Efficient Transferable Optimal Transport via Min-Sliced Transport Plans

Xinran Liu[*]
Harvard University
xinranliu@hsph.harvard.edu

Elaheh Akbari
Vanderbilt University
elaheh.akbari@vanderbilt.edu

Rocío Díaz Martín
Florida State University
rd25v@fsu.edu

Navid NaderiAlizadeh
Duke University
navid.naderi@duke.edu

Soheil Kolouri
Vanderbilt University
soheil.kolouri@vanderbilt.edu

Abstract

Optimal Transport (OT) plans provide correspondences between distributions supporting alignment tasks in various domains. Sliced transport plans have been recently proposed as a computationally efficient alternative to OT plans. These methods optimize a one-dimensional projection (slice) to obtain a conditional transport plan that minimizes the transport cost in the ambient space. Despite their efficiency, it remains unclear whether learned slicers transfer to new distribution pairs under shift, an issue central to evolving data and repeated OT computations over related distributions. We study the min-Sliced Transport Plan (min-STP) framework and examine slicer transferability: can a slicer learned on one distribution pair produce effective transport plans for unseen pairs? Theoretically, we show that optimized slicers remain close under slight perturbations of the data distributions, enabling efficient transfer across related tasks. To further improve scalability, we introduce a minibatch formulation of min-STP and provide statistical rates on its accuracy. Empirically, we demonstrate that the transferable min-STP achieves strong one-shot matching performance and facilitates amortized training for point cloud and image analysis. Our code is available at `https://github.com/xinranliueva/Min-STP`.

1 Introduction

Many tasks in computer vision, natural language processing, biology, and operations research involve establishing correspondences between samples drawn from two distributions; a problem elegantly framed through the lens of optimal transport (OT) [50, 36]. OT provides a principled framework for determining the most efficient coupling between distributions while minimizing a transport cost, consistent with the principle of least effort (or least action) in transferring mass. Over the past decade, OT has been applied across a broad range of domains, including domain adaptation [12, 14], generative modeling [5, 18, 28], shape and image matching [40, 45, 6], word-embedding alignment [54, 22], single-cell analysis [43, 53], and resource allocation [19], among many others [25].

Despite its success, OT remains computationally expensive, with a complexity that scales cubically in the number of samples [36]. To address this bottleneck, a large body of research has sought scalable approximations, including entropic regularization [13, 3], low-rank and linearized formulations [41, 51], and hierarchical or multiscale methods [44, 17, 20]. Slicing-based techniques [8, 7, 26] form another efficient class of methods, reducing high-dimensional problems to a set of one-dimensional

[*]Work done while at Vanderbilt University.

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

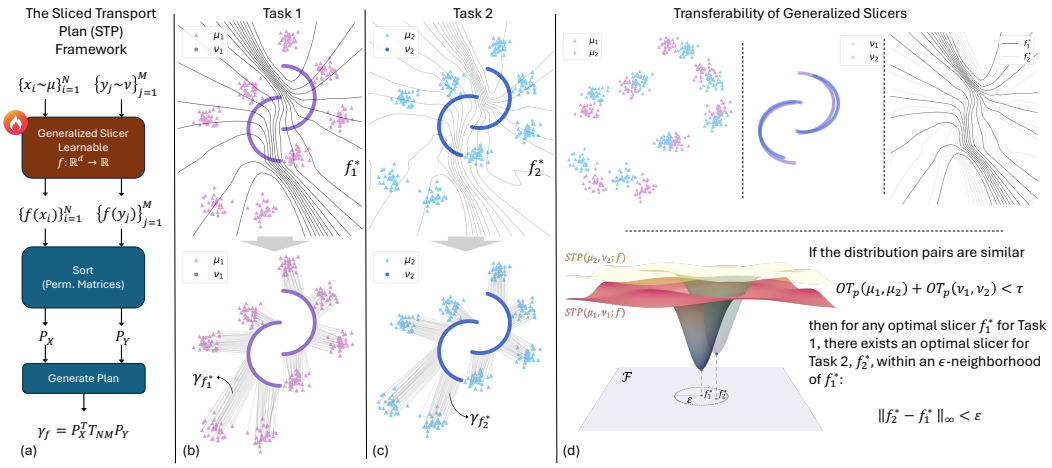

Figure 1: *Overview of the Sliced Transport Plan (STP) framework and our transferability results. (a–c) The STP framework computes transport plans γ_f using a generalized slicer $f : \mathbb{R}^d \to \mathbb{R}$, which projects high-dimensional samples onto one-dimensional marginals, sorts them via (soft) permutation matrices, and generates slice-wise transport plans that are lifted back to the ambient space to obtain a transportation plan between the input measures. The min-STP framework extends STP by learning an optimal slicer f^* that minimizes the transportation cost in the ambient space. (d) In this paper, we establish a* transferability theorem *showing that if two source-target distribution pairs (μ_1, ν_1) and (μ_2, ν_2) are close, then there exists an optimal slicer f_2^* for (μ_2, ν_2) in an ε-vicinity of f_1^*. This result enables efficient slicer reuse across related tasks, achieving amortized optimal transport.*

projections. While effective for computing OT distances, these methods do not recover explicit transportation plans. Recent extensions [39, 31, 30, 10] have begun to bridge this gap by constructing one-dimensional transportation plans along slices and lifting them to the ambient space to obtain transportation plans between input distributions.

However, most existing efficient OT methods are designed for single-instance problems. In many real-world applications, one must repeatedly solve OT between evolving or structurally similar distributions, where valuable information from prior solutions remains unused. Even when two pairs of distributions are nearly identical, current solvers recompute everything from scratch. This inefficiency is particularly limiting in domains such as developmental biology, where cell populations evolve gradually and successive distributions differ only slightly.

In this work, we move beyond isolated optimization and propose a formal framework for quantifying and leveraging the *transferability* of OT solutions. Our approach enables reusing information from previously solved instances, paving the way for adaptive, amortized OT across related problems. Building on differentiable generalized sliced plans [10], we revisit the min-generalized slice formulation and present a unified theoretical and algorithmic framework with numerically stable implementations. The *closed-form quantile function of the sum of Laplace distributions (LapSum)* [46] forms the backbone of both the theoretical and computational aspects of our method, allowing for efficient and differentiable evaluation of one-dimensional transport plans. Our central contribution is a *transferability theorem* that establishes a formal link between distributional similarity and slicer proximity: when two source-target distribution pairs are similar, their optimal slicers are correspondingly close. To the best of our knowledge, this result provides the first theoretical foundation for amortized slicer reuse. Finally, we demonstrate that the slicer can be trained stochastically on *mini-batches* ($n \ll N$), requiring only slice-and-sort operations on small subsets and enabling scalable optimization via stochastic gradient descent for large-scale datasets.

Contributions. We advance the theoretical foundations and practical scalability of large-scale OT problems through the following contributions:

1. Characterizing the transferability of optimized slicers under small source-target perturbations;

2. Deriving finite-sample statistical rates for slicers optimized using source-target mini-batches;

3. Demonstrating transferability on point cloud alignment, mean-flow generations and image translation.

2 Preliminaries

Definition 2.1. ***Optimal Transport (OT).*** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be endowed with a metric $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We use $\mathcal{P}(\mathcal{X})$ to denote the set of Borel probability measures defined on $\mathcal{X}$, and $\mathcal{P}_p(\mathcal{X})$ to denote the subset of probability measures with finite p^{th} moment ($p \in [1, \infty)$), i.e., $\mathcal{P}_p(\mathcal{X}) \coloneqq \{\mu \in \mathcal{P}(\mathcal{X}) \mid \int_{\mathcal{X}} c^p(x, x_0)\mathrm{d}\mu(x) < \infty$, for some $x_0 \in \mathcal{X}\}$. For $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$, the optimal transport distance (aka the p-Wasserstein distance) is defined as:*

$$\mathrm{OT}_p(\mu, \nu) \coloneqq \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c^p(x, y)\, \mathrm{d}\gamma(x, y) \right)^{\frac{1}{p}}, \tag{1}$$

where $\Gamma(\mu, \nu) \coloneqq \{\gamma \in \mathcal{P}_p(\mathcal{X} \times \mathcal{X}) \mid \pi^1_{\#}\gamma = \mu,\ \pi^2_{\#}\gamma = \nu\}$, π^i's denote the canonical projections, and $f_{\#}\mu$ denotes the pushforward of μ through f, which is formally defined as $f_{\#}\mu(A) = \mu(f^{-1}(A))$ for any Borel subset $A \subseteq \mathcal{X}$.

For $p \in [1, \infty)$, OT_p defines a proper metric between μ and ν, with many favorable geometric characteristics.

A prominent computational paradigm in optimal transport is the sliced-OT framework [38], which leverages the efficiency of OT [36, 6] solvers on one-dimensional measures by computing the OT distances between the one-dimensional marginals of d-dimensional measures. Owing to the injectivity of the Radon transform and the Cramér-Wold theorem (probabilistic route) or the Fourier Slice theorem (analytic viewpoint), comparing these marginals defines a valid metric between the original d-dimensional measures [32, 26].

Definition 2.2. ***Sliced OT.*** *Let $f : \mathbb{R}^d \times \mathbb{S}^{d-1} \to \mathbb{R}$ defined by the scalar projection $f(x, \theta) = \langle x, \theta\rangle$, where $\langle x, \theta\rangle$ denotes the standard inner product in $\mathbb{R}^d$. The sliced OT (SOT) problem between two probability measures $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$ is defined as:*

$$\mathrm{SOT}_p(\mu, \nu) \coloneqq \left(\int_{\mathbb{S}^{d-1}} \mathrm{OT}_p^p(f(\cdot, \theta)_{\#}\mu, f(\cdot, \theta)_{\#}\nu)\mathrm{d}\sigma(\theta) \right)^{\frac{1}{p}},$$

where $\sigma \in \mathcal{P}(\mathbb{S}^{d-1})$ is a probability measure with non-zero density everywhere on the d-dimensional unit sphere $\mathbb{S}^{d-1}$.

The notion of slicing was later extended to *generalized slices* by allowing $f(\cdot, \theta) : \mathbb{R}^d \times \mathbb{R}^{d'} \to \mathbb{R}$ to be a nonlinear parametric function. With a slight abuse of notation, we drop θ and write f when unambiguous. Under suitable assumptions on f [27, 11], it has been shown that SOT defines a proper metric between measures $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. Notably, extensive work has also been dedicated to the choice of σ, which is orthogonal to our discussion and is not discussed in this paper [34, 33, 35].

2.1 Sliced Transportation Plans

The central idea behind methods that estimate transportation plans via slicing is that the one-dimensional transportation plan between the pushforward measures $f_{\#}\mu$ and $f_{\#}\nu$, can be lifted back to define a coupling in the original space, denoted by γ_f. Following [30], in the case of discrete measures and injective projections f, this lifting is straightforward: the unique one-dimensional optimal plan calculated in the sliced space, $\sigma_f \in \Gamma(f_{\#}\mu, f_{\#}\nu)$, corresponds directly to a valid transportation plan for the original measures, $\gamma_f \in \Gamma(\mu, \nu)$. To the best of our knowledge, this idea of lifting via injective projections was first introduced by [39] as part of the Projected Wasserstein (PW) distance framework between empirical distributions and was later adopted in the influential work of [31]. When the projections $f_{\#}\mu$ and $f_{\#}\nu$ are not injective, however, the lifting becomes non-trivial and requires more careful handling, as discussed in [30].

Definition 2.3. ***Sliced transportation plan (STP).*** *Let γ_f denote the transportation plan between discrete probability measures $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, obtained by lifting the optimal 1D plan from slices via an injective map f defined on the supports of μ and ν. We define the sliced transportation plan dissimilarity between μ and ν as:*

$$\mathrm{STP}_p(\mu, \nu; f) \coloneqq \left(\sum_{x \in \mathrm{supp}(\mu)} \sum_{y \in \mathrm{supp}(\nu)} c^p(x, y)\, \gamma_f(x, y) \right)^{\frac{1}{p}}.$$

For general measures $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$, we formally write

$$\mathrm{STP}_p(\mu, \nu; f) \coloneqq \left(\int_{\mathcal{X} \times \mathcal{X}} c^p(x, y) \, \mathrm{d}\gamma_f(x, y) \right)^{\frac{1}{p}}$$

whenever γ_f is a well-defined lift coupling; that is, if σ_f denotes the unique optimal transport plan between $f_{\#}\mu$ and $f_{\#}\nu$, then γ_f is the uniquely determined element of $\Gamma(\mu, \nu)$ satisfying the property $(f, f)_{\#}\gamma_f = \sigma_f$. Such uniqueness is guaranteed, for instance, when f is injective.

When $f(\cdot) = \langle \cdot, \theta \rangle$, for a certain $\theta \in \mathbb{S}^{d-1}$, and the cost function is $c(x, y) = \|x - y\|_p$, STP_p coincides with the Sliced Wasserstein Generalized Geodesics (SWGG) distance [31], recently generalized for general measures [47].

Proposition 2.4. *(Proposition 2 in [10]) Let $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$. Let $f : \mathcal{X} \to \mathbb{R}$ be an injective map on $\mathcal{X}$. Then, for $p \geq 1$, $\mathrm{STP}_p(\cdot, \cdot\,; f)$ is a distance on $\mathcal{P}_p(\mathcal{X})$.*

Since γ_f is a valid coupling of μ and ν, though not necessarily optimal, the cost induced by γ_f upper bounds the infimum in equation 1, i.e.,

$$\mathrm{OT}_p(\mu, \nu) \leq \mathrm{STP}_p(\mu, \nu; f).$$

While STP_p provides an upper bound for OT_p, the gap between the two could be large due to the potential discrepancy between γ_f and an optimal transport plan γ^* for equation 1. This motivates the introduction of the min-STP_p, which coincides with the min-SWGG [31] and Differentiable Generalized Sliced Wasserstein Plan (DGSWP) defined by [10] in specific cases.

Definition 2.5. min-STP_p. *Let $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$. The minimum Sliced Transportation Plan (*$\min$-STP_p*) between μ and ν is defined as:*

$$\min\text{-}\mathrm{STP}_p(\mu, \nu) = \min_{f \in \mathcal{F}} \mathrm{STP}_p(\mu, \nu; f), \tag{2}$$

where $\mathcal{F}$ is a class of parametric functions.

Clearly, $\min$-STP_p satisfies:

$$\mathrm{OT}_p(\mu, \nu) \leq \min\text{-}\mathrm{STP}_p(\mu, \nu) \leq \mathrm{STP}_p(\mu, \nu; f).$$

Moreover, if $f : \mathcal{X} \to \mathbb{R}$ is injective on $\mathcal{X} \subset \mathbb{R}^d$, then $\min$-STP_p provides a pseudo-metric on $\mathcal{P}_p(\mathcal{X})$ ($p \geq 1$).

3 Method and Theoretical Results

3.1 Transferability

We first discuss the transferability of optimal slicers $f^\star : \mathcal{X} \to \mathbb{R}$ in problem 2. We show that when the source and target distributions experience small variations, the corresponding optimal slicer varies only marginally. This stability property ensures that $f^\star$ can be reliably transferred as a prior for the varied distributions.

Definition 3.1 (Perturbed slicers)**.** *Let $\mathsf{Lap}_d(\eta^2\Sigma)$ denote the symmetric multivariate Laplace distribution on $\mathbb{R}^d$ with mean 0 and covariance matrix $\eta^2\Sigma$, for a scale parameter $\eta > 0$ and a symmetric positive definite matrix $\Sigma \in \mathbb{R}^{d \times d}$. Given a measurable function $f : \mathcal{X} \to \mathbb{R}$, we define its perturbation by*

$$g_{\xi_\eta, f}(x) \;\coloneqq\; f(x) + \langle \xi_\eta, x \rangle, \text{ where } \xi_\eta \sim \mathsf{Lap}_d(\eta^2\Sigma).$$

Unless more than a single function f is involved, we will omit the subscript f and write g_{ξ_η}.

Throughout this paper, Σ stays fixed and we focus on varying scales η. This choice is motivated by LapSum [46], which applies Laplace perturbations for order-type smoothing, and it integrates naturally with LapSum's differentiable sorting mechanisms used in our implementations (see Section 3.2). Specifically, the perturbation ensures that, for any $\mu \in \mathcal{P}(\mathcal{X})$, with μ-probability one, the projection g_{ξ_η} is injective on the support of μ (see Lemma A.3 in the Appendix). This thereby enables the lift of a 1D coupling to the original space: Given $\mu, \nu \in \mathcal{P}(\mathcal{X})$, for each $\sigma \in \Gamma((g_{\xi_\eta})_{\#}\mu, (g_{\xi_\eta})_{\#}\nu) \subset \mathcal{P}(\mathbb{R} \times \mathbb{R})$, there exists one and only one (lifted) coupling $\gamma \in \Gamma(\mu, \nu) \subset \mathcal{P}(\mathcal{X} \times \mathcal{X})$ such that $(g_{\xi_\eta}, g_{\xi_\eta})_{\#}\gamma = \sigma$ (see Proposition A.4 in the Appendix).

Generalizing the classical projectors $f(x, \theta) = \langle x, \theta \rangle$ while preserving some of their fundamental properties, our assumptions are as follows:

Assumption 3.2. *Let $(\mathcal{X}, c)$ be a compact metric space in $\mathbb{R}^d$, and let $\mathcal{F}$ consist of all Lipschitz continuous functions $f : \mathcal{X} \to \mathbb{R}$ with global Lipschitz constant $L > 0$, and uniformly bounded by $M > 0$, i.e., $\mathcal{F} \coloneqq \{f \in C(\mathcal{X}) \mid Lip(f) \leq L, \|f\|_\infty \leq M\}$, where $Lip(f)$ denotes the Lipschitz constant of f. Specifically, we work with discrete measures and further require $\mathcal{X}$ to be discrete with at most countable samples.*

Under these assumptions, the slicer family $\mathcal{F}$ is compact in the uniform norm (as a consequence of the Arzelà–Ascoli theorem; see Lemma A.1 in the Appendix). Thus, minimizers of regular functionals are attained over $\mathcal{F}$. In particular, we next introduce an objective obtained by applying Laplace perturbations to STP, which has better regularity properties and recovers STP in the limit as the perturbations vanish.

Definition 3.3 (Expected lifted plan)**.** *Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$, and $f \in \mathcal{F}$ with perturbation g_{ξ_η} for some $\eta > 0$. Let σ_{ξ_η} denote the 1D optimal plan between $(g_{\xi_\eta})_\# \mu$ and $(g_{\xi_\eta})_\# \nu$, with corresponding (unique) lifted coupling $\gamma_{\xi_\eta} \in \Gamma(\mu, \nu)$, such that $(g_{\xi_\eta}, g_{\xi_\eta})_\# \gamma_{\xi_\eta} = \sigma_{\xi_\eta}$. We define the expected lifted plan and the smoothed objective by*

$$\bar{\gamma}_f^{(\eta)} \coloneqq \mathbb{E}_{\xi_\eta}\big[\,\gamma_{\xi_\eta}\,\big], \qquad J_\eta(\mu, \nu; f) \coloneqq \int_{\mathcal{X}^2} c(x, y)^p \, \mathrm{d}\bar{\gamma}_f^{(\eta)}(x, y). \tag{3}$$

Theorem 3.4. *[Properties]*

1. *Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$, $f \in \mathcal{F}$. Then, $\lim_{\eta \to 0} J_\eta(\mu, \nu; f) = \mathrm{STP}_p^p(\mu, \nu; f)$.*
2. *Fix $\eta > 0$ and $\mu_2, \nu_2 \in \mathcal{P}(\mathcal{X})$. Given $\varepsilon > 0$, there exists $\tau > 0$ such that for any $\mu_1, \nu_1 \in \mathcal{P}(\mathcal{X})$ with $\mathrm{OT}_p(\mu_1, \mu_2) + \mathrm{OT}_p(\nu_1, \nu_2) < \tau$, and every $f_1^\star \in \arg\min_{f \in \mathcal{F}} J_\eta(\mu_1, \nu_1; f)$, there exists $f_2^\star \in \arg\min_{f \in \mathcal{F}} J_\eta(\mu_2, \nu_2; f)$ with $\|f_2^\star - f_1^\star\|_\infty < \varepsilon$.*

For the first property, we refer to Proposition A.7 in the Appendix, which uses the following auxiliary inequalities: for $f, g \in \mathcal{F}$, $\kappa, \kappa' \in \mathcal{P}(\mathcal{X})$, and $\mu_i, \nu_i \in \mathcal{P}(\mathbb{R})$ with optimal couplings γ_i, $i = 1, 2$,

- $\mathrm{OT}_p\big(f_\# \kappa,\ f_\# \kappa'\big) \leq L \cdot \mathrm{OT}_p(\kappa, \kappa')$
- $\mathrm{OT}_p\big(f_\# \kappa,\ g_\# \kappa\big) \leq \|f - g\|_\infty$
- $\mathrm{OT}_p^p\big(\gamma_1, \gamma_2\big) \leq \mathrm{OT}_p(\mu_1, \mu_2)^p + \mathrm{OT}_p(\nu_1, \nu_2)^p$

(see Lemmas A.5 and A.6 in the Appendix).

The second property in Theorem 3.4 is *transferability*. It holds as an application of Berge's Maximum Theorem [2] using compactness of $\mathcal{F}$ (Model Assumption 3.2) together with the continuity of J_η. Indeed, Proposition A.8 in the Appendix establishes that J_η is continuous as a function on $(\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})) \times \mathcal{F}$, and its proof relies on the three inequalities listed above.

For an illustration, see Figure 1 (d). For a preliminary discussion on the relationship between τ and ε, we refer the reader to the Appendix.

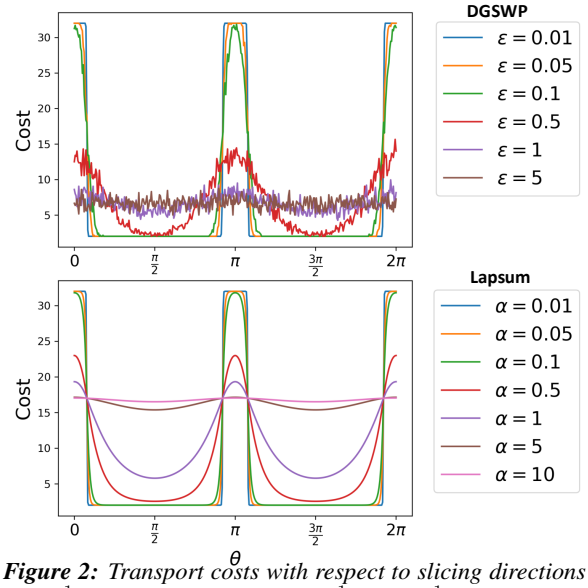

Figure 2: *Transport costs with respect to slicing directions $\theta \in \mathbb{S}^1$ for distribution pair $\mu = \frac{1}{2}\delta_{[0,0]} + \frac{1}{2}\delta_{[1,1]}$ and $\nu = \frac{1}{2}\delta_{[0,1]} + \frac{1}{2}\delta_{[1,0]}$, using DGSWP [10] and LapSum [46]. For gradient behaviors, please refer to E.*

3.2 Smoothed Sorting (LapSum)

An essential step in minimizing the STP objective is computing the one-dimensional optimal transport plan on $\mathbb{R}$. In one dimension, the OT problem reduces to sorting, which is inherently non-differentiable. To enable gradient-based optimization, a differentiable variant of the STP framework was introduced in [10].

The authors proposed the differentiable generalized sliced transport plan (DGSWP), which formulates the search for a sliced plan as a bilevel optimization problem and smooths the outer, non-differentiable STP objective by injecting Gaussian noise into the slice parameters and estimating gradients using Stein's lemma. While this yields a differentiable surrogate, it suffers in practice from high variance when naive Monte Carlo estimators are used, as noted by the authors in [10] and shown in Figure 2. To mitigate this, they employed control variates as a variance-reduction technique to stabilize training. Nevertheless, the computational overhead of repeatedly solving one-dimensional OT problems for every perturbed slicer remains a significant bottleneck.

We adopt a different approach to inject differentiability, which aligns with the smooth surrogate J_η proposed in Definition 3.3 and avoids the extra cost of repeated 1-D OT problems. This is achieved via LapSum, the differentiable sorting method introduced in [46]. Let $x_1, \ldots, x_n$ denote n scalars in $\mathbb{R}$ with the corresponding empirical cumulative distribution function (CDF)

$$\hat{F}_x(t) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{x_i \le t\}.$$

Then the rank of x_i and sorted k-th value can be written as

$$r_i \;=\; n\,\hat{F}_x(x_i), \quad x_{(k)} \;=\; \hat{F}_x^{-1}(k/n)\,, \;\; k = 1, \ldots, n.$$

LapSum injects i.i.d. Laplace noise $\xi_i \sim \mathsf{Lap}(0, \alpha)$ and smooths each of the hard functions $\mathbf{1}\{x_i \le t\}$ as the CDF of $\mathsf{Lap}(0, \alpha)$, denoted by Φ_α, resulting in a smoothed CDF $\hat{F}_{x,\alpha}(t) \;=\; \frac{1}{n}\sum_{i=1}^{n} \Phi_\alpha(t - x_i)$. This yields differentiable surrogates for rank and sorting: $\tilde{r}_i = n\hat{F}_{x,\alpha}(x_i)$ and $\tilde{x}_{(k)} = \hat{F}^{-1}_{x,\alpha}\big(\frac{k}{n}\big)$, with $\hat{F}_{x,\alpha} \to \hat{F}_x$ and $\tilde{x}_{(k)} \to \mathrm{sort}(x)$ as $\alpha \to 0$. The elegance of LapSum is that the inverse $\hat{F}^{-1}_{x,\alpha}$ can be calculated in closed-form and the algorithm has a complexity of only $\mathcal{O}(n \log n)$ (see Algorithm 2 in the Appendix for deriving LapSum-based soft permutation).

In our smooth surrogate in Definition 3.1, since the noise $\xi \sim \mathsf{Lap}_d(\eta^2\Sigma)$, each projection $\langle \xi, x\rangle$ is a 1D Laplace distribution with scale $b_x = \eta\sqrt{\frac{1}{2}\, x^\top \Sigma x}$ (see Lemma A.2 in the Appendix). In other words, we inject Laplacian noise into each of the one-dimensional samples and calculate the expectation of plans in one shot by LapSum. This closely connects to the probabilistic assumptions of LapSum's soft comparisons. In practice, we conveniently adopt a fixed global scale for the Laplace noise across all samples, which allows for absorbing the perturbations in the calculations of the 1D optimal transport plan using LapSum, notably with a time complexity of only $\mathcal{O}(n \log n)$ and $\mathcal{O}(n)$ memory requirement.

3.3 Mini-batch Training

Training a slicer network on large-scale datasets is computationally demanding because each update requires evaluating the full STP objective. We mitigate this cost by optimizing a mini-batched STP loss, which enables fast iterations. Theoretically, we establish a stability and convergence guarantee demonstrating how the mini-batch objective approaches the full-batch objective as the batch size becomes large.

Let $X = \{x_i\}_{i=1}^N$ and $Y = \{y_j\}_{j=1}^M$ be two sets of discrete samples, and define $[n] \coloneqq \{1, \ldots, n\}$. The discrete STP objective is given by $J_{N,M}(f) \coloneqq \sum_{i=1}^N \sum_{j=1}^M \gamma_{ij}\, c(x_i, y_j)^p$, where γ denotes the lifted transport plan by $f : \mathbb{R}^d \to \mathbb{R}$. Fix a batch size $B \le \min(N, M)$ and let $\binom{[n]}{B}$ represent the set of all possible batches of B indices drawn from $[n]$ without replacement. For $S \in \binom{[N]}{B}, T \in \binom{[M]}{B}$ and their corresponding batches $x_S = (x_i)_{i\in S}$ and $y_T = (y_j)_{j\in T}$, define the kernel with slicer $f : \mathbb{R}^d \to \mathbb{R}$ as $h_B(f; x_S, y_T) \coloneqq \frac{1}{B}\sum_{i=1}^B c\big(x^f_{(i)}, y^f_{(i)}\big)^p$, where $x^f_{(i)}, y^f_{(i)} \in \mathbb{R}^d$ denote the re-ordered x_S and y_T, paired by the lifted coupling from 1D coupling, i.e., $f(x^f_{(1)}) \le \cdots \le f(x^f_{(B)})$ and $f(y^f_{(1)}) \le \cdots \le f(y^f_{(B)})$ for the slicer $f : \mathbb{R}^d \to \mathbb{R}$. The mini-batch STP loss is then defined as

$$J_{N,M,B}(f) \coloneqq \frac{1}{\binom{N}{B}\binom{M}{B}} \sum_{S\in\binom{[N]}{B}} \sum_{T\in\binom{[M]}{B}} h_B\big(f;\, x_S,\, y_T\big).$$

For a fixed slicer f, it follows the same minibatch averaging construction as [15, Definition 3], with h_B serving as the batchwise transport cost. In practical implementations, the incomplete estimator draws K i.i.d. batch pairs $\{(S_k, T_k)\}_{k=1}^K$ uniformly from $\binom{[N]}{B} \times \binom{[M]}{B}$, and averages:

$$\overline{J}_{B,K}(f) \coloneqq \frac{1}{K}\sum_{k=1}^{K} h_B\big(f;\, x_{S_k},\, y_{T_k}\big).$$

Proposition 3.5. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, there exist constants C_1, C_2, C_3, R (which depend on the datasets and f) such that*

$$\begin{aligned}\big|\overline{J}_{B,K}(f) - J_{N,M}(f)\big| \;\le\; \frac{C_1}{B} \;+\; C_2\omega_X\big(\tfrac{1}{2}\sqrt{\tfrac{N-B}{NB}}\big) \\ + C_3\omega_Y\big(\tfrac{1}{2}\sqrt{\tfrac{M-B}{MB}}\big) \;+\; R\sqrt{\frac{\log(2/\delta)}{2K}}.\end{aligned}$$

where ω_X, ω_Y denote the quantile moduli of $f_{\#}\mu_N$ and $f_{\#}\nu_M$, which are non-decreasing, and satisfy $\omega_X(0) = \omega_Y(0) = 0$.

Proof sketch. The proof proceeds in two steps: (i) controlling the mini-batch bias $\big|J_{N,M,B}(f) - J_{N,M}(f)\big|$ and (ii) controlling the Monte Carlo fluctuation over K batches $\big|\overline{J}_{B,K}(f) - J_{N,M,B}(f)\big|$. The first term is deterministically bounded by a $\mathcal{O}(B^{-1})$ term from well-stratified batches, together with the quantile moduli ω_X, ω_Y capturing deviations from this ideal stratification. The second term utilizes the fact that $\mathbb{E}\overline{J}_{B,K}(f) = J_{N,M,B}(f)$. By the boundedness of finite discrete samples, applying Hoeffding's inequality gives the probabilistic bound $\mathcal{O}(K^{-1/2})$. Details can be found in Propositions B.2 and B.7 in the Appendix. □

4 Empirical Results

4.1 Implementations

We train the slicer with symmetric two-branch gradient flows (see Figure 11). For the projected samples $\{f(x_i)\}_{i=1}^N$ and $\{f(y_j)\}_{j=1}^M$, the differentiable LapSum operator produces soft permutation matrices $\tilde{P}_X, \tilde{P}_Y$. Let P_X, P_Y be the corresponding hard permutations and let $T_{N,M} \in \mathbb{R}^{N\times M}$ be the fixed optimal transport plan between one-dimensional sorted N points and sorted M points. As differentiating through both soft permutations simultaneously may amplify noise and result in high variance in training, we form two plans during backpropagation,

$$\gamma_1 = \tilde{P}_X^\top\, T_{N,M}\, P_Y, \qquad \gamma_2 = P_X^\top\, T_{N,M}\, \tilde{P}_Y,$$

and optimize the loss using their average, $\tilde{\gamma} \;:=\; \frac{1}{2}\big(\gamma_1 + \gamma_2\big)$, for more stable optimization. A more detailed discussion on the connection between the theoretical objective in 3.3 and this practical loss can be found in Appendix C. We empirically test the training scheme by comparing the transport plans produced by different training strategies, shown in Figure 12. The visualizations illustrate that the mini-batch $\min$-STP maintains high-quality matching even for smaller batch sizes, closely resembling both the optimal transport plan and the full-batch $\min$-STP solution.

4.2 Transferability under Gradual Drift

We evaluate slicer transferability on a sequence of gradually changing distribution pairs $\{(\mu_t, \nu_t)\}_{t=1}^7$ generated from a fixed two-moons plus eight-Gaussians template in $\mathbb{R}^2$. Each distribution has 256 samples. In each step, we apply a small rotation and zoom-out to produce the next distribution pair. For every task, we use a Set-Transformer [29] as the parameterized slicer to minimize the STP objective. For $t \ge 2$, we compare initial values of the STP objective under two initializations: (i) **random initializations** (fresh and random parameters), and (ii) **transferred / pretrained initializations**, which reuses the optimal slicer $f_{t-1}^\star$ from the previous task (μ_{t-1}, ν_{t-1}). As a lower bound, we also compute the OT cost between (μ_t, ν_t). Figure 3 shows that in all steps, the pretrained initialization starts from a substantially lower objective than random and sits noticeably closer to the OT reference, indicating that leveraging $f_{t-1}^\star$ places us at a much stronger starting point for the new task and the learned slicer captures geometry that persists under small rotations and scale changes.

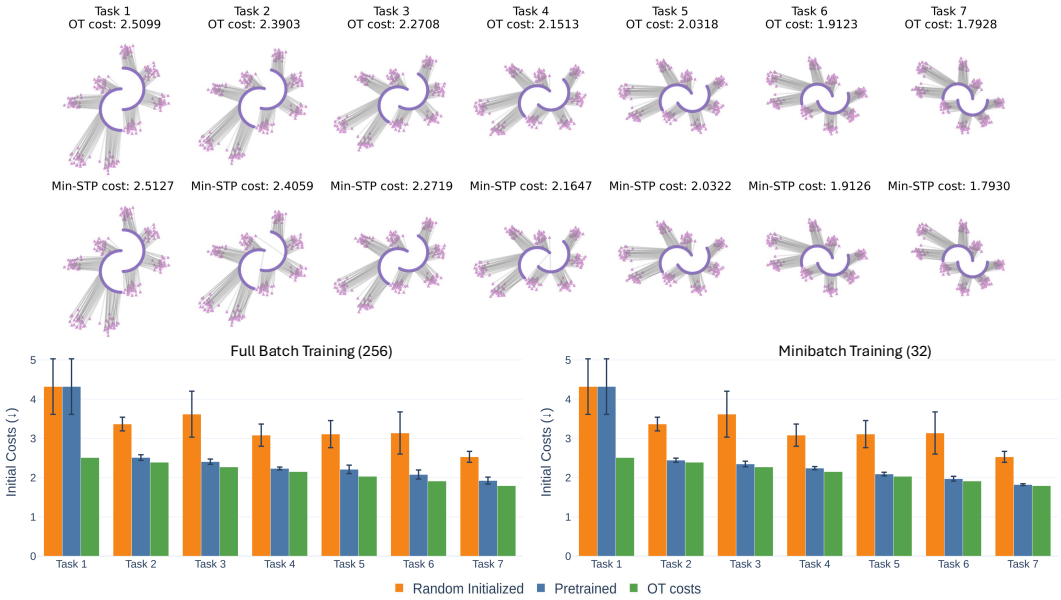

Figure 3: *(Top row) Generated tasks* $\{(\mu_t, \nu_t)\}_{t=1}^{7}$ *with OT plans and corresponding costs. (Middle row)* min-STP *plans and costs optimized with a generalized slicer (i.e., a neural network). (Bottom row) Initial slicer costs (mean over 5 runs) for a randomly initialized neural network, the optimal slicer for previous task, and the OT lower bound for full-batch (left) and mini-batch (right) training.*

4.3 Amortized Min-STP for Point Cloud Alignment

To further assess the transferability of min-STP, we investigate whether a single amortized slicer can generalize across a family of related distribution pairs. We use the ModelNet10 dataset [52], which contains ten object categories (bathtub, bed, chair, desk, dresser, monitor, night_stand, sofa, table, toilet), each represented as a point cloud of size 1024. For every ordered pair of categories, we construct training and test sets of distribution pairs by sampling point clouds from the corresponding classes. A single amortized slicer $f_a^\star$ is then learned by minimizing a global objective over all training pairs. The slicer is implemented as an MLP with 3 hidden layers of dimensions $[256, 512, 256]$. As input, we concatenate each 3D point with a context vector, obtained from a pretrained point cloud auto-encoder. More details can be found in Section F. We then calculate the correlations between the transport costs induced by $f_a^\star$ and the exact OT costs. We compare the correlations against per-pair min-STP, DGSWP [10], STP with a random slicer, and Wasserstein Wormhole [21] as a global embedding from point clouds into a Euclidean space, trained to preserve pairwise Wasserstein distances. Table 1 reports the mean$\pm$std Pearson correlations with OT across category pairs. Figure 4 shows an example pair (desk/sofa), with additional pairs in Figure 14 (Supplementary). Per-pair min-STP achieves higher mean train and test correlations than DGSWP, its direct counterpart since both optimize a slicer separately for each distribution pair. Using a single amortized slicer, min-STP achieves high correlations with exact OT while outperforming the random-slicer STP and Wasserstein Wormhole baselines, demonstrating effective generalization across related distributions.

Method	Train Corr.	Test Corr.
min-STP (Amortized)	$0.907^{\pm 0.048}$	$0.902^{\pm 0.054}$
min-STP (Per-pair)	$0.959^{\pm 0.046}$	$0.947^{\pm 0.073}$
DGSWP	$0.923^{\pm 0.023}$	$0.916^{\pm 0.023}$
STP (Random Slicer)	$0.702^{\pm 0.090}$	$0.719^{\pm 0.113}$
Wasserstein Wormhole	$0.831^{\pm 0.132}$	$0.770^{\pm 0.177}$

Table 1: *Mean and standard deviation of Pearson correlations with exact OT over all ModelNet10 category pairs. Per-pair* min-STP *achieves the highest correlations, while amortized* min-STP *remains competitive without requiring per-pair optimization. Note that unlike all the sliced plan-based methods, Wasserstein Wormhole does not produce a transportation plan.*

4.4 Min-STP based Flow

A promising application of optimal transport is its use in improving flow-based generative modeling [48] and single-step generation [28]. OT-MF [1] demonstrates that incorporating optimal transport into the MeanFlow algorithm [16] can enhance single-step generation performance across various tasks.

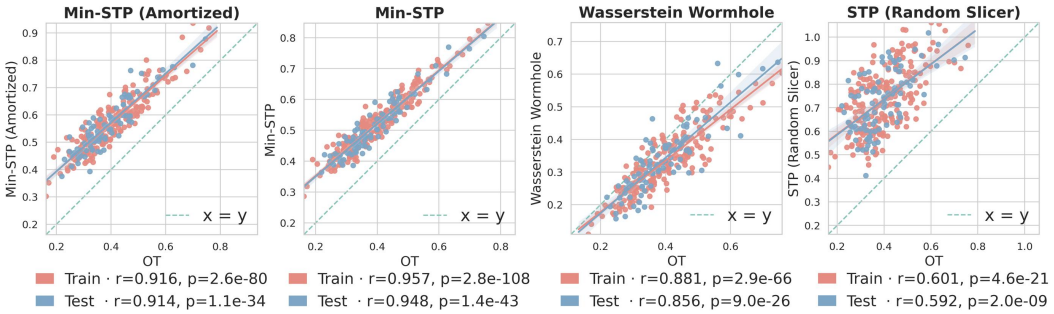

Figure 4: *Correlations for desk and sofa pairs in ModelNet10. Amortized* min-STP *matches the performance of per-pair* min-STP *and outperforms the two other baselines.*

Motivated by these findings, we aim to investigate the effectiveness of min-STP in a flow-based setting. Specifically, we adopt the experimental setup and datasets from [1] for point cloud generation, using the "Chair" class from the ShapeNet dataset [9]. Additional results and implementation details are available in section F.4.

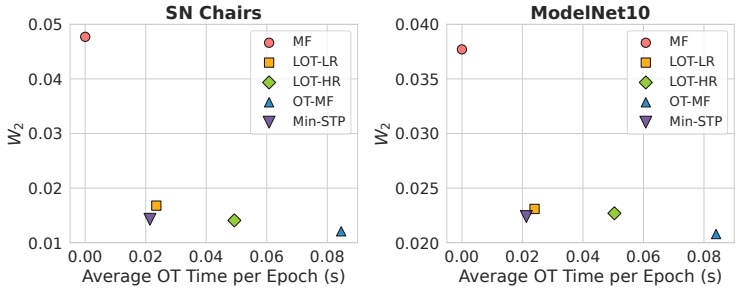

Figure 5: *Average OT compute time per epoch vs. the* $\mathrm{OT}_2(W_2)$ *distance for each method. For both metrics, lower is better.*

Figure 5 shows the average time required to compute OT per epoch vs. the 2-Wasserstein distance (W_2) for min-STP and the baseline methods. A lower average W_2 across test samples shows the ability of the method to generate point cloud samples that are closest to the ground truth. An ideal method should also achieve a lower computation time for OT to reduce the overall training cost. As shown in Figure 5, min-STP appears in the bottom-left region of the plot, demonstrating its ability to balance low computation time with low W_2, resulting in higher-quality generation with faster training. Figure 16 (in the appendix) shows generated samples (1-step) for ShapeNet-Chairs. min-STP produces samples that more closely resemble the ground truth while adding only a slight overhead during training.

4.5 Unpaired Image-to-Image Translation

In the unpaired image-to-image translation setting, we aim to learn a mapping from a source distribution μ to a target distribution ν. Here we use the face images from the FFHQ dataset [23] and work with 512-dimensional latent vectors encoded by ALAE [37]. Given a transport plan γ between distributions μ and ν over the latent vectors, we train an MLP T_θ,

$$\min_\theta \sum_{i,j} \|T_\theta(x_i) - y_j\|_2^2 \, \gamma_{i,j}, \qquad (4)$$



Figure 6: *Adult-to-Child translation visualizations with different methods.*

where the source samples $\{x_i\} \sim \mu$ and the target samples $\{y_j\} \sim \nu$. Training is performed in a mini-batch manner: at each iteration, we sample batches from μ and ν, compute a transport plan γ between the two batches, and update θ using the corresponding batch-wise objective.

We consider several transport plans, including min-STP (amortized), DGSWP with the amortized implementation [10], entropic optimal transport (Sinkhorn) [13], Expected Transport Plan (EST) [30], Progressive Entropic OT (ProgOT) [24], and Low-Rank OT (LOT) [42]. Since a transport plan is recomputed at every training iteration, this setting requires repeatedly solving transport problems on different mini-batches. For each method, we evaluate the resulting trained mapping T_{θ^*} using the 2–Wasserstein distance $W_2\big((T_{\theta^*})_{\#}\mu, \nu\big)$. The Adult-to-Child translation results are reported in Figure 7 and Figure 6: min-STP reaches $W_2 < 13$ in 41 s, versus 200 s for ProgOT and 377 s for OT, and ends close to them (12.61 vs. 12.21 and 12.02), whereas the remaining approximations (EST, Sinkhorn, LOT, and DGSWP) plateau above $W_2 = 14$. Once the plan spreads each source's mass over many targets; since the objective is minimized by the barycentric projection $T_\theta(x_i) = \sum_j \gamma_{ij} y_j / \sum_j \gamma_{ij}$, the network regresses toward an average face, and its translations become nearly identical across inputs (e.g. Sinkhorn in Figure 6).

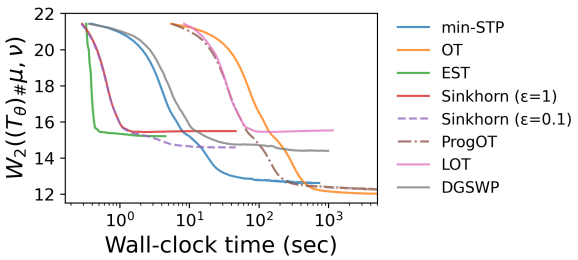

Figure 7: *Adult-to-Child translation performance (W_2) against wall-clock training time.*

5 Limitations and Future Work

The main limitation of this work is the restriction to discrete probability measures. In particular, the construction of the lifted transport plan relies on the injectivity of perturbed slicers on finite or countable supports. Although this setting covers the empirical measures used in our experiments, extending the framework and its guarantees to general continuous distributions requires further analysis. Within this setting, our transferability guarantees concern small distributional perturbations and gradual shifts. This opens an interesting direction for investigating how far slicer reuse can extend under larger domain shifts. From a computational perspective, we focus on per-iteration efficiency, while total training cost also depends on the number of iterations required for convergence. A more comprehensive assessment of computational efficiency therefore motivates a complementary study of how training dynamics and convergence vary across data scales, distributions, and optimization settings.

6 Conclusion

We studied the transferability of optimized slicers in the min-Sliced Transport Plan (min-STP) framework. Leveraging smoothing via Laplace distributions, our theoretical results establish that optimal slicers are stable under small distributional shifts, hence enabling transfer across related source-target tasks. To improve scalability, we also introduced a mini-batch formulation of min-STP and provided guarantees for its approximation accuracy. We empirically showed that transferred slicers maintain strong performance in point cloud alignment, flow-based generative modeling, and image translation. Our work highlights the promise of geometry-aware transport methods that generalize across domains and opens new avenues for scalable, transferable OT in dynamic, multi-domain settings.

Acknowledgments and Disclosure of Funding

This work was supported in part by NSF CAREER Award No. 2339898 and NSF DMS Award No. 2603773. The authors also gratefully acknowledge the computing resources provided by the NVIDIA Academic Grant Program.

References

[1] Elaheh Akbari, Ping He, Ahmadreza Moradipari, Yikun Bai, and Soheil Kolouri. Transport based mean flows for generative modeling, 2026.

[2] Charalambos D Aliprantis and Kim C Border. *Infinite dimensional analysis: a hitchhiker's guide*. Springer, 2006.

[3] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in Neural Information Processing Systems*, 2017.

[4] Brandon Amos, Giulia Luise, Samuel Cohen, and Ievgen Redko. Meta optimal transport. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[6] Yikun Bai, Bernhard Schmitzer, Matthew Thorpe, and Soheil Kolouri. Sliced optimal partial transport. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13681–13690, 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] Nicolas Bonnotte. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Université Paris Sud-Paris XI; Scuola normale superiore (Pise, Italie), 2013.

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

[10] Laetitia Chapel, Romain Tavenard, and Samuel Vaiter. Differentiable generalized sliced Wasserstein plans. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[11] Xiongjie Chen, Yongxin Yang, and Yunpeng Li. Augmented sliced wasserstein distances. *arXiv preprint arXiv:2006.08812*, 2020.

[12] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

[13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

[14] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 447–463, 2018.

[15] Kilian Fatras, Younes Zine, Szymon Majewski, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Minibatch optimal transport distances; analysis and applications. *arXiv preprint arXiv:2101.01792*, 2021.

[16] Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

[17] Samuel Gerber and Mauro Maggioni. Multiscale strategies for computing optimal transport. *Journal of Machine Learning Research*, 18(72):1–32, 2017.

[18] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[19] Wenxuan Guo, YoonHaeng Hur, Tengyuan Liang, and Chris Ryan. Online learning to transport via the minimal selection principle. In *Conference on Learning Theory*, pages 4085–4109. PMLR, 2022.

[20] Peter Halmos, Julian Gold, Xinhao Liu, and Benjamin Raphael. Hierarchical refinement: Optimal transport to infinity and beyond. In *Forty-second International Conference on Machine Learning*, 2025.

[21] Doron Haviv, Russell Zhang Kunes, Thomas Dougherty, Cassandra Burdziak, Tal Nawy, Anna Gilbert, and Dana Pe'Er. Wasserstein wormhole: Scalable optimal transport distance with transformers. *ArXiv*, pages arXiv–2404, 2024.

[22] Viet Huynh, He Zhao, and Dinh Phung. Otlda: A geometry-aware optimal transport approach for topic modeling. *Advances in Neural Information Processing Systems*, 33:18573–18582, 2020.

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[24] Parnian Kassraie, Aram-Alexandre Pooladian, Michal Klein, James Thornton, Jonathan Niles-Weed, and Marco Cuturi. Progressive entropic optimal transport solvers. *Advances in Neural Information Processing Systems*, 37:19561–19590, 2024.

[25] Abdelwahed Khamis, Russell Tsuchida, Mohamed Tarek, Vivien Rolland, and Lars Petersson. Scalable optimal transport methods in machine learning: A contemporary survey. *IEEE transactions on pattern analysis and machine intelligence*, 2024.

[26] Soheil Kolouri, Kimia Nadjahi, Shahin Shahrampour, and Umut Şimşekli. Generalized sliced probability metrics. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4513–4517. IEEE, 2022.

[27] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[28] Nikita Kornilov, Petr Mokrov, Alexander Gasnikov, and Alexander Korotin. Optimal flow matching: Learning straight trajectories in just one step. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 104180–104204. Curran Associates, Inc., 2024.

[29] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR, 2019.

[30] Xinran Liu, Rocio Diaz Martin, Yikun Bai, Ashkan Shahbazi, Matthew Thorpe, Akram Aldroubi, and Soheil Kolouri. Expected sliced transport plans. In *The Thirteenth International Conference on Learning Representations*, 2025.

[31] Guillaume Mahey, Laetitia Chapel, Gilles Gasso, Clément Bonet, and Nicolas Courty. Fast optimal transport through sliced generalized wasserstein geodesics. *Advances in Neural Information Processing Systems*, 36:35350–35385, 2023.

[32] Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. *Advances in Neural Information Processing Systems*, 33:20802–20812, 2020.

[33] Khai Nguyen and Nhat Ho. Energy-based sliced wasserstein distance. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[34] Khai Nguyen, Nhat Ho, Tung Pham, and Hung Bui. Distributional sliced-wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2021.

[35] Khai Nguyen, Shujian Zhang, Tam Le, and Nhat Ho. Sliced wasserstein with random-path projecting directions. In *International Conference on Machine Learning*, pages 37879–37899. PMLR, 2024.

[36] Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[37] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14104–14113, 2020.

[38] 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.

[39] Mark Rowland, Jiri Hron, Yunhao Tang, Krzysztof Choromanski, Tamas Sarlos, and Adrian Weller. Orthogonal estimation of wasserstein distances. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 186–195. PMLR, 2019.

[40] 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.

[41] Meyer Scetbon and Marco Cuturi. Low-rank optimal transport: Approximation, statistics and debiasing. *Advances in Neural Information Processing Systems*, 35:6802–6814, 2022.

[42] Meyer Scetbon, Marco Cuturi, and Gabriel Peyré. Low-rank sinkhorn factorization. In *International Conference on Machine Learning*, pages 9344–9354. PMLR, 2021.

[43] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[44] Bernhard Schmitzer. A sparse multiscale algorithm for dense optimal transport. *Journal of Mathematical Imaging and Vision*, 56(2):238–259, 2016.

[45] Zhengyang Shen, Jean Feydy, Peirong Liu, Ariel H Curiale, Ruben San Jose Estepar, Raul San Jose Estepar, and Marc Niethammer. Accurate point cloud registration with robust optimal transport. *Advances in Neural Information Processing Systems*, 34:5373–5389, 2021.

[46] Łukasz Struski, Michal B. Bednarczyk, Igor T. Podolak, and Jacek Tabor. Lapsum - one method to differentiate them all: Ranking, sorting and top-k selection. In *Forty-second International Conference on Machine Learning*, 2025.

[47] Eloi Tanguy, Laetitia Chapel, and Julie Delon. Sliced optimal transport plans. *arXiv preprint arXiv:2508.01243*, 2025.

[48] Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. Expert Certification.

[49] Minh-Phuc Truong and Khai Nguyen. Amortized optimal transport from sliced potentials. *arXiv preprint arXiv:2604.15114*, 2026.

[50] Cedric Villani. *Optimal transport: old and new*. Springer, 2009.

[51] Wei Wang, Dejan Slepčev, Saurav Basu, John A Ozolek, and Gustavo K Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101(2):254–269, 2013.

[52] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[53] Karren Dai Yang, Karthik Damodaran, Saradha Venkatachalapathy, Ali C Soylemezoglu, GV Shivashankar, and Caroline Uhler. Predicting cell lineages using autoencoders and optimal transport. *PLoS computational biology*, 16(4):e1007828, 2020.

[54] Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin M Solomon. Hierarchical optimal transport for document representation. *Advances in neural information processing systems*, 32, 2019.

A Transferability

1. Model assumption and properties.

Under Assumption 3.2, we take $\mathcal{X} \subset \mathbb{R}^d$ to be a compact metric space. Since $\mathcal{X}$ is compact, all probability measures on $\mathcal{X}$ automatically have finite p-th moments. Therefore,

$$\mathcal{P}_p(\mathcal{X}) = \mathcal{P}(\mathcal{X}).$$

Lemma A.1. *Consider the family $\mathcal{F}$ defined over the compact metric space $(\mathcal{X} \subset \mathbb{R}^d, d)$ under Assumption 3.2. Then $\mathcal{F}$ is compact in $(C(\mathcal{X}), \|\cdot\|_\infty)$.*

Proof. For any $\varepsilon > 0$, choose $\delta := \varepsilon/L$. If $d(x, y) < \delta$, then for all $f \in \mathcal{F}$, we have $|f(x) - f(y)| \leq Ld(x, y) < \varepsilon$. By uniform boundedness, $\|f\|_\infty < M$ for all $f \in \mathcal{F}$. For each $x \in \mathcal{X}$, the set $\{f(x) : f \in \mathcal{F}\} \subset [-M, M]$ is compact. By Arzelà–Ascoli theorem on the compact space $\mathcal{X}$, $\mathcal{F}$ is precompact in $\|\cdot\|_\infty$ (i.e., its closure is compact in the ambient space). Moreover, $\mathcal{F}$ is closed as uniform limits preserve the Lipschitz constant L and the sup bound M. Therefore, $\mathcal{F}$ is compact since $C(\mathcal{X})$ is complete under the uniform topology (due to the fact that $\mathcal{X}$ is compact). □

2. Perturbed version of STP.

We recall that a 1D random variable has a Laplace distribution $\mathsf{Lap}_1(0, b)$ if its probability density function is $\frac{1}{2b}e^{-\frac{|x|}{b}}$. For the symmetric multivariate Laplace distribution $\mathsf{Lap}_d(0, \Sigma)$ in $\mathbb{R}^d$, where $\Sigma \in R^{d\times d}$ is a symmetric positive definite matrix, a typical characterization is given through its characteristic function: $\left(1 + \frac{1}{2}\, t^\top \Sigma t\right)^{-1}$, $t \in \mathbb{R}^d$.

Lemma A.2 (Noisy perturbations are Laplace)**.** *Let $\xi \sim \mathsf{Lap}_d(\eta^2\Sigma)$. For any fixed $x \in \mathbb{R}^d$,*

$$\langle \xi, x\rangle \;\sim\; \mathsf{Lap}_1\big(0,\; b_x\big), \qquad b_x = \eta\,\sqrt{\tfrac{1}{2}\, x^\top \Sigma x}.$$

In particular, with $\Sigma = 2I_d$ one gets $\langle \xi, x\rangle \sim \mathsf{Lap}_1(0, \eta\|x\|_2)$.

Proof. The characteristic function of ξ is $\phi_\xi(t) = \left(1 + \frac{1}{2}\,\eta^2\, t^\top \Sigma t\right)^{-1}$, $t \in \mathbb{R}^d$. For $Y := \langle \xi, x\rangle$, $\phi_Y(t) = \phi_\xi(tx) = \left(1 + \frac{1}{2}\,\eta^2\, t^2\, x^\top \Sigma x\right)^{-1}$, which is the characteristic function of a univariate Laplace with 0 mean and scale $b_x = \eta\sqrt{(x^\top \Sigma x)/2}$. □

Lemma A.3 (Injectivity of projection)**.** *Fix $\eta > 0$, let $\mu \in \mathcal{P}(\mathcal{X})$ be purely atomic with at most countable support,i.e. $\mu = \sum_{i\in I} a_i\, \delta_{x_i}$, $a_i > 0$, $\sum_{i\in I} a_i = 1$, where I is finite or countable and the support $\{x_i\}_{i\in I} \subset \mathcal{X}$. Let $f \in \mathcal{F}$, and let g_{ξ_η} be as in Definition 3.1. Then for $\mathsf{Lap}_d(\eta^2\Sigma)$-a.e. ξ_η, the map g_{ξ_η} is injective on $supp(\mu) = \{x_i : i \in I\}$. In particular, for $\mathsf{Lap}_d(\eta^2\Sigma)$-a.e. ξ_η, there exists a Borel set $A_{\xi_\eta} \subseteq \mathcal{X}$ with $\mu(A_{\xi_\eta}) = 1$ such that $g_{\xi_\eta}|_{A_{\xi_\eta}} : A_{\xi_\eta} \to g_{\xi_\eta}(A_{\xi_\eta})$ is injective.*

Proof. For any two distinct support points $x_i \neq x_j$,

$$g_{\xi_\eta}(x_i) = g_{\xi_\eta}(x_j) \iff \langle \xi_\eta, x_i - x_j\rangle = f(x_j) - f(x_i),$$

which defines an affine hyperplane in $\mathbb{R}^d$ for ξ_η. Since the Laplace perturbation law is absolutely continuous with respect to Lebesgue measure, each such event has probability zero. Since there are only countably many distinct pairs of support points, the union of all such events still has probability zero. Hence, for almost every perturbation ξ_η, the values $\{g_{\xi_\eta}(x_i)\}_{i\in I}$ are all distinct, so g_{ξ_η} is injective on $\mathrm{supp}(\mu)$. Taking $A_{\xi_\eta} = \mathrm{supp}(\mu)$, we obtain a Borel full-μ-measure set on which g_{ξ_η} is injective.

□

Unless stated otherwise, we assume the probability measures are discrete with finite, or countable & bounded support in all of the following results.

Proposition A.4 (Uniqueness of the lift)**.** *Fix $\eta > 0$, $\mu, \nu \in \mathcal{P}(\mathcal{X})$ and $f \in \mathcal{F}$. Fix ξ_η, g_{ξ_η} as in Definition 3.1 and set*

$$\alpha_{\xi_\eta} := (g_{\xi_\eta})_{\#}\mu, \qquad \beta_{\xi_\eta} := (g_{\xi_\eta})_{\#}\nu.$$

Then for every $\sigma \in \Gamma(\alpha_{\xi_\eta}, \beta_{\xi_\eta})$, if $\tilde{\gamma}, \hat{\gamma} \in \Gamma(\mu, \nu)$ satisfy

$$(g_{\xi_\eta}, g_{\xi_\eta})_{\#}\tilde{\gamma} \;=\; \sigma \;=\; (g_{\xi_\eta}, g_{\xi_\eta})_{\#}\hat{\gamma},$$

then $\tilde{\gamma} = \hat{\gamma}$.

Proof. Let $A_{\xi_\eta}, B_{\xi_\eta}$ be Borel sets as in Lemma A.3 with $\mu(A_{\xi_\eta}) = 1 = \nu(B_{\xi_\eta})$ so that $g_{\xi_\eta}|_{A_{\xi_\eta}} : A_{\xi_\eta} \to g_{\xi_\eta}(A_{\xi_\eta})$ and $g_{\xi_\eta}|_{B_{\xi_\eta}} : B_{\xi_\eta} \to g_{\xi_\eta}(B_{\xi_\eta})$ are bijections with inverses $j^{g_{\xi_\eta},\mu}, j^{g_{\xi_\eta},\nu}$, respectively. By Lusin–Souslin theorem, such inverses are Borel measurable functions.

First note $\tilde{\gamma}(A_{\xi_\eta} \times B_{\xi_\eta}) = 1 = \hat{\gamma}(A_{\xi_\eta} \times B_{\xi_\eta})$ because $\mu(A_{\xi_\eta}) = \nu(B_{\xi_\eta}) = 1$. Now, for any Borel $E \subset A_{\xi_\eta} \times B_{\xi_\eta}$, the map

$$(g_{\xi_\eta}, g_{\xi_\eta})|_{A_{\xi_\eta} \times B_{\xi_\eta}} : \; A_{\xi_\eta} \times B_{\xi_\eta} \;\longrightarrow\; g_{\xi_\eta}(A_{\xi_\eta}) \times g_{\xi_\eta}(B_{\xi_\eta})$$

is a Borel bijection with Borel inverse $\Phi := (j^{g_{\xi_\eta},\mu}, j^{g_{\xi_\eta},\nu})$. Hence

$$\begin{aligned}\tilde{\gamma}(E) &= \tilde{\gamma}\big(\Phi\big((g_{\xi_\eta}, g_{\xi_\eta})(E)\big)\big) = \sigma\big((g_{\xi_\eta}, g_{\xi_\eta})(E)\big) \\ &= \hat{\gamma}\big(\Phi\big((g_{\xi_\eta}, g_{\xi_\eta})(E)\big)\big) = \hat{\gamma}(E),\end{aligned}$$

where we used $(g_{\xi_\eta}, g_{\xi_\eta})_{\#}\tilde{\gamma} = \sigma = (g_{\xi_\eta}, g_{\xi_\eta})_{\#}\hat{\gamma}$. Therefore, $\tilde{\gamma}$ and $\hat{\gamma}$ coincide on $A_{\xi_\eta} \times B_{\xi_\eta}$. Since $\tilde{\gamma}(A_{\xi_\eta} \times B_{\xi_\eta}) = \hat{\gamma}(A_{\xi_\eta} \times B_{\xi_\eta}) = 1$, for any bounded Borel function φ on $\mathcal{X}^2$,

$$\begin{aligned}\int \varphi \,\mathrm{d}\tilde{\gamma} &= \int \varphi\, \mathbf{1}_{A_{\xi_\eta} \times B_{\xi_\eta}} \,\mathrm{d}\tilde{\gamma} \\ &= \int \varphi\, \mathbf{1}_{A_{\xi_\eta} \times B_{\xi_\eta}} \,\mathrm{d}\hat{\gamma} = \int \varphi \,\mathrm{d}\hat{\gamma}.\end{aligned}$$

Thus, we conclude $\tilde{\gamma} = \hat{\gamma}$. □

Notation: From now on, we will assume

$$c(x, y) = \|x - y\|_p,$$

and for $\mu, \nu \in \mathcal{P}(\mathcal{X})$ we write $OT_p(\mu, \nu) = W_p(\mu, \nu)$ to refer to the p-Wasserstein distance defined in 1.

Lemma A.5. *For $f, g \in \mathcal{F}$ and $\kappa, \kappa' \in \mathcal{P}(\mathcal{X})$,*

(1) $W_p\big(f_{\#}\kappa,\; f_{\#}\kappa'\big) \;\le\; L\, W_p(\kappa, \kappa')$

(2) $W_p\big(f_{\#}\kappa,\; g_{\#}\kappa\big) \le \|f - g\|_\infty$

Proof. (1) Let $\gamma \in \Gamma(\kappa, \kappa')$ be any coupling between κ and κ'. Then $(f, f)_{\#}\gamma \in \Gamma(f_{\#}\kappa, f_{\#}\kappa')$, therefore

$$W_p^p\big(f_{\#}\kappa,\; f_{\#}\kappa'\big) \;\le\; \int_{\mathbb{R}^2} |u - v|^p \,\mathrm{d}((f, f)_{\#}\gamma)(u, v);$$

and by L-Lipschitzness,

$$\begin{aligned}&\int_{\mathbb{R}^2} |u - v|^p \,\mathrm{d}((f, f)_{\#}\gamma)(u, v) \\ &= \int_{\mathcal{X}^2} |f(x) - f(y)|^p \,\mathrm{d}\gamma(x, y) \\ &\le L^p \int_{\mathcal{X}^2} c(x, y)^p \,\mathrm{d}\gamma(x, y).\end{aligned}$$

Since this holds for any coupling $\gamma \in \Gamma(\kappa, \kappa')$,

$$W_p^p\big(f_\#\kappa,\, f_\#\kappa'\big) \le L^p \inf_{\gamma\in\Gamma(\kappa,\kappa')} \int_{\mathcal{X}^2} c(x,y)^p \mathrm{d}\gamma(x,y)$$
$$= L^p\, W_p^p(\kappa,\kappa'),$$

which proves the claim (1).

(2) Consider the identity coupling $\gamma_{id} \in \Gamma(\kappa,\kappa)$. Then $(f,g)_\#\gamma_{id} \in \Gamma(f_\#\kappa, g_\#\kappa)$ and

$$W_p^p\big(f_\#\kappa,\, g_\#\kappa\big) \le \int |f(x)-g(x)|^p \,\mathrm{d}\gamma_{id}(x,x)$$
$$= \int |f(x)-g(x)|^p \,\mathrm{d}\kappa(x) \le \|f-g\|_\infty^p.$$

Taking the p-th root gives the desired bound. □

Lemma A.6. *For 1D probability measures $\mu_i, \nu_i \in \mathcal{P}(\mathbb{R})$ with optimal couplings γ_i, $i = 1, 2$,*

$$W_p\big(\gamma_1,\gamma_2\big) \;\le\; \big(W_p(\mu_1,\mu_2)^p + W_p(\nu_1,\nu_2)^p\big)^{1/p},$$

where the ground metric on $\mathbb{R}^2$ is defined as $\|(s,t)-(s',t')\|_p = (|s-s'|^p + |t-t'|^p)^{1/p}$ for $s,t,s',t' \in \mathbb{R}$.

Proof. Let Q_{μ_i} and Q_{ν_i} denote the quantile functions of μ_i, ν_i. Then the 1D p-Wasserstein distances between μ_i and ν_i are:

$$W_p^p(\mu_i,\nu_i) = \int_0^1 |Q_{\mu_i}(t) - Q_{\nu_i}(t)|^p \mathrm{d}t,$$

and the corresponding optimal transport plan $\gamma_i = (Q_{\mu_i}, Q_{\nu_i})_\#\lambda_{[0,1]}$, where $\lambda_{[0,1]}$ is the Lebesgue measure on $[0,1]$. Define $\widetilde{\gamma} := (Q_{\mu_1}, Q_{\nu_1}, Q_{\mu_2}, Q_{\nu_2})_\#\lambda_{[0,1]} \in \Gamma(\gamma_1,\gamma_2)$, then

$$W_p^p(\gamma_1,\gamma_2) \le \int_{(\mathbb{R}^2)^2} \|(u_1,v_1)-(u_2,v_2)\|_p^p \,\mathrm{d}\widetilde{\gamma}$$
$$= \int_{(\mathbb{R}^2)^2} |u_1-u_2|^p + |v_1-v_2|^p \,\mathrm{d}\widetilde{\gamma}$$
$$= \int_{[0,1]} |Q_{\mu_1}(s) - Q_{\mu_2}(s)|^p + |Q_{\nu_1}(s) - Q_{\nu_2}(s)|^p \,\mathrm{d}s$$
$$= W_p^p(\mu_1,\mu_2) + W_p^p(\nu_1,\nu_2).$$

Taking p-th root gives the claim. □

Proposition A.7 (Recovery of STP as $\eta \to 0$ for injective slicer)**.** *Assume $f \in \mathcal{F}$ and **injective**. Let $\alpha := f_\#\mu$, $\beta := f_\#\nu$ and let σ_f be the 1D optimal coupling between α and β. Define the (unique) lift $\gamma_f \in \Gamma(\mu,\nu)$ such that*

$$(f,f)_\#\gamma_f = \sigma_f,$$

and recall

$$\mathit{STP}_p^p(\mu,\nu;f) := \int_{\mathcal{X}^2} c(x,y)^p \,\mathrm{d}\gamma_f(x,y).$$

Let $J_\eta(\mu,\nu;f)$ be the smoothed objective from Definition 3.3, then

$$\lim_{\eta\to 0} J_\eta(\mu,\nu;f) \;=\; \mathit{STP}_p^p(\mu,\nu;f).$$

Proof. Write $\xi_\eta = \eta Z$ with $Z \sim \mathsf{Lap}_d(\Sigma)$, so that $g_{\xi_\eta}(x) = f(x) + \eta\langle Z, x\rangle$. For each instance Z, define the 1D marginals

$$\alpha_{\eta,Z} := (g_{\xi_\eta})_\#\mu, \qquad \beta_{\eta,Z} := (g_{\xi_\eta})_\#\nu,$$

and their optimal plan $\sigma_{\eta,Z}$. Let $\gamma_{\eta,Z}$ be the (unique) lift of $\sigma_{\eta,Z}$ given by Proposition A.4:

$$(g_{\xi_\eta}, g_{\xi_\eta})_\#\gamma_{\eta,Z} = \sigma_{\eta,Z}, \qquad \gamma_{\eta,Z} \in \Gamma(\mu,\nu).$$

By Definition 3.3, $J_\eta(\mu,\nu;f) = \mathbb{E}_Z\big[\int c^p(x,y)\,\mathrm{d}\gamma_{\eta,Z}\big]$. Since $\mathcal{X}$ is compact, let $R := \sup_{x\in\mathcal{X}}\|x\|_p < \infty$, and by Hölder's inequality,

$$\|g_{\xi_\eta} - f\|_\infty \;\le\; \eta\,\|Z\|_q\,R;$$

where $\frac{1}{p}+\frac{1}{q}=1$. By Lemma A.5 part (2),

$$\begin{aligned} &W_p(\alpha_{\eta,Z},\alpha) \;\le\; \|g_{\xi_\eta} - f\|_\infty \;\le\; \eta\|Z\|_q R,\\ &\text{similarly, } W_p(\beta_{\eta,Z},\beta) \;\le\; \eta\|Z\|_q R. \end{aligned}$$

Hence, as $\eta\to 0$, we have $\alpha_{\eta,Z}\to\alpha$ and $\beta_{\eta,Z}\to\beta$ in the one-dimensional W_p-distance for each fixed Z. By Lemma A.6, as $\eta\to 0$, we have

$$W_p^p\big(\sigma_{\eta,Z},\sigma_f\big) \;\le\; W_p^p(\alpha_{\eta,Z},\alpha) + W_p^p(\beta_{\eta,Z},\beta) \;\to\; 0,$$

for each fixed Z.

Thus, $\sigma_{\eta,Z} \rightharpoonup \sigma_f$ weakly as $\eta\to 0$ (for each Z).

Let $\eta_n\to 0$ with corresponding $\{\gamma_{\eta_n,Z}\}_n$. Since $\mathcal{X}^2$ is compact, $\mathcal{P}(\mathcal{X}^2)$ is also compact, and so there exists a subsequence $\gamma_{\eta_{n_m},Z}\rightharpoonup\gamma^\star$ for some probability measure $\gamma^\star$ on $\mathcal{X}^2$. For any bounded $\varphi\in C(\mathbb{R}^2)$,

$$\begin{aligned} &\int \varphi\,\mathrm{d}(g_{\xi_{\eta_{n_m}}}, g_{\xi_{\eta_{n_m}}})_{\#}\gamma_{\eta_{n_m},Z}\\ &= \int \varphi\big(g_{\xi_{\eta_{n_m}}}(x), g_{\xi_{\eta_{n_m}}}(y)\big)\,\mathrm{d}\gamma_{\eta_{n_m},Z}(x,y)\\ &= \int \varphi\,\mathrm{d}\sigma_{\eta_{n_m},Z}. \end{aligned}$$

Because $g_{\xi_{\eta_{n_m}}}\to f$ uniformly on $\mathcal{X}$ and $\gamma_{\eta_{n_m},Z}\rightharpoonup\gamma^\star$, the LHS tends to $\int\varphi\big(f(x),f(y)\big)\,\mathrm{d}\gamma^\star(x,y)$. The RHS tends to $\int\varphi\,\mathrm{d}\sigma_f$. Hence,

$$(f,f)_{\#}\gamma^\star = \sigma_f.$$

Since f is injective and continuous on compact $\mathcal{X}$, it is a continuous bijection from the compact space $\mathcal{X}$ onto $f(\mathcal{X})\subset\mathbb{R}$ (Hausdorff), and so it is a homeomorphism: its inverse $j^f : f(\mathcal{X})\to\mathcal{X}$ is continuous. Then, the lift that solves $(f,f)_{\#}\gamma=\sigma_f$ is *unique* and equals $\gamma_f := (j^f,j^f)_{\#}\sigma_f$. In particular, $\gamma^\star=\gamma_f$. As the subsequence is arbitrary, we conclude that for *each fixed* Z,

$$\gamma_{\eta,Z} \;\rightharpoonup\; \gamma_f, \qquad \text{as } \eta\to 0.$$

Thus, since c^p is bounded and continuous on the compact $\mathcal{X}^2$,

$$\int c^p\,\mathrm{d}\gamma_{\eta,Z} \underset{\eta\to0}{\longrightarrow} \int c^p\,\mathrm{d}\gamma_f \quad \text{for each fixed } Z.$$

Moreover,

$$0 \;\le\; \int c^p(x,y)\,\mathrm{d}\gamma_{\eta,Z}(x,y) \;\le\; (\mathrm{diam}(\mathcal{X}))^p \;<\infty,$$

so by Dominated Convergence (with respect to Z),

$$\begin{aligned} J_\eta(\mu,\nu;f) = \mathbb{E}_Z\Big[\int c^p\,\mathrm{d}\gamma_{\eta,Z}\Big] &\underset{\eta\to0}{\longrightarrow} \mathbb{E}_Z\Big[\int c^p\,\mathrm{d}\gamma_f\Big]\\ &= \int c^p\,\mathrm{d}\gamma_f\\ &= \mathrm{STP}_p^p(\mu,\nu;f). \end{aligned}$$

□

3. Continuity of J_η.

Proposition A.8 (Continuity of J_η)**.** *Fix $\eta>0$ and let J_η be given by Definition 3.3. If $(\mu_n,\nu_n,f_n)\to(\mu,\nu,f)$ with $W_p(\mu_n,\mu)\to 0$, $W_p(\nu_n,\nu)\to 0$ and $\|f_n-f\|_\infty\to 0$ (as $n\to\infty$), then*

$$J_\eta(\mu_n,\nu_n;f_n) \underset{n\to\infty}{\longrightarrow} J_\eta(\mu,\nu;f).$$

Proof. We write $\xi_\eta = \eta Z$ with $Z \sim \mathsf{Lap}_d(\Sigma)$, and for each $n \in \mathbb{N}$ and Z let

$$g_{n,Z}(x) := f_n(x) + \eta\langle Z, x\rangle, \quad g_Z(x) := f(x) + \eta\langle Z, x\rangle.$$

Set the 1D pushforwards

$$\begin{aligned} \alpha_{n,Z} &:= (g_{n,Z})_\# \mu_n, \qquad \beta_{n,Z} := (g_{n,Z})_\# \nu_n, \\ \alpha_Z &:= (g_Z)_\# \mu, \qquad \beta_Z := (g_Z)_\# \nu, \end{aligned}$$

and their optimal couplings $\sigma_{n,Z}$ and σ_Z. By Lemma A.3, for $\mathbb{P}$–a.e. Z, the lifts $\gamma_{n,Z} \in \Gamma(\mu_n, \nu_n)$ and $\gamma_Z \in \Gamma(\mu, \nu)$ are uniquely defined by

$$(g_{n,Z}, g_{n,Z})_\# \gamma_{n,Z} = \sigma_{n,Z}, \qquad (g_Z, g_Z)_\# \gamma_Z = \sigma_Z.$$

By Definition 3.3,

$$\begin{aligned} J_\eta(\mu_n, \nu_n; f_n) &= \mathbb{E}_Z\left[\int_{\mathcal{X}^2} c(x,y)^p \,\mathrm{d}\gamma_{n,Z}(x,y)\right], \\ J_\eta(\mu, \nu; f) &= \mathbb{E}_Z\left[\int c^p \,\mathrm{d}\gamma_Z\right]. \end{aligned}$$

Since $f_n \to f$ uniformly and $x \mapsto \langle Z, x\rangle$ is $\|Z\|_q$–Lipschitz (with $1/p + 1/q = 1$), we have $g_{n,Z} \to g_Z$ uniformly and $\mathrm{Lip}(g_{n,Z}) \le L + \eta\|Z\|_q$. By Lemma A.5,

$$\begin{aligned} &W_p(\alpha_{n,Z}, \alpha_Z) \\ &\le W_p\big((g_{n,Z})_\#\mu_n, (g_{n,Z})_\#\mu\big) + W_p\big((g_{n,Z})_\#\mu, (g_Z)_\#\mu\big) \\ &\le (L + \eta\|Z\|_q)\, W_p(\mu_n, \mu) + \|g_{n,Z} - g_Z\|_\infty \xrightarrow[n\to\infty]{} 0, \end{aligned}$$

and similarly $W_p(\beta_{n,Z}, \beta_Z) \xrightarrow[n\to\infty]{} 0$. By Lemma A.6,

$$\begin{aligned} W_p\big(\sigma_{n,Z}, \sigma_Z\big) \le& \big(W_p(\alpha_{n,Z}, \alpha_Z)^p + W_p(\beta_{n,Z}, \beta_Z)^p\big)^{1/p} \\ &\xrightarrow[n\to\infty]{} 0 \qquad \text{for each fixed } Z. \end{aligned}$$

For each $n \in \mathbb{N}_0$ (with $n = 0$ denoting (μ, ν, f) and $n \ge 1$ denoting (μ_n, ν_n, f_n)), Lemma A.3 yields a set $G_n \subset \mathbb{R}^d$ with $\mathbb{P}(Z \in G_n) = 1$ such that $g_{n,Z}$ is μ_n–a.e. and ν_n–a.e. injective for all $Z \in G_n$. Let $G := \bigcap_{n=0}^\infty G_n$; then $\mathbb{P}(Z \in G) = 1$ and all lifts above are defined uniquely for $Z \in G$.

Fix $Z \in G$. Since $\mathcal{X}$ is compact, let $\gamma_{n_m,Z} \rightharpoonup \bar\gamma$ be a convergent subsequence. Write $\Phi_n(x,y) := (g_{n,Z}(x), g_{n,Z}(y))$ and $\Phi(x,y) := (g_Z(x), g_Z(y))$. Then $\|\Phi_n - \Phi\|_\infty \to 0$, and for any bounded function $\varphi \in C(\mathbb{R}^2)$,

$$\int \varphi \,\mathrm{d}\sigma_{n_m,Z} = \int \varphi \circ \Phi_{n_m} \,\mathrm{d}\gamma_{n_m,Z} \xrightarrow[m\to\infty]{} \int \varphi \circ \Phi \,\mathrm{d}\bar\gamma,$$

while the LHS tends to $\int \varphi \,\mathrm{d}\sigma_Z$. Hence $(g_Z, g_Z)_\#\bar\gamma = \sigma_Z$. Since $Z \in G$, g_Z is μ–a.e. and ν–a.e. injective, so the lift of σ_Z is *unique*; thus $\bar\gamma = \gamma_Z$. As the subsequence is arbitrary, the whole sequence converges:

$$\gamma_{n,Z} \rightharpoonup \gamma_Z \qquad (Z \in G).$$

In particular, since c^p is bounded and continuous on compact $\mathcal{X}^2$,

$$\int c^p \,\mathrm{d}\gamma_{n,Z} \longrightarrow \int c^p \,\mathrm{d}\gamma_Z \qquad \text{for each } Z \in G.$$

Moreover, $0 \le \int c^p \,\mathrm{d}\gamma_{n,Z} \le (\operatorname{diam}\mathcal{X})^p$ for all n, Z. Hence, by dominated convergence (indeed, bounded convergence) with respect to Z,

$$\begin{aligned} J_\eta(\mu_n, \nu_n; f_n) =& \mathbb{E}_Z\left[\int c^p \,\mathrm{d}\gamma_{n,Z}\right] \\ &\xrightarrow[n\to\infty]{} \mathbb{E}_Z\left[\int c^p \,\mathrm{d}\gamma_Z\right] = J_\eta(\mu, \nu; f). \end{aligned}$$

□

4. Compactness and application of the Berge's Maximum Theorem.

We first review the Berge's Maximum Theorem [2].

Definition A.9 (Correspondence)**.** *A correspondence φ from a set X into a set Y assigns to each $x \in X$ a subset $\varphi(x)$ of Y ('point-to-set' assignment). We denote the correspondence as $\varphi : X \rightrightarrows Y$.*

Just as functions have inverses, so do correspondences. Here, we are only concerned with a "strong inverse": For $\varphi : X \rightrightarrows Y$, the *upper inverse* φ^u or *strong inverse* is defined by

$$\varphi^u(A) \;=\; \{\, x \in X \mid \varphi(x) \subset A \,\}.$$

Recall that a neighborhood of a set A is any set B for which there is an open set V satisfying $A \subset V \subset B$. Any open set V that satisfies $A \subset V$ is called an *open neighborhood* of A.

Definition A.10 (Upper Hemicontinuity)**.** *A set-valued correspondence $\varphi : X \rightrightarrows Y$ between topological spaces is upper hemicontinuous at the point x if for every open neighborhood U of $\varphi(x)$, the upper inverse image $\varphi^u(U)$ is a neighborhood of x in X. As with functions, we say φ is upper hemicontinuous on X, abbreviated* u.h.c.*, if it is upper hemicontinuous at every point of X.*

Theorem A.11 (**Berge's Maximum Theorem.**)**.** *Let $\varphi : X \rightrightarrows Y$ be a continuous correspondence with nonempty compact values, and suppose $f : X \times Y \to \mathbb{R}$ is continuous. Define the "value function" $m : X \to \mathbb{R}$ by*

$$m(x) \;:=\; \sup_{y \in \varphi(x)} f(x, y),$$

and the correspondence $\mu : X \rightrightarrows Y$ of maximizers by

$$s(x) \;:=\; \{\, y \in \varphi(x) \mid f(x, y) = m(x) \,\}.$$

Then, the value function m is continuous, and the "argmax" correspondence μ is upper hemicontinuous with non-empty and compact values. As a consequence, the sup *may be replaced by* max*.*

Proposition A.12. *Define the "argmin" correspondence $\mathcal{S} : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathcal{F}$*

$$\mathcal{S}(\mu, \nu) := \arg\min_{f \in \mathcal{F}} J_\eta(\mu, \nu; f),$$

and the minimum value function $v : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_{\geq 0}$

$$v(\mu, \nu) := \min_{f \in \mathcal{F}} J_\eta(\mu, \nu; f).$$

Then:

1. *v is continuous on $\Omega = (\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}), W_p + W_p)$;*
2. *for each $\mu, \nu \in \mathcal{P}(\mathcal{X})$, the set $\mathcal{S}(\mu, \nu)$ is nonempty and compact (in $\|\cdot\|_\infty$);*
3. *$\mathcal{S}$ is upper hemicontinuous (u.h.c.) at every (μ, ν).*

Proof. By Proposition A.8, the map $J_\eta : \Omega \times \mathcal{F} \to \mathbb{R}$ is continuous on $\Omega \times \mathcal{F}$. As $\mathcal{F}$ is non-empty and compact, applying Berge's Maximum Theorem A.11 [2] with constant feasible correspondence $\varphi(\mu, \nu) \equiv \mathcal{F}$, the conclusion follows. Indeed, in the notation of Theorem A.11, consider $X = \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$, $Y = \mathcal{F}$, $f = -J_\eta$, and constant $\varphi(\mu, \nu) = \mathcal{F}$ for each $(\mu, \nu) \in X$. □

5. Transferability of the optimal slicer.

Theorem A.13. *Given $\varepsilon > 0$, there exists $\tau > 0$ such that if $W_p(\mu_1, \mu_2) + W_p(\nu_1, \nu_2) < \tau$, then for every $f_1^\star \in \arg\min_{f \in \mathcal{F}} J_\eta(\mu_1, \nu_1; f)$, there exists $f_2^\star \in \arg\min_{f \in \mathcal{F}} J_\eta(\mu_2, \nu_2; f)$ in the ε neighborhood of $f_1^\star$, i.e., $\|f_2^\star - f_1^\star\|_\infty < \varepsilon$.*

Proof. Fix $(\mu_2, \nu_2) \in \Omega$ and $\varepsilon > 0$. We will keep the notations in Proposition A.12 and denote $\mathcal{S}(\mu, \nu) := \arg\min_{f \in \mathcal{F}} J_\eta(\mu, \nu; f)$. Consider the open neighborhood of $\mathcal{S}(\mu_2, \nu_2)$

$$U_\varepsilon := \bigcup_{g \in \mathcal{S}(\mu_2, \nu_2)} \Big\{ f \in \mathcal{F} :\; \|f - g\|_\infty < \varepsilon \Big\}.$$

By upper hemicontinuity of $\mathcal{S}$ at (μ_2, ν_2) (Proposition A.12 and the definition of u.h.c. for correspondences), there exists $\tau > 0$ such that

$$d_\Omega\big((\mu_1,\nu_1),(\mu_2,\nu_2)\big) = W_p(\mu_1,\mu_2) + W_p(\nu_1,\nu_2) < \tau \\ \implies \quad \mathcal{S}(\mu_1,\nu_1) \subset U_\varepsilon.$$

Hence for any $f_1^\star \in \mathcal{S}(\mu_1,\nu_1)$ there exists $f_2^\star \in \mathcal{S}(\mu_2,\nu_2)$ such that $\|f_2^\star - f_1^\star\|_\infty < \varepsilon$.

□

A.1 Quantitative ε–τ control.

We now quantify the ε–τ relation in Theorem A.13. To do so, we make explicit how the smoothing scale η and a bad-event with probability δ come into play. For simplicity, assume $p = 2$ and $c(x,y) = \|x-y\|_2$.

We expect that this analysis will motivate several directions for future research. Without appealing to the Berge's Maximum Theorem, we will show that can essentially say the following: For $i = 1, 2$, let

$$\mathcal{S}_i := \arg\min_{f\in\mathcal{F}} J_\eta(\mu_i,\nu_i;f)$$

denote the set of minimizers, which are non-empty and compact under our assumptions. Given $\varepsilon > 0$, let $U_\varepsilon(\mathcal{S}_i)$ denote the subset of functions in $\mathcal{F}$ whose distance to $\mathcal{S}_i$ is strictly less than ε. Consider the gaps

$$m_\varepsilon^{(i)} \;:=\; \min_{f\in\mathcal{F}\setminus U_\varepsilon(\mathcal{S}_i)} J_\eta(\mu_i,\nu_i;f) \;-\; \min_{g\in\mathcal{F}} J_\eta(\mu_i,\nu_i;g).$$

Then if $\mathcal{X}$ is finite, then with *high probability* one can choose $\tau \le m_\varepsilon^{(1)}/C$, for a constant C depending on η, and obtain $\mathcal{S}_2 \subset U_\varepsilon(\mathcal{S}_1)$. Similarly, if $\tau \le m_\varepsilon^{(2)}/C$, then $\mathcal{S}_1 \subset U_\varepsilon(\mathcal{S}_2)$, which in particular implies that for any $f_1^* \in \mathcal{S}_1$, there exists $f_2 \in \mathcal{S}_2$ with $\|f_2^\star - f_1^\star\|_\infty < \varepsilon$.

Below we present the technical details, precisions, and future open directions.

Assumption A.14. *Let $\Sigma = 2I_d$ and $Z \sim \mathsf{Lap}_d(\Sigma)$. We write $\xi_\eta = \eta Z$ and slicers $g_{\xi_\eta}(x) = f(x) + \eta\langle Z, x\rangle$ for $f \in \mathcal{F}$. To obtain a uniform inverse-Lipschitz bound for the slicers via a union bound, we will assume $\mathcal{X}$ is finite with $|\mathcal{X}| = B$; then there are $B' = \binom{B}{2}$ directions $u_j \in \mathbb{S}^{d-1}$ associated with the pairs of elements in $\mathcal{X}$.*

Step 1: A Lipschitz test for $d^2(\cdot,\cdot)$. First, let $diam(\mathcal{X}) =: D < \infty$ (since, in general, we assume $\mathcal{X}$ compact), and notice that the following inequalities

$$\begin{aligned}
&\big|\|x_1-y_1\|_2^2 - \|x_2-y_2\|_2^2\big| \\
&= |\langle (x_1-y_1)+(x_2-y_2), (x_1-y_1)-(x_2-y_2)\rangle| \\
&\le \big(\|x_1-y_1\|_2 + \|x_2-y_2\|_2\big)\,\|(x_1-y_1)-(x_2-y_2)\|_2 \\
&\le 2D\,\big(\|x_1-x_2\|_2 + \|y_1-y_2\|_2\big) \\
&\le \underbrace{2\sqrt{2}D}_{=:D'}\sqrt{\|x_1-x_2\|_2^2 + \|y_1-y_2\|_2^2} \\
&= D'\big\|(x_1,y_1)-(x_2,y_2)\big\|_2
\end{aligned}$$

imply that the function $d^2/D' : \mathbb{R}^d\times\mathbb{R}^d \to \mathbb{R}$ is 1-Lipschitz in $\mathbb{R}^d\times\mathbb{R}^d$ with the standard Euclidean product metric. Thus, for any $\gamma_1,\gamma_2 \in \mathcal{P}(\mathcal{X}\times\mathcal{X})$, by using duality,

$$\begin{aligned}
W_2(\gamma_1,\gamma_2) &\ge W_1(\gamma_1,\gamma_2) \\
&= \sup_{Lip(\psi)\le 1} \int \psi \,\mathrm{d}(\gamma_1-\gamma_2) \\
&\ge \frac{1}{D'}\int c(x,y)^2\,\mathrm{d}(\gamma_1-\gamma_2). \qquad (5)
\end{aligned}$$

Step 2: Uniform inverse-Lipschitz for the slicers on a high-probability event. Now, we want to uniformly control the inverse-Lipschitz constant (or co-Lipschitz reciprocal) of the injective functions

g_{ξ_η}. We recall that, for any fixed $\eta > 0$, injectivity of the slicers g_{ξ_η} is guaranteed for a.e. ξ_η (Lemma A.3), allowing us to later take $\eta \to 0$. We fix $\eta > 0$. For any distinct $x, x' \in \mathcal{X}$, let $u := (x - x')/\|x - x'\|_2 \in \mathbb{S}^{d-1}$, and from the reverse triangle inequality,

$$\begin{aligned} |g_{\xi_\eta}(x) - g_{\xi_\eta}(x')| &\geq ||\langle \xi_\eta, x - x' \rangle| - |f(x) - f(x')|| \\ &\geq \big(\eta |\langle Z, u \rangle| - L\big) \|x - x'\|_2. \end{aligned} \tag{6}$$

To beat the Lipschitz term L with high probability, we will use the extra assumptions A.14:

1. $\Sigma = 2I_d$,
2. $|\mathcal{X}| = B < \infty$.

Assuming $\Sigma = 2I_d$, by Lemma A.2 one has $\langle Z, u \rangle \sim \mathsf{Lap}_1(0, 1)$ (i.e, we have isotropic variance –it does not depend on the direction u). Thus,

$$\mathbb{P}\big(|\langle Z, u \rangle| < t\big) = 1 - e^{-t}, \qquad t \geq 0.$$

The extra assumption of $\mathcal{X}$ being **finite** with B number of elements, gives at most $B' = \binom{B}{2}$ directions $\{u_j\}_{j=1}^{B'} \subset \mathbb{S}^{d-1}$. We can control $\mathbb{P}\big(\min_{1 \leq j \leq B'} |\langle Z, u_j \rangle| < t\big)$ by the union bound $\sum_{j=1}^{B'} \mathbb{P}(|\langle Z, u_j \rangle| < t) = B'(1 - e^{-t})$. Let $0 < \delta < 1$ such that $t_{\delta,B} := -\ln\big(1 - \frac{\delta}{B'}\big) > L/\eta$. So,

$$\mathbb{P}\Big(\underbrace{\min_{1 \leq j \leq B'} |\langle Z, u_j \rangle| \geq t_{\delta,B}}_{=:G \text{ (“good event”)}} \Big) \geq 1 - B'(1 - e^{-t_{\delta,B}}) =: 1 - \delta.$$

Hence, if we denote by K_Z the inverse-Lipschitz constant of g_{ξ_η}, from 6, on the “good event”, with probability at least $1 - \delta$, we get

$$K_Z \leq \frac{1}{\eta t_{\delta,B} - L}.$$

In particular,

$$\sqrt{\mathbb{E}_Z(K_Z^2 \mathbf{1}_G)} \leq \frac{1}{\eta t_{\delta,B} - L} \quad \text{with } \mathbb{P}(G) \geq 1 - \delta. \tag{7}$$

Therefore, the inverse-Lipschitz constant K_Z of the slicers g_{ξ_η} is globally bounded on a high-probability event G.

Step 3: Bound for the change of J_η with a fixed $f \in \mathcal{F}$. Given $\mu_1, \mu_2, \nu_1, \nu_2 \in \mathcal{P}(\mathcal{X})$, define $\alpha_i^{(Z)} := (g_{\xi_\eta})_{\#}\mu_i$ and $\beta_i^{(Z)} := (g_{\xi_\eta})_{\#}\nu_i$. Let $\sigma_i^{(Z)}$ be the *unique* 1D optimal plan between $\alpha_i^{(Z)}, \beta_i^{(Z)}$, with corresponding *unique* lifted coupling $\gamma_i^{(Z)} \in \Gamma(\mu_i, \nu_i)$ such that $(g_{\xi_\eta}, g_{\xi_\eta})_{\#}\gamma_i^{(Z)} = \sigma_i^{(Z)}$, for $i = 1, 2$. Finally, define, as before,

$$J_\eta(\mu_i, \nu_i; f) := \int c(x, y)^p \, \mathrm{d}\bar{\gamma}_i^{(Z)}(x, y),$$

where $\bar{\gamma}_i^{(Z)} = \mathbb{E}_Z[\gamma_i^{(Z)}]$ for $i = 1, 2$.

Using the above steps, for any $f \in \mathcal{F}$:

$$\begin{aligned} &|J_\eta(\mu_1, \nu_1; f) - J_\eta(\mu_2, \nu_2; f)| \\ &= \left| \mathbb{E}_Z \left(\int c(x, y)^2 \, \mathrm{d}(\gamma_1^{(Z)} - \gamma_2^{(Z)}) \right) \right| \\ &\leq \mathbb{E}_Z \left(\left| \int c(x, y)^2 \, \mathrm{d}(\gamma_1^{(Z)} - \gamma_2^{(Z)}) \right| \mathbf{1}_G \right) \\ &+ \mathbb{E}_Z \left(\left| \int c(x, y)^2 \, \mathrm{d}(\gamma_1^{(Z)} - \gamma_2^{(Z)}) \right| \mathbf{1}_{G^c} \right) \\ &\leq \mathbb{E}_Z \left(\left| \int c(x, y)^2 \, \mathrm{d}(\gamma_1^{(Z)} - \gamma_2^{(Z)}) \right| \mathbf{1}_G \right) \\ &+ 2 diam(\mathcal{X})^2 \delta. \end{aligned} \tag{8}$$

Let us control the first term in 8: Using inequality 5 from Step 1, followed by Lemma A.5 (part 1) with K_Z given from Step 2, and then Lemma A.6, we have

$$\begin{aligned}
&\mathbb{E}_Z\left(\left|\int c(x,y)^2\,\mathrm{d}(\gamma_1^{(Z)}-\gamma_2^{(Z)})\right|\mathbf{1}_G\right) \qquad (9)\\
&\le D'\,\mathbb{E}_Z\left(W_2(\gamma_1^{(Z)},\gamma_2^{(Z)})\,\mathbf{1}_G\right)\\
&\le D'\,\mathbb{E}_Z\left(K_Z\,W_2(\sigma_1^{(Z)},\sigma_2^{(Z)})\,\mathbf{1}_G\right)\\
&\le D'\,\mathbb{E}_Z\left(K_Z\mathbf{1}_G\big(W_2^2(\alpha_1^{(Z)},\alpha_2^{(Z)})+W_2^2(\beta_1^{(Z)},\beta_2^{(Z)})\big)^{\frac{1}{2}}\right)\\
&\le D'\,\mathbb{E}_Z(K_Z\mathbf{1}_G\,(L+\eta\|Z\|_2))\big(W_2^2(\mu_1,\mu_2)+W_2^2(\nu_1,\nu_2)\big)^{\frac{1}{2}}.
\end{aligned}$$

We recall that the terms $\big(W_2^2(\mu_1,\mu_2)+W_2^2(\nu_1,\nu_2)\big)^{\frac{1}{2}}$ and $W_2(\mu_1,\mu_2)+W_2(\nu_1,\nu_2)$ are comparable, since

$$\sqrt{a^2+b^2}\le a+b\le\sqrt{2}\sqrt{a^2+b^2}\qquad\text{for } a,b\ge 0.$$

Moreover, by Cauchy-Schwartz inequality,

$$\begin{aligned}
&\mathbb{E}_Z(K_Z\mathbf{1}_G(L+\eta\|Z\|_2))\\
&\le(\mathbb{E}_Z(K_Z^2\mathbf{1}_G))^{\frac{1}{2}}(\mathbb{E}_Z((L+\eta\|Z\|_2)^2))^{\frac{1}{2}}.
\end{aligned}$$

Again, by Cauchy-Schwartz inequality $\mathbb{E}_Z(\|Z\|_2)\le(\mathbb{E}_Z(\|Z\|_2^2))^{\frac{1}{2}}$, and using $\mathbb{E}_Z(\|Z\|_2^2)=\mathrm{tr}(\Sigma)$ one can estimate

$$\begin{aligned}
\mathbb{E}_Z((L+\eta\|Z\|_2)^2)&\le L^2+2L\eta\mathrm{tr}(\Sigma)^{\frac{1}{2}}+\eta^2\mathrm{tr}(\Sigma)\\
&=(L+\eta(\mathrm{tr}(\Sigma))^{\frac{1}{2}})^2,
\end{aligned}$$

which, in particular, if $\Sigma=2I_d$, then $(\mathrm{tr}(\Sigma))^{\frac{1}{2}}=\sqrt{2d}$. Therefore, replacing the above observations in 9 together with 7, we have

$$\begin{aligned}
&\mathbb{E}_Z\left(\left|\int c(x,y)^2\,\mathrm{d}(\gamma_1^{(Z)}-\gamma_2^{(Z)})\right|\mathbf{1}_G\right)\\
&\le C(\eta,\delta)\,\big(W_2(\mu_1,\mu_2)+W_2(\nu_1,\nu_2)\big),
\end{aligned}$$

where

$$C(\eta,\delta):=2\sqrt{2}\,diam(\mathcal{X})\,\frac{L+\eta\sqrt{2d}}{\eta t_{\delta,B}-L}. \qquad (10)$$

As a conclusion,

$$\begin{aligned}
&|J_\eta(\mu_1,\nu_1;f)-J_\eta(\mu_2,\nu_2;f)| \qquad (11)\\
&\le C(\eta,\delta)\,\big(W_2(\mu_1,\mu_2)+W_2(\nu_1,\nu_2)\big)+2diam(\mathcal{X})^2\delta.
\end{aligned}$$

Step 4: From value stability to argmin transfer.

If $W_2(\mu_1,\mu_2)+W_2(\nu_1,\nu_2)<\tau$, inequality 11 implies

$$\left|\min_{f\in\mathcal{F}}J_\eta(\mu_1,\nu_1;f)-\min_{f\in\mathcal{F}}J_\eta(\mu_2,\nu_2;f)\right|<\underbrace{C(\eta,\delta)\tau+2D^2\delta}_{=:C'}$$

Indeed, if we define $v_i:=\min_{f\in\mathcal{F}}J_\eta(\mu_i,\nu_i;f)$, and $\mathcal{S}_i:=\mathcal{S}(\mu_i,\nu_i)=\arg\min_{f\in\mathcal{F}}J_\eta(\mu_i,\nu_i;f)$ (non-empty and compact), for $i=1,2$, from inequality 11 we have, in particular, that for any $f_1^\star\in\mathcal{S}_1$,

$$v_2\le J_\eta(\mu_2,\nu_2;f_1^\star)<J_\eta(\mu_1,\nu_1;f_1^\star)+C'=v_1+C',$$

and similarly (interchanging roles), for any $f_2^\star\in\mathcal{S}_2$,

$$v_1\le J_\eta(\mu_1,\nu_1;f_2^\star)<J_\eta(\mu_2,\nu_2;f_2^\star)+C'=v_2+C',$$

getting $|v_1-v_2|<C'=C(\eta,\delta)\tau+2diam(\mathcal{X})^2\delta$.

Step 5: Choice of τ, η, δ given $\varepsilon > 0$. Similarly as in the proof of Theorem A.13, we define

$$U_\varepsilon(\mathcal{S}_1) := \{f \in \mathcal{F} \mid \|f - \mathcal{S}_1\|_\infty < \varepsilon\},$$

where $\|f - \mathcal{S}_1\|_\infty$ denotes the L^∞-distance between f and the compact set $\mathcal{S}_1$. Define the "margin" around $\mathcal{S}_1$ by

$$m_\varepsilon^{(1)} := \inf_{f \in \mathcal{F} \setminus U_\varepsilon} J_\eta(\mu_1, \nu_1; f) - \min_{g \in \mathcal{F}} J_\eta(\mu_1, \nu_1; g).$$

Notice that $m_\varepsilon^{(1)} > 0$: indeed, by definition $m_\varepsilon^{(1)} \geq 0$, and if $m_\varepsilon^{(1)} = 0$, it would imply that there exists $\widetilde{f} \in \mathcal{F} \setminus U_\varepsilon(\mathcal{S}_1)$ such that $J_\eta(\mu_1, \nu_1; \widetilde{f}) = v_1$, i.e., $\widetilde{f} \in \mathcal{S}_1$ contradicting $\|\widetilde{f} - \mathcal{S}_1\|_\infty \geq \varepsilon > 0$ (where the existence of minimizer $\widetilde{f}$ relies on the compactness of $\mathcal{F}$ and the continuity of J_η). which is positive by compactness of $\mathcal{F}$ and continuity of J_η. Choose $\delta \in (0, 1)$ and $\eta > 0$ so that $\eta t_{\delta,B} - L > 0$ and the bad-event penalty obeys

$$2diam(\mathcal{X})^2\, \delta \leq m_\varepsilon^{(1)}/4. \tag{12}$$

Then choose

$$0 < \tau \leq \frac{m_\varepsilon^{(1)}}{4\, C(\eta, \delta)}. \tag{13}$$

As a consequence, we obtain

$$\left| \min_{f \in \mathcal{F}} J_\eta(\mu_1, \nu_1; f) - \min_{f \in \mathcal{F}} J_\eta(\mu_2, \nu_2; f) \right| < \frac{m_\varepsilon^{(1)}}{2}.$$

By the usual margin argument (similarly as in the proof of Theorem A.13), this yields $\mathcal{S}_2 \subset U_\varepsilon(\mathcal{S}_1)$. See Figure 8 for a schematic visualization. In fact, if $f_2^\star \in \mathcal{S}_2$ but $f_2^\star \in \mathcal{F} \setminus U_\varepsilon(\mathcal{S}_1)$, by definition of $m_\varepsilon^{(1)}$,

$$v_1 + m_\varepsilon^{(1)} \leq J_\eta(\mu_1, \nu_1; f_2^\star) < v_2 + \frac{m_\varepsilon^{(1)}}{2} < v_1 + m_\varepsilon^{(1)},$$

that is, $m_\varepsilon^{(1)} < m_\varepsilon^{(1)}$, which yields a contradiction.

Changing the roles and defining the margin around $\mathcal{S}_2$

$$m_\varepsilon^{(2)} := \inf_{f \in \mathcal{F} \setminus U_\varepsilon} J_\eta(\mu_2, \nu_2; f) - \min_{g \in \mathcal{F}} J_\eta(\mu_2, \nu_2; g).$$

we have that under the choices:

1. $\delta \in (0, 1)$ and $\eta > 0$ so that $\binom{B}{2}(1 - e^{-L/\eta}) < \delta \leq m_\varepsilon^{(1)}/8diam(\mathcal{X})^2$, and
2. $\tau \leq \frac{m_\varepsilon^{(2)}}{4\, C(\eta, \delta)}$,

we get $\mathcal{S}_1 \subset U_\varepsilon(\mathcal{S}_2)$, where

$$U_\varepsilon(\mathcal{S}_2) := \{f \in \mathcal{F} \mid \|f - \mathcal{S}_2\|_\infty < \varepsilon\}.$$

This part reads as the thesis of Theorem A.13: for every $f_1^\star \in \mathcal{S}_1$, under this choice of parameters η, δ, τ depending on $\varepsilon > 0$, we have $\|f_2^\star - f_1^\star\|_\infty < \varepsilon$ for some $f_2^\star \in \mathcal{S}_2$.

Note: Notice that using these quantitative bounds, there exists a trade-off between δ and η: Indeed, the condition $\eta t_{\delta,B} > L$ (equivalently, $\delta > \binom{B}{2}(1 - e^{-L/\eta})$) ensures that the slicers g_{ξ_η} are uniformly inverse-Lipschitz on a high-probability event. Larger values of η make this condition easier to satisfy (allowing smaller δ), while smaller η forces δ to be larger. On the other hand, the final stability bound requires δ to be sufficiently small in order to control the bad-event penalty.

Comment: For a fixed pair of source-target (μ, ν)measures, J_η is Lipschitz with respect to $(\mathcal{F}, \|\cdot\|_\infty)$ with high probability. Here we repeat some arguments from step 3 but, instead of fixing $f \in \mathcal{F}$ and varying $\mu_i, \nu_i \in \mathcal{P}(X)$ $(i = 1, 2)$, we fix a pair of source-target measures $\mu, \nu\mathcal{P}(\mathcal{X})$ and consider two different arbitrary slicers $f_1, f_2 \in \mathcal{F}$.

Let $f_1, f_2 \in \mathcal{F}$. Fix a realization of Z and set

$$g^{(i)}(x) := g_{\xi_\eta}^{f_i}(x) = f_i(x) + \eta\langle Z, x\rangle \qquad i = 1, 2.$$

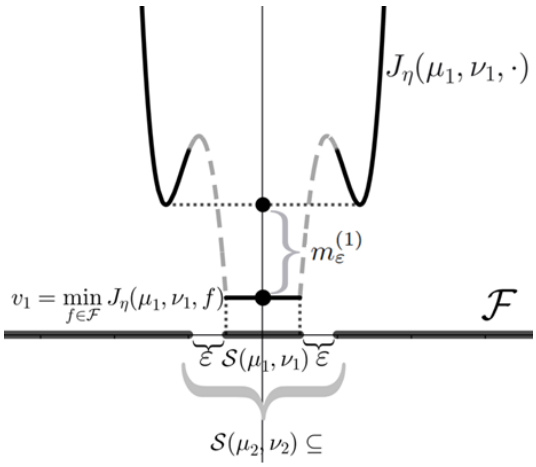

Figure 8: *Sketch illustrating that $\mathcal{S}(\mu_2, \nu_2) \subset U_\varepsilon(\mathcal{S}(\mu_1, \nu_1))$ if τ is chosen appropriately depending on the gap $m_\varepsilon^{(1)}$.*

Then for all $x \in \mathcal{X}$,

$$|g^{(1)}(x) - g^{(2)}(x)| = |f_1(x) - f_2(x)| \ \leq \ \|f_1 - f_2\|_\infty.$$

Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$.

For $i = 1, 2$, let $\alpha_{f_i}^{(Z)} := (g^{(i)})_\# \mu$ and $\beta_{f_i}^{(Z)} := (g^{(i)})_\# \nu$; let $\sigma_{f_i}^{(Z)}$ be the *unique* 1D optimal plan between $\alpha_{f_i}^{(Z)}, \beta_{f_i}^{(Z)}$, with corresponding *unique* lifted coupling $\gamma_{f_i}^{(Z)} \in \Gamma(\mu, \nu)$ such that $(g^{(i)}, g^{(i)})_\# \gamma_i^{(Z)} = \sigma_i^{(Z)}$.

By Lemma A.5, for the pushforwards of μ we obtain

$$W_2\big(\alpha_{f_1}^{(Z)}, \alpha_{f_2}^{(Z)}\big) \ \leq \ \|f_1 - f_2\|_\infty,$$

and similarly for the pushforwards of ν,

$$W_2\big(\beta_{f_1}^{(Z)}, \beta_{f_2}^{(Z)}\big) \ \leq \ \|f_1 - f_2\|_\infty.$$

Applying Lemma A.6, we get

$$\begin{aligned} W_2\big(\sigma_{f_1}^{(Z)}, \sigma_{f_2}^{(Z)}\big) &\leq \Big(W_2^2\big(\alpha_{f_1}^{(Z)}, \alpha_{f_2}^{(Z)}\big) \\ &\qquad + W_2^2\big(\beta_{f_1}^{(Z)}, \beta_{f_2}^{(Z)}\big) \Big)^{1/2} \\ &\leq \sqrt{2}\, \|f_1 - f_2\|_\infty. \end{aligned}$$

As a consequence of steps 1 and 2 and the above inequality,

$$\begin{aligned} &\big| J_\eta(\mu, \nu; f_1) - J_\eta(\mu, \nu; f_2) \big| \\ &\quad = \Big| \mathbb{E}_Z \Big[\int c^2 \, d\gamma_{f_1}^{(Z)} - \int c^2 \, d\gamma_{f_2}^{(Z)} \Big] \Big| \\ &\quad \leq \mathbb{E}_Z \Big[\Big| \int c^2 \, d\gamma_{f_1}^{(Z)} - \int c^2 \, d\gamma_{f_2}^{(Z)} \Big| \Big] \\ &\quad \leq 4 diam(\mathcal{X}) \, \|f_1 - f_2\|_\infty \, \mathbb{E}_Z[K_Z \mathbf{1}_G] + 2 diam(\mathcal{X})^2 \delta \\ &\quad \leq \frac{4 diam(\mathcal{X}) \, (1 - \delta)}{\eta t_{\delta, B} - L} \|f_1 - f_2\|_\infty \ + 2 diam(\mathcal{X})^2 \delta. \end{aligned}$$

Remark A.15. *A useful operational $\varepsilon - \tau$ transfer corollary is as follows: One has*

$$J_\eta(\mu_2, \nu_2; f_1^\star) - \min_{f \in \mathcal{F}} J_\eta(\mu_2, \nu_2; f) \leq 2 A_{\eta, \delta}(\Delta),$$

where we are denoting $\Delta := W_2(\mu_1, \mu_2) + W_2(\nu_1, \nu_2)$, $A_{\eta,\delta}(\Delta) := C(\eta, \delta)\Delta + 2\,\mathrm{diam}(\mathcal{X})^2\delta$. *Indeed, let* $v_i := \min_{f\in\mathcal{F}} J_\eta(\mu_i, \nu_i; f)$, $f_1^\star \in \arg\min_{f\in\mathcal{F}} J_\eta(\mu_1, \nu_1; f)$, *then from our bound 11 and the control on* $|v_1 - v_2|$ *from Step 4, we have*

$$\begin{aligned} 0 \le J_\eta(\mu_2, \nu_2; f_1^\star) - v_2 &= J_\eta(\mu_2, \nu_2; f_1^\star) - J_\eta(\mu_1, \nu_1; f_1^\star) + v_1 - v_2 \\ &\le |J_\eta(\mu_2, \nu_2; f_1^\star) - J_\eta(\mu_1, \nu_1; f_1^\star)| + |v_1 - v_2| \\ &\le 2A_{\eta,\delta}(\Delta). \end{aligned}$$

Therefore, in words, we can say that if δ *is very small so that the residual term* $4\,\mathrm{diam}(\mathcal{X})^2\delta$ *is negligible and can be omitted in* $A_{\eta,\delta}(\Delta)$, *then for all* $\varepsilon > 0$, *one can say that the slicer transferred from Task 1 is* ε*-suboptimal or better on Task 2 as long as* τ *is chosen as* $\tau < \frac{\varepsilon}{2C(\eta,\delta)}$. *However, there is a nuance on the* δ*-choice that is present in the constant* $C(\eta, \delta)$.

B Statistical Properties of Minibatch Training

In this section, we consider the minibatch training of the slicer over finite discrete samples. In each minibatch, samples are reweighted to have uniform mass.

B.1 Approximation and Concentration Bounds

Definition B.1 (Minibatch STP loss as a two-sample U-statistic)**.** *Let* $X = \{x_i\}_{i=1}^N$ *and* $Y = \{y_j\}_{j=1}^M$ *be two sets of discrete samples, and write* $[n] := \{1, \dots, n\}$. *Fix a batch size* $B \le \min(N, M)$. *Define the two-sample kernel for batches* $x_{i_1:i_B}$ *and* $y_{j_1:j_B}$ *with slicer* $f : \mathbb{R}^d \to \mathbb{R}$ *as*

$$h_B(f; x_{i_1:i_B}, y_{j_1:j_B}) := \frac{1}{B}\sum_{i=1}^{B} d\big(x_{(i)}^f,\, y_{(i)}^f\big)^p,$$

where $x_{(i)}^f,\, y_{(i)}^f \in \mathbb{R}^d$ *denote the re-ordered* $x_{i_1:i_B}$ *and* $y_{j_1:j_B}$, *paired by the lifted coupling from 1D optimal coupling, i.e.* $f(x_{(1)}^f) \le \cdots \le f(x_{(B)}^f)$ *and* $f(y_{(1)}^f) \le \cdots \le f(y_{(B)}^f)$ *for the slicer* $f : \mathbb{R}^d \to \mathbb{R}$.

The minibatch STP loss (U-statistic of order B, B*) for a slicer* $f : \mathbb{R}^d \to \mathbb{R}$ *is defined as*

$$J_{N,M,B}(f) := \frac{1}{\binom{N}{B}\binom{M}{B}} \sum_{S\in\binom{[N]}{B}} \sum_{T\in\binom{[M]}{B}} h_B\big(f;\, x_S,\, y_T\big),$$

where S *is the set of* B *indices drawn from* $[N]$ *without replacement and* T *is the set of* B *indices drawn from* $[M]$ *without replacement.* $x_S = (x_i)_{i\in S}$ *and* $y_T = (y_j)_{j\in T}$.

The incomplete estimator draws K *i.i.d. pairs* $\{(S_k, T_k)\}_{k=1}^K$ *uniformly from* $\binom{[N]}{B} \times \binom{[M]}{B}$, *and averages:*

$$\overline{J}_{B,K}(f) := \frac{1}{K}\sum_{k=1}^{K} h_B\big(f;\, x_{S_k},\, y_{T_k}\big).$$

In the following context, $(S, T) \sim \mathrm{Unif}\Big(\binom{[N]}{B} \times \binom{[M]}{B}\Big)$. *(i.e., uniform distribution over all size-*B *subsets.)*

Proposition B.2 (Incomplete vs. complete deviation)**.** *Fix the datasets* $X = \{x_i\}_{i=1}^N, Y = \{y_j\}_{j=1}^M$. *Let* $R := \max_{i,j} d(x_i, y_j)^p$. *Then* $\mathbb{E}_{S,T}[h_B\big(f;\, x_S, y_T\big)] = J_{N,M,B}(f)$ *and, for any* $\epsilon > 0$,

$$\Pr\big(\big|\overline{J}_{B,K}(f) - J_{N,M,B}(f)\big| \ge \epsilon\big) \;\le\; 2\exp\Big(-\,\frac{2K\epsilon^2}{R^2}\Big).$$

In particular,

$$\mathrm{Var}\big(\overline{J}_{B,K}(f)\big) \le \frac{R^2}{4K},$$

$$\mathbb{E}\big[\big|\overline{J}_{B,K}(f) - J_{N,M,B}(f)\big|\big] \le \frac{R}{2\sqrt{K}}.$$

Proof. By definition,

$$J_{N,M,B}(f) = \frac{1}{\binom{N}{B}\binom{M}{B}} \sum_{S \in \binom{[N]}{B}} \sum_{T \in \binom{[M]}{B}} h_B\big(f;\, x_S, y_T\big),$$

so taking expectation over a uniformly chosen (S, T) yields

$$\mathbb{E}_{S,T}\, h_B\big(f;\, x_S, y_T\big) = J_{N,M,B}(f).$$

Next, for each draw $\{(S_k, T_k)\}_{k=1}^K$,

$$\overline{J}_{B,K}(f) = \frac{1}{K} \sum_{k=1}^{K} h_B\big(f;\, x_{S_k}, y_{T_k}\big).$$

Since each summand in h_B is of the form $d(x^f_{(i)}, y^f_{(i)})^p \in [0, R]$, we have $0 \le h_B(f;\, x_S, y_T) \le R$ for all S, T. Hence, the variables $h_B\big(f;\, x_{S_k}, y_{T_k}\big)$ are i.i.d. and lie in $[0, R]$.

Applying Hoeffding's inequality to $\overline{J}_{B,K}(f)$ gives, for any $\epsilon > 0$,

$$\Pr\big(\big|\overline{J}_{B,K}(f) - \mathbb{E}\overline{J}_{B,K}(f)\big| \ge \epsilon\big) \le 2\exp\Big(-\frac{2K\epsilon^2}{R^2}\Big).$$

Using $\mathbb{E}\overline{J}_{B,K}(f) = \mathbb{E}_{S,T}\, h_B(f;\, x_S, y_T) = J_{N,M,B}(f)$ yields the stated tail bound.

For the variance bound, by independence,

$$\mathrm{Var}\big(\overline{J}_{B,K}(f)\big) = \frac{1}{K}\,\mathrm{Var}(h_B(f;\, x_S, y_T)) \le \frac{1}{K} \cdot \frac{R^2}{4} = \frac{R^2}{4K},$$

where we used that a random variable supported on an interval of length R has variance at most $R^2/4$. Finally, by Cauchy–Schwarz,

$$\mathbb{E}\big[\big|\overline{J}_{B,K}(f) - J_{N,M,B}(f)\big|\big] \le \sqrt{\mathrm{Var}\big(\overline{J}_{B,K}(f)\big)} \le \frac{R}{2\sqrt{K}}.$$

□

Lemma B.3 (McDiarmid's Inequality for two i.i.d. samples)**.** *Let $X, Y \subset \mathbb{R}^d$, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$. Let $F : X^N \times Y^M \to \mathbb{R}$ for $N, M \ge 1$. Assume for all $1 \le i \le N, 1 \le j \le M$, the change of F by substituting one coordinate is bounded, i.e. there exists a_i and b_j such that*

$$\big|F(\ldots, x_i, \ldots; y) - F(\ldots, x_i', \ldots; y)\big| \le a_i,$$
$$\big|F(x; \ldots, y_j, \ldots) - F(x; \ldots, y_j', \ldots)\big| \le b_j.$$

where $x \in X^N, y \in Y^M$ and $x_i, x_i' \in X, y_j, y_j' \in Y$. If $x_1, \ldots, x_N \overset{\text{i.i.d.}}{\sim} \mu$ and $y_1, \ldots, y_M \overset{\text{i.i.d.}}{\sim} \nu$ are all independent, then

$$\Pr\big(\big|F(x_{1:N}, y_{1:M}) - \mathbb{E}F(x_{1:N}, y_{1:M})\big| \ge \epsilon\big) \le 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^N a_i^2 + \sum_{j=1}^M b_j^2}\right).$$

Proof. Let $m := N + M$ and define the independent sequence $Z = \{z_l\}_{l=1}^m$ with $z_1 := x_1, \ldots, z_N := x_N, z_{N+1} := y_1, \ldots, z_m := y_M$. With a slight abuse of notation, we write $F(Z) := F(x_{1:N}, y_{1:M})$ and the Doob martingale $M_l := \mathbb{E}[F(Z) \mid z_1, \ldots, z_l]$ for $l = 1, \ldots, m$ and $M_0 = \mathbb{E}(F(Z))$, with increments $\Delta_l := M_l - M_{l-1}$ for $1 \le l \le m$.

Fix l and the prefix $(z_1, \ldots, z_{l-1})$. Consider $\varphi_l(z) := \mathbb{E}[F(z_1, \ldots, z_{l-1}, z, z_{l+1}, \ldots, z_m)|z_1, \ldots, z_{l-1}]$. By the bounded-differences assumption on the lth coordinate, $|\varphi_l(z) - \varphi_l(z')| \le \alpha_l$ for $z, z' \in X$ or $z, z' \in Y$, where

$$\alpha_l = \begin{cases} a_l, & l \le N, \\ b_{l-N}, & l > N. \end{cases}$$

Hence $\Delta_l = \varphi_l(z_l) - \mathbb{E}[\varphi_l(z_l) \mid z_1, \ldots, z_{l-1}]$ is centered and supported in an interval of length at most α_l. By Hoeffding's lemma,

$$\mathbb{E}\big[e^{\lambda \Delta_l} \mid z_1, \ldots, z_{l-1}\big] \le \exp\left(\frac{\lambda^2 \alpha_l^2}{8}\right).$$

Iterating and using the tower property,

$$\begin{aligned}\mathbb{E}\, e^{\lambda(F-\mathbb{E}F)} &= \mathbb{E}\, e^{\lambda \sum_{l=1}^m \Delta_l} \\ &\le \exp\left(\frac{\lambda^2}{8} \sum_{l=1}^m \alpha_l^2\right) \\ &= \exp\left(\frac{\lambda^2}{8}\left(\sum_{i=1}^N a_i^2 + \sum_{j=1}^M b_j^2\right)\right).\end{aligned}$$

By Chernoff's method,

$$\Pr(F - \mathbb{E}F \ge \epsilon) \le \exp\left(-\lambda\epsilon + \frac{\lambda^2}{8}\Big(\sum_i a_i^2 + \sum_j b_j^2\Big)\right).$$

Optimizing at $\lambda^\star = \dfrac{4\epsilon}{\sum_i a_i^2 + \sum_j b_j^2}$ gives

$$\Pr(F - \mathbb{E}F \ge \epsilon) \le \exp\left(-\frac{2\epsilon^2}{\sum_i a_i^2 + \sum_j b_j^2}\right).$$

Apply the same to $-(F - \mathbb{E}F)$ and we can obtain

$$\Pr\big(\big|F - \mathbb{E}F\big| \ge \epsilon\big) \le 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^N a_i^2 + \sum_{j=1}^M b_j^2}\right).$$

□

Proposition B.4 (Dataset → population deviation, fixed f)**.** *Let $x_{1:N} \overset{\text{i.i.d.}}{\sim} \mu$ and $y_{1:M} \overset{\text{i.i.d.}}{\sim} \nu$. Fix f and a batch size $B \le \min(N, M)$. Assume $d(x, y) \le D$ on $\mathcal{X}$ and set $R := D^p$. Define the population target $J_B(f) := \mathbb{E}\, h_B\big(f; x_{i_1:i_B}, y_{j_1:j_B}\big)$ with $x_{i_1:i_B} \sim \mu^{\otimes B}$ and $y_{j_1:j_B} \sim \nu^{\otimes B}$. Then, for any $\epsilon > 0$,*

$$\begin{aligned}&\Pr(|J_{N,M,B}(f) - J_B(f)| \ge \epsilon) \\ &\le 2\exp\left(-\frac{2\,\epsilon^2}{B^2 R^2\left(\frac{1}{N} + \frac{1}{M}\right)}\right).\end{aligned}$$

Proof. $J_{N,M,B}(f)$ is a two-sample U-statistic with kernel h_B, hence $\mathbb{E}\, J_{N,M,B}(f) = J_B(f)$.

View $J_{N,M,B}(f)$ as a function of the independent inputs $(x_1, \ldots, x_N, y_1, \ldots, y_M)$. Replacing a single x_i affects exactly the fraction $\binom{N-1}{B-1}/\binom{N}{B} = B/N$ of the summands; since each summand lies in $[0, R]$ we have a change of at most $(B/N)R$. Similarly, replacing a single y_j changes the value by at most $(B/M)R$. McDiarmid's inequality with bounded differences $a_i = (B/N)R$ for each x_i and $b_j = (B/M)R$ for each y_j yields

$$\begin{aligned}&\Pr\left(|J_{N,M,B}(f) - \mathbb{E}J_{N,M,B}(f)| \ge \epsilon\right) \\ &\le 2\exp\left(-\frac{2\,\epsilon^2}{\sum_i a_i^2 + \sum_j b_j^2}\right) \\ &= 2\exp\left(-\frac{2\,\epsilon^2}{B^2 R^2\left(\frac{1}{N} + \frac{1}{M}\right)}\right),\end{aligned}$$

and $\mathbb{E}J_{N,M,B}(f) = J_B(f)$ finishes the proof. □

Lemma B.5 (Subadditivity and scale control for moduli of continuity)**.** *Let $I \subset \mathbb{R}$ be an interval and $g : I \to \mathbb{R}$. Define the modulus of continuity*

$$\omega_g(t) \;:=\; \sup_{\substack{x,y\in I\\ |x-y|\le t}} |g(x)-g(y)|, \qquad t \ge 0.$$

Then $\omega_g(t)$ has the following properties

(i) (Monotonicity) $\omega_g(0) = 0$ and ω_g is nondecreasing.

(ii) (Subadditivity) For all $s, t \ge 0$,

$$\omega_g(s+t) \;\le\; \omega_g(s) + \omega_g(t).$$

(iii) (Scale-specific linear bound) For any $\eta > 0$ and all $t \ge 0$,

$$\omega_g(t) \;\le\; \Big(\tfrac{t}{\eta} + 1\Big)\, \omega_g(\eta).$$

Proof. (i) Trivial: the supremum over an empty displacement is 0; enlarging the admissible set in t cannot decrease the supremum.

(ii) Fix $x, y \in I$ with $|x - y| \le s + t$. Choose z on the segment $[x, y]$ so that $|x - z| \le s$ and $|z - y| \le t$ (possible since I is an interval). Then

$$|g(x) - g(y)| \;\le\; |g(x) - g(z)| + |g(z) - g(y)| \;\le\; \omega_g(s) + \omega_g(t).$$

Taking the supremum over such x, y gives $\omega_g(s + t) \le \omega_g(s) + \omega_g(t)$.

(iii) Write $t = m\eta + r$ with $m = \lfloor t/\eta \rfloor \in \mathbb{N}$ and $r \in [0, \eta)$. By (ii),

$$\begin{aligned}\omega_g(t) \;&=\; \omega_g(m\eta + r) \;\le\; m\,\omega_g(\eta) + \omega_g(r)\\ &\le\; (m+1)\,\omega_g(\eta) \;=\; \Big(\lceil \tfrac{t}{\eta} \rceil\Big)\,\omega_g(\eta),\end{aligned}$$

and since $\lceil t/\eta \rceil \le t/\eta + 1$, the second inequality follows. □

Corollary B.6 (Scale-specific linearization in expectation)**.** *Let $T \ge 0$ be any random variable with $\mathbb{E}T < \infty$. For any $\eta > 0$,*

$$\mathbb{E}\,\omega_g(T) \;\le\; \omega_g(\eta)\,\Big(\tfrac{\mathbb{E}T}{\eta} + 1\Big).$$

Proposition B.7 (Minibatch → full STP gap under bi-Lipschitz slicer)**.** *Let $X = \{x_i\}_{i=1}^N$, $Y = \{y_j\}_{j=1}^M$ and $f : \mathbb{R}^d \to \mathbb{R}$. Define $\mu_N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$, $\nu_M = \frac{1}{M}\sum_{j=1}^M \delta_{y_j}$ and let $\gamma = \{\gamma_{ij}\}$ be the lifted plan from the 1D optimal transport plan between $f_\#\mu_N$ and $f_\#\nu_M$, with $\gamma_{ij} \ge 0$, $\sum_j \gamma_{ij} = 1/N$, $\sum_i \gamma_{ij} = 1/M$. Recall the STP$_p^p$ objective. For convenience we denote it as $J_{N,M}$ in connection to the minibatch version:*

$$J_{N,M}(f) := \sum_{i=1}^N \sum_{j=1}^M \gamma_{ij}\, d(x_i, y_j)^p.$$

Assume there exist constants $c_X, c_Y > 0$ and $L_X, L_Y \ge 0$ such that

$$|f(x) - f(x')| \;\ge\; c_X\, d(x, x') \quad \textit{for all } x, x' \in X, \tag{14}$$

$$|f(y) - f(y')| \;\ge\; c_Y\, d(y, y') \quad \textit{for all } y, y' \in Y, \tag{15}$$

$$\big|d(x,y)^p - d(x',y')^p\big| \;\le\; L_X\, d(x, x') + L_Y\, d(y, y') \quad \textit{for all } x, x', y, y'. \tag{16}$$

Write the projected oscillations

$$\mathcal{O}_X(f) := \max_{x\in X} f(x) - \min_{x\in X} f(x),$$
$$\mathcal{O}_Y(f) := \max_{y\in Y} f(y) - \min_{y\in Y} f(y).$$

Then, deterministically (for the fixed datasets X, Y),

$$\big|\, J_{N,M,B}(f) - J_{N,M}(f) \,\big| \le \frac{1}{B}\bigg(\frac{L_X}{c_X}\,\mathcal{O}_X(f) + \frac{L_Y}{c_Y}\,\mathcal{O}_Y(f)\bigg) + 2\frac{L_X}{c_X}\omega_X\big(\frac{1}{2}\sqrt{\frac{N-B}{NB}}\big) + 2\frac{L_Y}{c_Y}\omega_Y\big(\frac{1}{2}\sqrt{\frac{M-B}{MB}}\big).$$

where ω_X, ω_Y are the quantile moduli for the pushforward empirical measures $f_{\#}\mu_N$ and $f_{\#}\nu_M$ (nondecreasing, zero at 0, step-like in the discrete case).

Proof. Order the samples by the slicer: $f(x^f_{(1)}) \le \cdots \le f(x^f_{(N)})$ and $f(y^f_{(1)}) \le \cdots \le f(y^f_{(M)})$. Define the stepwise-constant "quantile selectors"

$$X_f(u) := x^f_{(i)} \ \text{ for } u \in \big((i-1)/N,\, i/N\big],$$
$$Y_f(u) := y^f_{(j)} \ \text{ for } u \in \big((j-1)/M,\, j/M\big].$$

Then the full objective admits the integral form

$$J_{N,M}(f) \;=\; \int_0^1 d\big(X_f(u), Y_f(u)\big)^p \,\mathrm{d}u. \tag{17}$$

Partition $[0,1]$ into B equal sub-intervals $I_k = ((k-1)/B,\, k/B]$, and assume $u_k \in I_k$, $v_k \in I_k$ be the rank locations of the k-th order statistics of a size-B subset from X and Y respectively, i.e. the minibatch is constructed by picking one pair of points from each I_k. For any such subset pair (S, T),

$$h_B\big(f; x_S, y_T\big) \;=\; \frac{1}{B}\sum_{k=1}^{B} d\big(X_f(u_k),\, Y_f(v_k)\big)^p.$$

Fix a block $I_k = ((k-1)/B,\, k/B]$. Define the bivariate cost

$$\mathscr{D}_\times(u,v) \;:=\; d\big(X_f(u), Y_f(v)\big)^p, \qquad (u,v) \in [0,1]^2,$$

and the diagonal (comonotone) integrand

$$\mathscr{D}(u) \;:=\; \mathscr{D}_\times(u,u) \;=\; d\big(X_f(u), Y_f(u)\big)^p.$$

Let the bin average be

$$\overline{\mathscr{D}}_k \;:=\; B\int_{I_k} \mathscr{D}(u)\,du.$$

Then for $u_k, v_k \in I_k$,

$$\Big|\frac{1}{B}\,\mathscr{D}_\times(u_k, v_k) - \int_{I_k} \mathscr{D}(u)\,\mathrm{d}u\Big| = \frac{1}{B}\,\big|\mathscr{D}_\times(u_k, v_k) - \overline{\mathscr{D}}_k\big| \le\; \frac{1}{B}\,\mathcal{O}_{I_k\times I_k}(\mathscr{D}_\times),$$

where $\mathcal{O}_{I_k\times I_k}(\mathscr{D}_\times) := \sup_{(u,v),(u',v')\in I_k\times I_k} \big|\mathscr{D}_\times(u,v) - \mathscr{D}_\times(u',v')\big|$.

Next, bound the bin oscillation via the Lipschitz and bi-Lipschitz assumptions: for any $(u,v),(u',v') \in I_k \times I_k$,

$$\begin{aligned}
&\big|\mathscr{D}_\times(u,v) - \mathscr{D}_\times(u',v')\big| \\
&= \big|d(X_f(u), Y_f(v))^p - d(X_f(u'), Y_f(v'))^p\big| \\
&\le L_X\, d\big(X_f(u), X_f(u')\big) + L_Y\, d\big(Y_f(v), Y_f(v')\big) \\
&\le \frac{L_X}{c_X}\,\big|f(X_f(u)) - f(X_f(u'))\big| \\
&+ \frac{L_Y}{c_Y}\,\big|f(Y_f(v)) - f(Y_f(v'))\big|.
\end{aligned}$$

Taking supremum over $I_k \times I_k$ yields

$$\mathcal{O}_{I_k \times I_k}(\mathscr{D}_\times) \;\le\; \frac{L_X}{c_X}\,\mathcal{O}_{I_k}\big(f \circ X_f\big) + \frac{L_Y}{c_Y}\,\mathcal{O}_{I_k}\big(f \circ Y_f\big).$$

Summing over $k = 1, \dots, B$ and using monotonicity of $f \circ X_f$ and $f \circ Y_f$,

$$\sum_{k=1}^{B} \mathcal{O}_{I_k}\big(f \circ X_f\big) = \mathcal{O}_{[0,1]}\big(f \circ X_f\big) = \mathcal{O}_X(f),$$
$$\sum_{k=1}^{B} \mathcal{O}_{I_k}\big(f \circ Y_f\big) = \mathcal{O}_{[0,1]}\big(f \circ Y_f\big) = \mathcal{O}_Y(f).$$

Therefore,

$$\left| \frac{1}{B} \sum_{k=1}^{B} \mathscr{D}_\times(u_k, v_k) - \int_0^1 \mathscr{D}(u)\,\mathrm{d}u \right| \le \frac{1}{B}\left(\frac{L_X}{c_X}\,\mathcal{O}_X(f) + \frac{L_Y}{c_Y}\,\mathcal{O}_Y(f) \right),$$

and since $\int_0^1 \mathscr{D}(u)\,\mathrm{d}u = J_{N,M}(f)$ and $h_B(f; x_S, y_T) = \frac{1}{B}\sum_{k=1}^{B} \mathscr{D}_\times(u_k, v_k)$, the minibatch–full gap bound follows.

Now, if the samples are not stratified, we introduce another bound for the mismatch. Let $S \subset [N]$, $T \subset [M]$ be drawn uniformly without replacement (independently), with $|S| = |T| = B$, and let $X_{f,S}, Y_{f,T}$ be the corresponding batch selectors. Then

$$h_B(f; x_S, y_T) = \frac{1}{B} \sum_{k=1}^{B} d\big(X_{f,S}(u_k), Y_{f,T}(v_k)\big)^p, \qquad u_k, v_k \in I_k.$$

Add and subtract the full selectors at the same (u_k, v_k) and use the triangle inequality:

$$\begin{aligned} &\big|\, h_B(f; x_S, y_T) - J_{N,M}(f) \,\big| \\ &\le \left| \frac{1}{B} \sum_{k=1}^{B} d\big(X_f(u_k), Y_f(v_k)\big)^p - \int_0^1 d\big(X_f(u), Y_f(u)\big)^p\,\mathrm{d}u \right| \\ &+ \underbrace{\frac{1}{B} \sum_{k=1}^{B} \big|\, d\big(X_{f,S}(u_k), Y_{f,T}(v_k)\big)^p - d\big(X_f(u_k), Y_f(v_k)\big)^p \,\big|}_{\text{selector mismatch}}. \end{aligned}$$

The first term is already bounded by $\frac{1}{B}\big(\frac{L_X}{c_X}\mathcal{O}_X(f) + \frac{L_Y}{c_Y}\mathcal{O}_Y(f)\big)$.

For the mismatch term,

$$\begin{aligned} &\big|\, d\big(X_{f,S}(u_k), Y_{f,T}(v_k)\big)^p - d\big(X_f(u_k), Y_f(v_k)\big)^p \,\big| \\ &\le\; L_X\, d\big(X_{f,S}(u_k), X_f(u_k)\big) + L_Y\, d\big(Y_{f,T}(v_k), Y_f(v_k)\big) \\ &\le\; \frac{L_X}{c_X}\Delta_X(u_k) + \frac{L_Y}{c_Y}\Delta_Y(v_k), \end{aligned}$$

where $\Delta_X(u) := \big|f(X_{f,S}(u)) - f(X_f(u))\big|$ and similarly for Δ_Y. Let $Q_X(u), Q_Y(u)$ be the empirical quantile of $f_{\#}\mu_X$, and $Q_{X,S}(u), Q_{Y,T}(u)$ be the quantile of the subsampled distributions. Then $f \circ X_f(u) = Q_X(u)$, $f \circ X_{f,S}(u) = Q_{X,S}(u)$ and $\Delta_X(u) = \big|Q_X(u) - Q_{X,S}(u)\big|$ (and similarly for Y). Define the quantile modulus

$$\omega_X(\eta) := \sup_{\substack{u,v \in [0,1] \\ |u-v| \le \eta}} \big|Q_X(u) - Q_X(v)\big|$$
$$\omega_Y(\eta) := \sup_{\substack{u,v \in [0,1] \\ |u-v| \le \eta}} \big|Q_Y(u) - Q_Y(v)\big|$$

For each $k \in [B]$, let α_k (resp. β_k) be the full-data index of the k-th order in the subsample from X (resp. Y), and define the *actual* full-data levels

$$U_k := \frac{\alpha_k}{N}, \qquad V_k := \frac{\beta_k}{M}.$$

Fix deterministic references $u_k, v_k \in I_k = ((k-1)/B, k/B]$ (e.g. $u_k = v_k = k/B$). Since $f \circ X_{f,S}(u_k) = Q_X(U_k)$, for each k,

$$\begin{aligned}\big|Q_{X,S}(u_k) - Q_X(u_k)\big| &= \big|Q_X(U_k) - Q_X(u_k)\big| \\ &\le \omega_X\big(|U_k - u_k|\big),\end{aligned}$$

and analogously $\big|Q_{Y,T}(v_k) - Q_Y(v_k)\big| \le \omega_Y(|V_k - v_k|)$. Then

$$\begin{aligned}&\big|d(X_{f,S}(u_k), Y_{f,T}(v_k))^p - d(X_f(u_k), Y_f(v_k))^p\big| \\ &\le \frac{L_X}{c_X}\,\omega_X\big(|U_k - u_k|\big) + \frac{L_Y}{c_Y}\,\omega_Y\big(|V_k - v_k|\big).\end{aligned}$$

Averaging over $k = 1, \dots, B$ and over all subsets (S, T) (i.e. taking expectation w.r.t. the uniform law on size-B subsets), we obtain the bound

$$\begin{aligned}&\big|J_{N,M,B}(f) - J_{N,M}(f)\big| \\ &= \big|\mathbb{E}_{S,T} h_B(f; x_S, y_T) - J_{N,M}(f)\big| \\ &\le \mathbb{E}_{S,T}\big|\, h_B(f; x_S, y_T) - J_{N,M}(f)\big| \\ &\le \frac{1}{B}\left(\frac{L_X}{c_X}\,\mathcal{O}_X(f) + \frac{L_Y}{c_Y}\,\mathcal{O}_Y(f)\right) \\ &\quad + \frac{1}{B}\sum_{k=1}^{B}\left(\frac{L_X}{c_X}\,\mathbb{E}\,\omega_X\big(|U_k - u_k|\big) + \frac{L_Y}{c_Y}\,\mathbb{E}\,\omega_Y\big(|V_k - v_k|\big)\right).\end{aligned}$$

For $\mathbb{E}\omega_X\big(|U_k - u_k|\big)$ and $\sigma = \frac{1}{2}\sqrt{\frac{N-B}{NB}}$ we have by Corollary B.6,

$$\begin{aligned}\mathbb{E}\omega_X\big(|U_k - u_k|\big) &\le \Big(\frac{\mathbb{E}|U_k - u_k|}{\frac{1}{2}\sqrt{\frac{N-B}{NB}}} + 1\Big)\omega_X\Big(\frac{1}{2}\sqrt{\frac{N-B}{NB}}\Big) \\ &\le 2\omega_X\Big(\frac{1}{2}\sqrt{\frac{N-B}{NB}}\Big)\end{aligned}$$

where the second inequality comes from the fact that sampling B points without replacement from N points gives

$$Var(U_k) \le \frac{N-B}{4BN}$$

and consequently

$$\mathbb{E}|U_k - u_k| \le \frac{1}{2}\sqrt{\frac{N-B}{NB}}.$$

where $u_k = \mathbb{E}U_k$ as a convenience choice.

Similarly, $\mathbb{E}\omega_Y\big(|V_k - v_k|\big) \le 2\omega_Y\big(\frac{1}{2}\sqrt{\frac{M-B}{MB}}\big)$. Therefore, we get the final bound

$$\begin{aligned}&\big|J_{N,M,B}(f) - J_{N,M}(f)\big| \\ &\le \frac{1}{B}\left(\frac{L_X}{c_X}\,\mathcal{O}_X(f) + \frac{L_Y}{c_Y}\,\mathcal{O}_Y(f)\right) \\ &\quad + 2\frac{L_X}{c_X}\omega_X\Big(\frac{1}{2}\sqrt{\frac{N-B}{NB}}\Big) + 2\frac{L_Y}{c_Y}\omega_Y\Big(\frac{1}{2}\sqrt{\frac{M-B}{MB}}\Big).\end{aligned}$$

□

B.2 Uniform Control and Approximate Minimizers

The preceding concentration bounds are stated for a fixed slicer. To control a slicer learned from the sampled minibatches, we extend these bounds uniformly over the slicer class. We condition on the fixed datasets X and Y and consider the hard STP objectives, using a fixed deterministic rule to resolve ties consistently across the full datasets and their minibatches.

On finite datasets, the full-data objective $J_{N,M}(f)$, the complete minibatch objective $J_{N,M,B}(f)$, and its incomplete estimator $\overline{J}_{B,K}(f)$ depend on f only through the sample orderings it induces on X and Y. Let $Q \le N!M!$ denote the number of realizable pairs of orderings, and choose one representative $f_q \in \mathcal{F}$ for each, independently of the sampled minibatches. Define

$$\beta_B \coloneqq \max_{1\le q\le Q} \left|J_{N,M,B}(f_q) - J_{N,M}(f_q)\right|, \qquad R \coloneqq \max_{i,j} c(x_i, y_j)^p.$$

Applying Proposition B.2 to each representative with failure probability δ/Q and taking a union bound gives, with probability at least $1-\delta$,

$$\max_{1\le q\le Q} \left|\overline{J}_{B,K}(f_q) - J_{N,M,B}(f_q)\right| \le R\sqrt{\frac{\log(2Q/\delta)}{2K}}.$$

Since every slicer shares its objective values with one of the representatives, the triangle inequality yields

$$\begin{aligned}\sup_{f\in\mathcal{F}} \left|\overline{J}_{B,K}(f) - J_{N,M}(f)\right| &\le \beta_B + R\sqrt{\frac{\log(2Q/\delta)}{2K}} \\ &=: A_{B,K}(\delta).\end{aligned}$$

Under the assumptions of Proposition B.7 for each representative, β_B is controlled by the maximum of the corresponding deterministic bias bounds. For fixed datasets, the Monte Carlo term therefore retains its $K^{-1/2}$ rate, with the confidence factor $\log(2/\delta)$ replaced by $\log(2Q/\delta)$.

This uniform control also provides an approximate optimality guarantee for slicers selected using the sampled minibatches. Specifically, let $\widehat{f} \in \mathcal{F}$ satisfy

$$\overline{J}_{B,K}(\widehat{f}) \le \inf_{f\in\mathcal{F}} \overline{J}_{B,K}(f) + \rho, \qquad \rho \ge 0,$$

where ρ represents the empirical optimization error. On the same event,

$$\begin{aligned}J_{N,M}(\widehat{f}) &\le \overline{J}_{B,K}(\widehat{f}) + A_{B,K}(\delta) \\ &\le \inf_{f\in\mathcal{F}} \overline{J}_{B,K}(f) + \rho + A_{B,K}(\delta) \\ &\le \inf_{f\in\mathcal{F}} J_{N,M}(f) + \rho + 2A_{B,K}(\delta).\end{aligned}$$

Consequently,

$$0 \le J_{N,M}(\widehat{f}) - \inf_{f\in\mathcal{F}} J_{N,M}(f) \le 2A_{B,K}(\delta) + \rho.$$

Thus, uniform control over the finitely many realizable orderings translates approximation of the minibatch objective into a bound on full-data objective suboptimality, including for slicers that depend on the sampled minibatches.

C Connecting the Practical Loss to the Theoretical Objective

The implementation target is the exact hard lifted plan, and smoothing is needed to obtain a stable differentiable training objective. We connect the theoretical objective and the practical loss in two steps. For clarity, we consider empirical source and target measures μ and ν with the same number of uniformly weighted samples, so that their OT map exists. The corresponding argument for general plans follows.

Step 1.

$$\mathbb{E}_\xi\left[\widehat{F}^{-1}_{\nu,\xi} \circ \widehat{F}_{\mu,\xi}\right] \quad 3.3 \xrightarrow{\text{first-order approximation}} \left(\mathbb{E}_\xi[\widehat{F}_{\nu,\xi}]\right)^{-1} \circ \mathbb{E}_\xi[\widehat{F}_{\mu,\xi}] \quad \text{(LapSum)}.$$

In 1D, the OT map from the perturbed μ to the perturbed ν is the quantile-matching map $\widehat{T}_\xi = \widehat{F}^{-1}_{\nu,\xi} \circ \widehat{F}_{\mu,\xi}$, where $\widehat{F}_{\mu,\xi}$ and $\widehat{F}_{\nu,\xi}$ are the empirical CDFs. Def. 3.3 lifts the plan and then takes its expectation $\mathbb{E}_\xi[\widehat{F}^{-1}_{\nu,\xi} \circ \widehat{F}_{\mu,\xi}]$. The (normalized) F-Sum functions used by LapSum coincide exactly with the expectation of empirical CDFs of the perturbed projections: $\overline{F}_{\mu,\eta} := \mathbb{E}_\xi[\widehat{F}_{\mu,\xi}]$ and $\overline{F}_{\nu,\eta} := \mathbb{E}_\xi[\widehat{F}_{\nu,\xi}]$:

$$\mathbb{E}_\xi[\widehat{F}_{\mu,\xi}(t)] = \mathbb{E}_\xi\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{Z_i \le t\}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(Z_i \le t) = \frac{1}{n}\sum_{i=1}^{n} F_{\mathrm{Lap},b_i}\left(t - f(x_i)\right) =: \overline{F}_{\mu,\eta}(t).$$

where $Z_i = f(x_i) + \langle \xi, x_i \rangle$, whose marginal distribution is Laplace with scale b_i. They both describe expected portions of perturbed points below t.

Then these expected CDFs from Lapsum are used to construct $\widetilde{T}_{\mathrm{Lap}} = \overline{F}^{-1}_{\nu,\eta} \circ \overline{F}_{\mu,\eta} = (\mathbb{E}_\xi[\widehat{F}_{\nu,\xi}])^{-1} \circ \mathbb{E}_\xi[\widehat{F}_{\mu,\xi}]$ as the soft permutation.Thus, we use LapSum to make the approximation $\mathbb{E}_\xi[\widehat{F}^{-1}_{\nu,\xi} \circ \widehat{F}_{\mu,\xi}] \approx (\mathbb{E}_\xi[\widehat{F}_{\nu,\xi}])^{-1} \circ \mathbb{E}_\xi[\widehat{F}_{\mu,\xi}]$.

This can be viewed as a heuristic first-order approximation. Write $\widehat{F}_{\mu,\xi} = \overline{F}_{\mu,\eta} + \Delta_{\mu,\xi}$ and $\widehat{F}_{\nu,\xi} = \overline{F}_{\nu,\eta} + \Delta_{\nu,\xi}$, where $\mathbb{E}_\xi[\Delta_{\mu,\xi}] = \mathbb{E}_\xi[\Delta_{\nu,\xi}] = 0$. Let $\mathcal{T}(F_\mu, F_\nu) = F_\nu^{-1} \circ F_\mu$, then a first-order expansion around $(\overline{F}_{\mu,\eta}, \overline{F}_{\nu,\eta})$ gives $\mathcal{T}(\widehat{F}_{\mu,\xi}, \widehat{F}_{\nu,\xi}) = \mathcal{T}(\overline{F}_{\mu,\eta}, \overline{F}_{\nu,\eta}) + D_\mu\mathcal{T}[\Delta_{\mu,\xi}] + D_\nu\mathcal{T}[\Delta_{\nu,\xi}] + \mathcal{R}_\xi$, where $\mathcal{R}_\xi$ contains second- and higher-order terms. Taking expectations eliminates both linear terms because the CDF fluctuations are centered. Consequently, $\mathbb{E}_\xi[\widehat{F}^{-1}_{\nu,\xi} \circ \widehat{F}_{\mu,\xi}] = \overline{F}^{-1}_{\nu,\eta} \circ \overline{F}_{\mu,\eta} + \mathbb{E}_\xi[\mathcal{R}_\xi]$. Therefore, LapSum's expected map is first-order accurate, with higher-order residuals. Since both objectives integrate the same finite cost matrix against their respective lifted plans, the plan-level approximation immediately transfers to the losses. More importantly, for finite and distinct projected samples, both constructions converge to the same hard permutation as the perturbation and smoothing scales vanish.

Although a shared ξ would result in different scales b_i for different projected samples, we replace these sample-dependent scales with a common global scale for computational convenience. This introduces a homogeneous-scale approximation, which becomes negligible especially when both scales are small relative to the minimum separation between projected samples.

Step 2.

$$\begin{aligned}\widetilde{P}_X^{-1}\widetilde{P}_Y \quad \text{(LapSum, matrix form)} &\xrightarrow{\text{two-branch error control}} \tfrac{1}{2}\left[P_X^{-1}\widetilde{P}_Y + \left(P_Y^{-1}\widetilde{P}_X\right)^\top\right] \\ &= \tfrac{1}{2}\left(P_X^\top\widetilde{P}_Y + \widetilde{P}_X^\top P_Y\right).\end{aligned}$$

where the last equality uses $P^{-1} = P^\top$ for a hard permutation matrix P. And each branch calculates a transport plan from X to Y, the average is again a transport plan.

The single quantile map $\widetilde{P}_X^{-1}\widetilde{P}_Y$ (matrix form) from Lapsum permutations is still problematic, as for a hard permutation, $P^{-1} = P^\top$, but it does not hold for a soft permutation. Directly constructing the fully soft plan $\widetilde{P}_X^{-1}\widetilde{P}_Y$ via matrix inversion can amplify numerical and approximation errors. By making one permutation hard in each branch, we ensure symmetric gradient update for both distributions while reducing the error compounding towards the true hard plan $P_X^{-1}P_Y$.

D Time and Memory Complexity of the construction of a full soft permutation matrix

We decompose the time and memory for constructing a full soft permutation matrix as follows:

Component	Time	Parallel time	Memory
Two full LapSum permutations	$\mathcal{O}(B^2 \log B)$	$\mathcal{O}(B \log B)$	$\mathcal{O}(B^2)$
Two hard permutations	$\mathcal{O}(B \log B)$	$\mathcal{O}(B \log B)$	$\mathcal{O}(B)$
Transpose and plan composition	$\mathcal{O}(B^2)$	$\mathcal{O}(B)$	$\mathcal{O}(B^2)$
Overall	$\boldsymbol{\mathcal{O}(B^2 \log B)}$	$\boldsymbol{\mathcal{O}(B \log B)}$	$\boldsymbol{\mathcal{O}(B^2)}$

Table 2: *Time and memory complexity for a batch of size B. For LapSum, the soft top-k computations for the B different ranks are mutually independent and can therefore be evaluated concurrently using B parallel workers. For plan composition, multiplication by a hard permutation reduces to independent row or column reindexing.*

E Gradient Behaviors of Differentiable Sorting Surrogates

In complement to Figure 2, which compares the smoothed transport costs, we compare the gradients that DGSWP [10] and LapSum [46] provide for slicer optimization. We use a linear slicer with direction $(\cos\theta, \sin\theta)$, $\theta \in [0, 2\pi)$.

Variance reduction. Without the control variate of [10], the Stein estimator ($N = 20$) is noisy over the whole circle, including the flat regions where the true gradient is zero (Figure 9). The control variate removes this noise away from the ordering changes, but the estimate remains jagged near them unless N is increased.

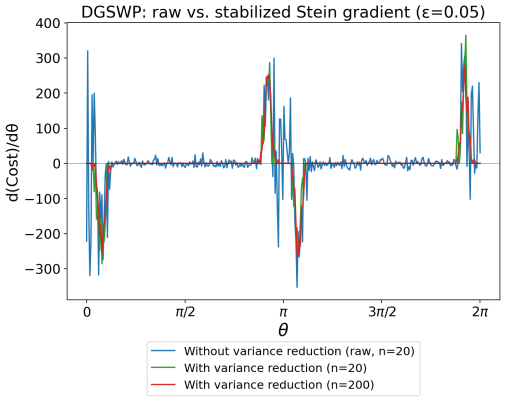

Figure 9: *DGSWP gradient estimates along the slicing angle θ ($\varepsilon = 0.05$). Without variance reduction ($N = 20$), the Stein estimator is noisy everywhere, including where the true gradient is zero. The control variate removes this noise away from the ordering changes ($N = 20$), and increasing to $N = 200$ further smooths the estimate near them.*

Smoothing scale. Even with $N = 200$, DGSWP faces a trade-off in ε (Figure 10, left). For $\varepsilon \leq 0.1$, the gradient is nonzero only in narrow neighborhoods of the ordering changes, so most initial directions receive no descent signal. Larger ε spreads the signal over the circle but with visible Monte Carlo noise, which dominates at $\varepsilon = 5$. LapSum (right) yields deterministic, smooth gradients for every α. Small α recovers the sharp, localized behavior of the hard objective, and larger α widens the region of nonzero gradient while keeping the same sign pattern; only very large α oversmooths and shrinks the magnitude. LapSum thus provides informative, noise-free gradients from a single evaluation, whereas DGSWP needs many OT solves per step and a careful choice of ε.

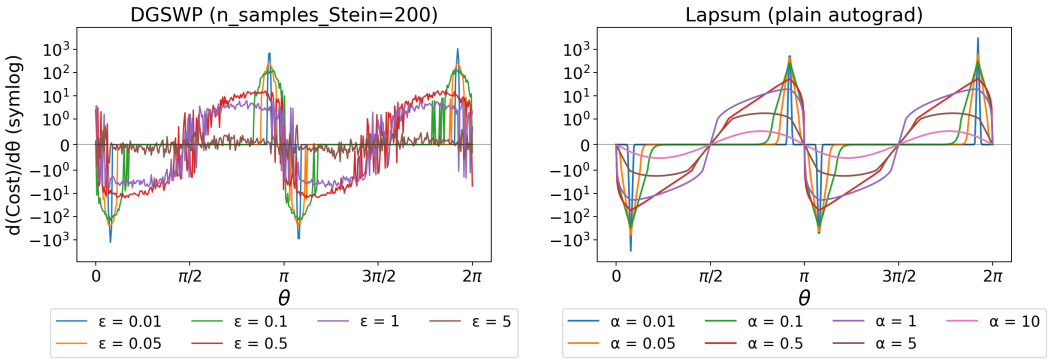

Figure 10: *Gradients along the slicing angle θ (symmetric-log scale). Left: DGSWP with $N = 200$ Stein samples for varying ε. Small ε gives gradients localized at ordering changes; larger ε spreads the signal but adds Monte Carlo noise, which dominates at $\varepsilon = 5$. Right: LapSum gradients by automatic differentiation for varying α. They are deterministic and smooth for all α, and the region of nonzero gradient widens as α grows.*

F Experiment Details

In this section, we provide comprehensive experimental settings and implementation specifics to ensure full clarity and reproducibility. Figure 11 shows the training pipeline throughout the experiment section.

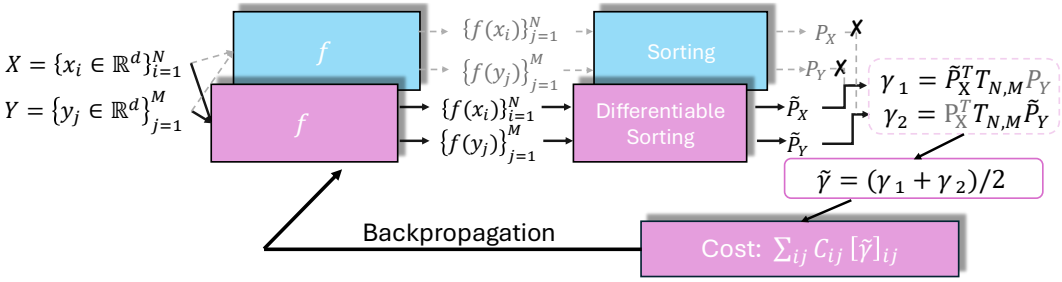

Figure 11: *Training with two-branch symmetric gradients through differentiable sorting. The slicer f projects X and Y to one-dimensional samples that are (soft) sorted to obtain $\tilde{P}_X$ and $\tilde{P}_Y$ alongside hard permutations P_X, P_Y. Two plans are constructed, $\gamma_1 = \tilde{P}_X^\top T_{N,M} P_Y$ and $\gamma_2 = P_X^\top T_{N,M} \tilde{P}_Y$, and the cost with their average is optimized $\tilde{\gamma} = \frac{1}{2}(\gamma_1 + \gamma_2)$ via $\sum_{i,j} c_{ij} [\tilde{\gamma}]_{ij}$.*

F.1 Full vs. Minibatch Comparison

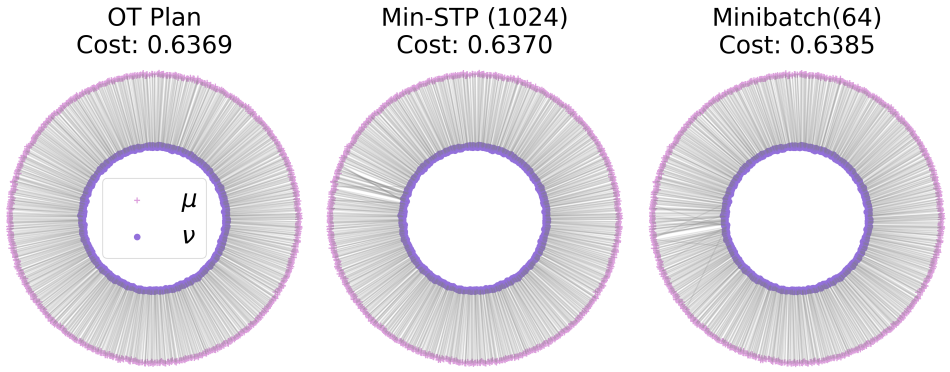

Figure 12: *Transport plans and costs under different training schemes, along with the optimal plan/cost. Each panel visualizes pointwise correspondences (gray segments) between two ring distributions, source μ and target ν, with $N = M = 1024$ points in each distribution. left: exact optimal transport (OT) plan. middle:* min-STP *trained with the full batch (all 1024 samples). right: mini-batch* min-STP *with batch sizes $B = 64$.*

F.2 Transferability under Gradual Drift

F.2.1 Model Configuration

We employ a Set-Transformer [29] architecture to process unordered point sets. The input consists of two-dimensional vectors (x, y), and the model outputs a scalar prediction for each element. The encoder is composed of two stacked Induced Set Attention Blocks (ISAB), each with hidden dimension 64, 4 attention heads, and 16 learned inducing points, along with layer normalization. The ISAB modules implement a two-stage attention mechanism in which a fixed set of inducing vectors first attends to the input set and the input subsequently attends back to these induced features, providing an efficient and scalable approximation of full self-attention. The decoder consists of two Set Attention Blocks (SAB) with the same hidden dimension and number of heads, enabling global context propagation across all elements in the encoded representation. Finally, a linear mapping layer projects the decoder output to a one-dimensional scalar. This design leverages the Set-Transformer's inherent permutation equivariance and expressive attention-based aggregation, making it well-suited for learning functions over sets. The configurations are summarized in Table 3

Configuration Parameter	Value
Input dimension	2
Output dimension	1
Hidden dimension	64
Number of inducing points (ISAB)	16
Number of attention heads	4
Encoder structure	ISAB $\rightarrow$ ISAB
Decoder structure	SAB $\rightarrow$ SAB
Output layer	Linear($64 \rightarrow 1$)

Table 3: *Summary of the SetTransformer configuration used 4.2.*

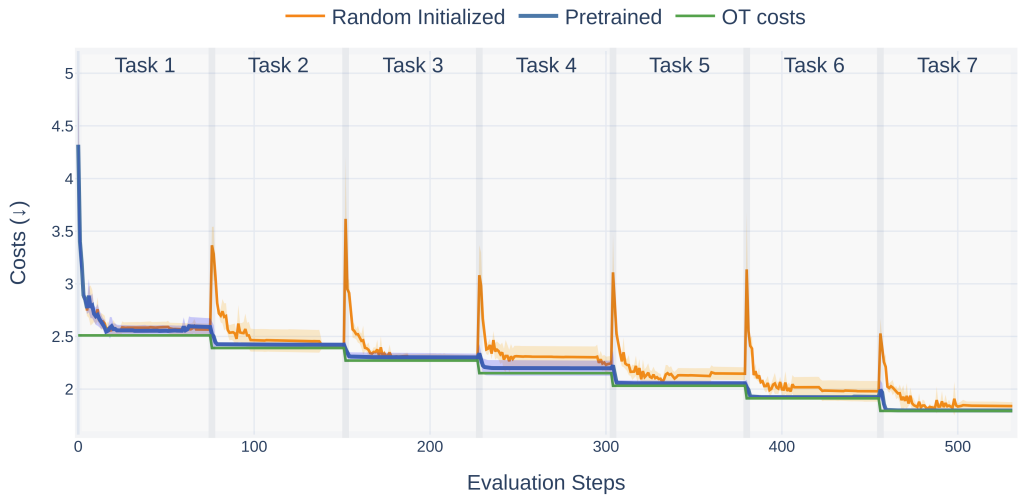

Figure 13: *We plot the transport costs (over 5 runs) evaluated throughout the entire task sequence for two initialization strategies: randomly initialized slicer (orange), pretrained slicer (blue), and the exact OT costs (green) as a lower bound. The pretrained model exhibits fast convergence at each new task and maintains stable performance, whereas random initialization produces pronounced spikes at task boundaries before stabilizing.*

F.2.2 Full Training Dynamics

Figure 13 shows the full evaluation trajectories for all seven tasks.

F.2.3 Ablation studies

To isolate the effects of minibatching, transferability, and Laplace noise, we generate two tasks similarly to the first two tasks, but with the sample size increased to 1024. The goal is to learn the optimal slicer for Task 2. We then compare the resulting transport cost at epochs 20 and 200 across different batch sizes, with and without transfer from task 1, and with and without Laplace noise / LapSum. The results are reported in the following table (averaged over 2 runs), where lower cost is better; for reference, the OT cost is 2.3688.

Overall, larger batch sizes tend to improve stability, transfer provides the strongest gain in the early stage, LapSum yields more consistent improvements across batch sizes and at later stages of training, and the combination of transfer and LapSum remains the closest to the OT cost most consistently.

F.3 Amortized Min-STP for Point Cloud Alignment

F.3.1 Dataset

We used the ModelNet10 dataset [52], a curated subset of the Princeton ModelNet collection containing 10 object categories and 5000 CAD models, with 4078 shapes for training and 922 for testing. Following standard practice, we applied the `NormalizeScale` transform to center each shape and scale it to the unit sphere, ensuring consistent global geometry across the dataset. We then sampled 1024 points uniformly from the mesh surface using the `SamplePoints(1024)` transform, yielding

Table 4: *Ablation results across different batch sizes and training iterations.*

	Iter. 20				Iter. 200			
Batchsize	128	256	512	1024	128	256	512	1024
No transfer & no LapSum	3.1418	3.1088	3.1156	3.0680	2.6466	2.6278	2.7207	2.5679
No transfer & LapSum	3.1176	3.0668	3.1121	3.0910	2.4459	2.4642	2.4588	2.4628
Transfer & no LapSum	2.8252	2.8006	2.6998	3.0888	2.5213	2.5796	2.5224	2.4447
Transfer & LapSum	2.5520	2.5616	2.5169	2.5355	2.3978	2.3819	2.3810	2.3797

point clouds $X \in \mathbb{R}^{1024\times 3}$ used to train the PointNet autoencoder. These normalized and uniformly sampled point clouds form the input for computing the context vectors used in our ModelNet10 alignment experiments.

Table 5: *Statistics of the ModelNet10 dataset. Each object category contains CAD meshes converted to point clouds. Following standard practice, we sample 1024 points per shape using uniform surface sampling.*

Category	**# Train**	**# Test**	**Total**
Bathtub	106	50	156
Bed	515	100	615
Chair	889	100	989
Desk	200	86	286
Dresser	200	86	286
Monitor	465	100	565
Night Stand	287	100	387
Sofa	680	100	780
Table	392	100	492
Toilet	344	100	444
Total	4,078	922	5,000

F.3.2 Context Vectors

To obtain a compact representation of each 3D shape in ModelNet10, we pretrain a PointNet-style autoencoder on point clouds $X \in \mathbb{R}^{N\times 3}$. The encoder ϕ applies a sequence of shared 1D convolutions and batch–normalization layers followed by global max pooling, producing a latent code $c_X = \phi(X) \in \mathbb{R}^{512}$. The decoder ψ maps c_X back to a reconstructed point set $\hat{X} \in \mathbb{R}^{N\times 3}$. We train the autoencoder end-to-end using the symmetric Chamfer distance between X and $\hat{X}$ as the reconstruction loss, which encourages the latent code to capture the global geometry and coarse part structure of each object while remaining invariant to point permutations. After training, we discard the decoder and keep the encoder ϕ. For every point cloud X used in our ModelNet10 alignment experiments, we compute its latent embedding c_X and use this vector as a *context vector* to condition the alignment model, allowing the learned transport plans to adapt to the specific shape instance while sharing parameters across the dataset.

Table 6: *Training hyperparameters used for the PointNet-style autoencoder*

Hyperparameter	**Value**
Latent dimension d_z	512
Batch size	8 (ModelNet10)
Optimizer	Adam
Learning rate	5×10^{-4}
Training epochs	1000
Point sampling	SamplePoints(1024)

F.3.3 Additional Results

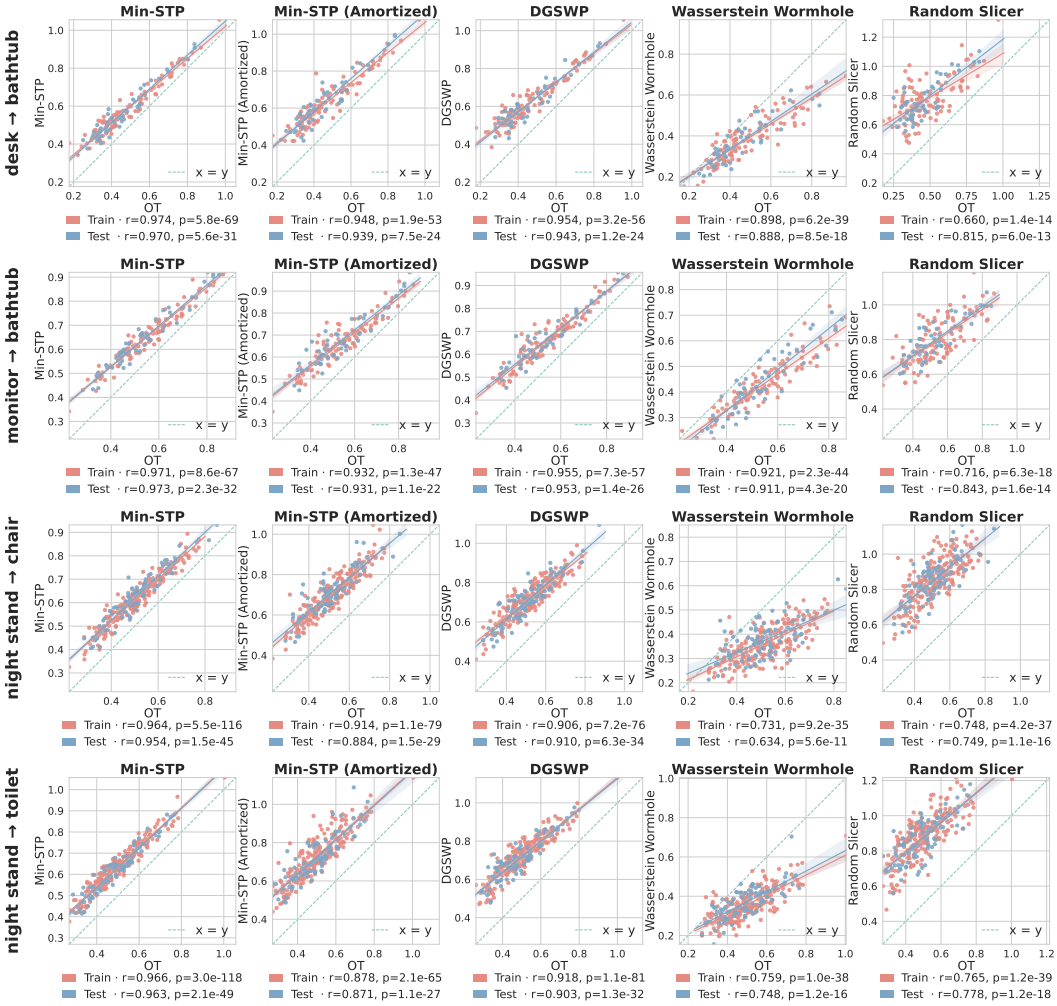

Figure 14: *Additional experiments on Point Cloud Alignment.*

F.3.4 More comparisons with amortized OT approaches

We further compare against Meta-OT [4] and amortized OT with sliced potentials [49] and report correlations across all category pairs in Table 7.

Method	Train Corr.	Test Corr.	Inference Time (ms)	Peak Mem. (MB)
min-STP (Amortized)	0.907 ± 0.048	0.902 ± 0.054	1.654 ± 0.213	41.2
Meta-OT [4]	0.956 ± 0.029	0.939 ± 0.043	5.570 ± 0.502	69.2
Sliced Potentials [49]	0.990 ± 0.005	0.989 ± 0.006	23.848 ± 1.785	49.0

Table 7: *Correlations with exact OT, inference time, and peak memory usage*

Sliced Potentials achieves the highest correlation on clean data, followed by Meta-OT, while min-STP provides substantially faster inference and the lowest peak memory usage. However, Sliced Potentials and Meta-OT are less robust than min-STP in the presence of outliers, as shown in Table 8.

Although Sliced Potentials and Meta-OT perform better in the absence of outliers, their performance deteriorates sharply as the number of outliers increases. This is because the permutation induced

by min-STP remains unchanged once the ordering is fixed, whereas these entropic potential-based approaches are sensitive to outlier magnitude: their dual potentials depend on both ordering and pairwise distances, and this sensitivity can be amplified when reconstructing the dense transport plan from dual potentials and pairwise distances.

Overall, these results reveal a clear trade-off. Meta-OT and Sliced Potentials provide higher correlations on clean data, whereas min-STP offers substantially lower inference cost and markedly greater robustness to outliers, which are common in real-world data. This supports min-STP as an efficient and stable amortized alternative, particularly in settings where point clouds may contain corrupted or extreme observations.

# Outliers (out of 1024 points)	min-STP (Amortized)	Sliced Potentials [49]	Meta-OT [4]
0	0.9167 ± 0.0005	0.9941 ± 0.0005	0.9645 ± 0.0008
5	0.9159 ± 0.0003	0.9587 ± 0.0002	0.4966 ± 0.0050
20	0.9114 ± 0.0005	0.8019 ± 0.0020	0.3616 ± 0.0062
50	0.8995 ± 0.0025	0.4304 ± 0.0086	0.3274 ± 0.0086
100	0.8722 ± 0.0010	-0.0086 ± 0.0148	0.3358 ± 0.0116

Table 8: *Correlations with exact OT under increasing outlier contamination for the desk–table pair. Results are averaged over three runs.*

F.4 Min-STP based Flow

For the flow matching experiments, we first train a slicer for each dataset using a SetTransformer [29] architecture with two self-attention blocks (SAB) in both the encoder and decoder. During training, we treat each point cloud as an individual target distribution and use an isotropic Gaussian as the source distribution. Similar to Section 4.2, we obtain a context vector from the latent space of a pretrained autoencoder and concatenate it with all source and target points.

Once the slicer is trained, we integrate it into the OT-MeanFlow framework to obtain a transport plan for each mini-batch. We follow the training setup and baselines provided in [1], and all experiments use single-step generation (NFE = 1). Additional results on the ModelNet10 [52] dataset are shown in figure 15. Similar to ShapeNet-Chairs, min-STP can generate point clouds that closely mimic the ground truth shapes.

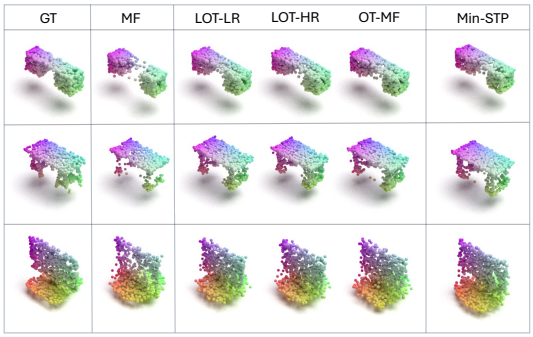

Figure 15: *Single step sample generation on ModelNet10*

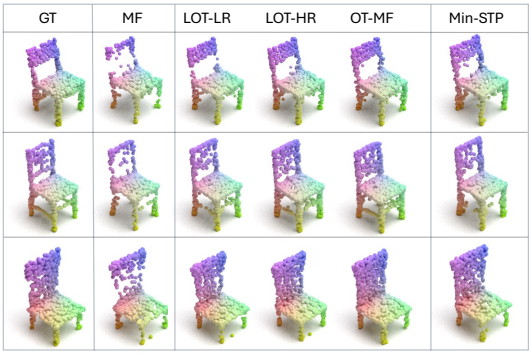

Figure 16: *Single-step sample generation on ShapeNet Chairs.*

F.4.1 Comparisons with DGSWP

Comparison with DGSWP. We provide comparisons against DGSWP [10]. We test three settings. For all DGSWP variants, we follow the defaults of the official implementation: the smoothed gradient is estimated from $N = 10$ perturbed copies of the slicer with smoothing scale $\varepsilon = 0.05$ and variance reduction, and the online slicers are optimized with Adam at a learning rate of 10^{-2}. The first setting follows the original DGSWP implementation: a small MLP maps each 3D point to a scalar and is trained online together with the flow model, with 1000 warm-up steps on the first batch followed by one update per training batch. The MLP slicer has two hidden layers of width 256 with SELU activations, as in the original flow-matching experiments of DGSWP. The second keeps this online training but replaces the MLP with the same set-based slicer used by Min-STP, conditioned on the shape context. The third uses the same set-based slicer, but pre-trains it with the DGSWP objective and then keeps it fixed during flow training, as Min-STP does. Together, these settings separate the effect of the slicer architecture and of pre-training from the effect of the training objective. As shown in Table 9, all DGSWP variants improve over MeanFlow without OT, but all remain well behind Min-STP. Using Min-STP's architecture inside DGSWP does not improve W_2 and makes the OT step about nine times slower, because each smoothed gradient update evaluates several perturbed copies of a much larger network. Pre-training brings the OT time down to that of Min-STP, but gives the worst W_2 among the DGSWP variants. These results suggest that Min-STP's advantage comes from how its slicer is trained, rather than from the slicer architecture or from pre-training alone.

Table 9: *DGSWP ablation on ShapeNet Chairs (single seed). ST denotes the set-based (SetTransformer) slicer used by Min-STP. Online variants update the slicer once per training batch without pre-training; the pre-trained variant learns the slicer before flow training and then keeps it fixed. W_2 is measured between generated and reference shapes on the validation split (lower is better). OT time is the average time, in seconds, spent computing the coupling for one training batch.*

Method	$W_2 \downarrow$	OT time (s)
MF (no OT)	0.0477	0
DGSWP (MLP, online)	0.0210	0.016
DGSWP (ST, online)	0.0218	0.144
DGSWP (ST, pre-trained)	0.0265	0.022
Min-STP	**0.0143**	0.021

F.4.2 Statistical Testing

We complement Figure 5 with a paired statistical comparison on ModelNet10, reported in Table 10. Each method is trained with the same three random seeds, and for every seed we compute the difference in W_2 between Min-STP and the baseline, as well as the relative change in average OT time per training batch. For DGSWP we choose the first setting (MLP, online). We report the mean over the three seeds with a 95% confidence interval based on the t-distribution, and the p-value of a two-sided t-test of whether the mean difference is zero. Min-STP achieves a significantly lower W_2 than MeanFlow without OT and than DGSWP. Compared to OT-MF, which solves the exact OT problem, Min-STP has a slightly higher W_2 but reduces the OT time by about 74%. Against LOT-LR

Table 10: *Paired comparison of Min-STP against each baseline for flow-based generation on ModelNet10 over three seeds, with 95% confidence intervals. ΔW_2: negative means Min-STP is better. Δ OT time: relative change in OT runtime; negative means Min-STP is faster.*

Min-STP vs.	ΔW_2 [95% CI]	p	Δ OT time (%) [95% CI]	p	Interpretation
MF	-0.01570 $[-0.01742, -0.01398]$	0.00065	N/A†	–	Much better W_2
OT-MF	$+0.00213$ $[+0.00099, +0.00328]$	0.015	-74.38 $[-74.74, -74.01]$	0.0000013	Much faster, but worse W_2
LOT-LR	-0.00043 $[-0.00101, +0.00014]$	0.083	-10.80 $[-12.53, -9.07]$	0.0014	Faster, with comparable W_2
LOT-HR	$+0.00027$ $[-0.00096, +0.00149]$	0.448	-57.38 $[-57.97, -56.78]$	0.0000058	Faster, with comparable W_2
DGSWP	-0.00547 $[-0.01029, -0.00065]$	0.040	$+31.34$ $[+28.54, +34.14]$	0.00043	Better W_2, but slower

†MF performs no OT solve, and its OT time is effectively zero. While the W_2 improvements over LOT-LR and LOT-HR are not statistically significant, Min-STP demonstrates clear and significant efficiency gains.

and LOT-HR, the differences in W_2 are not statistically significant, while Min-STP is about 11% and 57% faster, respectively. DGSWP is the only baseline with a lower OT time, about 24% below that of Min-STP, but this comes at a significantly worse W_2. Overall, Min-STP offers the best balance between generation quality and OT cost among the compared methods.

F.5 Additional Experiments on Unpaired Image-to-Image Translation

Additional results for Man-to-Woman image translation is available in Figure 17.



Figure 17: *Man-to-Woman translation with different methods*

F.5.1 Statistical Testing

We provide a paired comparison over three seeds on Man→Woman (Table 11) shows that min-STP attains a significantly lower cost than baselines Sinkhorn, EST, LOT, and DGSWP, and converges significantly faster than ProgOT, OT, and LOT, while ProgOT and OT reach only slightly lower costs; Sinkhorn and EST converge sooner, but to much higher costs. Amortized DGSWP, which follows the original implementation [2] in warm-starting a single slicer across minibatches rather than fitting a new one for each pair, reaches a significantly higher final W_2 than min-STP, with no significant difference in convergence time.

Table 11: *Paired comparison of min-STP against each baseline on MAN→WOMAN over three seeds, with 95% confidence intervals. Cost difference: negative means min-STP is better. Convergence time, the time to reach* 90% *of a method's total* W_2 *reduction: positive means min-STP is faster.*

Baseline	Δ Cost	Δ Time (s)
Sinkhorn ($\epsilon = 0.1$)	-2.31 $(-3.50, -1.12)$	-22.7 $(-34.0, -11.4)$
Sinkhorn ($\epsilon = 1.0$)	-2.93 $(-3.04, -2.81)$	-24.1 $(-32.4, -15.8)$
ProgOT	$+0.127$ $(0.007, 0.246)$	$+217.2$ $(145.4, 289.1)$
OT	$+0.703$ $(0.565, 0.841)$	$+451.1$ $(326.1, 576.2)$
LOT	-2.96 $(-3.07, -2.85)$	$+53.1$ $(46.4, 59.7)$
EST	-2.63 $(-2.74, -2.51)$	-24.7 $(-33.2, -16.3)$
DGSWP	-1.79 $(-2.00, -1.58)$	-2.3 $(-12.1, 7.5)$

[2] https://github.com/rtavenar/dgswp

G Algorithms

In this section we present the key algorithms used in our method, including the mini-batch training procedure and the soft permutation operator. The mini-batch training procedure is detailed in Algorithm 1, with its workflow illustrated in Figure 11. For soft permutations, we borrowed the official implementation [3] of LapSum [46]. However, the code was only for Soft Top-k Scores (see Algorithm 2), here we show how we are constructing a Soft Permutation Matrix from Soft Top-k Scores. Given a score vector $r \in \mathbb{R}^n$, we generate a soft permutation matrix by repeatedly applying the soft top-k operator and then differencing the resulting cumulative masks. The procedure is:

Algorithm 1 Mini-batch training with slicer $f_\theta : \mathbb{R}^d \to \mathbb{R}$

Input: Datasets $X = \{x_i\}_{i=1}^N$, $Y = \{y_j\}_{j=1}^M$; batch size B; mini-batches per epoch k; number of epochs E; α in LapSum; learning rate η.
Output: Optimal parameters θ
Precompute $C_{ij} \leftarrow c(x_i, y_j)$.
Set $a \leftarrow \frac{1}{N}\mathbf{1}_N$, $b \leftarrow \frac{1}{M}\mathbf{1}_M$.
for $e = 1$ **to** E **do**
  Set the batch gradient $G \leftarrow 0$.
  for $i = 1$ **to** k **do**
    Draw random permutations σ of $[N]$ and τ of $[M]$.
    Select first B indices $I \leftarrow \sigma[1:B]$, $J \leftarrow \tau[1:B]$.
    Set the pair of batches $X_B \leftarrow X[I]$, $Y_B \leftarrow Y[J]$.
    $u_x \leftarrow f_\theta(X_B)$, $u_y \leftarrow f_\theta(Y_B)$.
    $\tilde{P}_x \leftarrow \mathrm{LapSum}(u_x; \alpha)$ (2), $P_y \leftarrow \mathrm{HardSort}(u_y)$, $\gamma_1 \leftarrow \tilde{P}_x^\top P_y$.
    $P_x \leftarrow \mathrm{HardSort}(u_x)$, $\tilde{P}_y \leftarrow \mathrm{LapSum}(u_y; \alpha)$, $\gamma_2 \leftarrow P_x^\top \tilde{P}_y$.
    $\tilde{\gamma}_B \leftarrow \frac{1}{2}(\gamma_1 + \gamma_2)$
    $\mathcal{L}_i \leftarrow \frac{1}{B}\langle \tilde{\gamma}_B, C[I,J] \rangle$.
    $G \leftarrow G + \nabla_\theta \mathcal{L}_i$.
  end for
  $\theta \leftarrow \theta - \eta \cdot \frac{1}{k} G$.
end for
Return: θ

1. **Compute soft top-k masks for all k.** For each $k \in \{1, \ldots, n-1\}$, apply the differentiable soft top-k operator to a copy of r:

$$\mathrm{softk}_k = \mathrm{SoftTopK}(r, k, \alpha),$$

   producing a matrix $\mathrm{softk} \in \mathbb{R}^{(n-1)\times n}$ where each row is a softened indicator of the top-k elements.

2. **Pad with boundary rows.** Add a row of zeros at the top and a row of ones at the bottom:

$$R = \begin{bmatrix} \mathbf{0}^\top \\ \mathrm{softk} \\ \mathbf{1}^\top \end{bmatrix} \in \mathbb{R}^{(n+1)\times n}.$$

3. **Difference consecutive rows to obtain a soft permutation.** The soft permutation matrix is obtained by discrete differentiation:

$$P_{\ell i} \;=\; R_{\ell,i} - R_{\ell-1,i}, \qquad \ell = 1, \ldots, n.$$

   Each row of P corresponds to the (soft) probability that element i occupies rank ℓ in the sorted order of r.

The resulting $P \in \mathbb{R}^{n\times n}$ is a row-stochastic, differentiable relaxation of a permutation matrix, consistent with the ordering encoded by the soft top-k operator.

[3] `https://github.com/gmum/LapSum`

Algorithm 2 LapSum (SoftTopK)

Input: Sequence $(s_i)_{i=0}^{n-1}$, parameter $k \in (0, n)$
Sort s in decreasing order into $r = (r_i)_{i=0}^{n-1}$
Set $r_{-1} \leftarrow +\infty$, $r_n \leftarrow -\infty$
Initialize $a_{n-1} \leftarrow 0$, $b_0 \leftarrow 0$, $c_0 \leftarrow 0$
for $j = n-1$ **to** 0 **do**
    $a_{j-1} \leftarrow (1 + a_j)\, \exp(r_j - r_{j-1})$
end for
for $j = 0$ **to** $n-1$ **do**
    $b_{j+1} \leftarrow (1 + b_j)\, \exp(r_{j+1} - r_j)$
    $c_{j+1} \leftarrow 1 + c_j$
end for
Set $w_{-1} \leftarrow 0$, $w_n \leftarrow n$
for $j = 0$ **to** $n-1$ **do**
    $w_j \leftarrow \frac{1}{2} a_j\, \exp(r_{j+1} - r_j) \;-\; \frac{1}{2} b_{j+1} \;+\; c_{j+1}$.
end for
Find $j \in \{0, \ldots, n\}$ such that $k \in [w_{j-1}, w_j]$.
if $j = 0$ **then**
    $b \leftarrow r_0 - \log 2 - \log k + \log a_0$.
else if $0 < j < n$ **then**
    $b \leftarrow r_{j+1} + \log a_j - \log\Big(k - c_{j+1} + \sqrt{|k - c_{j+1}|^2 + a_j b_{j+1}\, e^{\,r_{j+1} - r_j}}\Big)$.
else
    $b \leftarrow r_{n-1} + \log 2 + \log(c_n - k) - \log b_n$.
end if
for $i = 0$ **to** $n-1$ **do**
    $p_i \leftarrow \mathrm{Lap}(s_i - b)$.
end for
Return: $p = (p_i)_{i=0}^{n-1}$.

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: Abstract and introduction are based on the findings in the main paper.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We have included the limitation section.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: Assumptions are provided.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Code repo is provided.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: Code repo is provided and openly accessible.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: In the appendix.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: In the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: A6000

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Absolutely.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [No]

    Justification: Not directly.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification:

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Credited.

    Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification:

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification:

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification:

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification:

Guidelines:

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