SLICED-REGULARIZED OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a new regularized optimal transport (OT) formulation, termed sliced-regularized optimal transport (SROT). Unlike entropic OT (EOT), which regularizes the transport plan toward an independent coupling, SROT regularizes it toward a smoothened sliced OT (SOT) plan. To the best of our knowledge, SROT is the first approach to leverage a version of SOT plan as a reference to improve classical OT. We provide a formal definition of SROT, derive its dual formulation, and provide a post-Bayesian interpretation of SROT. We then develop a Sinkhorn-style algorithm for efficient computation, retaining the same scalability advantages as EOT. By incorporating a scalable SOT plan as a prior, SROT yields more accurate approximations of the exact OT plan than EOT under the same level of regularization. Moreover, the resulting transport plan improves upon the reference SOT plan itself. We further introduce the corresponding OT divergence induced by SROT, named SROT divergence, and analyze its topological and computational properties. Finally, we validate our approach through experiments on synthetic datasets and color transfer tasks, demonstrating that SROT is better than both EOT and SOT in approximating exact OT. Additional experiments on gradient flows further highlight the advantages of SROT divergence.

1 INTRODUCTION

Optimal transport (OT) (Villani, 2003; 2009) and Wasserstein distance are fundamental mathematical tools with various applications in statistics, machine learning, and data sciences. In generative modeling, it has been used to enhance generative adversarial networks (Arjovsky et al., 2017; Genevay et al., 2018), flow-based models (Lipman et al., 2023; Pooladian et al., 2023; Tong et al., 2024), and drifting models (He et al., 2026). In addition, it is used to align source and target distributions in domain adaptation (Courty et al., 2017; Damodaran et al., 2018). Beyond these settings, it has found important applications in computational biology (Bunne et al., 2023; Schiebinger et al., 2019) as well as in image processing (Feydy et al., 2017), signal processing (Kolouri et al., 2017), computer graphics (Solomon et al., 2016; 2015), statistical inference (Bernton et al., 2019a;b; Nguyen et al., 2026b), dependency measurement (Catalano et al., 2021; 2024), among many others.

One challenge in OT is its computation. Entropic regularization (Cuturi, 2013) is the most widely used and principled approach to accelerate OT via approximation. It smooths the transport problem by adding an entropy penalty, yielding a strongly convex formulation known as entropic OT (EOT). In the discrete setting, this regularization guarantees a unique optimal transport plan and enables efficient computation through matrix scaling algorithms such as Sinkhorn–Knopp (Sinkhorn & Knopp, 1967), or equivalently via iterative Bregman projections (Benamou et al., 2015). Moreover, EOT gives rise to the Sinkhorn divergence (Genevay et al., 2018), which metrizes weak convergence of probability measures while remaining significantly more computationally efficient than the Wasserstein distance. Beyond reducing computational complexity to near-quadratic time, entropic regularization improves statistical rates for estimating both the transport cost and Sinkhorn divergence (Genevay et al., 2019) and the transport plan (Manole et al., 2024; Rigollet & Stromme, 2025).

There are several extensions of entropic OT (EOT) obtained by modifying the regularization term. For example, quadratically regularized OT (Lorenz et al., 2021) encourages diffuse but not overly entropic couplings by penalizing the squared mass of the transport plan. Sparsity-constrained OT (Liu et al., 2023) enforces sparsity in the transportation plan, promoting more localized matchings. Low-rank OT (Scetbon & Cuturi, 2022) exploits a low-rank structure in the optimal plan to improve computational efficiency and scalability. While these regularizers encode useful inductive biases

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

about desirable transportation patterns, they are typically non-informative with respect to the true optimal transport plan in the sense that they are not explicitly designed to favor couplings that are close to the unregularized OT solution. As a result, they may introduce bias away from the true OT geometry, potentially trading fidelity for computational or structural convenience. Moreover, recovering the true OT solution might require using very small regularization strengths, but this in turn can significantly degrade the computational advantages of the regularized formulations (Altschuler et al., 2017), as it leads to slower convergence of numerical solvers.

Regularized OT methods can be viewed as encouraging the transport plan to remain close to a prescribed reference (prior) plan under a chosen notion of discrepancy, both of which may be specified explicitly or implicitly. On the discrepancy side, entropic OT (EOT) (Cuturi, 2013) relies on the Kullback–Leibler (KL) divergence, more general Csiszár divergences are considered in (Chizat et al., 2018), and quadratic regularization based on the $\mathbb{L}_2$ norm is studied in (Lorenz et al., 2021). Regarding the choice of prior, EOT adopts the independent coupling, while alternative works explore sparse plans (Lorenz et al., 2021; Liu et al., 2023), low-rank structures (Scetbon & Cuturi, 2022), and Gaussian couplings (Freulon et al., 2025). As noted, these priors are typically non-informative for recovering the true optimal transport plan. However, designing an informative prior remains challenging, as the space of transport plans is highly complex and solving for the exact optimal plan is computationally demanding (Peyré et al., 2019).

We address this challenge by proposing to use a sliced OT (SOT) (Rabin et al., 2012; Nguyen, 2025) plan as the reference plan. SOT plans (Mahey et al., 2023; Liu et al., 2025; Tanguy et al., 2025; Chapel et al., 2026) have recently emerged as efficient proxies for OT plans due to their favorable computational properties. The key idea of SOT is to project high-dimensional probability measures onto one-dimension, where the OT problem admits a closed-form solution, yielding a one-dimensional transport plan. These one-dimensional plans are then lifted back to the original space, and the final transport plan is obtained by aggregating over projections. While being computationally efficient, SOT plans provide only proxy to the true OT and depend critically on the choice and design of projection functions (Chapel et al., 2026; Liu et al., 2025). Nevertheless, using an SOT plan as a reference can yield a more informative prior, inducing a different structure in the resulting new regularized OT problem and potentially improving both the quality of the prior and the final transport plan.

For related work, one-dimensional Kantorovich potentials derived from SOT have been used to initialize the potentials in EOT (Thornton & Cuturi, 2023), although the optimization objective remains that of EOT. In amortized settings involving multiple pairs of probability measures, SOT transportation costs and Kantorovich potentials have also been leveraged to predict their (entropic) OT counterparts (Nguyen et al., 2026a; Truong & Nguyen, 2026). However, to the best of our knowledge, no prior work has employed a SOT plan as a reference measure to define a new class of regularized OT problems. In summary, our contributions are threefold:

1. We propose *sliced-regularized optimal transport* (SROT), a novel regularized OT formulation that employs a smoothened SOT plan as the reference measure and the KL divergence as the discrepancy. We establish the dual formulation of SROT and provide a post-Bayesian interpretation. Furthermore, we derive a Sinkhorn-type algorithm for SROT, closely mirroring the structure of EOT.

2. We introduce the SROT divergence, analogous to the Sinkhorn divergence in EOT. We show that this divergence is symmetric, non-negative, and discriminative. In addition, we prove that it metrizes weak convergence of probability measures, similarly to the Sinkhorn divergence.

3. We validate our approach through experiments on synthetic datasets with diverse geometries, including Half Moons, Eight Gaussians, and Two Rings, demonstrating the effectiveness of SROT in recovering OT plans. We also perform extensive ablation studies on key hyperparameters, such as regularization strength, number of Sinkhorn iterations, and number of projections, highlighting the robustness of the method. These findings are further supported by a color transfer task. Finally, we present a gradient flow experiment illustrating the favorable behavior of the SROT divergence.

Organization. We begin by reviewing background on OT, EOT, and SOT in Section 2. Next, we discuss the definition of SROT, its duality and computation, and SROT divergence in 3. Section 4 presents discussed experiments. Finally, we conclude in Section 5. Additional materials, including technical proofs and experimental results are provided in the Appendices.

2 BACKGROUND

We begin by reviewing the background on OT, EOT, and SOT, and introduce the necessary notation.

Optimal Transport. Given two probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, and a ground metric $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, the OT (Villani, 2009) problem is defined as follows: $\pi^\star \in \arg\min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y)$, where $\Pi(\mu,\nu)$ is the set of admissible transportation plans (joint measures) between μ and ν, and $\pi^\star$ is the optimal transportation plan.

Entropic-Regularized Optimal Transport. EOT (Cuturi, 2013) smoothens the OT problem and it can be conveniently written in a single convex optimization problem as follows:

$$\pi_\varepsilon^\star = \underset{\pi \in \Pi(\mu,\nu)}{\arg\min} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi \mid \mu \otimes \nu), \tag{1}$$

where $\mathrm{KL}(\pi \mid \xi) = \int_{\mathcal{X}\times\mathcal{Y}} \left(\log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\xi}(x,y)\right)\right) \mathrm{d}\pi(x,y)$ and $\mu \otimes \nu$ is the product measure of μ and ν. The optimal plan of EOT is unique as the problem is strongly convex. Let

$$\mathrm{OT}_\varepsilon(\mu,\nu) = \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi \mid \mu \otimes \nu), \tag{2}$$

be the EOT functional, Sinkhorn divergence (Genevay et al., 2018) is defined as follows:

$$\mathcal{S}_\varepsilon(\mu,\nu) = \mathrm{OT}_\varepsilon(\mu,\nu) - \frac{1}{2}\mathrm{OT}_\varepsilon(\mu,\mu) - \frac{1}{2}\mathrm{OT}_\varepsilon(\nu,\nu). \tag{3}$$

Sinkhorn divergence helps to remove the bias created by entropic regularization. In particular, $\mathrm{OT}_\varepsilon(\mu,\nu) = 0 \not\Rightarrow \mu = \nu$ but $\mathcal{S}_\varepsilon(\mu,\nu) = 0 \Rightarrow \mu = \nu$.

Sliced Optimal Transport. SOT considers a function $\mathbb{P}_\theta^c : \mathcal{X} \cup \mathcal{Y} \to \mathbb{R}$ where $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is the ground metric and $\theta \sim \sigma(\theta) \in \mathcal{P}(\Theta)$ where Θ is the space of projection parameters. For example, when $c(x,y) = \|x - y\|_2$, we have $\theta \sim \mathcal{U}(\mathbb{S}^{d-1})$ and $\mathbb{P}_\theta^c = \langle \theta, x \rangle$ (Bonneel et al., 2015; Rabin et al., 2012). For other geometry, we might need to use other types of projections (Kolouri et al., 2019; Bonet et al., 2025; Nguyen & Mueller, 2026). We note that there might not always be a mapping from c to $\mathbb{P}_\theta^c$ as designing projection function for SOT is still an active area of research. For $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, the one-dimensional OT plan with $\mathbb{P}_\theta^c$ admits the following closed-form: $\underline{\pi}_\theta = (F^{-1}_{\mathbb{P}_\theta^c \sharp \mu}, F^{-1}_{\mathbb{P}_\theta^c \sharp \nu}) \sharp \mathcal{U}([0,1])$, where $F^{-1}_{\mathbb{P}_\theta^c \sharp \mu}$ and $F^{-1}_{\mathbb{P}_\theta^c \sharp \nu}$ are quantile functions respectively. With the one-dimensional OT plan $\underline{\pi}_\theta$, we can construct a lifted transportation plan (Muzellec & Cuturi, 2019; Tanguy et al., 2025) as follows:

$$\pi_\theta = \mu_{t_1} \otimes \nu_{t_2} \otimes \underline{\pi}_\theta, \tag{4}$$

where μ and ν are disintegrated as $\mu_{t_1} \otimes \nu_{t_2}$ with respect to $\underline{\pi}_\theta(t_1,t_2)$. In particular, we can write out $\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi_\theta(x,y) = \int_{\mathbb{R}\times\mathbb{R}} \int_{(\mathbb{P}_\theta^c)^{-1}(t_1)\times(\mathbb{P}_\theta^c)^{-1}(t_2)} c(x,y)\mathrm{d}\mu_{t_1} \otimes \nu_{t_2}(x,y)\mathrm{d}\underline{\pi}_\theta(t_1,t_2)$, which is the transportation cost of π_θ. The final transportation plan from SOT is then defined by averaging over all $\theta \sim \sigma(\theta)$ (Liu et al., 2025):

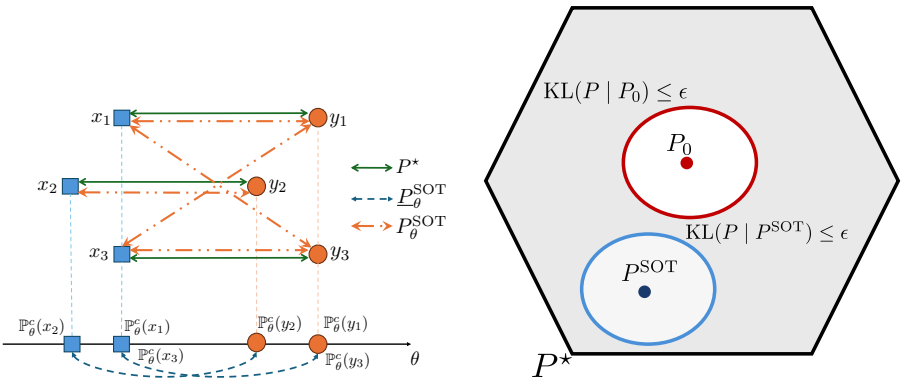
$$\pi^{\mathrm{SOT}} = \mathbb{E}_{\theta \sim \sigma(\theta)}[\pi_\theta], \tag{5}$$

where σ can be chosen to be uniform (Rowland et al., 2019) (with numerical approximation), Softmin of transportation cost of π_θ for $\theta \in \{\theta_1, \ldots, \theta_L\}$ ($L \geq 2$) (Liu et al., 2025), and searching for π_θ with the minimum cost (Mahey et al., 2023; Chapel et al., 2026).

Discrete Cases. In practice, we often work with discrete probability measures. In particular, we have $\mu = \sum_{i=1}^n \alpha_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m \beta_j \delta_{y_j}$ with $\sum_{i=1}^n \alpha_i = \sum_{j=1}^m \beta_j = 1$ and $\alpha_i > 0, \beta_j > 0 \forall i,j$. In this case, the EOT problem becomes:

$$P_\varepsilon^\star = \underset{P \in \Gamma(\boldsymbol{\alpha},\boldsymbol{\beta})}{\arg\min} \langle C, P \rangle + \varepsilon \mathrm{KL}(P \mid P_0), \tag{6}$$

where $\Gamma(\boldsymbol{\alpha},\boldsymbol{\beta}) = \{P \in \mathbb{R}_+^{n\times m} \mid P\mathbf{1} = \boldsymbol{\alpha}, P^\top \mathbf{1} = \boldsymbol{\beta}\}$ is the set of discrete plans, $P_0 = \boldsymbol{\alpha}\boldsymbol{\beta}^\top$, $\mathrm{KL}(P \mid P_0) = \sum_{i=1}^n \sum_{j=1}^m P_{ij} \log\left(\frac{P_{ij}}{P_{0,ij}}\right)$, and $C_{ij} = c(x_i, y_j)$. In this case, the SOT plan, denoted as P_θ^{SOT} equation 4 and P^{SOT} equation 5 can be obtained efficiently in a closed-form (Liu et al., 2025) based on sorting permutation of atoms of $\mathbb{P}_\theta^c \sharp \mu = \sum_{i=1}^n \alpha_i \delta_{\mathbb{P}_\theta^c(x_i)}$ and $\mathbb{P}_\theta^c \sharp \nu = \sum_{j=1}^m \beta_j \delta_{\mathbb{P}_\theta^c(y_j)}$ respectively. The key computational benefit comes from the fact that sorting only costs $\mathcal{O}(n \log n)$ and $\mathcal{O}(m \log m)$.



(a) An example of SOT plan (b) Intuition of SROT

Figure 1: (a) An example of SOT plan (adapting from (Nguyen, 2025)) and (b) intuition of SROT: restricting the feasible set of plans around P^{SOT} (there is a correspondence between ϵ and ε).

3 Sliced-Regularized Optimal Transport

In this section, we define primal and dual formulation of SROT in Section 3.1. We then propose the computational algorithm of SROT in Section 3.2. Finally, we introduce SROT divergence and discuss its topological properties in Section 3.3.

3.1 Primal and Dual Formulation

Definition 1 (SROT) *Let $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ be two probability measures, $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ a ground cost, $\varepsilon > 0$ a regularization parameter, the* sliced-regularized optimal transport *(SROT) is defined as follows:*

$$\pi^\star_{\varepsilon,\mathrm{SOT}} = \underset{\pi \in \Pi(\mu,\nu)}{\arg\min} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}\big(\pi \mid \pi^{\mathrm{SOT}}\big)\,, \tag{7}$$

where $\Pi(\mu,\nu)$ is the set of admissible transportation plans and π^{SOT} is the SOT reference plan.

Compared to EOT in equation 1, the reference coupling is replaced from $\mu \otimes \nu$ to π^{SOT}. As $\varepsilon \to 0$, the optimal plan satisfies $\pi^\star_{\varepsilon,\mathrm{SOT}} \to \pi^\star$, whereas as $\varepsilon \to \infty$, we obtain $\pi^\star_{\varepsilon,\mathrm{SOT}} \to \pi^{\mathrm{SOT}}$. To ensure that equation 7 cover enough feasible transport plans, we require $\pi \ll \pi^{\mathrm{SOT}}$, i.e., that π^{SOT} has full support on $\mathcal{X} \times \mathcal{Y}$. Since this property may not hold in general, a simple remedy is to introduce a smoothed reference plan $\pi^{\mathrm{SOT}}_\gamma = (1-\gamma)\cdot\pi^{\mathrm{SOT}} + \gamma\cdot\mu\otimes\nu$ for $\gamma \in [0,1]$. In practice, however, we find that setting $\gamma = 0$ i.e., using the unsmoothed SOT plan works well in approximating OT. However, to keep consistency between theory and practice, we can keep γ to be very small e.g., $\gamma = 1e-8$. For convenience, we omit γ from the notation and write $\pi^{\mathrm{SOT}}_\gamma$ simply as π^{SOT}, with the understanding that a "SOT plan" may refer to its smoothed version when needed. When $\mu = \sum_{i=1}^n \alpha_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m \beta_j \delta_{y_j}$ with $\sum_{i=1}^n \alpha_i = \sum_{j=1}^m \beta_j$, SROT problem becomes:

$$P^\star_{\varepsilon,\mathrm{SOT}} = \underset{P \in \Gamma(\boldsymbol{\alpha},\boldsymbol{\beta})}{\arg\min} \langle C, P\rangle + \varepsilon \sum_{i=1}^n \sum_{j=1}^m P_{ij} \log\left(\frac{P_{ij}}{P^{\mathrm{SOT}}_{ij}}\right), \tag{8}$$

where $\Gamma(\boldsymbol{\alpha},\boldsymbol{\beta})$ is the set of discrete plans and $C_{ij} = c(x_i, y_j)$. We demonstrate an example of a SOT plan in Figure 1(a) and the intuition of SROT in Figure 1(b) which is changing the center of feasible set of plans from independent plan to a SOT plan.

A Post-Bayesian Interpretation. To further explain why changing the reference plan is a natural idea, we discuss a post-Bayesian view of SROT. We consider a pair of random variables (X, Y) representing a probabilistic alignment on $\mathcal{X} \times \mathcal{Y}$, with $X \sim \mu \in \mathcal{P}(\mathcal{X})$ and $Y \sim \nu \in \mathcal{P}(\mathcal{Y})$. We introduce a prior $\pi_0 \in \Pi(\mu,\nu)$ on (X, Y) as our initial belief. After observing a ground metric c, we aim to update this belief. Following the generalized Bayes framework (Bissiri et al., 2016), noting that specifying a sampling model for the function c is non-trivial, we define a generalized likelihood

Algorithm 1 SOT Plan

Require: Projection function $\mathbb{P}_\theta^c$
1: Initialize $P^{\mathrm{SOT}} \leftarrow 0$
2: **for** $l = 1, \ldots, L$ **do**
3: Sample $\theta_l \sim \mathcal{U}(\mathbb{S}^{d-1})$
4: Obtain $P_{\theta_l}^{\mathrm{SOT}}$ and weights w_l
5: **end for**
6: $P^{\mathrm{SOT}} \leftarrow \sum_{l=1}^{L} \frac{w_l}{\sum_{l'=1}^{L} w_{l'}} P_{\theta_l}^{\mathrm{SOT}}$
7: **return** P^{SOT}

Algorithm 2 SROT Plan

Require: C, P^{SOT}, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, ε
1: $K \leftarrow P^{\mathrm{SOT}} \odot \exp(-C/\varepsilon)$
2: Initialize $\mathbf{u} \leftarrow \mathbf{1}, \mathbf{v} \leftarrow \mathbf{1}$
3: **while** not converged **do**
4: $\mathbf{u} \leftarrow \boldsymbol{\alpha} \oslash (K\mathbf{v})$
5: $\mathbf{v} \leftarrow \boldsymbol{\beta} \oslash (K^\top \mathbf{u})$
6: **end while**
7: **return** $P_{\varepsilon,\mathrm{SOT}} = \mathrm{diag}(\mathbf{u}) K \,\mathrm{diag}(\mathbf{v})$

as $\ell(c \mid X, Y) = \exp\left(-\frac{c(X,Y)}{\varepsilon}\right)$. The resulting generalized posterior is given by $\pi(X, Y \mid c) \propto \exp\left(-\frac{c(X,Y)}{\varepsilon}\right) \pi_0(X, Y)$. As shown in (Bissiri et al., 2016), this posterior is the solution to the following optimization problem: $\arg\min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y), \mathrm{d}\pi(x, y) + \varepsilon \mathrm{KL}\left(\pi \mid \pi_0\right)$, which is exactly the entropic OT problem with KL regularization and reference measure π_0. Therefore, when we use π^{SOT} as the reference plan, we effectively change the prior in this generalized Bayesian model. With a more informative prior, inference can become more concentrated. In SROT, the prior π^{SOT} depends implicitly on the observation c, making the approach an instance of empirical Bayes (Robbins, 1992), i.e., a data-dependent prior. While we can use any transportation plans as the prior plan including low-rank OT (Scetbon & Cuturi, 2022), SOT plans are fast to obtain. SROT serves as a bridge between the SOT literature and regularized OT. As the quality of SOT plans improves, the quality of the corresponding SROT plans also improves.

Duality. We now discuss the duality of SROT which we later use to derive our computational algorithms. Similar to EOT, SROT provides a strong duality.

Theorem 1 (Duality of SROT) *The strong dual of the minimization problem in equation 7 is*

$$\max_{f \in \mathcal{C}(\mathcal{X}), g \in \mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} f(x)\mathrm{d}\mu(x) + \int_{\mathcal{Y}} g(y)\mathrm{d}\nu(y) - \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{f(x) + g(y) - c(x, y)}{\varepsilon}\right) \mathrm{d}\pi^{\mathrm{SOT}}(x, y) \tag{9}$$

where $\mathcal{C}(\mathcal{X})$ and $\mathcal{C}(\mathcal{Y})$ are sets of continuous functions on $\mathcal{X}$ and $\mathcal{Y}$ respectively. The optimal transport plan can be recovered as follows:

$$\mathrm{d}\pi_\varepsilon^\star(x, y) = \exp\left(\frac{f^\star(x) + g^\star(y) - c(x, y)}{\varepsilon}\right) \mathrm{d}\pi^{\mathrm{SOT}}(x, y), \tag{10}$$

where $(f^\star, g^\star)$ are optimal dual potentials.

The proof of Theorem 1 is given in Appendix A.1. Compared to EOT, the optimal plan of SROT is integrated $\left(\frac{f^\star(x)+g^\star(y)-c(x,y)}{\varepsilon}\right)$ with respect to $\pi^{\mathrm{SOT}}(x, y)$ instead of $\mathrm{d}\mu(x)\mathrm{d}\nu(y)$. In the discrete case, $\mu = \sum_{i=1}^n \alpha_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m \beta_j \delta_{y_j}$, the dual problem reads

$$\max_{\mathbf{f} \in \mathbb{R}^n, \mathbf{g} \in \mathbb{R}^m} \langle \mathbf{f}, \boldsymbol{\alpha} \rangle + \langle \mathbf{g}, \boldsymbol{\beta} \rangle - \varepsilon \sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij}^{\mathrm{SOT}} \exp\left(\frac{\mathbf{f}_i + \mathbf{g}_j - C_{ij}}{\varepsilon}\right). \tag{11}$$

Let $C'_{ij} = C_{ij} - \varepsilon \log P_{ij}^{SOT}$, we can see that SROT solves EOT with an adjusted ground cost C' that increases transportation cost for pairs with high probabilities in P_{ij}^{SOT}. An optimal coupling is recovered as $P_{\varepsilon,\mathrm{SOT},ij}^\star = P_{ij}^{\mathrm{SOT}} \exp\left(\frac{\mathbf{f}_i^\star + \mathbf{g}_j^\star - C_{ij}}{\varepsilon}\right)$, where $(\mathbf{f}^\star, \mathbf{g}^\star)$ are optimal discrete potentials. We can see SROT with Sinkhorn solver as an automatic adjustment of the regularization strength ε for each pair of atoms. We recall that SROT is a completely new formulation of regularized OT which can admit different solvers and can be generalized beyond KL divergence regularization.

3.2 Computational Algorithm

From the dual problem equation 9, we can perform stochastic optimization as discussed in Genevay et al. (2016). Nevertheless, SOT shows the most benefit in the discrete settings where SOT plans can be obtained efficiently. Therefore, we now focus our computational discussion to discrete case. We first start with deriving Sinkhorn algorithm for SROT by performing gradient-based optimization for $\mathbf{f}$ and $\mathbf{g}$ in equation 11.

Proposition 1 *The maximization of the duality equation 11 over $(\mathbf{f}, \mathbf{g})$ can be performed via coordinate ascent, where each block update admits the closed-form expressions*

$$\mathbf{f}_i = \varepsilon \left[\log \alpha_i - \log \sum_{j=1}^{m} P^{\mathrm{SOT}}_{ij} \exp\left(\frac{\mathbf{g}_j - C_{ij}}{\varepsilon} \right) \right], \tag{12}$$

$$\mathbf{g}_j = \varepsilon \left[\log \beta_j - \log \sum_{i=1}^{n} P^{\mathrm{SOT}}_{ij} \exp\left(\frac{\mathbf{f}_i - C_{ij}}{\varepsilon} \right) \right]. \tag{13}$$

The proof of Proposition 1 is given in Appendix A.2. Let $K = P^{\mathrm{SOT}} \odot \exp(-C/\varepsilon)$, where all exponential and logarithm operations are understood elementwise. Then the coordinate updates can be written as $\mathbf{f} = \varepsilon \left[\log \boldsymbol{\alpha} - \log \left(K \exp(\mathbf{g}/\varepsilon) \right) \right], \qquad \mathbf{g} = \varepsilon \left[\log \boldsymbol{\beta} - \log \left(K^\top \exp(\mathbf{f}/\varepsilon) \right) \right]$, which is a log-stable Sinkhorn algorithm (Benamou et al., 2015). We would like to also discuss the Sinkhorn-style update via matrix scaling. Let $\mathbf{u} = \exp(\mathbf{f}/\varepsilon)$ and $\mathbf{v} = \exp(\mathbf{g}/\varepsilon)$, and $K = P^{\mathrm{SOT}} \odot \exp(-C/\varepsilon)$, the dual updates are equivalent to the multiplicative scaling iterations

$$\mathbf{u} = \boldsymbol{\alpha} \oslash (K\mathbf{v}), \qquad \mathbf{v} = \boldsymbol{\beta} \oslash (K^\top \mathbf{u}), \tag{14}$$

where $\oslash$ denotes elementwise division. The transport plan admits the factorized form

$$P_{\varepsilon,\mathrm{SOT}} = \mathrm{diag}(\mathbf{u}) K \, \mathrm{diag}(\mathbf{v}), \tag{15}$$

and converges to a unique fixed point, as in EOT (Sinkhorn, 1967), due to the smoothing of P^{SOT}. We summarize the algorithms for computing the SOT plan and the SROT plan in Algorithm 1 and Algorithm 2, respectively. In practice, we run $T > 0$ Sinkhorn iterations and apply early stopping when the maximum marginal violation falls below a prescribed tolerance. Since we recover the same scaling structure as in EOT, existing analyses of computational complexity directly apply (Altschuler et al., 2017). The only additional cost arises from computing the SOT plan, which is negligible compared to the iterative Sinkhorn procedure.

3.3 Sliced-Regularized Optimal Transport Divergence

We now discuss the transportation cost aspect of SROT. As in EOT, SROT provides an approximation of the optimal transport plan. Due to approximation error, the resulting transport cost is generally biased in the sense that it can be 0 for two different probability measures. To address this issue, we introduce a debiased version, which we refer to as the SROT divergence.

Definition 2 (SROT divergence) *Let $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ be two probability measures, $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ a ground cost, $\varepsilon > 0$ a regularization parameter, a fixed SOT plan π^{SOT} (independent copies of μ and ν), we define the SROT functional as follows:*

$$\mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y) + \varepsilon \mathrm{KL}\big(\pi \mid \pi^{\mathrm{SOT}}\big). \tag{16}$$

With the SROT functional, we define the SROT divergence as follows:

$$\mathcal{S}_{\varepsilon,\mathrm{SOT}}(\mu, \nu) = \mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \nu) - \frac{1}{2}\mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \mu) - \frac{1}{2}\mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\nu, \nu). \tag{17}$$

SROT is motivated by Sinkhorn divergence (Genevay et al., 2018) from EOT. It contains the first term $\mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \nu)$ for attraction and two terms $-\frac{1}{2}\mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \mu)$ and $-\frac{1}{2}\mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\nu, \nu)$ for repulsion. The SROT divergence can also be seen as the interpolation of Wasserstein distance when $\varepsilon \to 0$ and a version of the maximum mean discrepancy (MMD) (Gretton et al., 2012) when $\varepsilon \to \infty$ (please see (Feydy et al., 2019)).

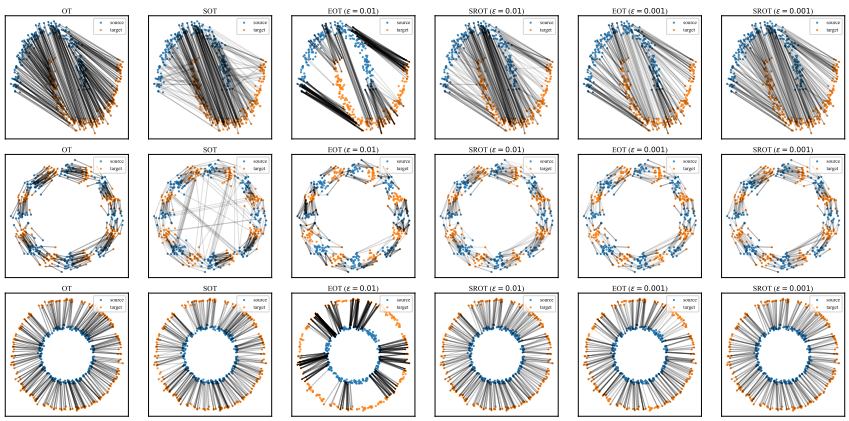


Figure 2: Visualization of transportation plans from OT, SOT, EOT, and SROT for synthetic datasets

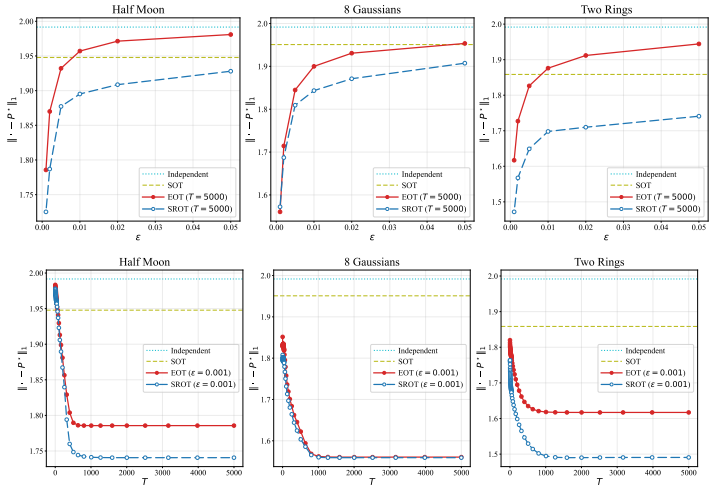


Figure 3: Ablation studies of varying the regularization strengths (ε) and Sinkhorn iterations (T)

Theorem 2 (Topological properties) *Let $\mathcal{X}, \mathcal{Y}$ be compact metric spaces and $c(x, y)$ a Lipschitz cost, assume $\pi^{\mathrm{SOT}}(x, y) \exp(-c(x, y)/\varepsilon)$ is positive definite, then $\mathcal{S}_{\varepsilon,\mathrm{SOT}}$ is symmetric, non-negative, and satisfies: $\mu = \nu \iff \mathcal{S}_{\varepsilon,\mathrm{SOT}}(\mu, \nu) = 0$, for any $\varepsilon > 0$. Moreover, it metrizes weak convergence: $\mu_n \rightharpoonup \mu \iff \mathcal{S}_{\varepsilon,\mathrm{SOT}}(\mu_n, \mu) \to 0$, for any $\varepsilon > 0$.*

The proof of Theorem 2 is given in Appendix A.3, which follows techniques in Feydy et al. (2019). Theorem 2 guarantees the usage of SROT divergence as a loss for estimating parameters.

4 Experiments

In this section, we aim to compare SROT with EOT in approximating OT. In particular, we focus on the Euclidean setting where $c(x, y) = \|x - y\|_2$ as it appears widely in practice. We conduct comparison on synthetic datasets in Section 4.1 and color transfer in Section 4.2. In addition, we compare SROT divergence with Sinkhorn divergence in gradient flow in Section 4.3. Additional experiments mentioned in the main paper is given in Appendix B. All experiments are conducted on a HP Omen 25L desktop.

4.1 Synthetic Data

We construct discrete measures μ and ν on $\mathbb{R}^d$ by sampling n points from three synthetic datasets with distinct geometries: *half moon*, consisting of two noisy, interleaved half-moons defining μ and ν; *8 Gaussians*, where each distribution is a mixture of eight isotropic Gaussian components; and *two rings*, where μ and ν are supported on concentric circles with different radii and small radial

Table 1: $\mathbb{L}_1$ error (mean $\pm$ std) versus exact OT across 132 pairs of images in color transfer.

ε	SOT	EOT	SROT
0.1	1.8463 ± 0.0285	1.9487 ± 0.0090	$\mathbf{1.8007 \pm 0.0303}$
0.01	1.8463 ± 0.0285	1.7770 ± 0.0519	$\mathbf{1.6212 \pm 0.0596}$
0.001	1.8463 ± 0.0285	1.1760 ± 0.3386	$\mathbf{1.1485 \pm 0.3317}$

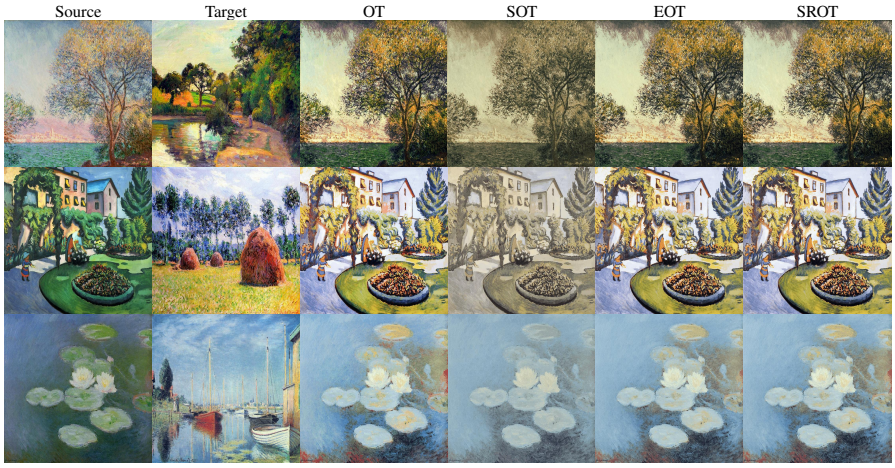


Figure 4: Color transfer results of OT, SOT, EOT, and SROT.

perturbations. For SOT, we adopt a uniform slicing distribution by default. Empirically, SROT with this choice performs robustly compared to alternative slicing strategies, even though some may provide a closer reference to the exact OT plan. We refer the reader to the ablation study in Figure 6 (Appendix B) for further details. We set the number of projections to $L = 100$ for SOT.

Visualization of transportation plans. We visualize the transportation plans obtained from exact OT, SOT, EOT, and SROT in Figure 2. For EOT and SROT, we report results with $\varepsilon \in \{0.01, 0.001\}$ and $T = 5000$ Sinkhorn iterations. The results show that SROT yields visually more accurate transport plans than EOT, particularly for larger values of ε. Moreover, SROT remains effective even when the SOT approximation is relatively poor, as illustrated in the 8-Gaussians example.

Regularization strength (ε). We fix the number of Sinkhorn iterations to $T = 5000$ and vary the entropic regularization parameter ε. For each ε, we compare EOT and SROT using the approximation error $\|\cdot - P^\star\|_1$, where $P^\star$ denotes the exact OT plan. We also include two horizontal baselines under the same metric: the independent product coupling and the SOT plan. The results are shown in the first row of Figure 3. We observe that the SOT plan is consistently closer to the OT plan than the independent coupling. Consequently, SROT outperforms EOT across nearly all choices of ε, with the exception of a single setting at very small ε in the 8 Gaussians case. As ε increases, EOT gradually approaches the independent coupling, while SROT converges toward the SOT plan.

Sinkhorn iterations (T). We fix $\varepsilon = 0.001$ and record the approximation error to the true OT plan as a function of the number of Sinkhorn iterations T, again comparing EOT and SROT. The results are shown in the second row of Figure 3. Overall, SROT achieves a lower approximation error than EOT at convergence, while both methods exhibit comparable convergence rates.

Computational Speed. We report the wall-clock runtime of SROT in Figure 7 in Appendix B. We find that the cost of computing the SOT reference plan is negligible compared to the runtime of the Sinkhorn algorithm, even when parallelization is employed for SOT (since the projections are independent). Consequently, the Sinkhorn iterations dominate the overall computational cost for both EOT and SROT, leading us to conclude that SROT is comparable to Sinkhorn in terms of efficiency.

4.2 Color Transfer

Color transfer is formulated as an OT problem by representing each image as a weighted point cloud in the normalized RGB space, $[0, 1]^3$. Each image is discretized into $K = 256$ colors via median-cut quantization without dithering, yielding palette centroids (atoms) and normalized bin frequencies

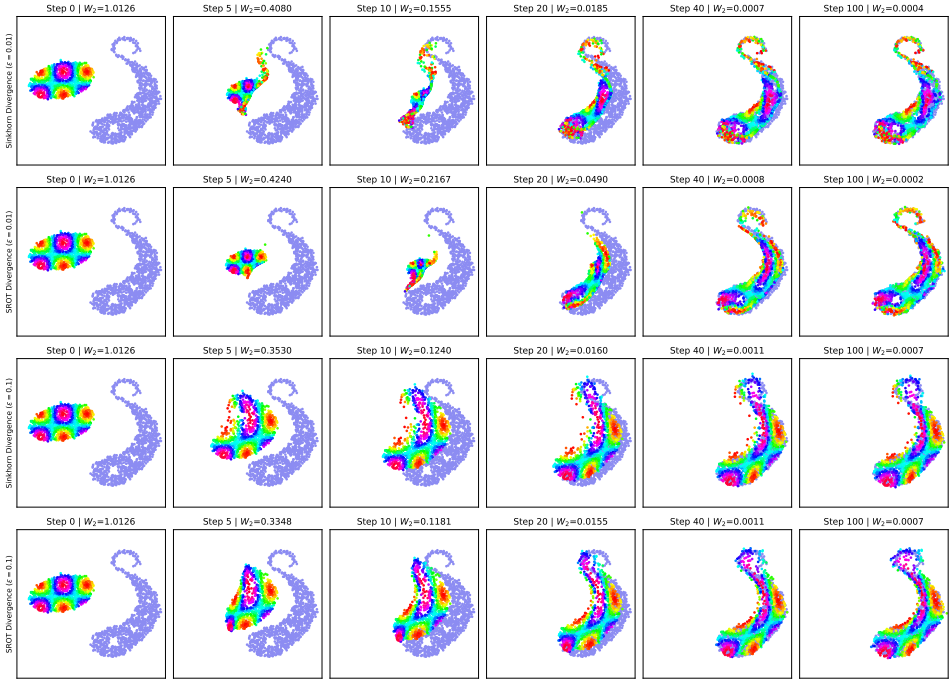


Figure 5: Gradient flows of Sinkhorn divergence and SROT divergence.

(weights). We compare three couplings: (i) SOT, (ii) EOT, and (iii) SROT. The entropic regularization parameter is swept over $\varepsilon \in \{10^{-3}, 10^{-2}, 10^{-1}\}$ with $T = 5000$ Sinkhorn iterations. For each image pair, the transferred image is obtained via barycentric projection, where each source color bin is mapped to a convex combination of target centroids using the row-normalized transport plan. Quantitative performance is reported as the mean and standard deviation of the $\mathbb{L}_1$ error between approximate and exact transport plans across 132 image pairs (Table 1). Qualitative results for three random pairs are shown in Figure 4. Overall, SROT consistently outperforms both SOT and EOT in color transfer, especially with large ε, in agreement with both quantitative and qualitative evaluations.

4.3 Gradient Flow

We follow the gradient-flow setup of Feydy et al. (2019); Santambrogio (2015) and consider the evolution of an empirical measure $\mu(t)$ toward a fixed target measure ν by minimizing a discrepancy functional $\mathcal{D}(\mu(t), \nu)$. In our implementation, both measures are represented by point clouds in $\mathbb{R}^2$ with equal masses: $\nu = \frac{1}{n}\sum_{j=1}^{n} \delta_{Y_j}, \mu(t) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(t)}$, where $n = 1000$. Starting from $X(0)$, we integrate $\dot{X}(t) = -n\nabla_{X(t)}\big[\mathcal{D}\big(\frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(t)}, \nu\big)\big]$ with an explicit Euler scheme: $X^{k+1} = X^k - \eta\nabla_X \mathcal{D}(X^k, Y), \eta = 0.05$. We run the flow for 100 iterations. We evaluate both the Sinkhorn divergence and SROT divergence using Wasserstein distance. We report the result for $\varepsilon \in [0.01, 0.1]$ in Figure 5 and for $\varepsilon = 1$ in Figure 8 in Appendix B. We observe that SROT divergence provides weaker repulsion than Sinkhorn divergence especially for small ε e.g., 0.01. The reason is that SOT plan is not uniform. For $\varepsilon \in (0.1, 1)$, SROT divergence makes the flow converge faster in the sense of Wasserstein distance,

5 Conclusion

We propose sliced-regularized optimal transport (SROT), a framework that leverages a smoothened SOT plan as an informative prior to improve regularized OT. SROT retains the computational efficiency of EOT via a Sinkhorn-style algorithm while providing more accurate approximations to the true OT plan. We establish its theoretical properties and introduce the SROT divergence. Experiments show that SROT is consistently better than both EOT and SOT across tasks such as synthetic matching, color transfer, and gradient flows. A limitation of SROT is its reliance on the quality of the prior SOT plan, which is still an active area of research, especially in high-dimensional settings. Future work includes studying statistical properties of SROT and its divergence (e.g., sample complexity and central limit behavior), and handling high-dimensionality.

AI USE STATEMENT

In this work, we used generative AI tools for editing the paper and for coding. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the code reproduce the results in the paper.

REFERENCES

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pp. 1964–1974, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(2):235–269, 2019a.

Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 2019b.

Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78 (5):1103–1130, 2016.

Clément Bonet, Lucas Drumetz, and Nicolas Courty. Sliced-Wasserstein distances and flows on Cartan-Hadamard manifolds. *Journal of Machine Learning Research*, 26(32):1–76, 2025.

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

Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia Del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature methods*, 20(11):1759–1768, 2023.

Marta Catalano, Antonio Lijoi, and Igor Prünster. Measuring dependence in the Wasserstein distance for Bayesian nonparametric models. *The Annals of Statistics*, 49(5):2916–2947, 2021.

Marta Catalano, Hugo Lavenant, Antonio Lijoi, and Igor Prünster. A Wasserstein index of dependence for random measures. *Journal of the American Statistical Association*, 119(547):2396–2406, 2024.

Laetitia Chapel, Romain Tavenard, and Samuel Vaiter. Differentiable generalized sliced Wasserstein plans. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=hKKbtN4cp9.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 3730–3739, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

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)*, pp. 447–463, 2018.

Jean Feydy, Benjamin Charlier, François-Xavier Vialard, and Gabriel Peyré. Optimal transport for diffeomorphic registration. In *Medical Image Computing and Computer Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pp. 291–299. Springer, 2017.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2681–2690. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/feydy19a.html.

Paul Freulon, Nikitas Georgakis, and Victor Panaretos. Entropic optimal transport beyond product reference couplings: the Gaussian case on Euclidean space. *arXiv preprint arXiv:2507.01709*, 2025.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1574–1583. PMLR, 2019.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Ping He, Om Khangaonkar, Hamed Pirsiavash, Yikun Bai, and Soheil Kolouri. Sinkhorn-drifting generative models. *arXiv preprint arXiv:2603.12366*, 2026.

Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59, 2017.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. In *Advances in Neural Information Processing Systems*, pp. 261–272, 2019.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Tianlin Liu, Joan Puigcerver, and Mathieu Blondel. Sparsity-constrained optimal transport. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=yHY9NbQJ5BP.

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. URL https://openreview.net/forum?id=P7O1Vt1BdU.

Dirk A Lorenz, Paul Manns, and Christian Meyer. Quadratically regularized optimal transport. *Applied Mathematics & Optimization*, 83(3):1919–1949, 2021.

Guillaume Mahey, Laetitia Chapel, Gilles Gasso, Clément Bonet, and Nicolas Courty. Fast optimal transport through sliced Wasserstein generalized geodesics. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 35350–35385, 2023.

Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *The Annals of Statistics*, 52(3):966–998, 2024.

Boris Muzellec and Marco Cuturi. Subspace detours: Building transport plans that are optimal on subspace projections. In *Advances in Neural Information Processing Systems*, pp. 6917–6928, 2019.

Khai Nguyen. An introduction to sliced optimal transport: foundations, advances, extensions, and applications. *Foundations and Trends® in Computer Graphics and Vision*, 17(3-4):171–391, 2025.

Khai Nguyen and Peter Mueller. Summarizing nonparametric Bayesian mixture posteriors–sliced optimal transport metrics for Gaussian mixtures. *Journal of Computational and Graphical Statistics*, (just-accepted):1–22, 2026.

Khai Nguyen, Hai Nguyen, and Nhat Ho. Fast estimation of Wasserstein distances via regression on sliced Wasserstein distances. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=oa7L4vcJ77.

Khai Nguyen, Yang Ni, and Peter Mueller. Vertical consensus inference for high-dimensional random partition. *arXiv preprint arXiv:2603.27864*, 2026b.

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

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky TQ Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *International Conference on Machine Learning*, pp. 28100–28127. PMLR, 2023.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pp. 435–446. Springer, One New York Plaza, Suite 4600, New York, NY 10004-1562, 2012.

Philippe Rigollet and Austin J Stromme. On the sample complexity of entropic optimal transport. *The Annals of Statistics*, 53(1):61–90, 2025.

Herbert E Robbins. An empirical Bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and Basic Theory*, pp. 388–394. Springer, 1992.

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*, pp. 186–195. PMLR, 2019.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

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

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.

Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

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

Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (TOG)*, 35(4):72, 2016.

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

James Thornton and Marco Cuturi. Rethinking initialization of the Sinkhorn algorithm. In *International Conference on Artificial Intelligence and Statistics*, pp. 8682–8698. PMLR, 2023.

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. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW. Expert Certification.

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

Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, One New York Plaza, Suite 4600, New York, NY 10004-1562, 2009.

A Proofs

A.1 Proof of Theorem 1

We introduce Lagrange multipliers $f \in \mathcal{C}(\mathcal{X})$ and $g \in \mathcal{C}(\mathcal{Y})$ for the marginal constraints. The Lagrangian is

$$\mathcal{L}(\pi, f, g) = \int_{\mathcal{X}\times\mathcal{Y}} \big[c(x,y) - f(x) - g(y)\big] \mathrm{d}\pi(x,y) + \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\pi^{\mathrm{SOT}}}(x,y)\right) \mathrm{d}\pi(x,y) + \int_{\mathcal{X}} f(x)\mathrm{d}\mu(x) + \int_{\mathcal{Y}} g(y)\mathrm{d}\nu(y). \tag{18}$$

Let $\rho = \frac{\mathrm{d}\pi}{\mathrm{d}\pi^{\mathrm{SOT}}}$, we have

$$\mathcal{L}(\pi, f, g) = \int_{\mathcal{X}\times\mathcal{Y}} \Big[(c(x,y) - f(x) - g(y))\rho(x,y) + \varepsilon\rho(x,y)\log\rho(x,y)\Big] \mathrm{d}\pi^{\mathrm{SOT}}(x,y) + \int_{\mathcal{X}} f(x)\mathrm{d}\mu(x) + \int_{\mathcal{Y}} g(y)\mathrm{d}\nu(y). \tag{19}$$

Let $h(x,y) = c(x,y) - f(x) - g(y)$ and we minimize pointwise over ρ:

$$\inf_{\rho\geq 0} \left[h\rho + \varepsilon\rho\log\rho\right]. \tag{20}$$

The first-order optimality condition is

$$h + \varepsilon(\log\rho + 1) = 0, \tag{21}$$

which yields

$$\rho^*(x,y) = \exp\left(-\frac{h(x,y)}{\varepsilon} - 1\right). \tag{22}$$

Substituting ρ^* into the objective gives

$$h\rho^* + \varepsilon\rho^* \log \rho^* = -\varepsilon\rho^* = -\varepsilon \exp\left(-\frac{h(x,y)}{\varepsilon} - 1\right). \tag{23}$$

Absorbing the constant factor e^{-1} into the dual potentials yields

$$\inf_{\rho \geq 0} \left[h\rho + \varepsilon\rho \log \rho\right] = -\varepsilon \exp\left(-\frac{h(x,y)}{\varepsilon}\right). \tag{24}$$

Therefore,

$$\begin{aligned}\inf_{\pi \geq 0} \mathcal{L}(\pi, f, g) = &\int_{\mathcal{X}} f(x) \mathrm{d}\mu(x) + \int_{\mathcal{Y}} g(y) \mathrm{d}\nu(y) \\ &- \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{f(x) + g(y) - c(x,y)}{\varepsilon}\right) \mathrm{d}\pi^{\mathrm{SOT}}(x,y).\end{aligned} \tag{25}$$

Maximizing over (f, g) yields the dual problem. Strong duality follows from convexity of the primal objective and the existence of a feasible $\pi \ll \pi^{\mathrm{SOT}}$. The optimal coupling satisfies

$$\frac{\mathrm{d}\pi^\star}{\mathrm{d}\pi^{\mathrm{SOT}}}(x,y) = \exp\left(\frac{f^*(x) + g^*(y) - c(x,y)}{\varepsilon}\right), \tag{26}$$

which completes the proof.

A.2 Proof of Proposition 1

Consider the dual objective

$$\mathcal{L}(\mathbf{f}, \mathbf{g}) = \langle \mathbf{f}, \boldsymbol{\alpha} \rangle + \langle \mathbf{g}, \boldsymbol{\beta} \rangle - \varepsilon \sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij}^{\mathrm{SOT}} \exp\left(\frac{f_i + g_j - C_{ij}}{\varepsilon}\right). \tag{27}$$

Taking derivatives with respect to f_i and g_j and setting them to 0 yields

$$\frac{\partial \mathcal{L}}{\partial f_i} = \alpha_i - \sum_{j=1}^{m} P_{ij}^{\mathrm{SOT}} \exp\left(\frac{f_i + g_j - C_{ij}}{\varepsilon}\right) = 0, \tag{28}$$

$$\frac{\partial \mathcal{L}}{\partial g_j} = \beta_j - \sum_{i=1}^{n} P_{ij}^{\mathrm{SOT}} \exp\left(\frac{f_i + g_j - C_{ij}}{\varepsilon}\right) = 0. \tag{29}$$

Rearranging the first equation, we have

$$\alpha_i = \sum_{j=1}^{m} P_{ij}^{\mathrm{SOT}} \exp\left(\frac{f_i + g_j - C_{ij}}{\varepsilon}\right) = e^{f_i/\varepsilon} \sum_{j=1}^{m} P_{ij}^{\mathrm{SOT}} \exp\left(\frac{g_j - C_{ij}}{\varepsilon}\right), \tag{30}$$

which leads to

$$f_i = \varepsilon \left[\log \alpha_i - \log \sum_{j=1}^{m} P_{ij}^{\mathrm{SOT}} \exp\left(\frac{g_j - C_{ij}}{\varepsilon}\right)\right]. \tag{31}$$

Similarly, rearranging the first equation, we have

$$g_j = \varepsilon \left[\log \beta_j - \log \sum_{i=1}^{n} P_{ij}^{\mathrm{SOT}} \exp\left(\frac{f_i - C_{ij}}{\varepsilon}\right)\right]. \tag{32}$$

We conclude the proof.

A.3 Proof of Theorem 2

We recall that we assume $\mathcal{X}, \mathcal{Y}$ are compact metric spaces, $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ is a symmetric Lipschitz cost (i.e. $c(x, y) = c(y, x)$). We define the *Gibbs kernel*

$$k_\varepsilon(x, y) := \pi^{\mathrm{SOT}}(x, y) \exp\big(-c(x, y)/\varepsilon\big), \tag{33}$$

which is positive definite on $\mathcal{X} \times \mathcal{Y}$ by assumption. This is the direct analogue of the assumption in Feydy et al. (2019). Note that by Benamou et al. (2015), the SROT functional can be written (up to an additive constant) as a KL projection onto $\Pi(\mu, \nu)$:

$$\mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \nu) = \varepsilon \min_{\pi \in \Pi(\mu,\nu)} \mathrm{KL}(\pi \mid k_\varepsilon). \tag{34}$$

We define the *SROT negentropy* (cf. Feydy et al. (2019), Definition 1) by

$$F_\varepsilon(\mu) := -\tfrac{1}{2}\mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \mu), \tag{35}$$

so that $\mathcal{S}_{\varepsilon,\mathrm{SOT}}(\mu, \nu) = \mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \nu) + F_\varepsilon(\mu) + F_\varepsilon(\nu)$.

We first establish an alternative variational representation of F_ε, following Feydy et al. (2019), Proposition 3. By the symmetry of the dual problem (when $\mu = \nu$), the optimal potentials satisfy $f = g$ and the dual collapses to an optimization over a single function. Performing a change of variables $\rho = \exp(f/\varepsilon)\mu$ in the dual problem, one obtains

$$F_\varepsilon(\mu) = \varepsilon \min_{\rho} \Big[\big\langle \mu, \log \tfrac{\mathrm{d}\mu}{\mathrm{d}\rho} \big\rangle + \tfrac{1}{2}\|\rho\|^2_{k_\varepsilon} \Big] - \tfrac{\varepsilon}{2}, \tag{36}$$

where $\|\rho\|^2_{k_\varepsilon} := \iint k_\varepsilon(x, y)\mathrm{d}\rho(x)\mathrm{d}\rho(y)$. The functional $\rho \mapsto \|\rho\|^2_{k_\varepsilon}$ is *strictly* convex because k_ε is a positive universal kernel. Combined with the strict convexity of $\mathrm{KL}(\mu \mid \cdot)$, the integrand in equation 36 is strictly convex in (μ, ρ) jointly. By a standard interpolation argument (see Feydy et al. (2019), Proposition 4), this implies that F_ε is strictly convex on $\mathcal{P}(\mathcal{X})$.

The SROT functional is weak continuous and differentiable with respect to each marginal separately, as a consequence of the Lipschitz regularity and uniform convergence of optimal dual potentials under weak convergence of measures Feydy et al. (2019), Propositions 12–13. Its *partial* gradient with respect to the first marginal is

$$\nabla_1 \mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \nu) = f^{\mu,\nu}, \tag{37}$$

where $f^{\mu,\nu}$ is the optimal first dual potential for $\mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \nu)$, defined on all of $\mathcal{X}$ via the Sinkhorn mapping. By symmetry, $\nabla_2 \mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\nu, \nu) = g^{\nu,\nu} = f^{\nu,\nu}$. Hence the gradient of F_ε satisfies

$$\nabla F_\varepsilon(\mu) = -\tfrac{1}{2}\nabla_1 \mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \mu) = -f^\mu, \tag{38}$$

where $f^\mu = f^{\mu,\mu}$ is the (unique, symmetric) optimal potential.

Symmetry. Since c is symmetric and π^{SOT} is symmetric, transposing any $\pi \in \Pi(\mu, \nu)$ yields a bijection $\Pi(\mu, \nu) \to \Pi(\nu, \mu)$ under which both the transport cost and $\mathrm{KL}(\pi \mid k_\varepsilon)$ are preserved. Therefore $\mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \nu) = \mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\nu, \mu)$, and symmetry of $\mathcal{S}_{\varepsilon,\mathrm{SOT}}$ follows immediately from the definition.

Non-negativity. We define the *symmetric Bregman divergence* of F_ε (see Bregman (1967); Feydy et al. (2019)):

$$H_\varepsilon(\mu, \nu) := \tfrac{1}{2}\langle \mu - \nu, \nabla F_\varepsilon(\mu) - \nabla F_\varepsilon(\nu) \rangle. \tag{39}$$

Since F_ε is strictly convex, $-F_\varepsilon$ is strictly concave. Hence, $H_\varepsilon(\mu, \nu) \geq 0$, with equality if and only if $\mu = \nu$.

Since $\mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \cdot)$ is convex (as a supremum of linear functionals in ν from the dual representation), we have the subgradient inequalities for its *partial* gradient in the second argument:

$$\mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \nu) \geq \mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \mu) + \langle \nu - \mu, \nabla_2 \mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \mu) \rangle, \tag{40}$$

$$\mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \nu) \geq \mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\nu, \nu) + \langle \mu - \nu, \nabla_1 \mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\nu, \nu) \rangle. \tag{41}$$

Using the envelope theorem and the symmetry of $\mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \mu)$, one identifies $\nabla_2 \mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \mu) = f^\mu = -\nabla F_\varepsilon(\mu)$, and similarly for the second inequality. Summing and dividing by 2:

$$\mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \nu) \geq \tfrac{1}{2}\mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\mu, \mu) + \tfrac{1}{2}\mathrm{OT}_{\varepsilon,\mathrm{SOT}}(\nu, \nu) + H_\varepsilon(\mu, \nu), \tag{42}$$

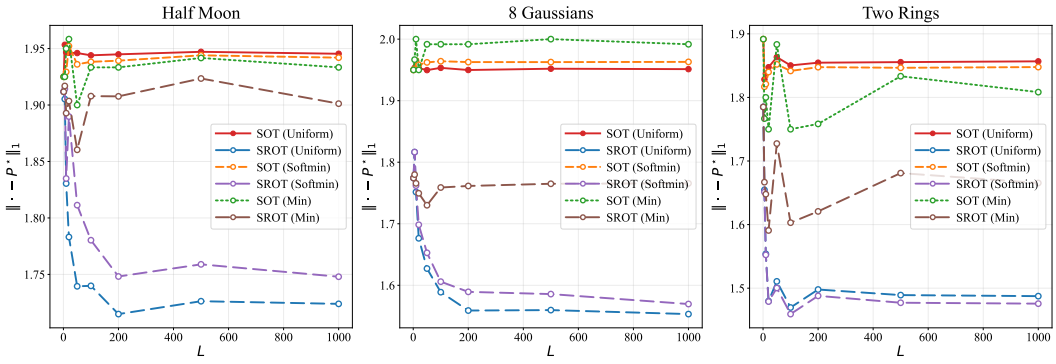


Figure 6: Ablation study of varying the number of projections L.

which gives

$$\mathcal{S}_{\varepsilon,\mathrm{SOT}}(\mu,\nu) \geq H_\varepsilon(\mu,\nu) \geq 0. \tag{43}$$

Identity of Indiscernibles. If $\mathcal{S}_{\varepsilon,\mathrm{SOT}}(\mu,\nu) = 0$, then $H_\varepsilon(\mu,\nu) = 0$, which implies $\mu = \nu$ since F_ε is strictly convex. The converse $\mu = \nu \Rightarrow \mathcal{S}_{\varepsilon,\mathrm{SOT}}(\mu,\nu) = 0$ is immediate.

Metrization of Weak Convergence. Let $\mu_n \rightharpoonup \mu$. By Proposition 13 of Feydy et al. (2019) applied to SROT (uniform convergence of optimal dual potentials under weak convergence of marginals), $\mathrm{OT}_{\varepsilon,\mathrm{SOT}}$ is weak* continuous and so is F_ε. Therefore each of the three terms in $\mathcal{S}_{\varepsilon,\mathrm{SOT}}(\mu_n,\mu)$ converges, giving $\mathcal{S}_{\varepsilon,\mathrm{SOT}}(\mu_n,\mu) \to 0$.

Suppose $\mathcal{S}_{\varepsilon,\mathrm{SOT}}(\mu_n,\mu) \to 0$. Since $\mathcal{X}$ is compact, $\mathcal{P}(\mathcal{X})$ is sequentially compact in the weak topology (Prokhorov's theorem). Every subsequence of (μ_n) therefore admits a further subsequence (μ_{n_k}) converging weakly to some μ_∞. By the forward direction,

$$\mathcal{S}_{\varepsilon,\mathrm{SOT}}(\mu_\infty,\mu) = \lim_{k\to\infty} \mathcal{S}_{\varepsilon,\mathrm{SOT}}(\mu_{n_k},\mu) = 0, \tag{44}$$

so $\mu_\infty = \mu$. Since every weakly convergent subsequence of (μ_n) must converge to the same limit μ, and $\mathcal{P}(\mathcal{X})$ is sequentially compact, the whole sequence satisfies $\mu_n \rightharpoonup \mu$.

B Additional Experiments

Number of projections (L) and SOT reference plan. We study how the number of projections L affects plan quality before and after Sinkhorn for SROT, with $(\varepsilon, T) = (0.001, 5000)$ fixed. For each L, a single random draw of L unit directions defines three reference couplings on the same slices: uniform averaging (SOT (Uniform)), a softmin variant (SOT (Softmin)), and the one-dimensional plan with minimum slice cost (SOT (Min)). For each reference, we report $\|\cdot - P^\star\|_1$ for both the initialization and the SROT solution. Results are shown in Figure 6. We observe that, at initialization, uniform averaging does not necessarily best approximate the true OT plan. Nevertheless, after Sinkhorn, SROT initialized with the uniform SOT is typically the most accurate, except in the two-rings case. This may be due to the greater smoothness of the corresponding SOT plan. For both uniform and softmin initializations, SROT improves as L increases, even when the initial plans do not. Overall, we recommend uniform SOT as a default choice, while noting that it may not be optimal and merits further exploration.

Computational speed. We report the wall-clock runtime of SROT in Figure 7 (Appendix B). The cost of computing the SOT reference plan is negligible (even with many projections) relative to the Sinkhorn iterations despite parallelization of SOT (since projections are independent). Consequently, the Sinkhorn procedure dominates the overall computational cost for both EOT and SROT, indicating that SROT is comparable to Sinkhorn in terms of efficiency. We also observe that early stopping is effective in avoiding redundant scaling once the current plan is sufficiently accurate.

Gradient flow. We report the result of the gradient flow from Sinkhorn divergence and SROT divergence with $\varepsilon = 1$ in Figure 8. Overall, we observe that SROT is slightly better than EOT.

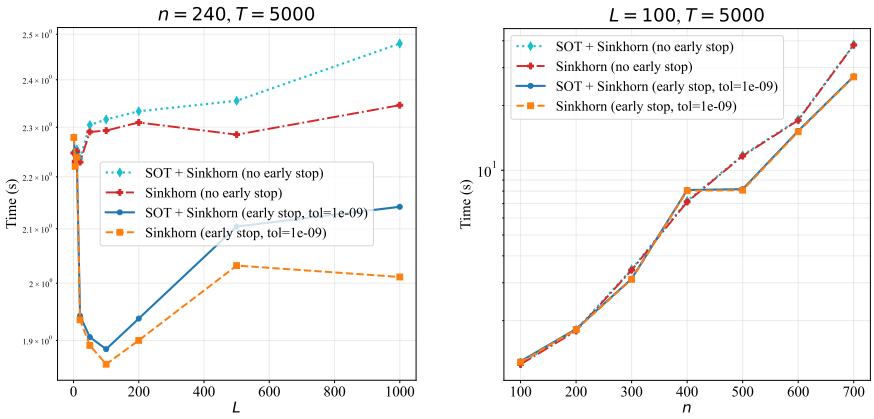


Figure 7: Computational speed measurement when varying the number of projections L and the number of atoms n.

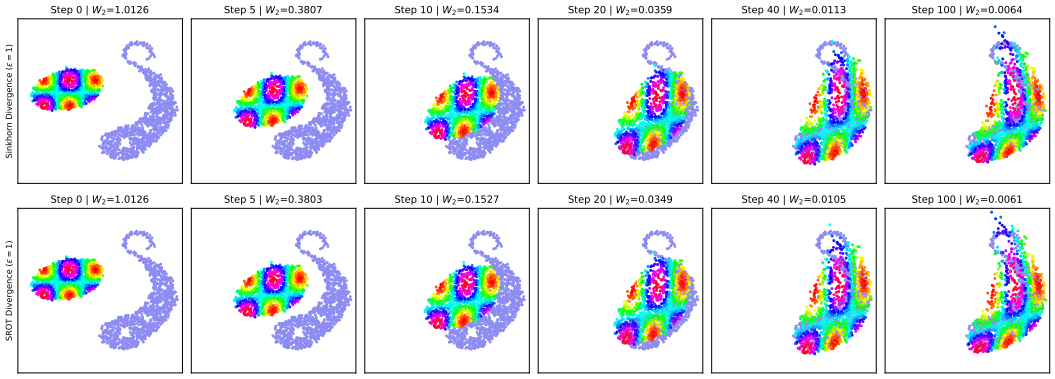


Figure 8: Gradient flows of Sinkhorn divergence and SR divergence with Wasserstein distance as netural evaluation metric.