Tree-Sliced Orlicz Integral Probability Metric

Tuan A. Hoang*
Center for AI Research
VinUniversity
tuan.ha3@vinuni.edu.vn

Trung-Khang Tran*
School of Computing
National University of Singapore
ttk0106@u.nus.edu

Viet-Hoang Tran*
Department of Mathematics
National University of Singapore
hoang.tranviet@u.nus.edu

Tan M. Nguyen
Department of Mathematics
National University of Singapore
tanmn@nus.edu.sg

Abstract

Tree-sliced distances have emerged as scalable alternatives to Sliced Wasserstein distances by projecting probability measures onto tree metric spaces and exploiting closed-form transport on trees. Existing tree-sliced constructions, however, are largely built around first-order Wasserstein geometry. While this choice yields efficient computation, it fixes the discrepancy to an L^1-type aggregation of subtree mass imbalances and provides limited control over the geometry used in downstream optimization. We propose Tree-Sliced Orlicz IPM (TS-Orlicz), a tree-sliced framework that replaces the tree-level W_1 discrepancy with an Orlicz integral probability metric. Building on the tractable formulation of Orlicz IPMs on trees, TS-Orlicz aggregates Orlicz-induced discrepancies over random tree systems while preserving the scalability of tree-sliced computation. By varying the underlying Orlicz function, the framework recovers L^p-type behavior and also supports non-polynomial, tail-sensitive geometries that emphasize large subtree discrepancies. We show that TS-Orlicz preserves key properties of tree-sliced constructions and admits efficient computation via closed-form special cases and one-dimensional scalar optimization. We further extend the framework to probability measures on hyperspheres. Experiments across Euclidean and spherical settings, including gradient flows, self-supervised learning, and diffusion-based generative modeling, show that Orlicz-induced tree-sliced discrepancies achieve competitive or improved performance over sliced and tree-sliced baselines while maintaining low computational cost.

1 Introduction

Many problems in machine learning can be formulated over the space of probability measures, requiring principled functionals to quantify discrepancies between distributions. Optimal Transport (OT) [68, 46] provides a canonical framework for this purpose, enabling geometry-aware comparisons between measures. As a result, OT has been widely adopted across machine learning [16, 11, 8, 41], statistics [34, 69, 71, 48, 29, 42, 39], computer vision and graphics [40, 24, 56, 52], data valuation [18, 20], and multimodal analysis [32, 44].

Despite its strong geometric properties, OT remains computationally expensive. For discrete measures with n support points, standard solvers scale as $\mathcal{O}(n^3 \log n)$, limiting applicability to large-scale problems. A common approach to mitigate this cost is the Sliced Wasserstein (SW) distance [50, 7],

*These authors contributed equally. Correspondence to: tuan.ha3@vinuni.edu.vn & hoang.tranviet@u.nus.edu.

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

which approximates OT by projecting measures onto one-dimensional subspaces, where transport admits a closed-form solution. Building on this idea, a large body of work has focused on improving SW through enhanced sampling strategies [35, 36], optimized projection directions [10], and alternative projection mechanisms [22, 6].

However, restricting projections to one dimension limits the ability of SW to capture complex geometric and topological structures in high-dimensional data. This limitation has motivated extensions of the slicing paradigm that replace linear projections with more expressive integration domains, including Euclidean subspaces [3, 43, 45], graphs [27], spheres [49, 60, 5], tree metric spaces [28, 25, 63], and hyperbolic spaces [6]. Among these, tree-based approaches have emerged as a particularly effective alternative. The Tree-Sliced Wasserstein (TSW) framework [63] replaces independent projections with structured tree systems, providing a richer representation of the data while retaining computational efficiency through fast OT solvers on trees [17, 28, 27]. Subsequent studies have further developed the TSW framework [59, 61, 66, 65], extended it to mixed-curvature spaces [47] and spherical domains [62], and introduced tree-sliced entropy partial transport for measures with unbalanced masses [64].

While TSW significantly improves upon classical slicing by leveraging structured tree systems, its formulation remains fundamentally tied to the closed-form expression of the tree-Wasserstein distance, which is only available for the first-order case $p = 1$. As a consequence, TSW inherits an L^1-type geometry along the tree, where discrepancies are aggregated linearly across edges. This restriction is limiting, as higher-order transport metrics with $p > 1$ are often preferred in modern learning settings. In particular, p-Wasserstein distances with $p > 1$ enjoy stronger geometric properties, including strict convexity and smoother objective landscapes [67, 54], which are known to facilitate more stable and efficient gradient-based optimization [46]. In contrast, the $p = 1$ case leads to non-smooth objectives and lacks strict convexity, making it less suitable for optimization-driven applications. Consequently, the reliance of TSW on first-order transport limits its ability to capture richer geometric variations and to provide well-behaved gradients for learning tasks.

A natural direction for extending beyond first-order transport geometry is to consider discrepancies induced by general convex functions. Orlicz spaces provide a principled generalization of L^p geometries by replacing power functions with a broader class of convex functions, enabling finer control over growth and regularity [1]. This flexibility has been used in machine learning to design Orlicz-norm losses that interpolate between ℓ_1 and ℓ_2 behavior and adapt to heterogeneous noise [4, 57]. In optimal transport, it leads to Orlicz-Wasserstein distances, where the ground cost is replaced by an Orlicz-induced metric [58, 19], and to Orlicz-regularized transport formulations [31]. These works show that Orlicz-based transport can go beyond polynomial growth and provide improved statistical or geometric behavior [12, 2]. However, this increased expressiveness comes at a high computational cost, as OW typically requires solving a coupled optimization over transport plans and scaling parameters. Recent work suggests that such expressive geometries can be incorporated into structured transport settings while retaining tractability. In particular, [26] introduces a generalized transport formulation on tree-structured domains that incorporates Orlicz-induced geometry while admitting an equivalent univariate optimization, thereby substantially reducing computational complexity. This result highlights that appropriate structural decompositions can bridge the gap between geometric expressiveness and computational efficiency.

Contributions. Building on the limitations of existing Tree-Sliced Wasserstein (TSW) formulations and recent progress in Orlicz IPMs on trees, we introduce a tree-sliced transport framework that moves beyond first-order geometry. Our main contribution is a scalable extension of TSW that incorporates Orlicz-induced transport geometries, enabling more flexible and adaptive aggregation of discrepancies along tree structures. This results in a richer class of tree-sliced metrics/discrepancies with improved regularity and geometric expressiveness, while retaining the computational efficiency of tree-sliced methods. The structure of the paper is as follows:

1. Section 2 reviews the tree-slicing background underlying our method. We recall the closed-form expression of W_1 on tree metric spaces and describe how tree systems and the associated Radon transform project measures from $\mathbb{R}^d$ onto trees to define TSW.
2. Section 3 develops the proposed TS-Orlicz framework. We review Orlicz IPMs on trees, introduce the Euclidean TS-Orlicz distance, establish its metricity under measure-distinguishing projections, Euclidean invariance under sampling-law preservation and projection equivariance, and connection to TSW, and analyze its practical computation via

closed-form special cases and one-dimensional scalar optimization. We then extend the construction to probability measures on the hypersphere, yielding the spherical variant STS-Orlicz.

3. Section 4 provides empirical evaluations on both Euclidean and spherical data, demonstrating the effectiveness and efficiency of the proposed methods.

The Appendix contains supporting material, including background theory, full derivations and proofs, experimental details with supplementary tables and figures, and a notation summary.

2 Background on Tree-Sliced Wasserstein Distance

In this section, we review the tree metric and tree-slicing background underlying our method.

Tree Metric Spaces. Let $\mathcal{T} = (V, E)$ be a tree with vertex set $V \subset \mathbb{R}^d$ and edge set E, where each edge $e \in E$ is assigned a positive length $w_e > 0$. We regard $\mathcal{T}$ as the metric graph consisting of its vertices together with all points along its edges. For any $x, y \in \mathcal{T}$, let $[x, y]$ denote the unique path connecting them. The tree distance $d_{\mathcal{T}}(x, y)$ is defined as the length of this path. Then $(\mathcal{T}, d_{\mathcal{T}})$ is a metric space. This metric structure induces a canonical Borel measure ω on $\mathcal{T}$, given by arc length, so that for any $x, y \in \mathcal{T}$,

$$\omega([x, y]) = d_{\mathcal{T}}(x, y).$$

Given a root $r \in \mathcal{T}$, the subtree rooted at $x \in \mathcal{T}$ is defined as

$$\Lambda(x) := \{y \in \mathcal{T} : x \in [r, y]\}.$$

Wasserstein distance. Let $\mu, \nu \in \mathcal{P}(\mathcal{T})$. Denote by $\Pi(\mu, \nu)$ the set of couplings π on $\mathcal{T} \times \mathcal{T}$ with marginals μ and ν. For $p \geq 1$, the p-Wasserstein distance is defined as

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

For $p = 1$, the distance admits a closed-form expression on trees [28]:

$$\mathrm{W}_1(\mu, \nu) = \int_{\mathcal{T}} |\mu(\Lambda(x)) - \nu(\Lambda(x))| \, \omega(dx).$$

Although the tree formula above gives an efficient way to compare measures that are already supported on a tree, most machine-learning data are represented in $\mathbb{R}^d$. To bridge this gap, we adopt the tree-slicing framework [63], which maps measures in $\mathbb{R}^d$ onto tree metric spaces. This enables the use of efficient tree-based formulations, including closed-form expressions, in high-dimensional settings.

Tree systems. A line in $\mathbb{R}^d$ is represented by a pair $(x, \theta) \in \mathbb{R}^d \times \mathbb{S}^{d-1}$, where x is a reference point and θ is a direction. A collection of k lines is therefore an element of $(\mathbb{R}^d \times \mathbb{S}^{d-1})^k$, denoted by $\mathcal{L} = \{l_i\}_{i=1}^k$, where each line l_i has the parametric form $x_i + t\theta_i$ for $t \in \mathbb{R}$.

A *tree system* augments this collection with a tree structure. We denote a tree system by $\mathcal{T} = \{l_i\}_{i=1}^k$. Its associated ground set is

$$\overline{\mathcal{T}} = \{(x, l_i) \in \mathbb{R}^d \times \mathcal{T} \ : \ x = x_i + t_x\theta_i, \ t_x \in \mathbb{R}\}.$$

This construction induces a continuous tree metric space. We denote by $\mathbb{T}$ the family of such tree systems and equip it with a probability measure σ induced by a sampling procedure.

Radon transform on tree systems. Let $\mathcal{C}(\mathbb{R}^d \times \mathbb{T}, \Delta_{k-1})$ denote the set of continuous splitting maps. For $f \in L^1(\mathbb{R}^d)$ and $\mathcal{T} \in \mathbb{T}$, define

$$\mathcal{R}^\alpha_{\mathcal{T}} f(x, l_i) = \int_{\mathbb{R}^d} f(y) \, \alpha(y, \mathcal{T})_i \, \delta\big(t_x - \langle y - x_i, \theta_i \rangle\big) \, dy.$$

This defines the operator $\mathcal{R}^\alpha : f \mapsto (\mathcal{R}^\alpha_{\mathcal{T}} f)_{\mathcal{T}\in\mathbb{T}}$, which maps functions on $\mathbb{R}^d$ to their representations over tree systems. Additionally, the operator $\mathcal{R}^\alpha$ is injective [63].

Tree-Sliced Wasserstein distance. The Tree-Sliced Wasserstein (TSW) distance compares measures by projecting them onto tree systems and computing transport costs on the induced tree metric spaces.

Formally, let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. For each $\mathcal{T} \in \mathbb{T}$, let $\mathcal{R}^\alpha_\mathcal{T}\mu$ and $\mathcal{R}^\alpha_\mathcal{T}\nu$ denote the induced measures. The TSW distance is defined as

$$\mathrm{TSW}(\mu, \nu) := \int_{\mathbb{T}} \mathrm{W}_1\big(\mathcal{R}^\alpha_\mathcal{T}\mu,\, \mathcal{R}^\alpha_\mathcal{T}\nu\big)\, d\sigma(\mathcal{T}).$$

Here, σ is a distribution over tree systems, and the splitting map α determines how mass is distributed across branches. A key property of this framework is that W_1 on each tree admits a closed-form expression, enabling efficient Monte Carlo approximation of the integral.

3 Tree-Sliced Orlicz Integral Probability Metric

We review Orlicz IPMs on trees, introduce TS-Orlicz, and analyze its theoretical properties and computation.

3.1 Orlicz Integral Probability Metric

While Tree-Sliced Wasserstein (TSW) exploits this structure to achieve efficient computation, it is restricted to first-order transport costs, inducing an L^1-type geometry along the tree. This restriction limits the expressiveness of the resulting discrepancy. To address this limitation, we extend the formulation beyond the L^1 setting by introducing a generalized tree-sliced transport framework based on Orlicz IPMs. This allows the use of a broader class of convex cost functions, leading to more flexible and expressive geometries on the space of probability measures.

A class of convex functions. We consider the family of N-functions [1, Section 8.2], which are a particular class of convex functions defined on $\mathbb{R}_+$. In what follows, a strictly increasing and convex function $\Phi : [0, \infty) \to [0, \infty)$ is called an N-function if it satisfies

$$\lim_{t\to 0} \frac{\Phi(t)}{t} = 0 \quad \text{and} \quad \lim_{t\to+\infty} \frac{\Phi(t)}{t} = +\infty.$$

Some examples of N-functions include: **(i)** $\Phi(t) = t^p$ with $1 < p < \infty$; **(ii)** $\Phi(t) = \exp(t) - t - 1$; **(iii)** $\Phi(t) = \exp(t^p) - 1$ with $1 < p < \infty$; and **(iv)** $\Phi(t) = (1 + t)\log(1 + t) - t$ [1, Section 8.2].

Orlicz functional space. Let Φ be an N-function and let ω be the canonical length measure on $\mathcal{T}$, as in Section 2. We use this measure throughout the tree-based Orlicz construction below. We define $L_\Phi(\mathcal{T}, \omega)$ as the linear hull of the set of all Borel measurable functions $f : \mathcal{T} \to \mathbb{R}$ such that: $\int_\mathcal{T} \Phi\left(|f(x)|\right) \omega(dx) < \infty$. Then, $L_\Phi(\mathcal{T}, \omega)$ is a space equipped with the Luxemburg norm:

$$||f||_{L_\Phi} := \inf\left\{t > 0 | \int_\mathcal{T} \Phi\left(\frac{|f(x)|}{t}\right) \omega(dx) \le 1\right\}.$$

For nonzero functions, the above infimum is attained [1, Section 8.9].

Orlicz IPM on Trees. Let Φ be an N-function with complementary function Ψ defined by

$$\Psi(t) := \sup_{a\ge 0}\{at - \Phi(a)\}, \qquad t \ge 0.$$

Let $\mathcal{T}$ be a rooted tree equipped with its canonical length measure ω. We define the tree-based Orlicz–Sobolev space $WL_\Psi(\mathcal{T}, \omega)$ as the set of continuous functions $f : \mathcal{T} \to \mathbb{R}$ such that there exists $h \in L_\Psi(\mathcal{T}, \omega)$ satisfying

$$f(x) - f(z_0) = \int_{[z_0,x]} h(y)\, \omega(dy), \qquad \forall x \in \mathcal{T},$$

where z_0 denotes the root and $[z_0, x]$ is the unique path from z_0 to x. The function h is unique ω-almost everywhere and is denoted by f'. For $\mu, \nu \in \mathcal{P}(\mathcal{T})$, the Orlicz IPM is defined as

$$\mathcal{S}_\Phi(\mu, \nu) := \sup\left\{\left|\int_\mathcal{T} f\, d\mu - \int_\mathcal{T} f\, d\nu\right| :\ f \in WL_\Psi(\mathcal{T}, \omega),\ \|f'\|_{L_\Psi} \le 1\right\}.$$

Computation via a univariate optimization problem. A key advantage of the Orlicz IPM is that, on rooted trees, it admits an equivalent univariate formulation [26]. Specifically, for $\mu, \nu \in \mathcal{P}(\mathcal{T})$,

$$\mathcal{S}_\Phi(\mu,\nu) = \inf_{a>0} \frac{1}{a}\left(1 + \int_{\mathcal{T}} \Phi\big(a\,|h(x)|\big)\,\omega(dx)\right),$$

where $h(x) := \mu(\Lambda(x)) - \nu(\Lambda(x))$, and $\Lambda(x)$ denotes the subtree rooted at x. For discrete measures supported on tree nodes, this becomes

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

where γ_e denotes the set of nodes whose path from the root contains the edge e. Notably, an important special case is obtained by choosing the power N-function

$$\Phi(t) = \frac{(p-1)^{p-1}}{p^p}\,t^p, \qquad p \in (1,\infty).$$

For this choice of Φ, the Orlicz IPM admits the closed-form expression,

$$\mathcal{S}_\Phi(\mu,\nu) = \left(\int_{\mathcal{T}} |\mu(\Lambda(x)) - \nu(\Lambda(x))|^p\,\omega(dx)\right)^{1/p}.$$

3.2 Tree-Sliced Orlicz Integral Probability Metric

Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. For a tree system $\mathcal{T} \in \mathbb{T}$ and a splitting map α, the Radon transform $\mathcal{R}^\alpha_{\mathcal{T}}$ maps μ and ν to the induced measures $\mu_{\mathcal{T}}$ and $\nu_{\mathcal{T}}$ on the corresponding tree. The Orlicz IPM $\mathcal{S}_\Phi(\mu_{\mathcal{T}}, \nu_{\mathcal{T}})$ is then computed on each tree and aggregated over the space of tree systems.

Definition 3.1 (Tree-Sliced Orlicz IPM)**.** Let Φ be an N-function. For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, the *Tree-Sliced Orlicz Integral Probability Metric* (TS-Orlicz) is defined as

$$\text{TS-Orlicz}_\Phi(\mu,\nu) := \int_{\mathbb{T}} \mathcal{S}_\Phi\big(\mu_{\mathcal{T}}, \nu_{\mathcal{T}}\big)\,d\sigma(\mathcal{T}).$$

We next establish the basic theoretical properties of Tree-Sliced Orlicz IPM.

Metricity and invariance. Let $\mathrm{E}(d) = \mathrm{T}(d) \rtimes \mathrm{O}(d)$ denote the Euclidean group of rigid transformations on $\mathbb{R}^d$. Each element $g \in \mathrm{E}(d)$ acts on a point $x \in \mathbb{R}^d$ by

$$gx = Qx + v, \qquad Q \in \mathrm{O}(d),\ v \in \mathbb{R}^d.$$

This action naturally extends from points in $\mathbb{R}^d$ to tree systems by applying the same rigid transformation to the underlying tree representation. We call a splitting map $\alpha \in \mathcal{C}(\mathbb{R}^d \times \mathbb{T}, \Delta_{k-1})$ $\mathrm{E}(d)$-*invariant* if

$$\alpha(gx, g\mathcal{T}) = \alpha(x,\mathcal{T}), \qquad x \in \mathbb{R}^d,\ \mathcal{T} \in \mathbb{T},\ g \in \mathrm{E}(d).$$

Throughout this paper, we use the $\mathrm{E}(d)$-invariant splitting map α introduced in [59], for which the associated Radon transform is injective.

Theorem 3.2. *Let Φ be an N-function. Assume that the tree-level Orlicz IPMs are well-defined extended metrics and that their evaluations are measurable in $\mathcal{T}$. Then:*

(i) TS-Orlicz is an extended pseudometric *on $\mathcal{P}(\mathbb{R}^d)$, allowing infinite values and possibly vanishing between distinct measures.*

(ii) If the sampled projections distinguish measures, namely,

$$\mathcal{R}^\alpha_{\mathcal{T}}\mu = \mathcal{R}^\alpha_{\mathcal{T}}\nu \quad \text{for } \sigma\text{-almost every } \mathcal{T} \quad \Longrightarrow \quad \mu = \nu,$$

then TS-Orlicz is an extended metric*. It is a* finite metric *on any class of measures for which its defining integral is finite for every pair.*

(iii) Let $g \in E(d)$. Suppose that $\mathcal{T} \mapsto g\mathcal{T}$ preserves σ, induces a root- and length-measure-preserving tree isometry, and satisfies

$$\mathcal{R}^{\alpha}_{g\mathcal{T}}(g_{\#}\mu) = g_{\#}(\mathcal{R}^{\alpha}_{\mathcal{T}}\mu)\,,$$

where the right-hand pushforward uses the induced tree isometry. Then

$$\text{TS-Orlicz}_{\Phi}(g_{\#}\mu, g_{\#}\nu) = \text{TS-Orlicz}_{\Phi}(\mu, \nu).$$

The proof of Theorem 3.2 is given in Appendix C.1.

Under the assumptions of Theorem 3.2(iii), TS-Orlicz is invariant under translations and orthogonal transformations.

Connection with TSW. TS-Orlicz generalizes the TSW distance by replacing the L^1 discrepancy with a broader class of convex geometries. We next show that the standard TSW distance is recovered from the normalized power Orlicz functions as $p \to 1^+$.

Theorem 3.3. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be probability measures with finite support. Assume that, for σ-almost every sampled tree system $\mathcal{T}$, the projected measures $\mu_{\mathcal{T}}$ and $\nu_{\mathcal{T}}$ have their root-to-support paths contained in a finite-length subtree $\mathcal{T}_{\mu,\nu}$ of the induced tree, and that $\omega_{\mathcal{T}}(\mathcal{T}_{\mu,\nu}) \le C$ for some constant $C < \infty$. For $p \in (1, \infty)$, let Φ_p be the normalized power N-function defined by*

$$\Phi_p(t) := \frac{(p-1)^{p-1}}{p^p} t^p, \tag{1}$$

then the Tree-Sliced Orlicz IPM satisfies

$$\lim_{p \to 1^+} \text{TS-Orlicz}_{\Phi_p}(\mu, \nu) = \text{TSW}(\mu, \nu). \tag{2}$$

Proof of Theorem 3.3 is deferred to Appendix C.2. Theorem 3.3 establishes that TS-Orlicz provides a continuous interpolation between L^p-type geometries and the standard TSW distance.

Remark 3.4. The finite-length assumption in Theorem 3.3 is generally mild in practice as real-world data are typically normalized or restricted to a bounded domain, and empirical measures have finite support. Hence, for finite trees sampled over the bounded projected range, the subtree supporting the projected samples satisfies this assumption.

Practical computation of TS-Orlicz. We approximate the integral over tree systems by Monte Carlo sampling. Given $\mathcal{T}_1, \ldots, \mathcal{T}_L \sim \sigma$ from $\mathbb{T}$, we compute

$$\widehat{\text{TS-Orlicz}}_{\Phi}(\mu, \nu) = \frac{1}{L} \sum_{i=1}^{L} \widehat{\mathcal{S}}_{\Phi}(\mu_{\mathcal{T}_i}, \nu_{\mathcal{T}_i}).$$

Theorem 3.5 (Monte Carlo approximation)**.** *Assume that the sampled trees are i.i.d. and*

$$\sigma^2_{\Phi} := \text{Var}_{\mathcal{T} \sim \sigma}\left[\mathcal{S}_{\Phi}(\mu_{\mathcal{T}}, \nu_{\mathcal{T}})\right] < \infty.$$

With exact tree-level evaluations, the Monte Carlo estimator is unbiased, converges almost surely to TS-Orlicz as $L \to \infty$, and satisfies

$$\mathbb{E}\left[\left(\widehat{\text{TS-Orlicz}}_{\Phi} - \text{TS-Orlicz}_{\Phi}\right)^2\right] = \frac{\sigma^2_{\Phi}}{L}.$$

For each sampled $\mathcal{T}$, let $\mu_{\mathcal{T}}$ and $\nu_{\mathcal{T}}$ be the projected measures induced by $\mathcal{R}^{\alpha}_{\mathcal{T}}$. For every $e \in E_{\mathcal{T}}$, define $\Delta_{\mathcal{T}}(e) := \mu_{\mathcal{T}}(\gamma_e) - \nu_{\mathcal{T}}(\gamma_e)$. Using the discrete tree formula, the tree-level Orlicz IPM is

$$\mathcal{S}_{\Phi}(\mu_{\mathcal{T}}, \nu_{\mathcal{T}}) = \inf_{a > 0} F_{\mathcal{T}}(a), \qquad F_{\mathcal{T}}(a) := \frac{1}{a}\left(1 + \sum_{e \in E_{\mathcal{T}}} w_e \Phi(a|\Delta_{\mathcal{T}}(e)|)\right).$$

Thus, the only remaining computation is a scalar optimization over a for $\inf_{a>0} F_{\mathcal{T}}(a)$. For the computational procedure, we restrict attention to differentiable N-functions Φ, which covers the standard choices used in practice and all functions considered in our experiments. We solve the scalar problem through its first-order condition. Let $z_e(a) := a|\Delta_{\mathcal{T}}(e)|$ and define

$$G_{\mathcal{T}}(a) := \sum_{e \in E_{\mathcal{T}}} w_e \left[z_e(a)\Phi'(z_e(a)) - \Phi(z_e(a))\right] - 1.$$

Algorithm 1 Tree-level Orlicz IPM log-space bisection

Require: $\{\Delta_{\mathcal{T}}(e), w_e\}_{e \in E_{\mathcal{T}}}$, Φ differentiable, $0 < a_{\min} < a_{\max}$, $\rho > 1$, $\varepsilon_\eta, \varepsilon_G > 0$, $I_{\max}$.
Ensure: Approximation $\widehat{\mathcal{S}}_\Phi(\mu_{\mathcal{T}}, \nu_{\mathcal{T}})$.
if $\Delta_{\mathcal{T}}(e) = 0$ for all $e \in E_{\mathcal{T}}$ **then return** 0 **end if**
while $G_{\mathcal{T}}(a_{\min}) > 0$ **do** $a_{\min} \leftarrow a_{\min}/\rho$ **end while**
while $G_{\mathcal{T}}(a_{\max}) < 0$ **do** $a_{\max} \leftarrow \rho a_{\max}$ **end while**
$\eta_{\min} \leftarrow \log a_{\min}$, $\eta_{\max} \leftarrow \log a_{\max}$
for $i = 1, \ldots, I_{\max}$ **do**
    $\eta \leftarrow (\eta_{\min} + \eta_{\max})/2$, $a \leftarrow \exp(\eta)$, $G \leftarrow G_{\mathcal{T}}(a)$
    if $|G| \le \varepsilon_G$ **or** $\eta_{\max} - \eta_{\min} \le \varepsilon_\eta$ **then break end if**
    if $G < 0$ **then** $\eta_{\min} \leftarrow \eta$ **else** $\eta_{\max} \leftarrow \eta$ **end if**
end for
$\widehat{a}_{\mathcal{T}} \leftarrow \exp((\eta_{\min} + \eta_{\max})/2)$; **return** $F_{\mathcal{T}}(\widehat{a}_{\mathcal{T}})$

For a differentiable N-function Φ, the map $z \mapsto z\Phi'(z) - \Phi(z)$ is continuous and nondecreasing, with limits 0 and $+\infty$ as $z \to 0^+$ and $z \to \infty$, respectively. Thus, whenever some edge has $w_e > 0$ and $|\Delta_{\mathcal{T}}(e)| > 0$, $G_{\mathcal{T}}$ is continuous and nondecreasing, with

$$G_{\mathcal{T}}(0^+) = -1, \qquad G_{\mathcal{T}}(a) \to +\infty \quad \text{as } a \to \infty.$$

A root therefore exists, and geometric interval expansion finds a valid bracket. Since $F'_{\mathcal{T}}(a) = G_{\mathcal{T}}(a)/a^2$, every root is a global minimizer. The root is unique when $z \mapsto z\Phi'(z) - \Phi(z)$ is strictly increasing, as holds for all functions used in our experiments; otherwise, $F_{\mathcal{T}}$ is constant on the root interval. See Lemma A.10 in Appendix A.2.

If all edge imbalances vanish, then $G_{\mathcal{T}} \equiv -1$ and $\inf_{a>0} F_{\mathcal{T}}(a) = 0$, so Algorithm 1 returns 0 directly. Otherwise, we approximate a root by bisection in log-space, $\eta = \log a$, which ensures positivity and controls relative accuracy, and return

$$\widehat{\mathcal{S}}_\Phi(\mu_{\mathcal{T}}, \nu_{\mathcal{T}}) = F_{\mathcal{T}}(\widehat{a}_{\mathcal{T}}).$$

Algorithm 1 solves the scalar optimality equation $G_{\mathcal{T}}(a) = 0$ by bisection in the log-domain. We use the change of variables $\eta = \log a$ because $a > 0$ is a scale parameter whose optimal value can vary by several orders of magnitude across trees and Orlicz functions. Bisection in η-space therefore controls the relative uncertainty of a, rather than an absolute error on a. Since $G_{\mathcal{T}}$ is monotone, the sign of $G_{\mathcal{T}}(a)$ determines which side of the current point contains the root. If $G_{\mathcal{T}}(a) < 0$, the root lies to the right and the lower endpoint is updated; otherwise, the root lies to the left and the upper endpoint is updated. The stopping criterion $\eta_{\max} - \eta_{\min} \le \varepsilon_\eta$ controls the relative uncertainty of a, while $|G_{\mathcal{T}}(a)| \le \varepsilon_G$ controls the residual of the optimality condition. The choice of the initial search interval $[a_{\min}, a_{\max}]$ is discussed in Appendix E.7. In our experiments, we use a small fixed $I_{\max}$, so the solver adds only a constant-factor overhead. Its effect on accuracy and downstream performance is studied in Appendix E.6.

Complexity. Let n be the number of support points, L the number of sampled tree systems, and k the number of lines in each tree system. Projection and sorting cost $O(Lkdn + Lkn \log n)$, as in standard tree-sliced methods. Each binary-search step evaluates $G_{\mathcal{T}}$ by one pass over the tree edges, costing $O(Lkn)$ across all sampled tree systems. With at most $I_{\max}$ binary-search iterations, the total computational complexity is

$$O(Lkdn + Lkn \log n + I_{\max} Lkn).$$

As shown empirically in Appendix E.6, a small fixed value of $I_{\max}$ is sufficient in practice. Thus, the asymptotic complexity remains $O(Lkdn + Lkn \log n)$.

3.3 Spherical Tree-Sliced Orlicz Integral Probability Metric

The TS-Orlicz$_\Phi$ construction extends naturally to probability measures on the hypersphere, i.e., $\mu, \nu \in \mathcal{P}(\mathbb{S}^d)$. This extension leverages spherical tree systems [62], which enable projections of measures onto tree metric spaces via a spherical Radon transform.

Given a tree system $\mathcal{T} \in \mathbb{T}$, the projection induces measures $\mu_{\mathcal{T}}$ and $\nu_{\mathcal{T}}$ on the corresponding tree. The Orlicz IPM is then computed on each tree and aggregated over the space of tree systems.

Table 1: Average Wasserstein distance over 5 runs between source and target distributions (scaled by 10^{-1} for the 30D Gaussian case). All methods use 100 projection directions.

Methods	25 Gaussians Iteration					Gaussian 30D Iteration				
	500	1000	1500	2000	2500	500	1000	1500	2000	2500
SW	4.03e-2	1.83e-2	1.09e-2	8.40e-3	7.14e-3	2.93	2.87	2.80	2.72	2.64
MaxSW	7.99e-2	7.22e-2	6.77e-2	6.39e-2	6.18e-2	2.57	2.41	2.35	2.34	2.31
SWGG	2.04e-2	8.24e-4	3.35e-6	3.93e-6	2.86e-6	2.72	2.74	2.74	2.74	2.74
LCVSW	8.04e-3	5.92e-3	5.58e-3	5.04e-3	4.91e-3	2.85	2.71	2.58	2.45	2.33
TSW-SL	4.90e-3	1.06e-6	8.90e-7	8.25e-7	7.32e-7	2.48	2.29	2.16	2.05	1.97
Db-TSW	9.17e-4	9.68e-7	7.94e-7	7.48e-7	7.68e-7	**2.22**	1.69	1.16	6.52e-1	2.77e-1
Spatial-TSW	**3.57e-5**	1.68e-6	1.59e-6	1.44e-6	1.27e-6	2.30	**1.59**	**1.08**	6.82e-1	3.96e-1
TS-Orlicz$_{\Phi_{0,1.1}}$	5.63e-4	**8.92e-7**	**7.64e-7**	**7.45e-7**	**6.43e-7**	**2.22**	1.66	1.10	5.91e-1	2.29e-1
TS-Orlicz$_{\Phi_{0,2}}$	1.51e-3	4.41e-6	3.23e-6	2.53e-6	2.57e-6	**2.22**	1.68	**1.08**	**5.37e-1**	**1.76e-1**
TS-Orlicz$_{\Phi_1}$	8.99e-4	1.92e-5	1.42e-5	1.40e-5	1.39e-5	2.23	1.70	1.13	5.96e-1	2.16e-1
TS-Orlicz$_{\Phi_2}$	1.19e-3	7.84e-5	8.68e-5	9.11e-5	7.45e-5	2.23	1.69	1.09	5.58e-1	1.93e-1

Definition 3.6 (Spherical Tree-Sliced Orlicz IPM)**.** Let Φ be an N-function. For $\mu, \nu \in \mathcal{P}(\mathbb{S}^d)$, the Spherical Tree-Sliced Orlicz IPM is defined as

$$\text{STS-Orlicz}_\Phi(\mu, \nu) := \int_{\mathbb{T}} \mathcal{S}_\Phi\big(\mu_\mathcal{T}, \nu_\mathcal{T}\big)\, d\sigma(\mathcal{T}). \tag{3}$$

The spherical construction inherits the metric properties of its Euclidean counterpart under the corresponding measurability, measure-distinguishing, and finiteness conditions. Invariance under $O(d+1)$ additionally requires equivariant projections and preservation of the tree-sampling distribution and tree length measure. For normalized L^p-type Orlicz functions, STS-Orlicz recovers STSW as $p \to 1^+$. Computation uses the same Monte Carlo estimator and tree-level solver as in the Euclidean setting. Details are given in Appendix D.2.

4 Experimental Results

We evaluate how the Orlicz function Φ affects Euclidean and spherical gradient flows, diffusion-based image generation, and spherical self-supervised learning.

N-function choices. Across all experiments, we consider a unified family of N-functions Φ, including power functions $\Phi_{0,p}(t) = \dfrac{(p-1)^{p-1}t^p}{p^p}$ for $p \in \{1.1, 2\}$, exponential-type functions $\Phi_1(t) = e^t - t - 1$ and $\Phi_2(t) = e^{t^2} - 1$. This selection allows us to compare different geometric regimes, ranging from near-L^1 behavior to heavier-tailed exponential growth. Additional Orlicz growth regimes and sensitivity results are reported in Appendix E.5.

4.1 Euclidean data

Gradient flow on $\mathbb{R}^d$. We evaluate our approach on a gradient flow problem, where the objective is to evolve a distributional path $\{\mu_t\}_{t\geq 0}$ that minimizes a discrepancy $D(\mu_t, \nu)$ between an initial source $\mu_0 = \mathcal{N}(0, I)$ and a fixed target distribution ν. The dynamics follow the gradient flow

$$\partial_t \mu_t = -\nabla D(\mu_t, \nu),$$

where $\nabla D(\mu_t, \nu)$ denotes the functional gradient of the chosen distance.

Experiments are conducted on the *25 Gaussians* and *Gaussian 30D* datasets. We compare our methods against a diverse set of baselines, including Sliced Wasserstein (SW) variants—such as vanilla SW [7], MaxSW [10], LCVSW [37], and SWGG [33]—as well as tree-based methods including TSW-SL [63], Db-TSW [59], Spatial-TSW [61]. Performance is measured by tracking the Wasserstein distance to the target distribution throughout optimization, up to 2500 iterations. Detailed quantitative results are reported in Table 1.

Table 2: A comparison of DDGAN models on the CIFAR-10 unconditional generation benchmark, showing Fréchet Inception Distance (FID) scores and per-epoch training times.

Model	FID↓	Time/Epoch (s)↓
DDGAN [72]	3.64	49
SW-DD [38]	2.90	50
DSW-DD [38]	2.88	225
EBSW-DD [38]	2.87	51
RPSW-DD [38]	2.82	51
IWRPSW-DD [38]	2.70	51
TSW-SL-DD [63]	2.83	74
Db-TSW-DD [59]	2.60	74
Db-TSW-DD$^{\perp}$ [59]	2.53	75
Circular-TSW-DD [62]	2.33	74
TS-Orlicz$_{\Phi_1}$-DD (ours)	$\underline{2.320 \pm 0.019}$	74
TS-Orlicz$_{\Phi_2}$-DD (ours)	$\mathbf{2.294 \pm 0.026}$	74

Table 3: Accuracy of the linear classifier on encoded (E) features and projected (P) features on $\mathbb{S}^9$ over 5 independent runs.

Method	Acc. E(%)↑	Acc. P(%)↑
Hypersphere	80.18 ± 0.20	74.76 ± 0.03
SimCLR	79.53 ± 1.04	73.61 ± 0.32
SSW	70.22 ± 0.26	64.40 ± 0.22
S3W	78.38 ± 0.29	73.60 ± 0.50
RI-S3W (5)	79.65 ± 0.18	73.77 ± 0.97
ARI-S3W (5)	79.60 ± 0.17	74.25 ± 0.74
STSW	80.36 ± 0.17	77.14 ± 0.24
Spatial-STSW	80.32 ± 0.12	$\underline{77.33} \pm 0.21$
STS-Orlicz$_{\Phi_{0,2}}$	$\underline{80.56} \pm 0.18$	$\mathbf{77.47} \pm 0.21$
STS-Orlicz$_{\Phi_1}$	$\mathbf{80.58} \pm 0.10$	77.28 ± 0.11
STS-Orlicz$_{\Phi_2}$	80.55 ± 0.14	77.28 ± 0.13

Table 4: Log of the Wasserstein distance between source and target distributions over 10 runs on a mixture of 12 vMFs.

Methods	Epoch				
	50	100	150	200	250
SSW	-2.445 ± 0.054	-2.817 ± 0.044	-2.953 ± 0.047	-3.024 ± 0.054	-3.073 ± 0.055
S3W	-2.025 ± 0.038	-2.216 ± 0.045	-2.294 ± 0.061	-2.297 ± 0.060	-2.301 ± 0.067
RI-S3W (1)	-2.093 ± 0.025	-2.495 ± 0.029	-2.706 ± 0.024	-2.833 ± 0.025	-2.928 ± 0.030
RI-S3W (5)	-2.439 ± 0.030	-2.812 ± 0.025	-2.989 ± 0.022	-3.102 ± 0.026	-3.192 ± 0.037
ARI-S3W (30)	-2.636 ± 0.031	-3.007 ± 0.033	-3.197 ± 0.049	-3.357 ± 0.080	-3.501 ± 0.121
LSSOT	-2.480 ± 0.028	-2.645 ± 0.026	-2.639 ± 0.015	-2.619 ± 0.021	-2.599 ± 0.012
STSW	$\mathbf{-3.265 \pm 0.043}$	$\underline{-3.652 \pm 0.174}$	-3.831 ± 0.204	-3.966 ± 0.216	-4.002 ± 0.190
Spatial-STSW	-3.183 ± 0.038	-3.511 ± 0.064	-3.681 ± 0.103	-3.796 ± 0.148	-3.863 ± 0.197
STS-Orlicz$_{\Phi_1}$	$\underline{-3.237 \pm 0.041}$	$\mathbf{-3.685 \pm 0.176}$	$\mathbf{-3.888 \pm 0.248}$	$\underline{-4.005 \pm 0.267}$	$\underline{-4.044 \pm 0.266}$
STS-Orlicz$_{\Phi_2}$	-3.234 ± 0.030	-3.625 ± 0.093	$\underline{-3.854 \pm 0.168}$	$\mathbf{-4.029 \pm 0.236}$	$\mathbf{-4.140 \pm 0.268}$

Table 1 shows that TS-Orlicz variants attain the lowest final mean distances among the methods compared in that table on both Euclidean benchmarks. On *25 Gaussians*, TS-Orlicz$_{\Phi_{0,1.1}}$ reaches 6.43×10^{-7} at iteration 2500, compared with 7.32×10^{-7} for TSW-SL, and has the lowest reported mean at every checkpoint from iteration 1000 onward. On *Gaussian 30D*, TS-Orlicz$_{\Phi_{0,2}}$ has a final reported mean of 1.76×10^{-1}, compared with 2.77×10^{-1} for Db-TSW, using the table's scaling convention. The preferred Orlicz function thus differs between the two benchmarks; additional choices are compared in Appendix E.5.

Diffusion models. We evaluate our method in the context of unconditional image generation using Denoising Diffusion Generative Adversarial Networks (DDGANs) [72]. Following the framework of [38], we incorporate our proposed TS-Orlicz distance into the Augmented Generalized Mini-batch Energy (AGME) objective. We compare against a range of DDGAN variants based on Sliced and Tree-Sliced Wasserstein distances. Detailed experimental configurations are provided in Appendix E.3.

Table 2 reports the lowest mean FID over 5 runs for TS-Orlicz$_{\Phi_2}$-DD at 2.294 ± 0.026, followed by TS-Orlicz$_{\Phi_1}$-DD at 2.320 ± 0.019. Both means are below the reported FIDs of Circular-TSW-DD (2.33) and Db-TSW-DD$^{\perp}$ (2.53). Both TS-Orlicz variants have a reported training time of 74 seconds per epoch, matching Circular-TSW-DD and close to the 75 seconds of Db-TSW-DD$^{\perp}$. These results suggest that the Orlicz choice can improve the reported FID while retaining comparable per-epoch runtime.

4.2 Evaluation on Spherical Data

Self-supervised learning (SSL). Contrastive learning objectives are commonly decomposed into two complementary components: an alignment term, which pulls together representations of similar inputs, and a uniformity term, which encourages embeddings to be evenly distributed on the representation space [70]. Following this framework, and building on [5], we replace the Gaussian kernel typically used in the uniformity term with our proposed distance. The resulting objective takes the form

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} \|z_i^A - z_i^B\|_2^2 + \frac{\lambda}{2}\Big(\text{STS-Orlicz}_\Phi(z^A, \nu) + \text{STS-Orlicz}_\Phi(z^B, \nu)\Big),$$

where $\nu = \mathcal{U}(\mathbb{S}^d)$ denotes the uniform distribution on the unit sphere, $z^A, z^B \in \mathbb{R}^{n\times(d+1)}$ are embeddings of two augmented views, and $\lambda > 0$ balances alignment and uniformity. We follow the training protocol of [5] and [60, 62], using a ResNet18 [14] encoder trained on CIFAR-10 [23] for 200 epochs. After pretraining, a linear classifier is fitted on top of the frozen encoder to evaluate representation quality.

Table 3 reports different winners for encoded and projected features over five independent runs. STS-Orlicz$_{\Phi_1}$ has the highest mean encoded accuracy at $80.58 \pm 0.10\%$, while STS-Orlicz$_{\Phi_{0,2}}$ has the highest mean projected accuracy at $77.47 \pm 0.21\%$ and an encoded accuracy of $80.56 \pm 0.18\%$. STS-Orlicz$_{\Phi_2}$ obtains $80.55 \pm 0.14\%$ encoded and $77.28 \pm 0.13\%$ projected accuracy. All three variants have higher mean encoded accuracy than STSW and Spatial-STSW; for projected features, only the quadratic variant has a higher mean than both tree-based baselines. The improvements over these tree-based baselines are small relative to the reported variability. All three variants also have higher mean encoded and projected accuracies than the SSW and S3W variants listed in the table.

Gradient flow on the sphere. We consider a gradient flow problem on $\mathbb{S}^d$, where the objective is to transport a source distribution μ toward a target distribution ν by minimizing a discrepancy $d(\mu, \nu)$. We compare our method against several spherical transport distances, including SSW [5], S3W [60], LSSOT [30], STSW [62] and Spatial-STSW [61]. Following prior work [5, 60, 62], the target distribution is defined as a mixture of 12 von Mises–Fisher (vMF) components with 2400 samples. Optimization is performed using projected gradient descent on the sphere [5], with full-batch updates.

Table 4 reports $\log W_2$ at epochs 50, 100, 150, 200, and 250 over 10 independent runs. STSW has the lowest mean at epoch 50; STS-Orlicz$_{\Phi_1}$ has the lowest means at epochs 100 and 150, and STS-Orlicz$_{\Phi_2}$ at epochs 200 and 250. At the final epoch, STS-Orlicz$_{\Phi_2}$ and STS-Orlicz$_{\Phi_1}$ attain -4.140 ± 0.268 and -4.044 ± 0.266, respectively, compared with -4.002 ± 0.190 for STSW, -3.073 ± 0.055 for SSW, and -3.501 ± 0.121 for ARI-S3W (30). Both Orlicz variants therefore have lower final means than STSW, although the differences are small relative to the reported variability.

5 Conclusion

We introduced Tree-Sliced Orlicz IPM (TS-Orlicz) and its spherical variant, STS-Orlicz, as flexible extensions of the Tree-Sliced Wasserstein framework. Our method replaces the tree-level W_1 discrepancy with an Orlicz IPM, enabling tree-sliced distances beyond the standard L^1-type geometry. By revisiting Orlicz IPMs on tree metric spaces, we leveraged their tractable univariate formulation and derived efficient closed-form special cases or scalar optimization. This yields a family of tree-sliced discrepancies that can recover L^p-type behavior and incorporate more tail-sensitive geometries while retaining the computational efficiency of existing tree-sliced methods. We established metricity under the stated separation and finiteness conditions, invariance under sampling-law preservation and equivariant root- and length-preserving tree isometries, and the connection between TS-Orlicz and TSW. Empirical results on Euclidean and spherical benchmarks show that TS-Orlicz and STS-Orlicz improve over recent sliced and tree-sliced baselines in gradient flows, self-supervised learning, and diffusion model training, with comparable runtime. A limitation of the current framework is that the choice of Orlicz function remains task-dependent and may require tuning for different data geometries. Future work may study adaptive selection of Orlicz functions, extend the framework to unbalanced measures, and develop tree-sliced Orlicz methods that recover explicit transport maps.

Acknowledgments and Disclosure of Funding

This paper is dedicated to Professor Richard G. Baraniuk on the occasion of his 60th birthday.

This research is supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG2-TC-2023-012-SGIL). This research is also supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2023) (A-8002040-00-00, A-8002039-00-00), the NUS Presidential Young Professorship Award (A-0009807-01-00), the NUS Artificial Intelligence Institute–Seed Funding (A-8003062-00-00), and the Cross Faculty Grant 2025, CFG25-012 (A-8004460-00-00).

References

[1] Robert A Adams and John JF Fournier. *Sobolev spaces*, volume 140. Elsevier, 2003.

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

[3] David Alvarez-Melis, Tommi Jaakkola, and Stefanie Jegelka. Structured optimal transport. In *International conference on artificial intelligence and statistics*, pages 1771–1780. PMLR, 2018.

[4] Alexandr Andoni, Chengyu Lin, Ying Sheng, Peilin Zhong, and Ruiqi Zhong. Subspace embedding and linear regression with orlicz norm. In *International Conference on Machine Learning*, pages 224–233. PMLR, 2018.

[5] Clément Bonet, Paul Berg, Nicolas Courty, François Septier, Lucas Drumetz, and Minh-Tan Pham. Spherical sliced-wasserstein. *arXiv preprint arXiv:2206.08780*, 2022.

[6] Clément Bonet, Laetitia Chapel, Lucas Drumetz, and Nicolas Courty. Hyperbolic sliced-wasserstein via geodesic and horospherical projections. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pages 334–370. PMLR, 2023.

[7] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

[8] Charlotte Bunne, Laetitia Papaxanthos, Andreas Krause, and Marco Cuturi. Proximal optimal transport modeling of population dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 6511–6528. PMLR, 2022.

[9] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[10] Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced wasserstein distance and its use for gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10648–10656, 2019.

[11] Jiaojiao Fan, Isabel Haasler, Johan Karlsson, and Yongxin Chen. On the complexity of the optimal transport problem with graph-structured cost. In *International conference on artificial intelligence and statistics*, pages 9147–9165. PMLR, 2022.

[12] Aritra Guha, Nhat Ho, and XuanLong Nguyen. On excess mass behavior in gaussian mixture models with orlicz-wasserstein distances. In *International Conference on Machine Learning*, pages 11847–11870. PMLR, 2023.

[13] Allen Hatcher. *Algebraic topology*. Cambridge University Press, 2005.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[16] Xinru Hua, Truyen Nguyen, Tam Le, Jose Blanchet, and Viet Anh Nguyen. Dynamic flows on curved space generated by labeled data. *arXiv preprint arXiv:2302.00061*, 2023.

[17] Piotr Indyk and Nitin Thaper. Fast image retrieval via embeddings. In *3rd international workshop on statistical and computational theories of vision*, volume 1. Nice, France, 2003.

[18] Hoang Anh Just, Feiyang Kang, Jiachen T Wang, Yi Zeng, Myeongseob Ko, Ming Jin, and Ruoxi Jia. Lava: Data valuation without pre-specified learning algorithms. *arXiv preprint arXiv:2305.00054*, 2023.

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

[20] Samuel Kessler, Tam Le, and Vu Nguyen. Sava: Scalable learning-agnostic data valuation. *arXiv preprint arXiv:2406.01130*, 2024.

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[22] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

[23] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[24] Hugo Lavenant, Sebastian Claici, Edward Chien, and Justin Solomon. Dynamical optimal transport on discrete surfaces. *ACM Transactions on Graphics (TOG)*, 37(6):1–16, 2018.

[25] Tam Le and Truyen Nguyen. Entropy partial transport with tree metrics: Theory and practice. In *International Conference on Artificial Intelligence and Statistics*, pages 3835–3843. PMLR, 2021.

[26] Tam Le, Truyen Nguyen, and Kenji Fukumizu. Generalized sobolev transport for probability measures on a graph. *arXiv preprint arXiv:2402.04516*, 2024.

[27] Tam Le, Truyen Nguyen, Dinh Phung, and Viet Anh Nguyen. Sobolev transport: A scalable metric for probability measures with graph metrics. In *International Conference on Artificial Intelligence and Statistics*, pages 9844–9868. PMLR, 2022.

[28] Tam Le, Makoto Yamada, Kenji Fukumizu, and Marco Cuturi. Tree-sliced variants of wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

[29] Lang Liu, Soumik Pal, and Zaid Harchaoui. Entropy regularized optimal transport independence criterion. In *International Conference on Artificial Intelligence and Statistics*, pages 11247–11279. PMLR, 2022.

[30] Xinran Liu, Yikun Bai, Rocío Díaz Martín, Kaiwen Shi, Ashkan Shahbazi, Bennett A Landman, Catie Chang, and Soheil Kolouri. Linear spherical sliced optimal transport: A fast metric for comparing spherical data. *arXiv preprint arXiv:2411.06055*, 2024.

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

[32] Manh Luong, Khai Nguyen, Nhat Ho, Reza Haf, Dinh Phung, and Lizhen Qu. Revisiting deep audio-text retrieval through the lens of transportation. *arXiv preprint arXiv:2405.10084*, 2024.

[33] Guillaume Mahey, Laetitia Chapel, Gilles Gasso, Clement Bonet, and Nicolas Courty. Fast optimal transport through sliced generalized wasserstein geodesics. In *NeurIPS*, 2023.

[34] Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in neural information processing systems*, 32, 2019.

[35] Kimia Nadjahi, Alain Durmus, Pierre E Jacob, Roland Badeau, and Umut Simsekli. Fast approximation of the sliced-wasserstein distance using concentration of random projections. *Advances in Neural Information Processing Systems*, 34:12411–12424, 2021.

[36] Khai Nguyen, Nicola Bariletto, and Nhat Ho. Quasi-monte carlo for 3d sliced wasserstein. *arXiv preprint arXiv:2309.11713*, 2023.

[37] Khai Nguyen and Nhat Ho. Sliced wasserstein estimation with control variates. *arXiv preprint arXiv:2305.00402*, 2023.

[38] Khai Nguyen, Shujian Zhang, Tam Le, and Nhat Ho. Sliced wasserstein with random-path projecting directions. *arXiv preprint arXiv:2401.15889*, 2024.

[39] Tin D Nguyen, Brian L Trippe, and Tamara Broderick. Many processors, little time: Mcmc for partitions via optimal transport couplings. In *International Conference on Artificial Intelligence and Statistics*, pages 3483–3514. PMLR, 2022.

[40] Trung Nguyen, Quang-Hieu Pham, Tam Le, Tung Pham, Nhat Ho, and Binh-Son Hua. Point-set distances for learning representations of 3d point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10478–10487, 2021.

[41] Vu Nguyen, Tam Le, Makoto Yamada, and Michael A Osborne. Optimal transport kernels for sequential and parallel neural architecture search. In *International Conference on Machine Learning*, pages 8084–8095. PMLR, 2021.

[42] Sloan Nietert, Ziv Goldfeld, and Rachel Cummings. Outlier-robust optimal transport: Duality, structure, and statistical analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 11691–11719. PMLR, 2022.

[43] Jonathan Niles-Weed and Philippe Rigollet. Estimation of wasserstein distances in the spiked transport model. *Bernoulli*, 28(4):2663–2688, 2022.

[44] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bridging vision and language spaces with assignment prediction. *arXiv preprint arXiv:2404.09632*, 2024.

[45] François-Pierre Paty and Marco Cuturi. Subspace robust wasserstein distances. In *International conference on machine learning*, pages 5072–5081. PMLR, 2019.

[46] Gabriel Peyré and Marco Cuturi. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019.

[47] Duy-Tung Pham, Viet-Hoang Tran, Thieu Vo, and Tan Nguyen. Mixed-curvature tree-sliced wasserstein distance. In *International Conference on Learning Representations*, volume 2026, pages 110044–110071, 2026.

[48] Thong Pham, Shohei Shimizu, Hideitsu Hino, and Tam Le. Scalable counterfactual distribution estimation in multivariate causal models. In *Causal Learning and Reasoning*, pages 1118–1140. PMLR, 2024.

[49] Michael Quellmalz, Robert Beinert, and Gabriele Steidl. Sliced optimal transport on the sphere. *Inverse Problems*, 39(10):105005, 2023.

[50] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International conference on scale space and variational methods in computer vision*, pages 435–446. Springer, 2011.

[51] Malempati M. Rao and Zhong Dao Ren. Theory of orlicz spaces. 1991.

[52] Mahdi Saleh, Shun-Cheng Wu, Luca Cosmo, Nassir Navab, Benjamin Busam, and Federico Tombari. Bending graphs: Hierarchical shape matching using gated optimal transport. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11757–11767, 2022.

[53] Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018.

[54] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser, Cham, 2015.

[55] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

[56] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (ToG)*, 34(4):1–11, 2015.

[57] Zhao Song, Ruosong Wang, Lin Yang, Hongyang Zhang, and Peilin Zhong. Efficient symmetric norm regression via linear sketching. *Advances in Neural Information Processing Systems*, 32, 2019.

[58] Karl-Theodor Sturm. Generalized orlicz spaces and wasserstein distances for convex–concave scale functions. *Bulletin des sciences mathématiques*, 135(6-7):795–802, 2011.

[59] Hoang V. Tran, Khoi N. M. Nguyen, Trang Pham, Thanh T. Chu, Tam Le, and Tan M. Nguyen. Distance-based tree-sliced wasserstein distance. *arXiv preprint arXiv:2503.11050*, 2025.

[60] Huy Tran, Yikun Bai, Abihith Kothapalli, Ashkan Shahbazi, Xinran Liu, Rocio Diaz Martin, and Soheil Kolouri. Stereographic spherical sliced wasserstein distances. *arXiv preprint arXiv:2402.02345*, 2024.

[61] Thanh Tran, Viet-Hoang Tran, Thanh Chu, Trang Pham, Laurent El Ghaoui, Tam Le, and Tan M Nguyen. Tree-sliced wasserstein distance with nonlinear projection. *arXiv preprint arXiv:2505.00968*, 2025.

[62] Viet-Hoang Tran, Thanh T Chu, Khoi NM Nguyen, Trang Pham, Tam Le, and Tan M Nguyen. Spherical tree-sliced wasserstein distance. *arXiv preprint arXiv:2503.11249*, 2025.

[63] Viet-Hoang Tran, Trang Pham, Tho Tran, Tam Le, and Tan M Nguyen. Tree-sliced wasserstein distance on a system of lines. *arXiv preprint arXiv:2406.13725*, 2024.

[64] Viet-Hoang Tran, Thanh Tran, Thanh Chu, Tam Le, and Tan Nguyen. Tree-sliced entropy partial transport. *Advances in Neural Information Processing Systems*, 38:16503–16560, 2026.

[65] Viet-Hoang Tran, Thanh Q Tran, Thanh Chu, Duy-Tung Pham, Trung-Khang Tran, Tam Le, and Tan Nguyen. Tree-sliced sobolev ipm. In *International Conference on Learning Representations*, volume 2026, pages 33846–33890, 2026.

[66] Viet-Hoang Tran, Thanh Q Tran, Thanh Chu, Trung-Khang Tran, Duy-Tung Pham, Tam Le, and Tan Nguyen. Revisiting tree-sliced wasserstein distance through the lens of the fermat–weber problem. In *International Conference on Learning Representations*, volume 2026, pages 62831–62860, 2026.

[67] Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

[68] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

[69] Jie Wang, Rui Gao, and Yao Xie. Two-sample test with kernel projected wasserstein distance. *arXiv preprint arXiv:2102.06449*, 2021.

[70] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.

[71] Jonathan Weed and Quentin Berthet. Estimation of smooth densities in wasserstein distance. In *conference on Learning Theory*, pages 3118–3119. PMLR, 2019.

[72] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021.

Table of Notation

$\mathbb{R}^d$	d-dimensional Euclidean space
$\|\cdot\|_2$	Euclidean norm
$\mathbb{S}^{d-1}$	$(d-1)$-dimensional hypersphere
(V, E)	vertex and edge set of a tree
e	edge in a tree
w_e	length/weight of edge e
r or z_0	root of a tree
$L^1(X)$	space of Lebesgue integrable functions on X
$\mathcal{P}(X)$	space of probability measures on X
$\mathcal{M}(X)$	space of measures on X
μ, ν	measures
$\delta(\cdot)$	1-dimensional Dirac delta function
$\mathcal{U}(\mathbb{S}^{d-1})$	uniform distribution on $\mathbb{S}^{d-1}$
$\mathcal{C}(X, Y)$	space of continuous maps from X to Y
$d(\cdot, \cdot)$	metric in metric space
$d_{\mathcal{T}}(\cdot, \cdot)$	tree metric
$\mathrm{E}(d)$	Euclidean group on $\mathbb{R}^d$
W_p	p-Wasserstein distance
SW_p	Sliced p-Wasserstein distance
Λ	(rooted) subtree
γ_e	rooted subtree associated with edge e
$\mathcal{T}$	tree system
L	number of Monte Carlo samples
k	number of lines in a system of lines or a tree system
$\mathcal{R}^\alpha$	Radon Transform on Systems of Lines
Δ_{k-1}	$(k-1)$-dimensional standard simplex
α	splitting map
ξ, ζ, c	tuning parameter
$\mathbb{T}$	space of tree systems
σ	distributions on (components of) space of tree systems
$\mathcal{N}$	normal (Gaussian) distribution
$\mathcal{U}$	uniform distribution
δ	Dirac delta distribution
Φ	N-function defining the Orlicz geometry
Ψ	complementary N-function of Φ
S_Φ	tree-level Orlicz integral probability metric
$h_{\mu,\nu}$	subtree mass discrepancy $\mu(\Lambda(\cdot)) - \nu(\Lambda(\cdot))$
ω	canonical length measure on a tree

Appendix of "Tree-Sliced Orlicz Integral Probability Metric"

Table of Contents

A Background of Orlicz Integral Probability Metrics on Tree Metric Spaces **17**

A.1 Setting and Notation 17

A.2 Orlicz Integral Probability Metrics on Tree Metric Spaces 18

B Tree-Sliced Wasserstein Background on Euclidean Spaces **20**

C Theoretical Proofs **22**

C.1 Proof of Theorem 3.2 22

C.2 Proof for Theorem 3.3 24

C.3 Proof of Theorem 3.5 25

C.4 Proof for Theorem A.7 25

C.5 Proof for Proposition A.9 26

D Spherical Tree-Sliced Orlicz Integral Probability Metric **27**

D.1 Background: Spherical Tree-Sliced Wasserstein Distances 27

D.2 Spherical Tree-Sliced Orlicz Integral Probability Metric 29

D.3 Theoretical Proofs for Spherical Tree-Sliced Orlicz Integral Probability Metric . . . 31

E Experimental Details **33**

E.1 Runtime and Memory Analysis 33

E.2 Gradient Flow 33

E.3 Diffusion Models 35

E.4 Self-Supervised Learning 36

E.5 Effect of the Orlicz Function 37

E.6 Sensitivity to Log-Space Bisection Steps 38

E.7 Choice of the scalar search interval. 39

E.8 Hardware Settings 40

E.9 Broader Societal Impacts 40

A Background of Orlicz Integral Probability Metrics on Tree Metric Spaces

A.1 Setting and Notation

Let $\mathcal{T} = (V, E)$ be a rooted tree with vertex set V, edge set E, and root $z_0 \in V$. Each edge $e \in E$ is assigned a positive length $w_e > 0$. We use the continuous tree convention: $\mathcal{T}$ contains not only the vertices in V, but also every point lying along each edge.

The tree is equipped with the metric $d_{\mathcal{T}}$, where $d_{\mathcal{T}}(x, y)$ is the length of the unique path connecting x and y. We denote this path by $[x, y]$. Let ω be the canonical length measure on $\mathcal{T}$, so that

$$\omega([x, y]) = d_{\mathcal{T}}(x, y), \qquad \forall x, y \in \mathcal{T}.$$

For any $x \in \mathcal{T}$, define the rooted subtree

$$\Lambda(x) := \{y \in \mathcal{T} : x \in [z_0, y]\}.$$

For any edge $e \in E$, define

$$\gamma_e := \{y \in \mathcal{T} : e \subset [z_0, y]\}.$$

We denote by $\mathcal{P}(\mathcal{T})$ the space of Borel probability measures on $\mathcal{T}$. For $\mu, \nu \in \mathcal{P}(\mathcal{T})$, define

$$h_{\mu,\nu}(x) := \mu(\Lambda(x)) - \nu(\Lambda(x)).$$

A.2 Orlicz Integral Probability Metrics on Tree Metric Spaces

By a continuous function on $\mathcal{T}$, we mean a function $f : \mathcal{T} \to \mathbb{R}$ that is continuous with respect to the topology induced by the tree metric. We denote the set of such functions by $C(\mathcal{T})$.

Definition A.1 (Tree derivative)**.** Let $f : \mathcal{T} \to \mathbb{R}$ be continuous. We say that f admits a tree derivative with respect to ω if there exists a measurable function $g : \mathcal{T} \to \mathbb{R}$ such that

$$f(x) - f(z_0) = \int_{[z_0,x]} g(y)\, \omega(dy), \qquad \forall x \in \mathcal{T}. \tag{4}$$

Such a function g, when it exists, is unique ω-almost everywhere. We call g the tree derivative of f and denote it by f'.

We next introduce the class of convex functions used to define Orlicz geometry.

Definition A.2 (N-function)**.** A function $\Phi : [0, \infty) \to [0, \infty)$ is called an N-function if it is strictly increasing, convex, and satisfies

$$\lim_{t\to 0} \frac{\Phi(t)}{t} = 0, \qquad \lim_{t\to\infty} \frac{\Phi(t)}{t} = +\infty. \tag{5}$$

Typical examples of N-functions include

$$\Phi(t) = t^p, \qquad 1 < p < \infty,$$
$$\Phi(t) = \exp(t) - t - 1,$$
$$\Phi(t) = \exp(t^p) - 1, \qquad 1 < p < \infty,$$

and

$$\Phi(t) = (1 + t)\log(1 + t) - t.$$

Definition A.3 (Orlicz functional space)**.** Let Φ be an N-function. The Orlicz space $L^\Phi(\mathcal{T}, \omega)$ is the linear hull of all Borel measurable functions $g : \mathcal{T} \to \mathbb{R}$ satisfying

$$\int_{\mathcal{T}} \Phi(|g(x)|)\, \omega(dx) < \infty. \tag{6}$$

This space is equipped with the Luxemburg norm

$$\|g\|_{L^\Phi} := \inf\left\{c > 0 : \int_{\mathcal{T}} \Phi\left(\frac{|g(x)|}{c}\right) \omega(dx) \le 1\right\}. \tag{7}$$

Under the standard assumptions for Orlicz spaces, the infimum in the Luxemburg norm is attained for nonzero functions.

We now define the tree-based function space that controls derivatives through an Orlicz norm.

Definition A.4 (Tree-based Orlicz-Sobolev space)**.** Let Φ be an N-function and let ω be the length measure on $\mathcal{T}$. A continuous function $f : \mathcal{T} \to \mathbb{R}$ is said to belong to the tree-based Orlicz-Sobolev space $WL^\Phi(\mathcal{T}, \omega)$ if there exists a function $g \in L^\Phi(\mathcal{T}, \omega)$ such that

$$f(x) - f(z_0) = \int_{[z_0,x]} g(y)\, \omega(dy), \qquad \forall x \in \mathcal{T}. \tag{8}$$

The function g is unique in $L^\Phi(\mathcal{T}, \omega)$ up to ω-almost everywhere equivalence. We denote it by f' and call it the tree derivative of f.

The space $WL^{\Phi}(\mathcal{T}, \omega)$ depends on the root z_0, the measure ω, and the N-function Φ. We suppress these dependencies when no confusion arises.

The Orlicz IPM below uses the complementary function of a given N-function.

Definition A.5 (Complementary N-function)**.** Let Φ be an N-function. Its complementary function $\Psi : [0, \infty) \to [0, \infty)$ is defined by

$$\Psi(t) = \sup_{a \geq 0} \{at - \Phi(a)\}, \qquad t \geq 0. \tag{9}$$

The pair (Φ, Ψ) satisfies Young's inequality:

$$ab \leq \Phi(a) + \Psi(b), \qquad a, b \geq 0. \tag{10}$$

Common complementary pairs include

$$\Phi(t) = \frac{t^p}{p}, \qquad \Psi(t) = \frac{t^q}{q}, \qquad \frac{1}{p} + \frac{1}{q} = 1, \quad 1 < p < \infty, \tag{11}$$

and

$$\Phi(t) = \exp(t) - t - 1, \qquad \Psi(t) = (1 + t)\log(1 + t) - t. \tag{12}$$

For $\Phi(t) = \exp(t^p) - 1$ with $1 < p < \infty$, the complementary N-function also exists, although its explicit expression is less simple.

We now define the tree-level Orlicz IPM used in our construction. The definition follows a dual viewpoint: the critic function is constrained by the Orlicz norm of its derivative along tree paths.

Definition A.6 (Tree-based Orlicz IPM)**.** Let Φ be an N-function and let Ψ be its complementary function. For $\mu, \nu \in \mathcal{P}(\mathcal{T})$, define

$$\mathcal{S}_{\Phi}(\mu, \nu) := \sup \left\{ \left| \int_{\mathcal{T}} f(x)\, \mu(dx) - \int_{\mathcal{T}} f(x)\, \nu(dx) \right| : f \in WL^{\Psi}(\mathcal{T}, \omega),\ \|f'\|_{L^{\Psi}} \leq 1 \right\}. \tag{13}$$

This is an integral probability metric over a tree-based Orlicz-Sobolev critic class. The function Φ determines the induced geometry through the complementary constraint on f'.

Although Definition A.6 is variational, the tree structure yields a one-dimensional optimization problem. The following representation is a specialization of the graph-based Orlicz IPM formulation of [26] to the case where the underlying graph is a tree.

Theorem A.7 (Univariate representation; adapted from [26])**.** *Let Φ be an N-function and let Ψ be its complementary function. For any $\mu, \nu \in \mathcal{P}(\mathcal{T})$,*

$$\mathcal{S}_{\Phi}(\mu, \nu) = \inf_{a > 0} \frac{1}{a} \left(1 + \int_{\mathcal{T}} \Phi(a |h_{\mu,\nu}(x)|)\, \omega(dx) \right), \tag{14}$$

where

$$h_{\mu,\nu}(x) = \mu(\Lambda(x)) - \nu(\Lambda(x)). \tag{15}$$

This theorem reduces the infinite-dimensional optimization over critic functions to a scalar optimization over $a > 0$. The univariate formula simplifies when the input measures are supported on vertices.

Corollary A.8 (Discrete tree formula)**.** *Assume that $\omega(\{x\}) = 0$ for every $x \in \mathcal{T}$. Suppose that $\mu, \nu \in \mathcal{P}(\mathcal{T})$ are supported on the vertex set V. Then*

$$\mathcal{S}_{\Phi}(\mu, \nu) = \inf_{a > 0} \frac{1}{a} \left[1 + \sum_{e \in E} w_e \Phi\big(a |\bar{h}_{\mu,\nu}(e)|\big) \right], \tag{16}$$

where

$$\bar{h}_{\mu,\nu}(e) := \mu(\gamma_e) - \nu(\gamma_e), \qquad e \in E. \tag{17}$$

This formula follows because $h_{\mu,\nu}(x)$ is constant along the interior of each edge when μ and ν are supported on vertices.

We now state the special case where the N-function is a normalized power function. The normalized power case below follows as a direct special case of the univariate representation; see also [26].

Proposition A.9 (Power N-functions)**.** *Let*

$$\Phi_p(t) := \frac{(p-1)^{p-1}}{p^p} t^p, \qquad 1 < p < \infty. \tag{18}$$

Then the tree-based Orlicz IPM admits the closed-form expression

$$\mathcal{S}_{\Phi_p}(\mu,\nu) = \left(\int_{\mathcal{T}} |h_{\mu,\nu}(x)|^p \omega(dx) \right)^{1/p}. \tag{19}$$

Equivalently, if μ and ν are supported on the vertices of $\mathcal{T}$, then

$$\mathcal{S}_{\Phi_p}(\mu,\nu) = \left(\sum_{e \in E} w_e |\mu(\gamma_e) - \nu(\gamma_e)|^p \right)^{1/p}. \tag{20}$$

Lemma A.10. *Let Φ be a differentiable N-function and define*

$$q(z) := z\Phi'(z) - \Phi(z), \qquad z > 0.$$

Then q is continuous and nondecreasing, with

$$\lim_{z \to 0^+} q(z) = 0, \qquad \lim_{z \to \infty} q(z) = +\infty.$$

If Φ is strictly convex, then q is strictly increasing.

Proof. Since Φ is convex and differentiable, Φ' is continuous and nondecreasing on $(0,\infty)$. Hence q is continuous. For $0 < z_1 < z_2$, convexity gives

$$\Phi(z_2) - \Phi(z_1) \le \Phi'(z_2)(z_2 - z_1),$$

and therefore

$$q(z_2) - q(z_1) \ge z_1\big(\Phi'(z_2) - \Phi'(z_1)\big) \ge 0.$$

If Φ is strictly convex, Φ' is strictly increasing, so the last inequality is strict.

Since $\Phi(0) = 0$, the supporting-line inequalities at z give

$$0 \le q(z) \le \Phi(2z) - 2\Phi(z) \le \Phi(2z).$$

As $\Phi(2z) \to 0$ when $z \to 0^+$, we obtain $q(z) \to 0$.

For $z > 1$, convexity also yields

$$\Phi(1) \ge \Phi(z) + \Phi'(z)(1 - z),$$

so

$$q(z) \ge \Phi'(z) - \Phi(1).$$

Finally, $\Phi'(z) \ge \Phi(z)/z \to +\infty$ by the N-function condition. Thus $q(z) \to +\infty$. □

B Tree-Sliced Wasserstein Background on Euclidean Spaces

A line in $\mathbb{R}^d$ is specified by a pair $(x,\theta) \in \mathbb{R}^d \times \mathbb{S}^{d-1}$, where x is a reference point and θ is a direction. This representation induces the parametrization $x + t\theta$ for $t \in \mathbb{R}$. A system of $k \ge 1$ lines is then an element of the product space $(\mathbb{R}^d \times \mathbb{S}^{d-1})^k$, written as $\mathcal{L} = \{(x_i,\theta_i)\}_{i=1}^k$. The system is connected when the union of its constituent lines forms a connected subset of $\mathbb{R}^d$. To impose a tree structure, we remove selected intersection points so that any pair of points in the resulting configuration is linked by a unique simple path. The resulting object, denoted $\mathcal{T} = \{l_i\}_{i=1}^k$, satisfies this single-path property. Using the remaining intersections, we construct a topological tree by gluing line segments

of $\mathbb{R}$ via disjoint union and quotient topology [13], which yields a space endowed with a valid tree metric.

We now introduce a Radon transform adapted to tree systems. Let $L^1(\mathbb{R}^d)$ be the space of Lebesgue-integrable functions on $\mathbb{R}^d$. For a fixed tree system $\mathcal{T} = \{l_i\}_{i=1}^k$, let $\bar{\mathcal{T}}$ denote the union of all points lying on these lines. A function $f : \bar{\mathcal{T}} \to \mathbb{R}$ is integrable over $\bar{\mathcal{T}}$ if

$$\|f\|_{\mathcal{T}} := \sum_{i=1}^{k} \int_{\mathbb{R}} |f(t_x, l_i)| \, dt_x < \infty, \tag{21}$$

and we denote the corresponding space by $L^1(\mathcal{T})$. The standard simplex is given by

$$\Delta_{k-1} = \left\{ a = (a_1, \ldots, a_k) \in \mathbb{R}^k \;\middle|\; a_i \geq 0, \; \sum_{i=1}^{k} a_i = 1 \right\}. \tag{22}$$

Let $C(\mathbb{R}^d \times \mathbb{T}_k^d, \Delta_{k-1})$ denote the space of continuous splitting maps $\alpha : \mathbb{R}^d \times \mathbb{T}_k^d \to \Delta_{k-1}$. For $f \in L^1(\mathbb{R}^d)$ and α in this class, we define the projection operator

$$\mathcal{R}_{\mathcal{T}}^{\alpha} f(x, l_i) = \int_{\mathbb{R}^d} f(y) \, \alpha(y, \mathcal{T})_i \, \delta\big(t_x - \langle y - x_i, \theta_i \rangle\big) \, dy, \tag{23}$$

where (x_i, θ_i) defines the line l_i and δ is the Dirac delta. This operator satisfies $\mathcal{R}_{\mathcal{T}}^{\alpha} f \in L^1(\mathcal{T})$ and the bound $\|\mathcal{R}_{\mathcal{T}}^{\alpha} f\|_{\mathcal{T}} \leq \|f\|_1$, so it is well-defined as a mapping $\mathcal{R}_{\mathcal{T}}^{\alpha} : L^1(\mathbb{R}^d) \to L^1(\mathcal{T})$. Aggregating over all tree systems $\mathcal{T} \in \mathbb{T}_k^d$, we define the Radon transform on tree systems by

$$\mathcal{R}^{\alpha} : L^1(\mathbb{R}^d) \to \prod_{\mathcal{T} \in \mathbb{T}_k^d} L^1(\mathcal{T}), \qquad f \mapsto \big(\mathcal{R}_{\mathcal{T}}^{\alpha} f\big)_{\mathcal{T} \in \mathbb{T}_k^d}. \tag{24}$$

If the splitting map α is invariant under the Euclidean group $\mathrm{E}(d)$, then the transform $\mathcal{R}^{\alpha}$ is injective [59].

Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be probability measures. Given a tree-structured system of lines $\mathcal{T} \in \mathbb{T}$ and a splitting map α that is invariant under the Euclidean group $\mathrm{E}(d)$, we consider the pushforward measures $\mathcal{R}_{\mathcal{T}}^{\alpha}\mu$ and $\mathcal{R}_{\mathcal{T}}^{\alpha}\nu$ supported on $\mathcal{T}$. Since each $\mathcal{T}$ is equipped with a tree metric $d_{\mathcal{T}}$, we evaluate their discrepancy using the 1-Wasserstein distance

$$W_{d_{\mathcal{T}},1}\big(\mathcal{R}_{\mathcal{T}}^{\alpha}\mu, \mathcal{R}_{\mathcal{T}}^{\alpha}\nu\big). \tag{25}$$

Averaging this quantity over the space of tree systems yields the Tree-Sliced Wasserstein (TSW) distance:

$$\mathrm{TSW}(\mu, \nu) := \int_{\mathbb{T}} W_{d_{\mathcal{T}},1}\big(\mathcal{R}_{\mathcal{T}}^{\alpha}\mu, \mathcal{R}_{\mathcal{T}}^{\alpha}\nu\big) \, d\sigma(\mathcal{T}), \tag{26}$$

where σ is a probability measure over $\mathbb{T}$. Although the notation suppresses explicit dependence on α, $\mathbb{T}$, and σ, the resulting metric depends on all three components. In the special case where each tree system reduces to a single line, the construction recovers the classical sliced Wasserstein distance.

We now describe a class of Euclidean-invariant splitting maps. For any $x \in \mathbb{R}^d$ and $\mathcal{T} \in \mathbb{T}$, define the distance from x to a line l by

$$d(x, \mathcal{T})_l := \inf_{y \in l} \|x - y\|_2. \tag{27}$$

This quantity is invariant under the action of $\mathrm{E}(d)$. A practical choice for α is obtained via a softmax weighting over these distances:

$$\alpha(x, \mathcal{T}) = \mathrm{softmax}\big(\{-\xi \cdot d(x, \mathcal{T})_l\}_{l \in \mathcal{T}}\big), \tag{28}$$

where $\xi > 0$ controls the concentration of the weights across lines. Invariance of the averaged discrepancy holds under the assumptions of Theorem 3.2(iii).

C Theoretical Proofs

C.1 Proof of Theorem 3.2

Proof. Recall that

$$\text{TS-Orlicz}_\Phi(\mu,\nu) = \int_{\mathbb{T}} \mathcal{S}_\Phi(\mathcal{R}^\alpha_\mathcal{T}\mu, \mathcal{R}^\alpha_\mathcal{T}\nu)\ d\sigma(\mathcal{T}).$$

By the measurability assumption and nonnegativity of the tree-level discrepancies, this integral is well-defined with values in $[0,\infty]$.

Extended pseudometricity. Since $\mathcal{S}_\Phi$ is a tree-level extended metric, its values are nonnegative and

$$\mathcal{S}_\Phi(\mathcal{R}^\alpha_\mathcal{T}\mu, \mathcal{R}^\alpha_\mathcal{T}\mu) = 0.$$

Integrating over $\mathbb{T}$ gives

$$\text{TS-Orlicz}_\Phi(\mu,\nu) \geq 0, \qquad \text{TS-Orlicz}_\Phi(\mu,\mu) = 0.$$

Similarly, tree-level symmetry implies

$$\begin{aligned}\text{TS-Orlicz}_\Phi(\mu,\nu) &= \int_{\mathbb{T}} \mathcal{S}_\Phi(\mathcal{R}^\alpha_\mathcal{T}\mu, \mathcal{R}^\alpha_\mathcal{T}\nu)\ d\sigma(\mathcal{T}) \\ &= \int_{\mathbb{T}} \mathcal{S}_\Phi(\mathcal{R}^\alpha_\mathcal{T}\nu, \mathcal{R}^\alpha_\mathcal{T}\mu)\ d\sigma(\mathcal{T}) \\ &= \text{TS-Orlicz}_\Phi(\nu,\mu).\end{aligned}$$

For any $\mu,\nu,\rho \in \mathcal{P}(\mathbb{R}^d)$, the tree-level triangle inequality gives

$$\begin{aligned}\mathcal{S}_\Phi(\mathcal{R}^\alpha_\mathcal{T}\mu, \mathcal{R}^\alpha_\mathcal{T}\nu) \leq\ & \mathcal{S}_\Phi(\mathcal{R}^\alpha_\mathcal{T}\mu, \mathcal{R}^\alpha_\mathcal{T}\rho) \\ & + \mathcal{S}_\Phi(\mathcal{R}^\alpha_\mathcal{T}\rho, \mathcal{R}^\alpha_\mathcal{T}\nu)\,.\end{aligned}$$

Integrating this inequality yields

$$\text{TS-Orlicz}_\Phi(\mu,\nu) \leq \text{TS-Orlicz}_\Phi(\mu,\rho) + \text{TS-Orlicz}_\Phi(\rho,\nu).$$

All terms are nonnegative, so this inequality remains valid when one or more terms equal $+\infty$. Thus, TS-Orlicz is an extended pseudometric, proving (i).

Extended and finite metricity. Suppose that the sampled projections distinguish measures as assumed in (ii), and that

$$\text{TS-Orlicz}_\Phi(\mu,\nu) = 0.$$

The integrand is nonnegative, so

$$\mathcal{S}_\Phi(\mathcal{R}^\alpha_\mathcal{T}\mu, \mathcal{R}^\alpha_\mathcal{T}\nu) = 0 \qquad \text{for } \sigma\text{-almost every } \mathcal{T}.$$

By tree-level metricity, this implies

$$\mathcal{R}^\alpha_\mathcal{T}\mu = \mathcal{R}^\alpha_\mathcal{T}\nu \qquad \text{for } \sigma\text{-almost every } \mathcal{T}.$$

The measure-distinguishing assumption therefore gives $\mu = \nu$. Together with (i), this proves that TS-Orlicz is an extended metric.

On any class of measures for which

$$\int_{\mathbb{T}} \mathcal{S}_\Phi(\mathcal{R}^\alpha_\mathcal{T}\mu, \mathcal{R}^\alpha_\mathcal{T}\nu)\ d\sigma(\mathcal{T}) < \infty$$

for every pair μ,ν in the class, all metric values are finite. Its restriction to this class is consequently a finite metric, proving (ii).

Invariance. Fix $g \in E(d)$ satisfying the assumptions in (iii). For each tree system $\mathcal{T}$, denote the induced tree isometry by

$$g_{\mathcal{T}} : \mathcal{T} \longrightarrow g\mathcal{T}.$$

By assumption, this map preserves the root and satisfies

$$(g_{\mathcal{T}})_{\#}\omega_{\mathcal{T}} = \omega_{g\mathcal{T}}.$$

We first verify that this isometry preserves the tree-level Orlicz IPM. Let $f \in WL_{\Psi}(\mathcal{T}, \omega_{\mathcal{T}})$ and define

$$f^g := f \circ g_{\mathcal{T}}^{-1}.$$

If z_0 is the root of $\mathcal{T}$, then for every $x \in \mathcal{T}$,

$$\begin{aligned} f^g(g_{\mathcal{T}}x) - f^g(g_{\mathcal{T}}z_0) &= f(x) - f(z_0) \\ &= \int_{[z_0,x]} f'(y)\, \omega_{\mathcal{T}}(dy) \\ &= \int_{[g_{\mathcal{T}}z_0, g_{\mathcal{T}}x]} f'(g_{\mathcal{T}}^{-1}u)\, \omega_{g\mathcal{T}}(du). \end{aligned}$$

Here, the final equality follows because the root-preserving isometry maps $[z_0, x]$ onto $[g_{\mathcal{T}}z_0, g_{\mathcal{T}}x]$ and preserves the length measure. Consequently,

$$(f^g)' = f' \circ g_{\mathcal{T}}^{-1} \qquad \omega_{g\mathcal{T}}\text{-almost everywhere.}$$

For every $t > 0$, change of variables gives

$$\int_{g\mathcal{T}} \Psi\left(\frac{|(f^g)'(u)|}{t}\right) \omega_{g\mathcal{T}}(du) = \int_{\mathcal{T}} \Psi\left(\frac{|f'(y)|}{t}\right) \omega_{\mathcal{T}}(dy).$$

By the definition of the Luxemburg norm,

$$\|(f^g)'\|_{L_{\Psi}(g\mathcal{T}, \omega_{g\mathcal{T}})} = \|f'\|_{L_{\Psi}(\mathcal{T}, \omega_{\mathcal{T}})}.$$

Applying the same argument to $g_{\mathcal{T}}^{-1}$ shows that $f \mapsto f^g$ is a bijection between the admissible critic classes on the two trees.

Moreover, for probability measures η, ζ on $\mathcal{T}$,

$$\int_{g\mathcal{T}} f^g\, d(g_{\mathcal{T}})_{\#}\eta - \int_{g\mathcal{T}} f^g\, d(g_{\mathcal{T}})_{\#}\zeta = \int_{\mathcal{T}} f\, d\eta - \int_{\mathcal{T}} f\, d\zeta.$$

Taking the supremum over the corresponding admissible critics therefore yields

$$\mathcal{S}_{\Phi}((g_{\mathcal{T}})_{\#}\eta, (g_{\mathcal{T}})_{\#}\zeta) = \mathcal{S}_{\Phi}(\eta, \zeta).$$

Using projection equivariance, we obtain

$$\begin{aligned} &\mathcal{S}_{\Phi}\big(\mathcal{R}^{\alpha}_{g\mathcal{T}}(g_{\#}\mu), \mathcal{R}^{\alpha}_{g\mathcal{T}}(g_{\#}\nu)\big) \\ &\quad = \mathcal{S}_{\Phi}((g_{\mathcal{T}})_{\#}\mathcal{R}^{\alpha}_{\mathcal{T}}\mu, (g_{\mathcal{T}})_{\#}\mathcal{R}^{\alpha}_{\mathcal{T}}\nu) \\ &\quad = \mathcal{S}_{\Phi}(\mathcal{R}^{\alpha}_{\mathcal{T}}\mu, \mathcal{R}^{\alpha}_{\mathcal{T}}\nu)\,. \end{aligned}$$

Finally, since $\mathcal{T} \mapsto g\mathcal{T}$ preserves σ,

$$\begin{aligned} \text{TS-Orlicz}_{\Phi}(g_{\#}\mu, g_{\#}\nu) &= \int_{\mathbb{T}} \mathcal{S}_{\Phi}(\mathcal{R}^{\alpha}_{\mathcal{T}}(g_{\#}\mu), \mathcal{R}^{\alpha}_{\mathcal{T}}(g_{\#}\nu))\; d\sigma(\mathcal{T}) \\ &= \int_{\mathbb{T}} \mathcal{S}_{\Phi}\big(\mathcal{R}^{\alpha}_{g\mathcal{T}}(g_{\#}\mu), \mathcal{R}^{\alpha}_{g\mathcal{T}}(g_{\#}\nu)\big)\; d\sigma(\mathcal{T}) \\ &= \int_{\mathbb{T}} \mathcal{S}_{\Phi}(\mathcal{R}^{\alpha}_{\mathcal{T}}\mu, \mathcal{R}^{\alpha}_{\mathcal{T}}\nu)\; d\sigma(\mathcal{T}) \\ &= \text{TS-Orlicz}_{\Phi}(\mu, \nu). \end{aligned}$$

This proves (iii). □

C.2 Proof for Theorem 3.3

Proof. For each tree system $\mathcal{T} \in \mathbb{T}$, let $\mu_{\mathcal{T}}$ and $\nu_{\mathcal{T}}$ denote the projected measures induced by the Radon transform. Define the subtree mass discrepancy

$$h_{\mathcal{T}}(x) := \mu_{\mathcal{T}}(\Lambda(x)) - \nu_{\mathcal{T}}(\Lambda(x)). \tag{29}$$

Let $\mathcal{T}_{\mu,\nu} \subseteq \mathcal{T}$ denote the finite region of the induced tree that contains the root-to-support paths of $\mu_{\mathcal{T}}$ and $\nu_{\mathcal{T}}$. By assumption, $\omega_{\mathcal{T}}(\mathcal{T}_{\mu,\nu}) \leq C$ for some constant $C < \infty$ independent of $\mathcal{T}$. Since $h_{\mathcal{T}}(x) = 0$ outside $\mathcal{T}_{\mu,\nu}$, the power case can be written as

$$\mathcal{S}_{\Phi_p}(\mu_{\mathcal{T}}, \nu_{\mathcal{T}}) = \left(\int_{\mathcal{T}_{\mu,\nu}} |h_{\mathcal{T}}(x)|^p \, \omega_{\mathcal{T}}(dx) \right)^{1/p}. \tag{30}$$

For every $x \in \mathcal{T}_{\mu,\nu}$, since $\mu_{\mathcal{T}}$ and $\nu_{\mathcal{T}}$ are probability measures, we have

$$|h_{\mathcal{T}}(x)| = |\mu_{\mathcal{T}}(\Lambda(x)) - \nu_{\mathcal{T}}(\Lambda(x))| \leq 1. \tag{31}$$

Hence,

$$\lim_{p \to 1^+} |h_{\mathcal{T}}(x)|^p = |h_{\mathcal{T}}(x)|. \tag{32}$$

Moreover, for all $p \in (1, \infty)$,

$$0 \leq |h_{\mathcal{T}}(x)|^p \leq 1. \tag{33}$$

Since $\omega_{\mathcal{T}}(\mathcal{T}_{\mu,\nu}) \leq C$, the constant function 1 is integrable on $\mathcal{T}_{\mu,\nu}$. Therefore, by the dominated convergence theorem, for every fixed $\mathcal{T}$,

$$\lim_{p \to 1^+} \int_{\mathcal{T}_{\mu,\nu}} |h_{\mathcal{T}}(x)|^p \, \omega_{\mathcal{T}}(dx) = \int_{\mathcal{T}_{\mu,\nu}} |h_{\mathcal{T}}(x)| \, \omega_{\mathcal{T}}(dx). \tag{34}$$

Taking the p-th root preserves the limit because $p \to 1^+$. Therefore,

$$\begin{aligned} \lim_{p \to 1^+} \mathcal{S}_{\Phi_p}(\mu_{\mathcal{T}}, \nu_{\mathcal{T}}) &= \int_{\mathcal{T}_{\mu,\nu}} |h_{\mathcal{T}}(x)| \, \omega_{\mathcal{T}}(dx) \\ &= \int_{\mathcal{T}} |h_{\mathcal{T}}(x)| \, \omega_{\mathcal{T}}(dx) \\ &= \mathrm{W}_{d_{\mathcal{T}},1}(\mu_{\mathcal{T}}, \nu_{\mathcal{T}}), \end{aligned} \tag{35}$$

where the last equality follows from the closed-form expression of W_1 on tree metric spaces.

By definition of $\text{TS-Orlicz}_{\Phi_p}$,

$$\text{TS-Orlicz}_{\Phi_p}(\mu, \nu) = \int_{\mathbb{T}} \mathcal{S}_{\Phi_p}(\mu_{\mathcal{T}}, \nu_{\mathcal{T}}) \, d\sigma(\mathcal{T}). \tag{36}$$

Furthermore, for every $p > 1$,

$$\begin{aligned} \mathcal{S}_{\Phi_p}(\mu_{\mathcal{T}}, \nu_{\mathcal{T}}) &= \left(\int_{\mathcal{T}_{\mu,\nu}} |h_{\mathcal{T}}(x)|^p \, \omega_{\mathcal{T}}(dx) \right)^{1/p} \\ &\leq \omega_{\mathcal{T}}(\mathcal{T}_{\mu,\nu})^{1/p} \leq \max\{1, C\}. \end{aligned} \tag{37}$$

Thus, $\mathcal{S}_{\Phi_p}(\mu_{\mathcal{T}}, \nu_{\mathcal{T}})$ is dominated by the integrable constant $\max\{1, C\}$. Applying the dominated convergence theorem over $\mathbb{T}$ gives

$$\begin{aligned} \lim_{p \to 1^+} \text{TS-Orlicz}_{\Phi_p}(\mu, \nu) &= \int_{\mathbb{T}} \lim_{p \to 1^+} \mathcal{S}_{\Phi_p}(\mu_{\mathcal{T}}, \nu_{\mathcal{T}}) \, d\sigma(\mathcal{T}) \\ &= \int_{\mathbb{T}} \mathrm{W}_{d_{\mathcal{T}},1}(\mu_{\mathcal{T}}, \nu_{\mathcal{T}}) \, d\sigma(\mathcal{T}) \\ &= \mathrm{TSW}(\mu, \nu). \end{aligned} \tag{38}$$

□

C.3 Proof of Theorem 3.5

Proof. By definition,

$$\mathbb{E}_{\mathcal{T}\sim\sigma}\big[\mathcal{S}_\Phi(\mu_\mathcal{T},\nu_\mathcal{T})\big] = \text{TS-Orlicz}_\Phi(\mu,\nu).$$

Thus, with exact tree-level evaluations, the Monte Carlo estimator is unbiased. Independence of the sampled trees gives

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{\text{TS-Orlicz}}_\Phi - \text{TS-Orlicz}_\Phi\right)^2\right] &= \text{Var}\left(\widehat{\text{TS-Orlicz}}_\Phi\right)\\ &= \frac{1}{L^2}\sum_{i=1}^{L}\text{Var}(\mathcal{S}_\Phi(\mu_{\mathcal{T}_i},\nu_{\mathcal{T}_i})) = \frac{\sigma_\Phi^2}{L}.\end{aligned}$$

Finally, finite variance implies integrability, so almost-sure convergence follows from the strong law of large numbers. □

C.4 Proof for Theorem A.7

Proof. For completeness, we include the full derivation following the framework of [26].

By Definition A.6, we have

$$\mathcal{S}_\Phi(\mu,\nu) = \sup_{f\in\mathcal{F}}\left|\int_\mathcal{T} f(x)\,\mu(dx) - \int_\mathcal{T} f(x)\,\nu(dx)\right|, \tag{39}$$

where

$$\mathcal{F} := \left\{f \in WL^\Psi(\mathcal{T},\omega) : \|f'\|_{L^\Psi} \le 1\right\}. \tag{40}$$

For any $f\in\mathcal{F}$, the definition of $WL^\Psi(\mathcal{T},\omega)$ gives

$$f(x) - f(z_0) = \int_{[z_0,x]} f'(y)\,\omega(dy), \qquad x\in\mathcal{T}. \tag{41}$$

Therefore,

$$\begin{aligned}\int_\mathcal{T} f(x)\,\mu(dx) &= \int_\mathcal{T}\left[f(z_0) + \int_{[z_0,x]} f'(y)\,\omega(dy)\right]\mu(dx)\\ &= f(z_0) + \int_\mathcal{T}\int_\mathcal{T} \mathbf{1}_{\{y\in[z_0,x]\}} f'(y)\,\omega(dy)\,\mu(dx)\\ &= f(z_0) + \int_\mathcal{T} f'(y)\left[\int_\mathcal{T}\mathbf{1}_{\{y\in[z_0,x]\}}\mu(dx)\right]\omega(dy)\\ &= f(z_0) + \int_\mathcal{T} f'(y)\mu(\Lambda(y))\,\omega(dy),\end{aligned} \tag{42}$$

where the third equality follows from Fubini's theorem and the last equality follows from

$$\Lambda(y) = \{x\in\mathcal{T} : y\in[z_0,x]\}. \tag{43}$$

By the same argument,

$$\int_\mathcal{T} f(x)\,\nu(dx) = f(z_0) + \int_\mathcal{T} f'(y)\nu(\Lambda(y))\,\omega(dy). \tag{44}$$

Subtracting the two identities yields

$$\int_\mathcal{T} f\,d\mu - \int_\mathcal{T} f\,d\nu = \int_\mathcal{T} f'(y)\big[\mu(\Lambda(y)) - \nu(\Lambda(y))\big]\,\omega(dy). \tag{45}$$

Let

$$h_{\mu,\nu}(y) := \mu(\Lambda(y)) - \nu(\Lambda(y)). \tag{46}$$

Then

$$\mathcal{S}_\Phi(\mu,\nu) = \sup_{f\in\mathcal{F}} \left| \int_{\mathcal{T}} f'(y) h_{\mu,\nu}(y)\, \omega(dy) \right|. \tag{47}$$

We now rewrite the supremum over functions as a supremum over their tree derivatives. Clearly,

$$\{f' : f \in \mathcal{F}\} \subset \{g \in L^\Psi(\mathcal{T},\omega) : \|g\|_{L^\Psi} \le 1\}. \tag{48}$$

Conversely, for any $g \in L^\Psi(\mathcal{T},\omega)$ with $\|g\|_{L^\Psi} \le 1$, define

$$f_g(x) := \int_{[z_0,x]} g(y)\, \omega(dy). \tag{49}$$

Then $f_g \in WL^\Psi(\mathcal{T},\omega)$, $f_g' = g$, and $\|f_g'\|_{L^\Psi} \le 1$. Hence,

$$\{f' : f \in \mathcal{F}\} = \{g \in L^\Psi(\mathcal{T},\omega) : \|g\|_{L^\Psi} \le 1\}. \tag{50}$$

Consequently,

$$\mathcal{S}_\Phi(\mu,\nu) = \sup_{\substack{g\in L^\Psi(\mathcal{T},\omega)\\ \|g\|_{L^\Psi}\le 1}} \left| \int_{\mathcal{T}} g(y) h_{\mu,\nu}(y)\, \omega(dy) \right|. \tag{51}$$

Since Ψ is the complementary N-function of Φ, the above supremum is the Orlicz norm of $h_{\mu,\nu}$ associated with Φ. More precisely, by [51, Proposition 10, p. 81],

$$\mathcal{S}_\Phi(\mu,\nu) = \|h_{\mu,\nu}\|_\Phi, \tag{52}$$

where the Orlicz norm is defined by

$$\|h_{\mu,\nu}\|_\Phi := \sup\left\{ \int_{\mathcal{T}} |h_{\mu,\nu}(x) g(x)|\, \omega(dx) : \int_{\mathcal{T}} \Psi(|g(x)|)\, \omega(dx) \le 1 \right\}. \tag{53}$$

Finally, applying Rao and Ren [51, Theorem 13, p. 69], we obtain

$$\|h_{\mu,\nu}\|_\Phi = \inf_{a>0} \frac{1}{a} \left(1 + \int_{\mathcal{T}} \Phi\big(a|h_{\mu,\nu}(x)|\big)\, \omega(dx) \right). \tag{54}$$

Combining the last two displays gives

$$\mathcal{S}_\Phi(\mu,\nu) = \inf_{a>0} \frac{1}{a} \left(1 + \int_{\mathcal{T}} \Phi\big(a|h_{\mu,\nu}(x)|\big)\, \omega(dx) \right), \tag{55}$$

which proves the theorem. □

C.5 Proof for Proposition A.9

Proof. By Theorem A.7,

$$\mathcal{S}_{\Phi_p}(\mu,\nu) = \inf_{a>0} \frac{1}{a} \left(1 + \int_{\mathcal{T}} \Phi_p(a|h_{\mu,\nu}(x)|)\, \omega(dx) \right). \tag{56}$$

Substituting

$$\Phi_p(t) = \frac{(p-1)^{p-1}}{p^p} t^p, \tag{57}$$

we obtain

$$\mathcal{S}_{\Phi_p}(\mu,\nu) = \inf_{a>0} \left[\frac{1}{a} + \frac{(p-1)^{p-1}}{p^p} a^{p-1} \int_{\mathcal{T}} |h_{\mu,\nu}(x)|^p \omega(dx) \right]. \tag{58}$$

Let

$$A := \int_{\mathcal{T}} |h_{\mu,\nu}(x)|^p \omega(dx). \tag{59}$$

Then

$$\mathcal{S}_{\Phi_p}(\mu,\nu) = \inf_{a>0}\left[\frac{1}{a} + \frac{(p-1)^{p-1}}{p^p} A a^{p-1}\right]. \tag{60}$$

If $A = 0$, the infimum is zero; if $A = \infty$, it is infinite. We henceforth assume $0 < A < \infty$. Differentiating with respect to a,

$$-\frac{1}{a^2} + \frac{(p-1)^p}{p^p} A a^{p-2} = 0. \tag{61}$$

Hence the minimizer satisfies

$$a^p = \left(\frac{p}{p-1}\right)^p \frac{1}{A}, \qquad \text{i.e.} \qquad a^\star = \frac{p}{p-1} A^{-1/p}. \tag{62}$$

Substituting $a^\star$ back gives

$$\begin{aligned} \mathcal{S}_{\Phi_p}(\mu,\nu) &= \frac{p-1}{p} A^{1/p} + \frac{1}{p} A^{1/p} \\ &= A^{1/p}. \end{aligned} \tag{63}$$

Therefore,

$$\mathcal{S}_{\Phi_p}(\mu,\nu) = \left(\int_{\mathcal{T}} |h_{\mu,\nu}(x)|^p \omega(dx)\right)^{1/p}. \tag{64}$$

If μ and ν are supported on the vertices of $\mathcal{T}$, then $h_{\mu,\nu}(x)$ is constant along the interior of each edge, and thus

$$\int_{\mathcal{T}} |h_{\mu,\nu}(x)|^p \omega(dx) = \sum_{e\in E} w_e |\mu(\gamma_e) - \nu(\gamma_e)|^p. \tag{65}$$

This yields the discrete expression. □

D Spherical Tree-Sliced Orlicz Integral Probability Metric

D.1 Background: Spherical Tree-Sliced Wasserstein Distances

We next summarize the spherical version of the tree-sliced construction. [62] formulate this extension through three main ingredients: Spherical Tree Systems, the Spherical Radon Transform, and the resulting Spherical Tree-Sliced Wasserstein distance. The construction can be viewed as the hyperspherical counterpart of the Euclidean Tree Systems and Radon transforms developed in [63, 59]. We review these notions before introducing our method.

Hypersphere. The spherical construction is built on the d-dimensional unit hypersphere $\mathbb{S}^d$. Formally, $\mathbb{S}^d$ is the set of unit-norm vectors in $\mathbb{R}^{d+1}$:

$$\mathbb{S}^d := \left\{x \in \mathbb{R}^{d+1} : \|x\|_2 = 1\right\}. \tag{66}$$

We endow $\mathbb{S}^d$ with its geodesic distance. For two points $a, b \in \mathbb{S}^d$, this distance is the length of the shortest path on the spherical surface, equivalently the angle between a and b:

$$d_{\mathbb{S}^d}(a,b) = \arccos\left(\langle a,b\rangle_{\mathbb{R}^{d+1}}\right). \tag{67}$$

Stereographic projection provides a link between the spherical domain and a Euclidean hyperplane. Fix a pole $x \in \mathbb{S}^d$, and let $H_x := x^\perp = \{z \in \mathbb{R}^{d+1} : \langle x, z\rangle = 0\}$ be the orthogonal hyperplane through the origin. The stereographic projection from x maps every point $y \in \mathbb{S}^d \setminus \{x\}$ to H_x by

$$\begin{aligned} \varphi_x : \mathbb{S}^d \setminus \{x\} &\longrightarrow H_x, \\ y &\longmapsto -\frac{\langle x,y\rangle}{1-\langle x,y\rangle} x + \frac{1}{1-\langle x,y\rangle} y. \end{aligned} \tag{68}$$

In particular, $\varphi_x(-x) = 0$. By assigning $\varphi_x(x) = \infty$, the projection extends to a map from $\mathbb{S}^d$ to $H_x \cup \{\infty\}$.

Components of spherical tree systems. The basic geometric primitive of a spherical tree system is a spherical ray. It plays the same role as a line component in the Euclidean tree-sliced construction, but its support lies on the hypersphere. The construction uses stereographic projection to move between the curved spherical domain and the orthogonal hyperplane $H_x = x^\perp$. Given a root $x \in \mathbb{S}^d$, we first specify a Euclidean ray in H_x, and then map this ray back to $\mathbb{S}^d$ through the inverse stereographic projection.

For a unit direction $y \in H_x \cap \mathbb{S}^d$, the closed spherical ray rooted at x is defined by

$$r_y^x = \varphi_x^{-1}\left(\{t \cdot y : t \geq 0\} \cup \{\infty\}\right). \tag{69}$$

The endpoint $t = 0$ maps to $-x$, and $t = \infty$ maps to the root x. The inverse image is a closed semicircle parameterized by $s \mapsto \cos(s)x + \sin(s)y$, where $s \in [0, \pi]$ is its geodesic distance from x.

A spherical tree system is the abstract star formed from k labeled copies of these closed rays, identifying only their roots. More precisely,

$$\mathcal{T} = \big(\{1, \ldots, k\} \times [0, \pi]\big)/\!\sim, \qquad (i, 0) \sim (j, 0),$$

with no other identifications. For representatives (i, s) and (j, t), the tree distance is $|s - t|$ if $i = j$, and $s + t$ otherwise. Although the spherical images of all endpoints (i, π) equal $-x$, these are distinct leaves in the abstract tree. Thus the tree has unique paths and total length $k\pi$; its metric is the path metric on the labeled branches, not the ambient spherical distance between their images. We write $\mathbb{T}_k^d$ for the family of these trees, equipped with a sampling law σ.

Sampling spherical tree systems. We follow the sampling procedure of [62]. The root x is first drawn uniformly from the hypersphere, $x \sim \mathcal{U}(\mathbb{S}^d)$. To construct the ray directions, we sample independent initial vectors $v_i \sim \mathcal{U}(\mathbb{S}^d)$, independently of x, and set

$$y_i = \frac{v_i - \langle x, v_i\rangle x}{\|v_i - \langle x, v_i\rangle x\|_2}.$$

The zero-projection event has probability zero for $d \geq 1$; a vector is resampled if this occurs. Each direction is unit and orthogonal to x. Uniform sampling and orthogonal projection commute in law with every $g \in O(d + 1)$, so the resulting law satisfies $g_\# \sigma = \sigma$. On the labeled star, $(i, s) \mapsto (i, s)$ between the trees with parameters $(x, y_1, \ldots, y_k)$ and $(gx, gy_1, \ldots, gy_k)$ preserves the root and branch lengths.

Spherical Radon transform on spherical tree systems. Let $L^1(\mathbb{S}^d)$ denote the space of functions integrable with respect to spherical surface measure, denoted by dy. For a spherical tree system $\mathcal{T} \in \mathbb{T}_k^d$, we define $L^1(\mathcal{T})$ as the space of integrable functions $f : \mathcal{T} \to \mathbb{R}$ with finite tree norm

$$\|f\|_{\mathcal{T}} := \sum_{i=1}^{k} \int_0^\pi \left|f(t, r_{y_i}^x)\right| \, dt. \tag{70}$$

A splitting map specifies how mass on the sphere is distributed across the rays of a sampled tree. Formally, it is a continuous map $\alpha \in C(\mathbb{S}^d \times \mathbb{T}_k^d, \Delta_{k-1})$. Given such α, the spherical Radon transform associated with $\mathcal{T}$ maps a function $f \in L^1(\mathbb{S}^d)$ to a function $\mathcal{R}_{\mathcal{T}}^\alpha f \in L^1(\mathcal{T})$. At the point $(t, r_{y_i}^x)$ on the i-th ray, it is given by

$$\mathcal{R}_{\mathcal{T}}^\alpha f(t, r_{y_i}^x) = \int_{\mathbb{S}^d} f(y)\, \alpha(y, \mathcal{T})_i\, \delta(t - \arccos\langle x, y\rangle) \, dy. \tag{71}$$

Here, δ denotes the Dirac delta. This operator is well-defined as a map from $L^1(\mathbb{S}^d)$ to $L^1(\mathcal{T})$. For a Borel probability measure μ, its measure-valued extension is specified by

$$\mathcal{R}_{\mathcal{T}}^\alpha \mu(A) = \sum_{i=1}^{k} \int_{\mathbb{S}^d} \alpha(y, \mathcal{T})_i\, \mathbf{1}_A([i, \arccos\langle x, y\rangle]) \, \mu(dy),$$

where $[i, s]$ is the corresponding point of the quotient tree. In particular, mass at $-x$ is split among the distinct leaves $[i, \pi]$, while all contributions at x accumulate at the common root. By collecting the

outputs over all spherical tree systems, one obtains the global transform $\mathcal{R}^\alpha$, which maps a function on the hypersphere to its family of tree-based projections. When the splitting map α is invariant under the orthogonal group $O(d+1)$, this transform is injective [62].

Spherical Tree-Sliced Wasserstein distance. For probability measures $\mu, \nu \in \mathcal{P}(\mathbb{S}^d)$, the transform $\mathcal{R}^\alpha_\mathcal{T}$ induces projected measures $\mathcal{R}^\alpha_\mathcal{T}\mu$ and $\mathcal{R}^\alpha_\mathcal{T}\nu$ on the spherical tree $\mathcal{T}$. Since $\mathcal{T}$ is equipped with a tree metric $d_\mathcal{T}$, the discrepancy between the projected measures can be computed through the tree 1-Wasserstein distance. The Spherical Tree-Sliced Wasserstein distance averages these tree-level discrepancies over the space of spherical trees:

$$\mathrm{STSW}(\mu,\nu) := \int_{\mathbb{T}^d_k} W_1(\mathcal{R}^\alpha_\mathcal{T}\mu, \mathcal{R}^\alpha_\mathcal{T}\nu)\, d\sigma(\mathcal{T}). \tag{72}$$

With the sampling law above and an $O(d+1)$-invariant splitting map, the projected measures are equivariant: g preserves $\arccos\langle x, y\rangle$ and the splitting weights, and acts by the root- and length-preserving isometry of the labeled stars. Thus the averaged STSW discrepancy is $O(d+1)$-invariant. It is a metric when the sampled projections distinguish measures, as in the separation condition of Theorem D.2(ii).

Invariant splitting maps. The invariance property above uses both the invariant sampling law and an invariant splitting map, together with the induced tree isometries. The group $O(d+1)$ contains all linear transformations of $\mathbb{R}^{d+1}$ that preserve the Euclidean norm and therefore leave $\mathbb{S}^d$ unchanged. A splitting map $\alpha \in C(\mathbb{S}^d \times \mathbb{T}^d_k, \Delta_{k-1})$ is $O(d+1)$-invariant if

$$\alpha(gy, g\mathcal{T}) = \alpha(y, \mathcal{T}), \qquad \forall g \in O(d+1). \tag{73}$$

A convenient way to obtain an invariant splitting rule is to define the ray weights through quantities that are themselves invariant under rotations and reflections of the sphere. Following [62], for a point $y \in \mathbb{S}^d$ and a spherical tree $\mathcal{T} \in \mathbb{T}^d_k$, we define an invariant distance vector $\beta(y, \mathcal{T}) \in \mathbb{R}^k$. Its i-th coordinate measures the relative proximity of y to the i-th ray and is given by

$$\beta\big(y, \mathcal{T}^x_{y_1,\dots,y_k}\big)_i = \begin{cases} 0, & \text{if } y = \pm x, \\ \arccos\left(\frac{\langle y, y_i\rangle}{\sqrt{1-\langle x,y\rangle^2}}\right)\sqrt{1-\langle x,y\rangle^2}, & \text{if } y \neq \pm x. \end{cases} \tag{74}$$

The splitting map is then obtained by applying a softmax to the negative scaled distances:

$$\alpha(y,\mathcal{T})_i = \mathrm{softmax}\left(\{-\xi\,\beta(y,\mathcal{T})_j\}_{j=1}^k\right)_i. \tag{75}$$

This choice assigns larger mass to rays that are closer to y, while the invariance of β ensures that the resulting splitting rule respects the $O(d+1)$-symmetry of the spherical construction.

D.2 Spherical Tree-Sliced Orlicz Integral Probability Metric

The proposed TS-Orlicz framework extends to probability measures supported on the hypersphere by replacing the Euclidean tree-sliced projection with its spherical counterpart. Let $\mu, \nu \in \mathcal{P}(\mathbb{S}^d)$. Instead of using tree systems generated by Euclidean lines, the spherical construction projects the measures onto spherical tree systems whose branches are spherical rays.

For a sampled spherical tree $\mathcal{T}$, the Spherical Radon Transform $\mathcal{R}^\alpha_\mathcal{T}$ induces projected measures $\mu_\mathcal{T} := \mathcal{R}^\alpha_\mathcal{T}\mu$ and $\nu_\mathcal{T} := \mathcal{R}^\alpha_\mathcal{T}\nu$ on the tree metric space $\mathcal{T}$. We then evaluate the tree-level Orlicz IPM $\mathcal{S}_\Phi(\mu_\mathcal{T}, \nu_\mathcal{T})$. The spherical distance is obtained by aggregating these Orlicz-induced discrepancies over the distribution σ of spherical tree systems.

Definition D.1 (Spherical Tree-Sliced Orlicz IPM)**.** Let Φ be an N-function. For $\mu, \nu \in \mathcal{P}(\mathbb{S}^d)$, the Spherical Tree-Sliced Orlicz IPM is defined as

$$\text{STS-Orlicz}_\Phi(\mu,\nu) := \int_{\mathbb{T}} \mathcal{S}_\Phi(\mu_\mathcal{T}, \nu_\mathcal{T})\, d\sigma(\mathcal{T}). \tag{76}$$

Metricity and invariance of STS-Orlicz. The following result gives the spherical counterpart of Theorem 3.2.

Theorem D.2 (Metricity and invariance)**.** *Let Φ be an N-function. Assume that the tree-level Orlicz IPMs are well-defined extended metrics and that their evaluations are measurable in $\mathcal{T}$. Then:*

(i) STS-Orlicz is an extended pseudometric *on $\mathcal{P}(\mathbb{S}^d)$, allowing infinite values and possibly vanishing between distinct measures.*

(ii) If the sampled spherical projections distinguish measures, namely,

$$\mathcal{R}^{\alpha}_{\mathcal{T}}\mu = \mathcal{R}^{\alpha}_{\mathcal{T}}\nu \quad \text{for } \sigma\text{-almost every } \mathcal{T} \quad \Longrightarrow \quad \mu = \nu,$$

then STS-Orlicz is an extended metric*. It is a* finite metric *on any class of measures for which its defining integral is finite for every pair.*

(iii) Let $g \in O(d+1)$. Suppose that $\mathcal{T} \mapsto g\mathcal{T}$ preserves σ, induces a root- and length-measure-preserving tree isometry, and satisfies

$$\mathcal{R}^{\alpha}_{g\mathcal{T}}(g_{\#}\mu) = g_{\#}(\mathcal{R}^{\alpha}_{\mathcal{T}}\mu)\,,$$

where the right-hand pushforward uses the induced tree isometry. Then

$$\text{STS-Orlicz}_{\Phi}(g_{\#}\mu, g_{\#}\nu) = \text{STS-Orlicz}_{\Phi}(\mu, \nu).$$

The proof is given in Appendix D.3.

Connection to STSW. STS-Orlicz extends the Spherical Tree-Sliced Wasserstein distance by replacing the tree-level W_1 discrepancy with an Orlicz IPM on each projected spherical tree. For normalized power N-functions, the tree-level Orlicz IPM becomes an L^p-type aggregation of subtree mass discrepancies. As $p \to 1^+$, this tree-level discrepancy converges to the tree 1-Wasserstein distance. Therefore, the spherical construction recovers STSW in the near-L^1 limit.

Theorem D.3. *Let $\mu, \nu \in \mathcal{P}(\mathbb{S}^d)$, and let Φ_p be the normalized power N-function*

$$\Phi_p(t) := \frac{(p-1)^{p-1}}{p^p} t^p, \qquad p \in (1, \infty). \tag{77}$$

Then

$$\lim_{p \to 1^+} \textit{STS-Orlicz}_{\Phi_p}(\mu, \nu) = \textit{STSW}(\mu, \nu). \tag{78}$$

The proof is identical to the Euclidean recovery result in Theorem 3.3, with the integration taken over spherical tree systems.

Computation of Spherical Tree-Sliced Orlicz IPM. The integral over spherical tree systems is approximated by Monte Carlo sampling. Given independently sampled spherical trees $\mathcal{T}_1, \ldots, \mathcal{T}_L \sim \sigma$, we estimate STS-Orlicz$_\Phi$ as

$$\widehat{\text{STS-Orlicz}}_{\Phi}(\mu, \nu) = \frac{1}{L} \sum_{i=1}^{L} \widehat{S}_{\Phi}(\mu_{\mathcal{T}_i}, \nu_{\mathcal{T}_i})\,. \tag{79}$$

Here, $\mu_{\mathcal{T}_i}$ and $\nu_{\mathcal{T}_i}$ are the projected measures induced by the Spherical Radon Transform, and $\widehat{S}_\Phi$ is evaluated using the same tree-level scalar optimization described in Section 3.2.

The Monte Carlo guarantees in Theorem 3.5 also apply to STS-Orlicz when the spherical trees are sampled i.i.d. and the tree-level discrepancy has finite variance.

These guarantees concern exact tree-level evaluations; the scalar-solver error is controlled separately. For L sampled trees, k rays per tree, ambient dimension d, and n support points, the projection and sorting steps cost

$$O(Lkdn + Lkn \log n). \tag{80}$$

The Orlicz evaluation adds one scalar optimization on each sampled tree. With at most $I_{\max}$ log-space bisection iterations, this step costs

$$O(I_{\max} Lkn). \tag{81}$$

Therefore, the total computational complexity is

$$O(Lkdn + Lkn \log n + I_{\max} Lkn). \tag{82}$$

In practice, $I_{\max}$ is a small fixed constant, so the additional tree-level Orlicz optimization only introduces a constant-factor overhead. The leading asymptotic scaling therefore matches STSW while replacing the tree-level W_1 cost with an Orlicz IPM.

D.3 Theoretical Proofs for Spherical Tree-Sliced Orlicz Integral Probability Metric

Proof of Theorem D.2

Proof. Recall that

$$\text{STS-Orlicz}_\Phi(\mu,\nu) = \int_{\mathbb{T}} \mathcal{S}_\Phi(\mu_\mathcal{T}, \nu_\mathcal{T})\, d\sigma(\mathcal{T}),$$

where $\mu_\mathcal{T} = \mathcal{R}^\alpha_\mathcal{T}\mu$ and $\nu_\mathcal{T} = \mathcal{R}^\alpha_\mathcal{T}\nu$. By measurability and nonnegativity of the integrand, this integral is well-defined in $[0,\infty]$.

Extended pseudometricity. Tree-level nonnegativity and vanishing on the diagonal imply

$$\text{STS-Orlicz}_\Phi(\mu,\nu) \geq 0, \qquad \text{STS-Orlicz}_\Phi(\mu,\mu) = 0.$$

Similarly, tree-level symmetry gives

$$\begin{aligned}\text{STS-Orlicz}_\Phi(\mu,\nu) &= \int_{\mathbb{T}} \mathcal{S}_\Phi(\nu_\mathcal{T}, \mu_\mathcal{T})\, d\sigma(\mathcal{T}) \\ &= \text{STS-Orlicz}_\Phi(\nu,\mu).\end{aligned}$$

For any $\mu,\nu,\rho \in \mathcal{P}(\mathbb{S}^d)$, the tree-level triangle inequality yields

$$\mathcal{S}_\Phi(\mu_\mathcal{T}, \nu_\mathcal{T}) \leq \mathcal{S}_\Phi(\mu_\mathcal{T}, \rho_\mathcal{T}) + \mathcal{S}_\Phi(\rho_\mathcal{T}, \nu_\mathcal{T}).$$

Integrating over $\mathbb{T}$ gives

$$\text{STS-Orlicz}_\Phi(\mu,\nu) \leq \text{STS-Orlicz}_\Phi(\mu,\rho) + \text{STS-Orlicz}_\Phi(\rho,\nu).$$

All terms are nonnegative, so this inequality also holds when a term equals $+\infty$. Thus, STS-Orlicz is an extended pseudometric, proving (i).

Extended and finite metricity. Suppose that the measure-distinguishing condition in (ii) holds and that $\text{STS-Orlicz}_\Phi(\mu,\nu) = 0$. Nonnegativity of the integrand implies

$$\mathcal{S}_\Phi(\mu_\mathcal{T}, \nu_\mathcal{T}) = 0 \qquad \text{for } \sigma\text{-almost every } \mathcal{T}.$$

By tree-level metricity,

$$\mathcal{R}^\alpha_\mathcal{T}\mu = \mathcal{R}^\alpha_\mathcal{T}\nu \qquad \text{for } \sigma\text{-almost every } \mathcal{T}.$$

The measure-distinguishing condition therefore gives $\mu = \nu$. Together with (i), this proves extended metricity. Restricting to any class on which the defining integral is finite for every pair yields a finite metric, proving (ii).

Invariance. Fix $g \in O(d{+}1)$ satisfying the assumptions in (iii). Denote the induced root-preserving tree isometry by $g_\mathcal{T} : \mathcal{T} \to g\mathcal{T}$. Preservation of the length measure gives

$$(g_\mathcal{T})_\# \omega_\mathcal{T} = \omega_{g\mathcal{T}}.$$

For an admissible critic f on $\mathcal{T}$, define $f^g = f \circ g_\mathcal{T}^{-1}$. As in the proof of Theorem 3.2, preservation of root-to-point paths and the length measure implies

$$(f^g)' = f' \circ g_\mathcal{T}^{-1} \quad \omega_{g\mathcal{T}}\text{-almost everywhere},$$

and hence

$$\|(f^g)'\|_{L_\Psi(g\mathcal{T},\omega_{g\mathcal{T}})} = \|f'\|_{L_\Psi(\mathcal{T},\omega_\mathcal{T})}.$$

The inverse isometry gives the inverse correspondence, so this transformation is a bijection between the admissible critic classes. Moreover,

$$\int_{g\mathcal{T}} f^g\, d(g_\mathcal{T})_\# \eta = \int_\mathcal{T} f\, d\eta$$

for each tree-supported measure η, whenever the integral is defined. Taking suprema over the corresponding critics yields

$$\mathcal{S}_\Phi((g_\mathcal{T})_\#\mu_\mathcal{T}, (g_\mathcal{T})_\#\nu_\mathcal{T}) = \mathcal{S}_\Phi(\mu_\mathcal{T}, \nu_\mathcal{T}).$$

By projection equivariance, it follows that

$$\mathcal{S}_\Phi\big(\mathcal{R}^\alpha_{g\mathcal{T}}(g_\#\mu), \mathcal{R}^\alpha_{g\mathcal{T}}(g_\#\nu)\big) = \mathcal{S}_\Phi(\mu_\mathcal{T}, \nu_\mathcal{T}).$$

Finally, preservation of σ under $\mathcal{T} \mapsto g\mathcal{T}$ gives

$$\begin{aligned}
\text{STS-Orlicz}_\Phi(g_\#\mu, g_\#\nu) &= \int_{\mathbb{T}} \mathcal{S}_\Phi(\mathcal{R}^\alpha_\mathcal{T}(g_\#\mu), \mathcal{R}^\alpha_\mathcal{T}(g_\#\nu))\ d\sigma(\mathcal{T}) \\
&= \int_{\mathbb{T}} \mathcal{S}_\Phi\big(\mathcal{R}^\alpha_{g\mathcal{T}}(g_\#\mu), \mathcal{R}^\alpha_{g\mathcal{T}}(g_\#\nu)\big)\ d\sigma(\mathcal{T}) \\
&= \int_{\mathbb{T}} \mathcal{S}_\Phi(\mu_\mathcal{T}, \nu_\mathcal{T})\, d\sigma(\mathcal{T}) \\
&= \text{STS-Orlicz}_\Phi(\mu, \nu).
\end{aligned}$$

This proves (iii). □

Proof for Theorem D.3

Proof. For each spherical tree system $\mathcal{T} \in \mathbb{T}$, let $\mu_\mathcal{T}$ and $\nu_\mathcal{T}$ be the projected measures induced by the Spherical Radon Transform. Define the subtree mass discrepancy

$$h_\mathcal{T}(x) := \mu_\mathcal{T}(\Lambda(x)) - \nu_\mathcal{T}(\Lambda(x)). \tag{83}$$

For the normalized power N-function Φ_p, the tree-level Orlicz IPM admits the closed-form expression

$$\mathcal{S}_{\Phi_p}(\mu_\mathcal{T}, \nu_\mathcal{T}) = \left(\int_\mathcal{T} |h_\mathcal{T}(x)|^p\, \omega_\mathcal{T}(dx)\right)^{1/p}. \tag{84}$$

Since $\mu_\mathcal{T}$ and $\nu_\mathcal{T}$ are probability measures, we have

$$0 \le |h_\mathcal{T}(x)| \le 1, \qquad \forall x \in \mathcal{T}. \tag{85}$$

Hence, for every fixed $\mathcal{T}$ and $x \in \mathcal{T}$,

$$\lim_{p\to 1^+} |h_\mathcal{T}(x)|^p = |h_\mathcal{T}(x)|. \tag{86}$$

Moreover,

$$0 \le |h_\mathcal{T}(x)|^p \le 1. \tag{87}$$

By dominated convergence on the tree $(\mathcal{T}, d_\mathcal{T})$ of finite total length, we obtain

$$\lim_{p\to 1^+} \int_\mathcal{T} |h_\mathcal{T}(x)|^p\, \omega_\mathcal{T}(dx) = \int_\mathcal{T} |h_\mathcal{T}(x)|\, \omega_\mathcal{T}(dx). \tag{88}$$

Since $p \to 1^+$, taking the p-th root preserves the limit. Therefore,

$$\begin{aligned}
\lim_{p\to 1^+} \mathcal{S}_{\Phi_p}(\mu_\mathcal{T}, \nu_\mathcal{T}) &= \int_\mathcal{T} |h_\mathcal{T}(x)|\, \omega_\mathcal{T}(dx) \\
&= \mathrm{W}_{d_\mathcal{T},1}(\mu_\mathcal{T}, \nu_\mathcal{T}),
\end{aligned} \tag{89}$$

where the last equality follows from the closed-form expression of W_1 on tree metric spaces [28].

It remains to pass the limit through the integral over spherical tree systems. By construction, each spherical tree consists of k rays parameterized by $[0, \pi]$, so its total tree length is finite and uniformly bounded:

$$\omega_\mathcal{T}(\mathcal{T}) = k\pi, \qquad \forall \mathcal{T} \in \mathbb{T}. \tag{90}$$

Thus, for $p > 1$,

$$\begin{aligned}
\mathcal{S}_{\Phi_p}(\mu_\mathcal{T}, \nu_\mathcal{T}) &= \left(\int_\mathcal{T} |h_\mathcal{T}(x)|^p\, \omega_\mathcal{T}(dx)\right)^{1/p} \\
&\le (\omega_\mathcal{T}(\mathcal{T}))^{1/p} = (k\pi)^{1/p} \le k\pi.
\end{aligned} \tag{91}$$

Since σ is a probability measure on $\mathbb{T}$, the constant $k\pi$ is σ-integrable. Applying dominated convergence over $\mathbb{T}$, we get

$$\begin{aligned}\lim_{p\to 1^+} \text{STS-Orlicz}_{\Phi_p}(\mu,\nu) &= \lim_{p\to 1^+} \int_{\mathbb{T}} \mathcal{S}_{\Phi_p}(\mu_{\mathcal{T}},\nu_{\mathcal{T}})\, d\sigma(\mathcal{T}) \\ &= \int_{\mathbb{T}} \lim_{p\to 1^+} \mathcal{S}_{\Phi_p}(\mu_{\mathcal{T}},\nu_{\mathcal{T}})\, d\sigma(\mathcal{T}) \\ &= \int_{\mathbb{T}} \mathrm{W}_{d_{\mathcal{T}},1}(\mu_{\mathcal{T}},\nu_{\mathcal{T}})\, d\sigma(\mathcal{T}) \\ &= \mathrm{STSW}(\mu,\nu). \end{aligned} \tag{92}$$

This proves the claim. □

E Experimental Details

E.1 Runtime and Memory Analysis

Implementation cost. Table 6 reports the main computational blocks of TS-Orlicz and STS-Orlicz. The additional component introduced by our method is the evaluation of the tree-level Orlicz IPM. For general N-functions, this amounts to solving a scalar monotone equation on each sampled tree system by the log-space binary-search procedure in Algorithm 1. Each binary-search step only requires a linear scan over the already-computed edge imbalances. Hence, with at most $I_{\max}$ iterations, this step costs $O(I_{\max}Lkn)$, which is lower-order relative to the projection and sorting pipeline when $I_{\max}$ is fixed. The dominant costs remain the standard tree-sliced operations: projection, distance-based splitting, and sorting.

Scaling with support size and dimension. We next evaluate whether the observed runtime and memory follow the complexity predicted in Table 6. We benchmark TS-Orlicz under three choices of the N-function,

$$\Phi_{0,p}(t) = \frac{(p-1)^{p-1}t^p}{p^p} \text{ with } p=2, \qquad \Phi_1(t) = e^t - t - 1, \qquad \Phi_2(t) = e^{t^2} - 1.$$

For all runs, we use $L = 100$ sampled tree systems and $k = 10$ lines per tree system. Experiments are performed on a single NVIDIA H100 GPU and averaged over 10 independent runs. We vary the number of support points and the ambient dimension as

$$n \in \{50, 1000, 2000, 3000, 4000, 5000\}, \qquad d \in \{100, 200, 500, 1000\}.$$

Figures 1–3 show the runtime and memory usage for the three Orlicz choices. Across all cases, the empirical scaling is close to linear in both n and d. This agrees with the leading projection and splitting terms, which scale as $O(Lknd)$, together with the sorting term $O(Lkn\log n)$. The memory curves also follow the expected linear trend, consistent with the storage requirement $O(Lkn + Lkd + nd)$. These results indicate that the scalar Orlicz solver does not alter the practical scaling behavior of the tree-sliced computation.

Comparison with sliced and tree-sliced baselines. Finally, Table 5 compares the runtime of TS-Orlicz with representative sliced and tree-sliced methods. Here N and M denote the numbers of support points in the two input measures. For all tree-sliced methods, we use $L = 2500$ tree systems with $k = 4$ lines per tree system, while SW uses 10000 one-dimensional projections. As expected, SW is faster in absolute runtime because it only requires independent one-dimensional projections. Among tree-sliced methods, however, TS-Orlicz has a runtime profile comparable to Db-TSW. The normalized power variant TS-Orlicz$_{\Phi_{0,2}}$ is nearly identical to Db-TSW, while the non-polynomial variants TS-Orlicz$_{\Phi_1}$ and TS-Orlicz$_{\Phi_2}$ introduce only a small overhead. For example, when number of supports is 2000, Db-TSW takes 159.347 ms, whereas TS-Orlicz$_{\Phi_1}$ and TS-Orlicz$_{\Phi_2}$ take 167.549 ms and 168.920 ms, respectively. This confirms that the scalar root-solving step is lightweight relative to the shared projection, splitting, and sorting operations.

E.2 Gradient Flow

Gradient flow on synthetic data. Table 1 reports the performance of our proposed methods compared with several baselines on the 25 Gaussians and Gaussian 30D datasets. We use $L = 25$ trees and

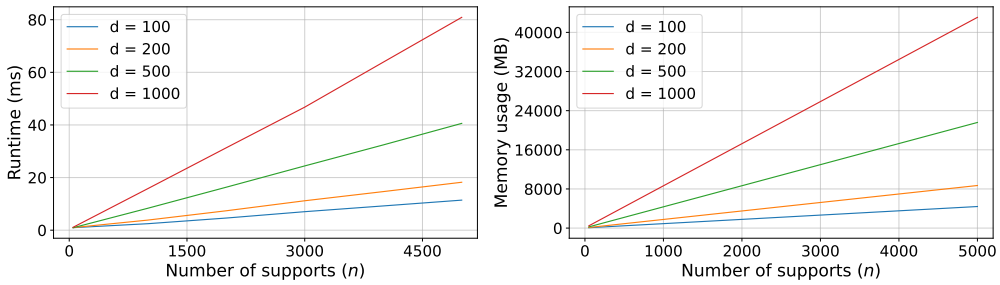

Figure 1: Runtime and memory usage of TS-Orlicz$_{\Phi_{0,2}}$

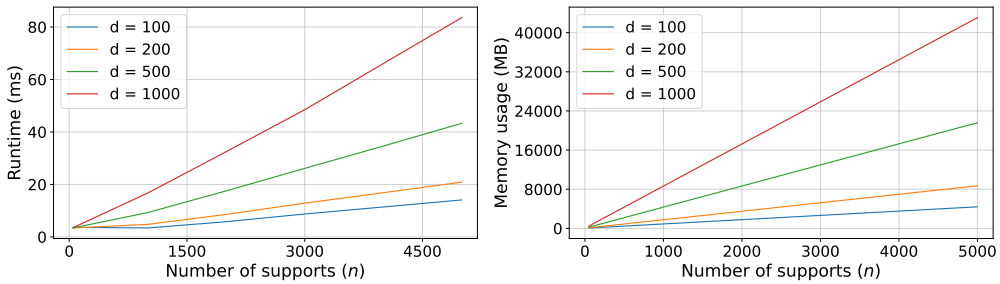

Figure 2: Runtime and memory usage of TS-Orlicz$_{\Phi_1}$.

Table 5: Runtime comparison between sliced, tree-sliced, and TS-Orlicz methods. We report mean runtime in milliseconds.

Number of supports	**SW**	**Db-TSW**	**TSW-SL**	**TS-Orlicz$_{\Phi_{0,2}}$**	**TS-Orlicz$_{\Phi_1}$**	**TS-Orlicz$_{\Phi_2}$**
50	0.375	4.683	4.682	4.362	4.942	4.994
100	0.407	8.762	8.765	8.473	9.279	9.359
200	0.826	16.795	16.762	16.541	17.778	17.943
300	0.931	24.846	24.822	24.614	26.354	26.592
400	1.125	32.661	32.595	32.347	34.581	34.921
500	1.186	40.778	40.778	40.485	43.248	43.647
1000	1.853	81.010	81.026	80.493	85.372	86.220
2000	3.330	159.347	159.393	158.439	167.549	168.920

$k = 4$ lines for tree-sliced methods and $L = 100$ projections for other sliced methods. We train for 2500 steps using the Adam optimizer with a global learning rate of 0.005. For the TSW variant, we use a learning rate of 0.005 for the 25 Gaussians dataset and 0.05 for the Gaussian 30D dataset, as in [59]. For TSW-SL, Db-TSW, Spatial-TSW and our TS-Orlicz variants, we use the same tree-sliced configuration. Each distribution has 500 samples.

Gradient flow on spherical data. We follow the standard von Mises–Fisher (vMF) mixture setup used in [5] and [60, 62]. For a mean direction $\mu \in \mathbb{S}^d$, the vMF density is

$$f(x; \mu, \kappa) = C_{d+1}(\kappa) \exp(\kappa \mu^\top x),$$

where $\kappa > 0$ controls the concentration and

$$C_{d+1}(\kappa) = \frac{\kappa^{(d-1)/2}}{(2\pi)^{(d+1)/2} I_{(d-1)/2}(\kappa)}$$

is the normalization constant in ambient dimension $d+1$, with $I_{(d-1)/2}$ denoting the modified Bessel function of the first kind.

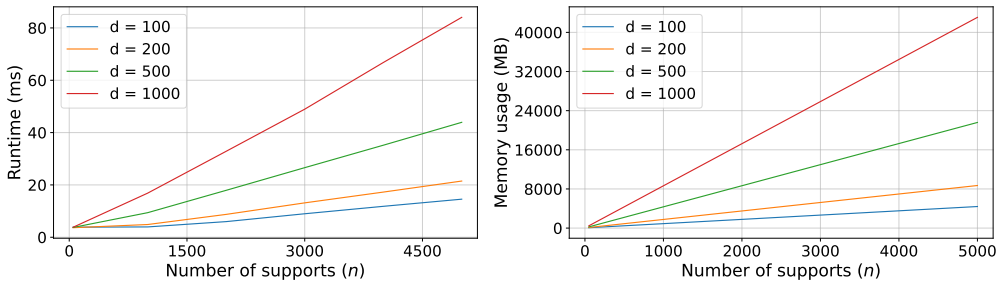

Figure 3: Runtime and memory usage of TS-Orlicz$_{\Phi_2}$.

Table 6: Complexity analysis of TS-Orlicz and STS-Orlicz. The tree-level Orlicz IPM for general N-functions is computed by log-space binary search with at most $I_{\max}$ iterations.

Distance	Operation	Description	Computation	Memory
TS-Orlicz	Sampling	Random sampling k-line tree systems	$O(Lkd)$	$O(Lkd)$
	Projection	Projecting n points onto L tree systems with k lines	$O(Lknd)$	$O(Lkn + Lkd + nd)$
	Distance-based weight splitting	Distance calculation and softmax over lines	$O(Lknd)$	$O(Lkn + Lkd + nd)$
	Sorting	Sorting projected coordinates on each line	$O(Lkn \log n)$	$O(Lkn)$
	Edge imbalance computation	Computing $\Delta_{\mathcal{T}}(e)$ over tree edges	$O(Lkn)$	$O(Lkn)$
	Root solving	Log-space binary search for the scalar Orlicz variable	$O(I_{\max} Lkn)$	$O(Lkn)$
	Total	General N-functions with binary search	$O(Lknd + Lkn \log n + I_{\max} Lkn)$	$O(Lkn + Lkd + nd)$
STS-Orlicz	Sampling	Random sampling spherical tree systems	$O(Lkd)$	$O(Lkd)$
	Projection	Projecting n spherical points onto sampled tree systems	$O(Lnd)$	$O(Ln + Ld + nd)$
	Distance-based weight splitting	Distance calculation and softmax over branches	$O(Lknd)$	$O(Lkn + Lkd + nd)$
	Sorting	Sorting projected coordinates	$O(Lkn \log n)$	$O(Lkn)$
	Edge imbalance computation	Computing $\Delta_{\mathcal{T}}(e)$ over tree edges	$O(Lkn)$	$O(Lkn)$
	Root solving	Log-space binary search for the scalar Orlicz variable	$O(I_{\max} Lkn)$	$O(Lkn)$
	Total	General N-functions with binary search	$O(Lknd + Lkn \log n + I_{\max} Lkn)$	$O(Lkn + Lkd + nd)$

The target distribution is a mixture of 12 vMF components with 2400 samples in total, i.e., 200 samples per component. We set $\kappa = 50$ and use the mean directions $\mu_i = v_i/\sqrt{1+\phi^2}$, where

$$v_1 = (-1, \phi, 0), \quad v_2 = (1, \phi, 0), \quad v_3 = (-1, -\phi, 0), \quad v_4 = (1, -\phi, 0),$$
$$v_5 = (0, -1, \phi), \quad v_6 = (0, 1, \phi), \quad v_7 = (0, -1, -\phi), \quad v_8 = (0, 1, -\phi),$$
$$v_9 = (\phi, 0, -1), \quad v_{10} = (\phi, 0, 1), \quad v_{11} = (-\phi, 0, -1), \quad v_{12} = (-\phi, 0, 1),$$

where

$$\phi = \frac{1+\sqrt{5}}{2}.$$

For tree-sliced spherical distances, we use $L = 200$ trees with $k = 5$ rays per tree. For the remaining sliced baselines, we use $L = 1000$ projections. ARI-S3W uses 30 rotations with a candidate pool of size 1000, while RI-S3W uses either 1 or 5 rotations. All methods are trained for 250 epochs with the Adam optimizer [21] and learning rate 0.01. For Tree-Sliced methods, we use a learning rate of 0.05.

Gradient Flow on Point Clouds. We evaluate our method on a point cloud interpolation task and compare it with SW [7], TSW-SL [63], Db-TSW [59] and Spatial-TSW [61]. The source and target point clouds are sampled from the ShapeNet Core-55 dataset [9]. Following the setup of [36], we use gradient approximation to perform Euler integration for 500 iterations with step size 0.01. Table 7 reports the Wasserstein distance between the interpolated point cloud and the target shape at iterations 100, 200, 300, 400, and 500, averaged over 5 runs. All sliced baselines use 100 projections, while tree-sliced methods use 25 trees with 4 lines per tree.

E.3 Diffusion Models

Diffusion Models. Diffusion models [55, 15] form a class of generative models that produce high-quality samples by learning to invert a gradual corruption process. Starting from a data sample x_0, the forward process adds Gaussian noise over T timesteps according to a fixed variance schedule $\{\beta_t\}_{t=1}^T$. Each noising step is defined by

$$q(x_t \mid x_{t-1}) = \mathcal{N}\Big(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I\Big).$$

The reverse process is learned with a neural network parameterized by θ, which maps a noisy sample x_t at timestep t to a distribution over the previous state x_{t-1}. This denoising transition is typically

Table 7: Comparison of methods across training steps.

Methods	Step 100	Step 200	Step 300	Step 400	Step 500
SW	**2.07e-04**	1.39e-04	1.31e-04	1.35e-04	1.28e-04
TSW-SL	3.12e-04	1.97e-04	2.08e-04	2.17e-04	2.19e-04
Spatial-TSW	4.09e-04	5.84e-05	2.07e-05	1.29e-05	1.16e-05
Db-TSW	3.87e-04	5.08e-05	**1.73e-05**	1.20e-05	9.06e-06
TS-Orlicz$_{\Phi_{0,1.1}}$	3.53e-04	**4.92e-05**	1.91e-05	**1.06e-05**	**6.91e-06**

modeled as

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I\big)\,.$$

Training proceeds by optimizing the evidence lower bound (ELBO), or equivalently by minimizing the Kullback–Leibler divergence between the learned reverse transition $p_\theta(x_{t-1} \mid x_t)$ and the true posterior $q(x_{t-1} \mid x_t)$.

Denoising Diffusion GANs. A major limitation of standard diffusion models is their slow sampling procedure, caused by the need to perform many sequential reverse steps. Denoising Diffusion GANs (DDGANs) [72] mitigate this bottleneck by replacing the conventional reverse denoising model with a conditional generative adversarial network at each timestep. This formulation permits more expressive reverse transitions, allowing the sampler to use far fewer steps, sometimes as few as four, while achieving substantial speedups with little degradation in sample quality.

The DDGAN training objective has also been improved beyond the original adversarial loss. [38] replace the standard adversarial objective with the Augmented Generalized Mini-batch Energy (AGME) objective. AGME is built upon the Generalized Mini-batch Energy (GME) objective [53], which compares two distributions through distances between empirical mini-batches. Following prior work, we use the GME objective with a base discrepancy D:

$$\mathrm{GME}_b^2(\mu, \nu) = 2\mathbb{E}[D(P_X, P_Y)] - \mathbb{E}[D(P_X, P_{X'})] - \mathbb{E}[D(P_Y, P_{Y'})]\,,$$

where P_X and P_Y are empirical measures constructed from mini-batches sampled from μ and ν, respectively. The geometry induced by the training objective is therefore determined by the choice of the base metric D. In this work, we instantiate D with Sliced Wasserstein (SW) baselines and our proposed Tree-Sliced Orlicz variants, allowing us to evaluate the effect of Orlicz-induced tree-sliced discrepancies within the DDGAN framework.

Implementation Details. Our experimental configuration follows [38] and [59] for the DDGAN architecture and core hyperparameters. All models are trained for 1800 epochs. For tree-sliced methods, including our TS-Orlicz variant, we use $L = 2500$ sampled trees, $k = 4$ lines per tree, and a tree-sampling standard deviation of 0.1, following [59]. For Sliced Wasserstein (SW) methods, we use $L = 10000$ projections, as in [38]. We adopt the same learning rates, setting $\mathrm{lr}_d = 1.25 \times 10^{-4}$ for the discriminator and $\mathrm{lr}_g = 1.6 \times 10^{-4}$ for the generator. Runtime is measured with batch size 256 on two NVIDIA H100 GPUs. For a fair comparison of computational cost, we additionally rerun each DDGAN variant for several epochs under the same hardware and training setup, and report the averaged per-epoch training time in Table 2. The baseline results are taken from previous reports, while our TS-Orlicz FID results are averaged over 5 runs under the same experimental protocol.

E.4 Self-Supervised Learning

Encoder. Following the standard spherical contrastive learning protocol [5, 60, 62], we train a ResNet18 encoder [14] on CIFAR-10 for 200 epochs with batch size 512. We use SGD with learning rate 0.05, momentum 0.9, and weight decay 10^{-3}. Positive pairs are generated using the augmentation pipeline in prior contrastive learning work [70, 5, 60], including resizing, random cropping, horizontal flipping, color jittering, and random grayscale conversion. For tree-sliced methods, including our STS-Orlicz variants, we use $L = 200$ trees. We use $k = 20$ rays per tree for STSW, Spatial-STSW, and all STS-Orlicz variants. For all non-tree sliced distances, we use $L = 200$ projections. We report the results in Table 3, with projected dimension $d = 9$. The regularization coefficients used for each projection dimension are reported in Table 8.

Linear Classifier. After pretraining, we freeze the encoder and train a linear classifier on the learned representations. Following the same evaluation protocol [5], the classifier is trained for 100 epochs

Table 8: Regularization coefficient λ across methods for projections on $\mathbb{S}^9$ in self-supervised learning.

	STS-Orlicz$_{\Phi_{0,2}}$	STS-Orlicz$_{\Phi_1}$	STS-Orlicz$_{\Phi_2}$	STSW	SSW	SW	S3W variants
$d = 9$	$\lambda = 9.0$	$\lambda = 7.0$	$\lambda = 5.0$	$\lambda = 10.0$	$\lambda = 20.0$	$\lambda = 1.0$	$\lambda = 0.5$

Table 9: Effect of the Orlicz function on Euclidean gradient flows. We report the average Wasserstein distance over 5 runs. Values for Gaussian 30D are scaled by 10^{-1}.

Method	**25 Gaussians Iteration**					**Gaussian 30D Iteration**				
	500	1000	1500	2000	2500	500	1000	1500	2000	2500
TS-Orlicz$_{\Phi_{0,1.1}}$	5.63e-4	**8.92e-7**	**7.64e-7**	**7.45e-7**	**6.43e-7**	2.22	1.66	1.10	5.91e-1	2.29e-1
TS-Orlicz$_{\Phi_{0,1.5}}$	8.72e-4	1.89e-6	1.52e-6	1.31e-6	1.31e-6	**2.20**	**1.61**	**1.00**	4.70e-1	1.38e-1
TS-Orlicz$_{\Phi_{0,1.6}}$	1.23e-3	2.35e-6	1.77e-6	1.64e-6	1.41e-6	2.21	1.63	1.04	5.05e-1	1.56e-1
TS-Orlicz$_{\Phi_{0,2}}$	1.51e-3	4.41e-6	3.23e-6	2.53e-6	2.57e-6	2.22	1.68	1.08	5.37e-1	1.76e-1
TS-Orlicz$_{\Phi_{0,4}}$	2.96e-3	7.49e-5	4.61e-5	3.99e-5	4.13e-5	2.38	2.13	1.89	1.65	1.40
TS-Orlicz$_{\Phi_{0,8}}$	9.03e-3	5.73e-3	4.80e-3	4.33e-3	4.23e-3	2.70	2.66	2.66	2.66	2.64
TS-Orlicz$_{\Phi_1}$	8.99e-4	1.92e-5	1.42e-5	1.40e-5	1.39e-5	2.23	1.70	1.13	5.96e-1	2.16e-1
TS-Orlicz$_{\Phi_2}$	1.19e-3	7.84e-5	8.68e-5	9.11e-5	7.45e-5	2.23	1.69	1.09	5.58e-1	1.93e-1
TS-Orlicz$_{\Phi_3}$	7.86e-4	2.69e-6	1.90e-6	1.71e-6	1.53e-6	2.22	1.64	1.03	4.94e-1	1.48e-1
TS-Orlicz$_{\Phi_4}$	**3.99e-4**	2.56e-6	2.08e-6	1.83e-6	1.67e-6	2.21	1.64	1.01	**4.59e-1**	**1.28e-1**

using Adam [21] with learning rate 10^{-3} and weight decay 0.2. The learning rate is decayed at epochs 60 and 80.

E.5 Effect of the Orlicz Function

We study the effect of the Orlicz function on Euclidean gradient-flow optimization. We keep the tree-sliced configuration fixed across all variants, using $L = 25$ sampled tree systems and $k = 4$ lines per tree. Each setting is averaged over 5 random seeds. We report the Wasserstein distance between the evolved source distribution and the target distribution at iterations $500, 1000, 1500, 2000, 2500$. For the Gaussian 30D benchmark, all reported values are scaled by 10^{-1}, following the convention in Table 1.

We consider the following Orlicz functions:

$$\Phi_{0,p}(t) = \frac{(p-1)^{p-1}}{p^p} t^p, \qquad p \in (1, \infty), \tag{93}$$

$$\Phi_1(t) = \exp(t) - t - 1, \tag{94}$$

$$\Phi_2(t) = \exp(t^2) - 1, \tag{95}$$

$$\Phi_3(t) = t \log(1 + t), \tag{96}$$

$$\Phi_4(t) = (1 + t) \log(1 + t) - t. \tag{97}$$

For the normalized power family $\Phi_{0,p}$, we evaluate $p \in \{1.1, 1.5, 1.6, 2, 4, 8\}$. This allows us to compare near-L^1, moderate-power, and high-power Orlicz geometries under the same tree-sliced projection mechanism. Table 9 shows that the choice of Φ changes the optimization behavior of TS-Orlicz in a task-dependent way. On the 25 Gaussians benchmark, the strongest late-stage performance is obtained by the near-L^1 power function $\Phi_{0,1.1}$, which reaches 6.43×10^{-7} at iteration 2500 and gives the best value from iteration 1000 onward. The nearby power functions $\Phi_{0,1.5}$ and $\Phi_{0,1.6}$ remain competitive, ending at 1.31×10^{-6} and 1.41×10^{-6}, respectively. The quadratic choice $\Phi_{0,2}$ is also effective, although its final value 2.57×10^{-6} is weaker than the lower-power variants. These results suggest that, for this low-dimensional multimodal gradient flow, mild power growth preserves the stable behavior of tree-sliced W_1-type geometry while still providing additional flexibility through the Orlicz formulation.

The behavior changes as the power exponent becomes large. The variants $\Phi_{0,4}$ and $\Phi_{0,8}$ converge much more slowly, reaching 4.13×10^{-5} and 4.23×10^{-3} at the final iteration. This does not imply

Table 10: Effect of the number of log-space bisection iterations $I_{\max}$ on Euclidean gradient-flow performance. We report the final Wasserstein distance at iteration 2500. Gaussian 30D values are scaled by 10^{-1}.

	25 Gaussians			**Gaussian 30D**		
Method	$I_{\max} = 6$	$I_{\max} = 10$	$I_{\max} = 20$	$I_{\max} = 6$	$I_{\max} = 10$	$I_{\max} = 20$
TS-Orlicz$_{\Phi_1}$	1.39e-5	1.40e-5	1.45e-5	2.16e-1	1.95e-1	2.08e-1
TS-Orlicz$_{\Phi_2}$	7.45e-5	7.85e-5	8.12e-5	1.93e-1	1.90e-1	1.98e-1
TS-Orlicz$_{\Phi_3}$	1.53e-6	1.57e-6	1.58e-6	1.48e-1	1.54e-1	1.64e-1
TS-Orlicz$_{\Phi_4}$	1.67e-6	1.64e-6	1.74e-6	1.28e-1	1.64e-1	1.49e-1

that high-growth Orlicz functions are universally undesirable. Rather, in this gradient-flow setting, high-order growth appears less stable for late-stage refinement. A possible explanation is that large powers emphasize large subtree discrepancies more strongly, which can help coarse corrections but may reduce the effectiveness of small local updates once the source distribution is already close to the target.

The non-polynomial functions provide additional evidence that the best geometry depends on the task and the stage of optimization. On 25 Gaussians, the logarithmic-growth functions Φ_3 and Φ_4 are stable and competitive, reaching 1.53×10^{-6} and 1.67×10^{-6}, respectively. The exponential-type functions Φ_1 and Φ_2 do not achieve the best final values in this particular gradient-flow task, but they still improve substantially over the initial distance and remain meaningful members of the Orlicz family. Their stronger growth can be useful in other settings where emphasizing larger discrepancies is beneficial, as also suggested by our results on downstream tasks.

The Gaussian 30D benchmark exhibits a different preference. During the early and middle stages, $\Phi_{0,1.5}$ gives the best values at iterations 500, 1000, and 1500. At later iterations, however, the entropy-logarithmic function $\Phi_4(t) = (1+t)\log(1+t) - t$ performs best, reaching 4.59×10^{-1} at iteration 2000 and 1.28×10^{-1} at iteration 2500. The logarithmic function $\Phi_3(t) = t\log(1+t)$ is also competitive, ending at 1.48×10^{-1}. This suggests that moderate-growth Orlicz functions can provide a favorable balance between sensitivity and stability in higher-dimensional Gaussian geometry.

Overall, this ablation shows that Φ acts as a geometry-control parameter rather than a universally ordered hyperparameter. Near-L^1 power functions are particularly effective for the 25 Gaussians gradient flow, while logarithmic-growth functions perform better in the late stage of Gaussian 30D. Very large power exponents are less stable in these gradient-flow experiments, but this should not be interpreted as a general weakness of stronger-growth Orlicz functions. In other settings, stronger growth can be beneficial; for example, in the DDGAN experiment in Table 2, the exponential-type choices Φ_1 and Φ_2 achieve lower mean FIDs than the L^1-type Db-TSW baselines. This supports using different Orlicz functions to adapt the tree-level geometry to the structure of the problem.

These results suggest that different Orlicz growth regimes induce different optimization behaviors and transport sensitivities, supporting the view of TS-Orlicz as a flexible family of adaptive transport geometries rather than a single fixed metric.

E.6 Sensitivity to Log-Space Bisection Steps

We further study the effect of the number of log-space bisection steps used in the tree-level scalar solver. Recall that for general Orlicz functions, the tree-level discrepancy is computed by solving a one-dimensional optimization problem. We approximate this problem by log-space bisection with a fixed maximum number of iterations $I_{\max}$. To quantify the numerical error, we compare the approximate value against a high-accuracy reference and compute

$$\mathrm{RelErr} = \frac{\left|\text{TS-Orlicz}_{\Phi} - \widehat{\text{TS-Orlicz}}_{\Phi}\right|}{|\text{TS-Orlicz}_{\Phi}| + 10^{-12}}. \tag{98}$$

Table 11 reports the relative error for $I_{\max} \in \{6, 10, 20\}$ on the Euclidean gradient-flow benchmarks. Even with only $I_{\max} = 6$, the approximation is already highly accurate. Across all tested Orlicz functions and both datasets, the maximum relative error remains below 1.4×10^{-3}. Increasing $I_{\max}$

Table 11: Sensitivity of the tree-level scalar solver to the number of log-space bisection steps $I_{\max}$. We report the mean and maximum relative error over evaluation checkpoints $\{500, 1000, 1500, 2000, 2500\}$.

		25 Gaussians			Gaussian 30D		
Method	**Error**	$I_{\max} = 6$	$I_{\max} = 10$	$I_{\max} = 20$	$I_{\max} = 6$	$I_{\max} = 10$	$I_{\max} = 20$
TS-Orlicz$_{\Phi_1}$	Mean	5.45e-4	2.13e-6	4.76e-8	1.66e-4	1.39e-6	2.78e-8
	Max	6.28e-4	2.32e-6	7.76e-8	2.28e-4	1.55e-6	4.49e-8
TS-Orlicz$_{\Phi_2}$	Mean	1.17e-3	4.80e-6	6.59e-8	7.69e-4	1.55e-6	3.76e-8
	Max	1.36e-3	4.97e-6	8.27e-8	8.37e-4	1.59e-6	7.30e-8
TS-Orlicz$_{\Phi_3}$	Mean	1.25e-4	4.80e-7	1.98e-8	3.47e-4	9.72e-7	2.09e-8
	Max	1.83e-4	6.22e-7	3.93e-8	4.87e-4	1.08e-6	4.40e-8
TS-Orlicz$_{\Phi_4}$	Mean	1.18e-4	4.54e-7	2.17e-8	1.15e-4	9.58e-7	3.39e-8
	Max	1.54e-4	6.34e-7	4.43e-8	1.87e-4	1.06e-6	4.81e-8

to 10 reduces the error to a maximum of 4.97×10^{-6}, while $I_{\max} = 20$ reduces the maximum to 8.27×10^{-8} across the reported settings.

We also evaluate whether this increased numerical accuracy affects downstream optimization. Table 10 shows that gradient-flow performance does not improve monotonically as $I_{\max}$ increases. On 25 Gaussians, for example, TS-Orlicz$_{\Phi_3}$ changes from 1.53×10^{-6} at $I_{\max} = 6$ to 1.58×10^{-6} at $I_{\max} = 20$, while TS-Orlicz$_{\Phi_4}$ remains within $[1.64, 1.74] \times 10^{-6}$. The Gaussian 30D benchmark also shows no consistent improvement, but some changes are larger. For instance, TS-Orlicz$_{\Phi_1}$ obtains final values 2.16×10^{-1}, 1.95×10^{-1}, and 2.08×10^{-1} for $I_{\max} = 6, 10, 20$, respectively. For TS-Orlicz$_{\Phi_4}$, the reported value rises from 1.28×10^{-1} at $I_{\max} = 6$ to 1.64×10^{-1} at $I_{\max} = 10$, an increase of 28.1%, before decreasing to 1.49×10^{-1} at $I_{\max} = 20$. Without variability estimates for these settings, the table does not establish that the downstream differences are negligible.

In these experiments, increasing $I_{\max}$ improves scalar-evaluation accuracy but does not systematically improve the final gradient-flow result. We use $I_{\max} = 6$ in the main experiments; its maximum reported relative error is 1.36×10^{-3}. This provides an empirical runtime–accuracy trade-off for the tested settings, rather than a general guarantee of downstream insensitivity to solver accuracy.

E.7 Choice of the scalar search interval.

The log-space bisection solver requires an initial interval $[a_{\min}, a_{\max}]$ for the auxiliary scale variable $a > 0$. This interval is used only for numerical bracketing and is not treated as a model hyperparameter. For the normalized power functions $\Phi_{0,p}$, the tree-level Orlicz IPM admits the closed-form expression in Proposition A.9; hence no scalar search is needed for these choices. The discussion below concerns the non-power Orlicz functions evaluated through scalar optimization.

For each sampled tree system $\mathcal{T}$, let

$$h_e := |\mu_{\mathcal{T}}(\gamma_e) - \nu_{\mathcal{T}}(\gamma_e)|, \qquad e \in E_{\mathcal{T}}, \tag{99}$$

denote the edge-wise subtree mass discrepancy. Since $\mu_{\mathcal{T}}$ and $\nu_{\mathcal{T}}$ are probability measures, we have $0 \le h_e \le 1$. If $h_e = 0$ for all $e \in E_{\mathcal{T}}$, then $S_\Phi(\mu_{\mathcal{T}}, \nu_{\mathcal{T}}) = 0$, and no scalar search is performed.

The scalar solver itself only requires Φ to be differentiable, since it uses the optimality function

$$G_{\mathcal{T}}(a) = \sum_{e \in E_{\mathcal{T}}} w_e \left[a h_e \Phi'(a h_e) - \Phi(a h_e) \right] - 1. \tag{100}$$

For the non-power Orlicz functions used with this solver in our experiments, the small-argument behavior is explicit. In particular,

$$\Phi(t) = c_\Phi t^2 + o(t^2), \qquad t \to 0, \tag{101}$$

with $c_\Phi = 1/2$ for $\Phi_1(t) = e^t - t - 1$ and $c_\Phi = 1$ for $\Phi_2(t) = e^{t^2} - 1$. The same quadratic initialization also applies to $\Phi_3(t) = t \log(1 + t)$ and $\Phi_4(t) = (1 + t) \log(1 + t) - t$, for which

$c_\Phi = 1$ and $c_\Phi = 1/2$, respectively. Substituting this local behavior into $G_\mathcal{T}$ gives the scale approximation

$$G_\mathcal{T}(a) \approx c_\Phi a^2 \sum_{e \in E_\mathcal{T}} w_e h_e^2 - 1. \tag{102}$$

This motivates the scale-adaptive initialization

$$a_0 = \left(c_\Phi \sum_{e \in E_\mathcal{T}} w_e h_e^2 + \varepsilon \right)^{-1/2}, \tag{103}$$

where $\varepsilon > 0$ is a small numerical constant. This local approximation is used only to initialize the numerical search interval; the final value is always computed from the original objective $F_\mathcal{T}(a)$.

Starting from a_0, we set

$$\eta_{\min}^{(0)} = \log a_0 - \log r, \qquad \eta_{\max}^{(0)} = \log a_0 + \log r, \tag{104}$$

where $r > 1$ is a fixed multiplicative bracket factor. Equivalently, the initial interval in the original scale is $[a_0/r, r a_0]$. Since $G_\mathcal{T}$ is monotone increasing, a valid bracket satisfies

$$G_\mathcal{T}(e^{\eta_{\min}}) \le 0 \le G_\mathcal{T}(e^{\eta_{\max}}). \tag{105}$$

If this condition is not initially satisfied, the interval is enlarged geometrically in log-space. Algorithm 1 describes the bracketing-and-bisection solver, which expands the interval until a valid bracket is obtained. In the actual implementation, we additionally cap the number of expansion steps and clip the resulting interval to broad numerical bounds

$$[\log a_{\text{low}}, \log a_{\text{high}}] \tag{106}$$

for stability. We use $a_{\text{low}} = 10^{-6}$ and $a_{\text{high}} = 10^4$ throughout.

These safeguards are purely numerical: they do not introduce additional model hyperparameters and do not change the objective being optimized. After the interval is initialized, the solver applies log-space bisection to the original optimality equation $G_\mathcal{T}(a) = 0$, and the final tree-level value is computed from the original objective $F_\mathcal{T}(a)$ rather than from the local approximation. Across all experiments, the same values of r, a_{low}, and a_{high} were used, and we did not observe failures of the scalar search interval.

E.8 Hardware Settings

Experiments used a single NVIDIA A100 GPU with 40GB of memory, except for the runtime and memory benchmarks, which used a single NVIDIA H100 GPU, and the denoising diffusion experiments, which used two NVIDIA H100 GPUs in parallel.

E.9 Broader Societal Impacts

This work is primarily methodological and theoretical. Its potential positive societal impact lies in improving scalable tools for comparing probability distributions in machine learning. Many applications, including generative modeling, representation learning, scientific simulation, computer vision, and data analysis, rely on meaningful distributional discrepancies. While optimal transport provides a powerful geometric framework, its computational cost can limit large-scale use. By introducing Tree-Sliced Orlicz IPM, this work provides a flexible class of tree-sliced discrepancies that captures richer geometric behavior while preserving computational efficiency. The proposed framework may benefit applications requiring efficient distribution comparison. In generative modeling, improved discrepancies may support more stable training and higher-quality synthetic data. In self-supervised learning, stronger distributional regularization may improve representation quality and downstream performance. More broadly, scalable distributional distances may help model complex high-dimensional, spherical, or structured empirical data more effectively. The Orlicz formulation further allows users to adapt the discrepancy to different data geometries, optimization objectives, or sensitivity requirements while remaining computationally practical through closed-form special cases or efficient scalar optimization.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction accurately reflect the paper's theoretical contributions and scope. They are supported by the main theorems and are stated under explicit assumptions.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Limitations of this work are discussed in Section 5.

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: We do provide the full set of assumptions and complete proofs for the main theorems.

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 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 conclusions of the paper.

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 open access to the data and code in Supplementary Material, with sufficient instructions to faithfully reproduce the main experimental results.

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: We specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results in Appendix E.

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: The paper reports suitably and correctly defined error bars.

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: We provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments.

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Broader Societal Impacts are discussed in Appendix E.9.

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: The paper does not release any data, pre-trained models, image generators, or scraped datasets. Therefore, safeguards for responsible release are not applicable.

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: The paper has credited all code and data that have been used.

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: The paper does not introduce new assets such as datasets, trained models, or software packages. Therefore, asset documentation is not applicable.

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: The paper does not involve crowdsourcing experiments or research with human subjects. Therefore, participant instructions, screenshots, and compensation details are not applicable.

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: The paper does not involve study participants or human-subject research. Therefore, participant risks, risk disclosure, and IRB approval are not applicable.

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: LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research.