrm{d}x) \le 1 \right\}.
\end{aligned}$$

Additionally, observe that by following Lemma A.5, for any $t > 0$, we have

$$2 \int_{\mathbb{G}} \Psi\left(\frac{|f'(x)|}{t} \right) \lambda(\mathrm{d}x) \le \int_{\mathbb{G}} \Psi\left(2 \, \frac{|f'(x)|}{t} \right) \lambda(\mathrm{d}x)$$

Consequently, we have

$$\left\{t>0 \mid 2\int_{\mathbb{G}}\Psi\left(\frac{|f'(x)|}{t}\right)\lambda(\mathrm{d}x)\leq 1\right\}\supseteq\left\{t>0 \mid \int_{\mathbb{G}}\Psi\left(2\,\frac{|f'(x)|}{t}\right)\lambda(\mathrm{d}x)\leq 1\right\}. \tag{93}$$

Notice that the infimum of a set is smaller than or equal to the infimum of its subset. Consequently, from Equations (92), and (93), we obtain

$$\|f'\|_{L_\Psi}\leq\|f'\|_{L_\Psi^{\hat{w}}}\leq\inf\left\{t>0 \mid \int_{\mathbb{G}}\Psi\left(2\,\frac{|f'(x)|}{t}\right)\lambda(\mathrm{d}x)\leq 1\right\}.$$

Moreover, we also have

$$\begin{aligned}\inf\left\{t>0 \mid \int_{\mathbb{G}}\Psi\left(2\,\frac{|f'(x)|}{t}\right)\lambda(\mathrm{d}x)\leq 1\right\}&=2\inf\left\{(t/2)>0 \mid \int_{\mathbb{G}}\Psi\left(\frac{|f'(x)|}{(t/2)}\right)\lambda(\mathrm{d}x)\leq 1\right\}\\&=2\,\|f'\|_{L_\Psi}.\end{aligned}$$

Thus, we obtain

$$\|f'\|_{L_\Psi}\leq\|f'\|_{L_\Psi^{\hat{w}}}\leq 2\,\|f'\|_{L_\Psi}.$$

With the technical assumption $f(z_0)=0$, then WL^1_Ψ is equal to $WL^1_{\Psi,0}$ (as assumed for GSI-M throughout our work). Thus, we have

$$\left\{f\in WL^1_\Psi,\|f'\|_{L_\Psi}\leq 1\right\}\supseteq\left\{f\in WL^1_{\Psi,0},\|f'\|_{L_\Psi^{\hat{w}}}\leq 1\right\}\supseteq\left\{f\in WL^1_\Psi,\|f'\|_{L_\Psi}\leq 1/2\right\}. \tag{94}$$

Additionally, for a positive scalar $c>0$, a pair of complementary N-function Φ, Ψ, let consider

$$\widetilde{\mathcal{GST}}_{\Phi,c}(\mu,\nu):=\sup_{f\in WL^1_\Psi,\|f'\|_{L_\Psi}\leq 1/c}\left|\int_{\mathbb{G}}f(x)\mu(\mathrm{d}x)-\int_{\mathbb{G}}f(x)\nu(\mathrm{d}x)\right|. \tag{95}$$

Following [32, Equations (12), (13) in §A.1], we can rewrite $\widetilde{\mathcal{GST}}_{\Phi,c}$ as follows:

$$\widetilde{\mathcal{GST}}_{\Phi,c}(\mu,\nu)=\sup_{g\in L_\Psi(\mathbb{G},\lambda):\,\|g\|_{L_\Psi}\leq 1/c}\left|\int_{\mathbb{G}}g(x)h(x)\,\lambda(\mathrm{d}x)\right|,$$

where $h(x):=\mu(\Lambda(x))-\nu(\Lambda(x))$ for all $x\in\mathbb{G}$. Consequently, let $\tilde{g}=cg$, we have

$$\widetilde{\mathcal{GST}}_{\Phi,c}(\mu,\nu)=\sup_{\tilde{g}\in L_\Psi(\mathbb{G},\lambda):\,\|\tilde{g}\|_{L_\Psi}\leq 1}\left|\int_{\mathbb{G}}\tilde{g}(x)\frac{h(x)}{c}\,\lambda(\mathrm{d}x)\right|$$

By applying [48, Proposition 10, pp.81], we obtain

$$\widetilde{\mathcal{GST}}_{\Phi,c}(\mu,\nu)=\left\|\frac{h}{c}\right\|_\Phi \tag{96}$$

where $\left\|\frac{h}{c}\right\|_\Phi$ is the Orlicz norm [48, Definition 2, pp.58], defined as

$$\left\|\frac{h}{c}\right\|_\Phi:=\sup\Big\{\int_{\mathbb{G}}\left|\frac{1}{c}h(x)g(x)\right|\lambda(\mathrm{d}x):\int_{\mathbb{G}}\Psi(|g(x)|)\lambda(\mathrm{d}x)\leq 1\Big\}. \tag{97}$$

Then, by applying [48, Theorem 13, pp.69], we have

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

Thus, together with (96), we obtain

$$\widetilde{\mathcal{GST}}_{\Phi,c}(\mu,\nu)=\frac{1}{c}\inf_{(k/c)>0}\frac{1}{(k/c)}\left(1+\int_{\mathbb{G}}\Phi((k/c)\,|h(x)|)\,\lambda(\mathrm{d}x)\right)=\frac{1}{c}\,\mathcal{GST}_\Phi(\mu,\nu). \tag{98}$$

Since the supremum of a set is larger than or equal to the supremum of its subset, then for $\mu,\nu\in\mathcal{P}(\mathbb{G})$, from Equations (94), (95), (98), and consider $c=2$, then we obtain

$$\mathcal{GST}_\Phi(\mu,\nu)\geq\widehat{\mathcal{GS}}(\mu,\nu)\geq\frac{1}{2}\,\mathcal{GST}_\Phi(\mu,\nu).$$

Hence, the proof is complete. ■

E.10 Proof for Proposition 4.6

Proof. For the limit case $\Phi(t) := t$, and graph $\mathbb{G}$ is a tree, then following [32, Remark 4.5 and Proposition 4.6], we have

$$\mathcal{GST}_{\Phi}(\mu,\nu) = \mathcal{OW}(\mu,\nu). \tag{99}$$

Additionally, following Proposition 4.5, we have

$$\frac{1}{2}\,\mathcal{GST}_{\Phi}(\mu,\nu) \le \widehat{\mathcal{GS}}_{\Phi}(\mu,\nu) \le \mathcal{GST}_{\Phi}(\mu,\nu). \tag{100}$$

Thus, from Equations (99), and (100), we obtain

$$\frac{1}{2}\,\mathcal{OW}(\mu,\nu) \le \widehat{\mathcal{GS}}_{\Phi}(\mu,\nu) \le \mathcal{OW}(\mu,\nu).$$

Hence, the proof is complete.

■

E.11 Proof for Lemma A.1

Proof. Consider an N-function Φ, $f \in WL^1_{\Phi,0}(\mathbb{G},\lambda)$, a nonnegative weight function $\hat{w}$, and $t>0$. Then we have

$$\begin{aligned}\int_{\mathbb{G}} \hat{w}(x)\,\Phi\!\left(\frac{1}{t}\,|f(x)|\right)\lambda(\mathrm{d}x) &= \int_{\mathbb{G}} \hat{w}(x)\,\Phi\!\left(\frac{1}{t}\left|\int_{\mathbb{G}} \mathbf{1}_{[z_0,x]}(y) f'(y)\lambda(\mathrm{d}y)\right|\right)\lambda(\mathrm{d}x)\\ &= \int_{\mathbb{G}} \hat{w}(x)\,\Phi\!\left(\frac{\lambda(\mathbb{G})}{t}\left|\frac{1}{\lambda(\mathbb{G})}\int_{\mathbb{G}} \mathbf{1}_{[z_0,x]}(y) f'(y)\lambda(\mathrm{d}y)\right|\right)\lambda(\mathrm{d}x). \end{aligned}\tag{101}$$

For $\alpha>0$, let $\bar{\Phi}_\alpha(s) := \Phi(\alpha s)$ for $s>0$. Then $\bar{\Phi}_\alpha$ is also a convex, non-decreasing function. Using this function, we can rewrite Equation (101) as follows:

$$\int_{\mathbb{G}} \hat{w}(x)\,\Phi\!\left(\frac{1}{t}\,|f(x)|\right)\lambda(\mathrm{d}x) = \int_{\mathbb{G}} \hat{w}(x)\,\bar{\Phi}_{\frac{\lambda(\mathbb{G})}{t}}\left(\left|\frac{1}{\lambda(\mathbb{G})}\int_{\mathbb{G}} \mathbf{1}_{[z_0,x]}(y) f'(y)\lambda(\mathrm{d}y)\right|\right)\lambda(\mathrm{d}x).$$

Then, by applying Jensen's inequality, we have

$$\begin{aligned}\int_{\mathbb{G}} \hat{w}(x)\,\Phi\!\left(\frac{1}{t}\,|f(x)|\right)\lambda(\mathrm{d}x) &\le \frac{1}{\lambda(\mathbb{G})}\int_{\mathbb{G}}\int_{\mathbb{G}} \hat{w}(x)\,\bar{\Phi}_{\frac{\lambda(\mathbb{G})}{t}}\left(\left|\mathbf{1}_{[z_0,x]}(y) f'(y)\right|\right)\lambda(\mathrm{d}y)\lambda(\mathrm{d}x)\\ &= \frac{1}{\lambda(\mathbb{G})}\int_{\mathbb{G}}\int_{\mathbb{G}} \hat{w}(x)\,\Phi\!\left(\frac{\lambda(\mathbb{G})}{t}\left|\mathbf{1}_{[z_0,x]}(y) f'(y)\right|\right)\lambda(\mathrm{d}y)\lambda(\mathrm{d}x).\end{aligned}$$

Due to $\Phi(0)=0$, we can rewrite the above expression as

$$\begin{aligned}\int_{\mathbb{G}} \hat{w}(x)\,\Phi\!\left(\frac{1}{t}\,|f(x)|\right)\lambda(\mathrm{d}x) &\le \frac{1}{\lambda(\mathbb{G})}\int_{\mathbb{G}} \hat{w}(x)\left(\int_{[z_0,x]} \Phi\left(\frac{\lambda(\mathbb{G})}{t}\,|f'(y)|\right)\lambda(\mathrm{d}y)\right)\lambda(\mathrm{d}x)\\ &= \frac{1}{\lambda(\mathbb{G})}\int_{\mathbb{G}} \hat{w}(x)\left(\int_{\mathbb{G}} \mathbf{1}_{[z_0,x]}(y)\,\Phi\left(\frac{\lambda(\mathbb{G})}{t}\,|f'(y)|\right)\lambda(\mathrm{d}y)\right)\lambda(\mathrm{d}x).\end{aligned}$$

Applying Fubini's theorem, we can interchange the order of the integrations. As a consequence, we obtain

$$\begin{aligned}\int_{\mathbb{G}} \hat{w}(x)\,\Phi\!\left(\frac{1}{t}\,|f(x)|\right)\lambda(\mathrm{d}x) &\le \frac{1}{\lambda(\mathbb{G})}\int_{\mathbb{G}}\left(\int_{\mathbb{G}} \hat{w}(x)\mathbf{1}_{[z_0,x]}(y)\lambda(\mathrm{d}x)\right)\Phi\left(\frac{\lambda(\mathbb{G})}{t}\,|f'(y)|\right)\lambda(\mathrm{d}y)\\ &= \frac{1}{\lambda(\mathbb{G})}\int_{\mathbb{G}} \Phi\left(\frac{\lambda(\mathbb{G})}{t}\,|f'(y)|\right)\bar{\lambda}_{\hat{w}}(\Lambda(y))\lambda(\mathrm{d}y).\end{aligned}$$

Hence, the proof is complete. ■

E.12 Proof for Lemma A.2

Proof. The conclusion of the lemma is trivial if $\|f\|_{L_\Phi} = 0$. Thus we only need to consider the case $\|f\|_{L_\Phi} > 0$ in the following proof. As a consequence and due to $f(z_0) = 0$, there must be $x \in \mathbb{G}$ such that $|f'(x)| > 0$, i.e. the set $\{x \in \mathbb{G} : |f'(x)| > 0\}$ is nonempty.

We will prove the result by contradiction. Specifically, suppose by contradiction that $0 < t < \frac{\|f\|_{L_\Phi}}{\lambda(\mathbb{G})}$. Then as $|f'(x)| = 0$ implies $\Phi\Big(\frac{|f'(x)|}{t}\Big) = 0$ and as Φ is strictly increasing, we have

$$\begin{aligned}\int_{\mathbb{G}} w_0(x)\,\Phi\left(\frac{|f'(x)|}{t}\right)\lambda(\mathrm{d}x) \qquad &= \int_{\mathbb{G},|f'|>0} w_0(x)\,\Phi\left(\frac{|f'(x)|}{t}\right)\lambda(\mathrm{d}x)\\ &> \int_{\mathbb{G},|f'|>0} w_0(x)\,\Phi\left(\frac{\lambda(\mathbb{G})}{\|f\|_{L_\Phi}}|f'(x)|\right)\lambda(\mathrm{d}x)\\ &= \int_{\mathbb{G}} w_0(x)\,\Phi\left(\frac{\lambda(\mathbb{G})}{\|f\|_{L_\Phi}}|f'(x)|\right)\lambda(\mathrm{d}x). \qquad (102)\end{aligned}$$

On the other hand, since

$$\|f\|_{L_\Phi} = \inf\left\{t > 0 \mid \int_{\mathbb{G}} \Phi\Big(\frac{|f(x)|}{t}\Big)\,\lambda(\mathrm{d}x) \le 1\right\}$$

we deduce that

$$\int_{\mathbb{G}} \Phi\left(\frac{|f(x)|}{\|f\|_{L_\Phi}}\right)\lambda(\mathrm{d}x) = 1.$$

Applying Lemma A.1 for $\hat{w} = 1$ and $t = \|f\|_{L_\Phi}$ to the above left hand side, we have

$$\begin{aligned}1 = \int_{\mathbb{G}} \Phi\left(\frac{|f(x)|}{\|f\|_{L_\Phi}}\right)\lambda(\mathrm{d}x) \qquad &\le \frac{1}{\lambda(\mathbb{G})}\int_{\mathbb{G}} \Phi\left(\frac{\lambda(\mathbb{G})}{\|f\|_{L_\Phi}}\,|f'(y)|\right)\lambda(\Lambda(y))\lambda(\mathrm{d}y)\\ &\le \frac{1}{\lambda(\mathbb{G})}\int_{\mathbb{G}} \Phi\left(\frac{\lambda(\mathbb{G})}{\|f\|_{L_\Phi}}\,|f'(y)|\right) w_0(y)\lambda(\mathrm{d}y).\end{aligned}$$

Thus, we obtain

$$\int_{\mathbb{G}} \Phi\left(\frac{\lambda(\mathbb{G})}{\|f\|_{L_\Phi}}\,|f'(x)|\right) w_0(x)\lambda(\mathrm{d}y) \ge \lambda(\mathbb{G}).$$

This together with the inequality in (102) yields

$$\int_{\mathbb{G}} w_0(x)\,\Phi\left(\frac{|f'(x)|}{t}\right)\lambda(\mathrm{d}x) > \lambda(\mathbb{G}),$$

which contradicts the assumption. Hence, the proof is complete. ■

E.13 Proof for Theorem A.3

Proof. We consider the Musielak norm with weight function $\frac{w_0}{\lambda(\mathbb{G})}$ for gradient function f'

$$\|f'\|_{L_\Phi^{w_0/\lambda(\mathbb{G})}} = \inf\left\{t > 0 \mid \int_{\mathbb{G}} \frac{w_0(x)}{\lambda(\mathbb{G})}\,\Phi\left(\frac{|f'(x)|}{t}\right)\lambda(\mathrm{d}x) \le 1\right\}.$$

In particular,

$$\int_{\mathbb{G}} w_0(x)\,\Phi\left(\frac{|f'(x)|}{\|f'\|_{L_\Phi^{w_0/\lambda(\mathbb{G})}}}\right)\lambda(\mathrm{d}x) \le \lambda(\mathbb{G}).$$

Therefore, by applying Lemma A.2 with $t := \|f\|_{L_\Phi^{w_0/\lambda(\mathbb{G})}}$, we conclude that

$$\|f\|_{L_\Phi} \leq \lambda(\mathbb{G})\, \|f'\|_{L_\Phi^{w_0/\lambda(\mathbb{G})}}.$$

Hence, the proof is complete. ■

E.14 Proof for Lemma A.4

Proof. For any N-function Φ, any Borel measurable function f on $\mathbb{G}$, and scalar $t > 0$, we have

$$\int_{\mathbb{G}} w_1(x)\, \Phi\bigg(\frac{|f(x)|}{t}\bigg)\, \lambda(\mathrm{d}x) \geq \int_{\mathbb{G}} w_2(x)\, \Phi\bigg(\frac{|f(x)|}{t}\bigg)\, \lambda(\mathrm{d}x).$$

Therefore, we obtain

$$\bigg\{t > 0 \mid \int_{\mathbb{G}} w_1(x)\, \Phi\bigg(\frac{|f(x)|}{t}\bigg)\, \lambda(\mathrm{d}x) \leq 1\bigg\} \subseteq \bigg\{t > 0 \mid \int_{\mathbb{G}} w_2(x)\, \Phi\bigg(\frac{|f(x)|}{t}\bigg)\, \lambda(\mathrm{d}x) \leq 1\bigg\}.$$

Notice that the infimum of a set is smaller than or equal to the infimum of its subset. As a consequence, we obtain

$$\|f\|_{L_\Phi^{w_1}} \geq \|f\|_{L_\Phi^{w_2}}.$$

The proof is complete.

■

E.15 Proof for Lemma A.5

Proof. Notice that Φ is a convex function and $\Phi(0) = 0$. Therefore, for any $0 < s \leq t$, we have

$$\begin{aligned}\Phi(s) \quad &= \Phi\Big(\frac{s}{t}t + \Big(1 - \frac{s}{t}\Big)\, 0\Big) \\ &\leq \frac{s}{t}\Phi(t) + \Big(1 - \frac{s}{t}\Big)\, \Phi(0) = \frac{s}{t}\Phi(t),\end{aligned}$$

which yields $\frac{\Phi(s)}{s} \leq \frac{\Phi(t)}{t}$. Thus, the function $t \mapsto \frac{\Phi(t)}{t}$ is nondecreasing on $(0, +\infty)$.

Consequently, since $k \geq 1$, we get

$$\frac{\Phi(t)}{t} \leq \frac{\Phi(kt)}{kt}.$$

That is, $k\Phi(t) \leq \Phi(kt)$. The proof is complete. ■

E.16 Proof for Lemma A.6

Following the definition of Musielak norm (Equation (3)), we have

$$\begin{aligned}\|kf\|_{L_\Phi^{\hat{w}}} \quad &= \inf\bigg\{t > 0 \mid \int_{\mathbb{G}} \hat{w}(x)\, \Phi\bigg(\frac{|kf(x)|}{t}\bigg)\, \lambda(\mathrm{d}x) \leq 1\bigg\} \\ &= k \inf\bigg\{\frac{t}{k} > 0 \mid \int_{\mathbb{G}} \hat{w}(x)\, \Phi\bigg(\frac{|f(x)|}{t/k}\bigg)\, \lambda(\mathrm{d}x) \leq 1\bigg\} \\ &= k\, \|f\|_{L_\Phi^{\hat{w}}}.\end{aligned}$$

The proof is complete.

E.17 Proof for Proposition A.7

Proof. Following [32, Remark A.1], for $\Phi(t) = \frac{(p-1)^{p-1}}{p^p} t^p$ with $1 < p < \infty$, we have its complementary N-function $\Psi(t) = t^q$ where q is the conjugate of p, i.e., $1/q + 1/p = 1$. Consequently, we have $L_\Psi = L^q$, and $L_\Psi^{\hat{w}} = L_{\hat{w}}^q$, where $L_{\hat{w}}^q$ is the weighted L^q space with weight function $\hat{w}$.

Additionally, let $\hat{f}(x) = \frac{\mu(\Lambda(x))-\nu(\Lambda(x))}{\hat{w}(x)}$ for $x \in \mathbb{G}$, by following Equation (68) in the proof of Theorem 3.5 in §E.2, we have

$$\begin{aligned}
\widehat{\mathcal{GS}}_{\Phi}(\mu,\nu) &= \sup_{g\in L_{\Psi}(\mathbb{G},\lambda):\, \|g\|_{L^{\hat{w}}_{\Psi}}\leq 1} \left| \int_{\mathbb{G}} \hat{w}(x)\hat{f}(x)g(x)\lambda(\mathrm{d}x) \right| \\
&= \sup_{g\in L^{q}(\mathbb{G},\lambda):\, \|g\|_{L^{q}_{\hat{w}}}\leq 1} \left| \int_{\mathbb{G}} \hat{w}(x)\hat{f}(x)g(x)\lambda(\mathrm{d}x) \right| \\
&= \left\| \hat{f} \right\|_{L^{p}_{\hat{w}}} \\
&= \left[\int_{\mathbb{G}} \hat{w}(x) \left| \hat{f}(x) \right|^{p} \lambda(\mathrm{d}x) \right]^{\frac{1}{p}} \\
&= \left[\int_{\mathbb{G}} \hat{w}(x)^{1-p} \left| \mu(\Lambda(x)) - \nu(\Lambda(x)) \right|^{p} \lambda(\mathrm{d}x) \right]^{\frac{1}{p}}.
\end{aligned}$$

Hence, the proof is complete.

■

E.18 Proof for Proposition A.8

Proof. Following Proposition 4.4, we have

$$\max(1, \lambda(\mathbb{G})^{-1})^{\frac{1-p}{p}}\, \hat{\mathcal{S}}_{p}(\mu,\nu) \leq \widehat{\mathcal{GS}}_{\Phi}(\mu,\nu) \leq \min(1, \lambda(\mathbb{G})^{-1})^{\frac{1-p}{p}}\, \hat{\mathcal{S}}_{p}(\mu,\nu). \tag{103}$$

Additionally, following [33, Theorem 4.2], we have

$$\left[\frac{\min(1, \lambda(\mathbb{G})^{p-1})}{1+\lambda(\mathbb{G})^{p}} \right]^{\frac{1}{p}} \mathcal{S}_{p}(\mu,\nu) \leq \hat{\mathcal{S}}_{p}(\mu,\nu) \leq \left[\max(1, \lambda(\mathbb{G})^{p-1}) \right]^{\frac{1}{p}} \mathcal{S}_{p}(\mu,\nu) \tag{104}$$

Therefore, from Equations (103), (104), we obtain

$$c_1\, \mathcal{S}_{p}(\mu,\nu) \leq \widehat{\mathcal{GS}}_{\Phi}(\mu,\nu) \leq c_2\, \mathcal{S}_{p}(\mu,\nu),$$

where $c_1 = \left[\frac{\min(1,\lambda(\mathbb{G})^{p-1}) \max(1,\lambda(\mathbb{G})^{-1})^{1-p}}{1+\lambda(\mathbb{G})^{p}} \right]^{\frac{1}{p}}$; $c_2 = \left[\min(1, \lambda(\mathbb{G})^{-1})^{1-p} \max(1, \lambda(\mathbb{G})^{p-1}) \right]^{\frac{1}{p}}$.

Hence, the proof is complete.

■

E.19 Proof for Proposition A.9

Proof. For N-function $\Phi(t) = \frac{(p-1)^{p-1}}{p^{p}} t^{p}$ with $1 < p < \infty$, following [32, Proposition 4.4], we have

$$\mathcal{GST}_{\Phi}(\mu,\nu) = \mathcal{ST}_{p}(\mu,\nu). \tag{105}$$

Additionally, following Proposition 4.5, we have

$$\frac{1}{2}\, \mathcal{GST}_{\Phi}(\mu,\nu) \leq \widehat{\mathcal{GS}}_{\Phi}(\mu,\nu) \leq \mathcal{GST}_{\Phi}(\mu,\nu). \tag{106}$$

Thus, from Equations (105), and (106), we obtain

$$\frac{1}{2}\, \mathcal{ST}_{p}(\mu,\nu) \leq \widehat{\mathcal{GS}}_{\Phi}(\mu,\nu) \leq \mathcal{ST}_{p}(\mu,\nu).$$

Hence, the proof is complete. ■

E.20 Proof for Proposition A.10

Proof. For the limit case $\Phi(t) := t$, and notice that $\lim_{p\to 1^+} \frac{(p-1)^{p-1}}{p^p} t^p = t$ [32, §A.7], then following [32, Proposition 4.4], by taking the limit $p \to 1^+$, the closed-form of generalized Sobolev transport (GST) is equal to the 1-order Sobolev transport (ST), i.e.,

$$\mathcal{GST}_{\Phi}(\mu, \nu) = \mathcal{ST}_1(\mu, \nu). \tag{107}$$

Additionally, suppose that graph $\mathbb{G}$ is a tree, then by following Le et al. [30], the 1-order ST is in turn equal to the 1-order Wasserstein.

$$\mathcal{ST}_1(\mu, \nu) = \mathcal{W}_1(\mu, \nu). \tag{108}$$

Therefore, from Equations (107), and (108), we obtain

$$\mathcal{GST}_{\Phi}(\mu, \nu) = \mathcal{W}_1(\mu, \nu). \tag{109}$$

Moreover, following Proposition 4.5, we have

$$\frac{1}{2}\,\mathcal{GST}_{\Phi}(\mu, \nu) \leq \widehat{\mathcal{GS}}_{\Phi}(\mu, \nu) \leq \mathcal{GST}_{\Phi}(\mu, \nu). \tag{110}$$

Thus, from Equations (109), and (110), we obtain

$$\frac{1}{2}\,\mathcal{W}_1(\mu, \nu) \leq \widehat{\mathcal{GS}}_{\Phi}(\mu, \nu) \leq \mathcal{W}_1(\mu, \nu).$$

Hence, the proof is complete.

■

E.21 Compute A_1 for $\mathcal{A}_{\Phi_1}$ in $\widehat{\mathcal{GS}}_{\Phi_1}$ (§A.3)

For $a > 0, b > 0$, we consider the term

$$A_1 := \beta \int_0^1 (at + b)\, \Phi_1\left(\frac{\alpha}{at + b}\right) \mathrm{d}t.$$

Set $u = at + b$, so $\mathrm{d}u = a\mathrm{d}t$ and $u \in [b, a + b]$, then we can rewrite

$$A_1 = \frac{\beta}{a} \int_b^{a+b} [u \exp(\alpha/u) - \alpha - u]\, \mathrm{d}u.$$

Next, we want to compute

$$B_{\Phi_1} = \int_b^{a+b} u \exp(\alpha/u) \mathrm{d}u.$$

First, we use the substitution by setting $v = \alpha/u$. Then $u = \alpha/v$ and $\mathrm{d}u = -\alpha v^{-2}\mathrm{d}v$.

The integrand becomes

$$\int u \exp(\alpha/u)\mathrm{d}u = \int \frac{\alpha}{v} \exp(v)[-\alpha v^{-2}\mathrm{d}v] = -\alpha^2 \int \exp(v) v^{-3}\mathrm{d}v. \tag{111}$$

Second, we use two integrations by parts to compute:

$$B_1 := \int \exp(v) v^{-3}\mathrm{d}v. \tag{112}$$

Applying the integration by part, we get

$$B_1 = \frac{1}{2} \int \exp(v) v^{-2}\mathrm{d}v - \frac{1}{2} \exp(v) v^{-2}. \tag{113}$$

Applying the integration by part for the integral in the right hand side, we obtain

$$\int \exp(v)v^{-2}\mathrm{d}v = \int \exp(v)v^{-1}\mathrm{d}v - \exp(v)v^{-1}. \tag{114}$$

Then, from Equations (112), (113), and (114), we obtain

$$\begin{aligned} B_1 &= \frac{1}{2}\left(Ei(v) - \exp(v)v^{-1}\right) - \frac{1}{2}\exp(v)v^{-2} + C &(115)\\ &= \frac{1}{2}Ei(v) - \frac{1}{2}\exp(v)v^{-1} - \frac{1}{2}\exp(v)v^{-2} + C. &(116)\end{aligned}$$

Thus, from Equations (111), and (115), and recall that $v = \alpha/u$, we have

$$\begin{aligned} \int u\exp(\alpha/u)\mathrm{d}u &= -\alpha^2 B_1\\ &= -\frac{\alpha^2}{2}Ei(\alpha/u) + \frac{\alpha^2}{2}\frac{\exp(\alpha/u)}{\alpha/u} + \frac{\alpha^2}{2}\frac{\exp(\alpha/u)}{(\alpha/u)^2} + C\\ &= \frac{1}{2}\left[u^2\exp(\alpha/u) + \alpha u\exp(\alpha/u) - \alpha^2 Ei(\alpha/u)\right] + C.\end{aligned}$$

Thus, an antiderivative for the integrand of A_1 is

$$\begin{aligned} F_{A_1} &:= \frac{\beta}{a}\int \left[u\exp(\alpha/u) - \alpha - u\right]\mathrm{d}u\\ &= \frac{\beta}{2a}\left(u^2\exp(\alpha/u) + \alpha u\exp(\alpha/u) - \alpha^2 Ei(\alpha/u) - 2\alpha u - u^2\right) + C.\end{aligned}$$

Hence, we have

$$A_1 = F_{A_1}(a+b) - F_{A_1}(b).$$

Or, we obtain

$$\begin{aligned} A_1 = \frac{\beta}{2a}\Bigg[&-\alpha^2\left[Ei\left(\frac{\alpha}{a+b}\right) - Ei\left(\frac{\alpha}{b}\right)\right] + \alpha\left[(a+b)\exp\left(\frac{\alpha}{a+b}\right) - b\exp\left(\frac{\alpha}{b}\right)\right]\\ &+\left[(a+b)^2\exp\left(\frac{\alpha}{a+b}\right) - b^2\exp\left(\frac{\alpha}{b}\right)\right] - 2\alpha a - \left[(a+b)^2 - b^2\right]\Bigg].\end{aligned}$$

Therefore, we have

$$\mathcal{A}_{\Phi_1} = A_1,$$

where $\beta = w_e, a = \frac{w_e}{\lambda(\mathbb{G})}, b = 1 + \frac{\lambda(\gamma_e)}{\lambda(\mathbb{G})}, \alpha = k|\bar{h}(e)|$.

Additionally, we next compute the derivative of A_1 w.r.t. k.

For simplicity, let

$$\begin{aligned} S_1(\alpha) := &-\alpha^2\left[Ei\left(\frac{\alpha}{a+b}\right) - Ei\left(\frac{\alpha}{b}\right)\right] + \alpha\left[(a+b)\exp\left(\frac{\alpha}{a+b}\right) - b\exp\left(\frac{\alpha}{b}\right)\right]\\ &+\left[(a+b)^2\exp\left(\frac{\alpha}{a+b}\right) - b^2\exp\left(\frac{\alpha}{b}\right)\right] - 2\alpha a - \left[(a+b)^2 - b^2\right],\end{aligned}$$

then we have

$$A_1 = \frac{\beta}{2a}S_1(\alpha).$$

Thus, we have

$$\frac{\mathrm{d}A_1}{\mathrm{d}k} = \frac{\mathrm{d}A_1}{\mathrm{d}\alpha}\frac{\mathrm{d}\alpha}{\mathrm{d}k}.$$

Notice that $\frac{\mathrm{d}}{\mathrm{d}z}Ei(z) = \frac{\exp(z)}{z}$, and use the chain rule, we have

For the first term of S_1, let $S_{11}(\alpha) := -\alpha^2\left[Ei\left(\frac{\alpha}{a+b}\right) - Ei\left(\frac{\alpha}{b}\right)\right]$, we have

$$\frac{\mathrm{d}S_{11}}{\mathrm{d}\alpha} = -2\alpha\left[Ei\left(\frac{\alpha}{a+b}\right) - Ei\left(\frac{\alpha}{b}\right)\right] - \alpha\left[\exp(\alpha/(a+b)) - \exp(\alpha/b)\right].$$

For second term of S_1, let $S_{12}(\alpha) := \alpha\left[(a+b)\exp\left(\frac{\alpha}{a+b}\right) - b\exp\left(\frac{\alpha}{b}\right)\right]$, we have

$$\frac{\mathrm{d}S_{12}}{\mathrm{d}\alpha} = (a+b)\exp\left(\frac{\alpha}{a+b}\right) - b\exp\left(\frac{\alpha}{b}\right) + \alpha\left[\exp(\alpha/(a+b)) - \exp(\alpha/b)\right].$$

For the third term of S_1, denote $S_{13}(\alpha) := (a+b)^2\exp\left(\frac{\alpha}{a+b}\right) - b^2\exp\left(\frac{\alpha}{b}\right)$, we have

$$\frac{\mathrm{d}S_{13}}{\mathrm{d}\alpha} = (a+b)\exp\left(\frac{\alpha}{a+b}\right) - b\exp\left(\frac{\alpha}{b}\right)$$

For the fourth term of S_1, we have $\frac{\mathrm{d}}{\mathrm{d}\alpha}(-2\alpha a) = -2a$.

The last term of S_1 is constant w.r.t. α.

Therefore, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}k}A_1 = \frac{\beta|\bar{h}(e)|}{a}\left\{-\alpha\left[Ei\left(\frac{\alpha}{a+b}\right) - Ei\left(\frac{\alpha}{b}\right)\right] + (a+b)\exp\left(\frac{\alpha}{a+b}\right) - b\exp\left(\frac{\alpha}{b}\right) - a\right\}.$$

Furthermore, we next compute the second-order derivative of A_1 w.r.t. k.

$$\frac{\mathrm{d}^2}{\mathrm{d}k^2}A_1 = \frac{\beta|\bar{h}(e)|}{a}\frac{\mathrm{d}}{\mathrm{d}k}Q_1(\alpha),$$

where we denote

$$Q_1(\alpha) := -\alpha\left[Ei\left(\frac{\alpha}{a+b}\right) - Ei\left(\frac{\alpha}{b}\right)\right] + (a+b)\exp\left(\frac{\alpha}{a+b}\right) - b\exp\left(\frac{\alpha}{b}\right) - a.$$

By using the chain rule, we have

$$\frac{\mathrm{d}^2}{\mathrm{d}k^2}A_1 = \frac{\beta|\bar{h}(e)|^2}{a}\frac{\mathrm{d}}{\mathrm{d}\alpha}Q_1(\alpha),$$

For the first term of Q_1, denote $Q_{11}(\alpha) := -\alpha\left[Ei\left(\frac{\alpha}{a+b}\right) - Ei\left(\frac{\alpha}{b}\right)\right]$, we have

$$\begin{aligned}\frac{\mathrm{d}Q_{11}}{\mathrm{d}\alpha} &= -\left[Ei\left(\frac{\alpha}{a+b}\right) - Ei\left(\frac{\alpha}{b}\right)\right] - \alpha\left[\frac{\exp(\alpha/(a+b))}{\alpha} - \frac{\exp(\alpha/b}{\alpha}\right] \\ &= -\left[Ei\left(\frac{\alpha}{a+b}\right) - Ei\left(\frac{\alpha}{b}\right)\right] - \left[\exp(\alpha/(a+b)) - \exp(\alpha/b\right].\end{aligned}$$

For the second term of Q_1, we have

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\left[(a+b)\exp\left(\frac{\alpha}{a+b}\right)\right] = \exp\left(\frac{\alpha}{a+b}\right).$$

For the third term of Q_1, we have

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\left[-b\exp(\alpha/b)\right] = -\exp(\alpha/b).$$

The last term of Q_1 is a constant w.r.t. α.

Therefore, we obtain

$$\frac{\mathrm{d}^2}{\mathrm{d}k^2}A_1 = -\frac{\beta|\bar{h}(e)|^2}{a}\left[Ei\left(\frac{\alpha}{a+b}\right) - Ei\left(\frac{\alpha}{b}\right)\right].$$

Hence, we complete the detailed derivation for A_1.

E.22 Compute A_2 for $\mathcal{A}_{\Phi_2}$ in $\widehat{\mathcal{GS}}_{\Phi_2}$ (§A.3)

For $a > 0, b > 0$, we consider

$$A_2 := \beta \int_0^1 (at + b)\, \Phi_2\Big(\frac{\alpha}{at+b}\Big)\, \mathrm{d}t.$$

Set $u = at + b$, so $\mathrm{d}u = a\mathrm{d}t$ and $u \in [b, a + b]$, then we can rewrite

$$A_2 = \frac{\beta}{a} \int_b^{a+b} \big[u \exp(\alpha^2/u^2) - u\big]\, \mathrm{d}u.$$

Next, we want to compute

$$B_{\Phi_2} = \int_b^{a+b} u \exp(\alpha^2/u^2)\mathrm{d}u.$$

First, we use the substitution by setting $v = \alpha^2/u^2$, so $v > 0$ on the interval. Then $\mathrm{d}v = -2\alpha^2 u^{-3}\mathrm{d}u$, or $\mathrm{d}u = -\frac{u^3}{2\alpha^2}\mathrm{d}v$.

The integrand becomes

$$\int u \exp(\alpha^2/u^2)\mathrm{d}u = -\int \frac{1}{2\alpha^2} u^4 \exp(v)\mathrm{d}v = -\frac{\alpha^2}{2}\int \exp(v)v^{-2}\mathrm{d}v,$$

since we have $u^4 = \alpha^4 v^{-2}$.

Second, following Equation (114), we have

$$B_2 := \int \exp(v)v^{-2}\mathrm{d}v = Ei(v) - \exp(v)v^{-1} + C.$$

Thus, we obtain

$$\begin{aligned}\int u \exp(\alpha^2/u^2)\mathrm{d}u \qquad &= -\frac{\alpha^2}{2}\left(Ei(v) - \exp(v)v^{-1}\right) + C \\ &= \frac{1}{2}\left[u^2 \exp(\alpha^2/u^2) - \alpha^2 Ei(\alpha^2/u^2)\right] + C,\end{aligned}$$

since $v = \alpha^2/u^2$, and consequently $\exp(v)v^{-1} = (u^2/\alpha^2)\exp(\alpha^2/u^2)$.

Thus, an antiderivative for the integrand of A_2 is

$$\begin{aligned}F_{A_2} \qquad &:= \frac{\beta}{a}\int \big[u \exp(\alpha^2/u^2) - u\big]\, \mathrm{d}u \\ &= \frac{\beta}{a}\left(\frac{1}{2}\left[u^2 \exp(\alpha^2/u^2) - \alpha^2 Ei(\alpha^2/u^2)\right] - \frac{u^2}{2}\right) + C.\end{aligned}$$

Hence, we have

$$A_2 = F_{A_2}(a + b) - F_{A_2}(b).$$

Or, we obtain

$$\begin{aligned}A_2 = \frac{\beta}{2a}\Bigg[&- \alpha^2\left[Ei\Big(\frac{\alpha^2}{(a+b)^2}\Big) - Ei\Big(\frac{\alpha^2}{b^2}\Big)\right] \\ &+ \left[(a+b)^2 \exp\Big(\frac{\alpha^2}{(a+b)^2}\Big) - b^2 \exp\Big(\frac{\alpha^2}{b^2}\Big)\right] - \left[(a+b)^2 - b^2\right]\Bigg].\end{aligned}$$

Therefore, we have

$$\mathcal{A}_{\Phi_2} = A_2,$$

where $\beta = w_e, a = \frac{w_e}{\lambda(\mathbb{G})}, b = 1 + \frac{\lambda(\gamma_e)}{\lambda(\mathbb{G})}, \alpha = k|\bar{h}(e)|$.

Additionally, we next compute the derivative of A_2 w.r.t. k.

For simplicity, let

$$\begin{aligned} S_2(\alpha) := -\alpha^2 \left[Ei\left(\frac{\alpha^2}{(a+b)^2}\right) - Ei\left(\frac{\alpha^2}{b^2}\right)\right] \\ + \left[(a+b)^2 \exp\left(\frac{\alpha^2}{(a+b)^2}\right) - b^2 \exp\left(\frac{\alpha^2}{b^2}\right)\right] - \left[(a+b)^2 - b^2\right], \end{aligned}$$

then we have

$$A_2 = \frac{\beta}{2a} S_2(\alpha).$$

Thus, we have

$$\frac{\mathrm{d}A_2}{\mathrm{d}k} = \frac{\mathrm{d}A_2}{\mathrm{d}\alpha}\frac{\mathrm{d}\alpha}{\mathrm{d}k}.$$

Notice that $\frac{\mathrm{d}}{\mathrm{d}z} Ei(z) = \frac{\exp(z)}{z}$, and use the chain rule, we have

$$\begin{aligned} \frac{\mathrm{d}S_2}{\mathrm{d}\alpha} &= -2\alpha \left[Ei\left(\frac{\alpha^2}{(a+b)^2}\right) - Ei\left(\frac{\alpha^2}{b^2}\right)\right] - \alpha^2 \left[\frac{2}{\alpha}\exp(\alpha^2/(a+b)^2) - \frac{2}{\alpha}\exp(\alpha^2/b^2)\right] \\ &\quad + 2\alpha \left[\exp(\alpha^2/(a+b)^2) - \exp(\alpha^2/b^2)\right] \\ &= -2\alpha \left[Ei\left(\frac{\alpha^2}{(a+b)^2}\right) - Ei\left(\frac{\alpha^2}{b^2}\right)\right]. \end{aligned}$$

Therefore, we obtain

$$\frac{\mathrm{d}A_2}{\mathrm{d}k} = -\frac{\beta\alpha|\bar{h}(e)|}{a}\left[Ei\left(\frac{\alpha^2}{(a+b)^2}\right) - Ei\left(\frac{\alpha^2}{b^2}\right)\right].$$

Furthermore, we next compute the second-order derivative of A_2 w.r.t. k.

$$\frac{\mathrm{d}^2}{\mathrm{d}k^2} A_2 = -\frac{\beta|\bar{h}(e)|}{a}\frac{\mathrm{d}}{\mathrm{d}k} Q_2(\alpha),$$

where we denote

$$Q_2(\alpha) := \alpha \left[Ei\left(\frac{\alpha^2}{(a+b)^2}\right) - Ei\left(\frac{\alpha^2}{b^2}\right)\right].$$

By using the chain rule, we have

$$\begin{aligned} \frac{\mathrm{d}^2}{\mathrm{d}k^2} A_2 &= -\frac{\beta|\bar{h}(e)|}{a}\Bigg\{ |\bar{h}(e)| \left[Ei\left(\frac{\alpha^2}{(a+b)^2}\right) - Ei\left(\frac{\alpha^2}{b^2}\right)\right] \\ &\quad + \alpha \left[\frac{2}{\alpha}\exp(\alpha^2/(a+b)^2) - \frac{2}{\alpha}\exp(\alpha^2/b^2)\right] |\bar{h}(e)| \Bigg\}. \end{aligned}$$

Therefore, we obtain

$$\frac{\mathrm{d}^2}{\mathrm{d}k^2} A_2 = -\frac{\beta|\bar{h}(e)|^2}{a}\left[Ei\left(\frac{\alpha^2}{(a+b)^2}\right) - Ei\left(\frac{\alpha^2}{b^2}\right) + 2\exp(\alpha^2/(a+b)^2) - 2\exp(\alpha^2/b^2)\right].$$

Hence, we complete the detailed derivation for A_2.