Score-based Generative Neural Networks for Large-Scale Optimal Transport: Further Experiments

Killian Steunou
Master 2 Mathématiques, Vision, Apprentissage : Computational Optimal Transport Course
École Normale Supérieure Paris-Saclay
killian.steunou@ens-paris-saclay.fr
January 15, 2025

Abstract

This paper investigates the integration of score-based generative models into regularized optimal transport for addressing the computational challenges of large-scale OT problems. Optimal transport often becomes computationally intractable for large datasets. Regularized OT, using methods like the Sinkhorn algorithm, introduces entropy-based regularization to enhance efficiency. However, these methods suffer from limitations, including averaging artifacts when deriving transport maps. To address this, [1] propose a hybrid approach combining score-based generative models with regularized OT. Their method follows the work of [2], using neural networks to approximate optimal dual variables and employs Langevin dynamics [3] for conditional sampling, enabling direct sampling from the Sinkhorn coupling without averaging artifacts. The contributions of this work are both theoretical and practical: they build on existing formulations of f-divergence-regularized OT and introduce a novel numerical framework, SCONES (Sinkorn Conditional Neural Sampling), for efficiently approximating and sampling OT plans. We validate their approach through experiments on toy distributions, comparing its performance against barycentric projection method from [2] in terms of accuracy and computational efficiency. Additionally, we analyze the effects of relevant hyperparameters parameters to understand their influence on the method's performance. This work extends the applicability of OT methods to large-scale, high-dimensional problems in machine learning.

1 Introduction

Optimal transport provides a powerful mathematical framework for comparing probability distributions by determining the most cost-efficient way to transform one distribution into another. Despite its theoretical appeal and versatility, OT computations often become intractable for large-scale, high-dimensional problems such as images due to their significant computational cost. These challenges arise from the intrinsic curse of dimensionality and the high complexity of solving the corresponding optimization problems.

To address these issues, regularized OT has emerged as a practical alternative, introducing entropy-based penalties to make computations more efficient [4]. Sinkhorn's algorithm, a widely used approach for entropy-regularized OT, replaces the original linear programming problem with a regularized problem that can be solved iteratively via matrix scaling. This modification allows the computational burden to scale less with data size. However, despite its success, the regularized framework faces limitations when applied to large-scale problems. Specifically, methods relying on barycentric projection to derive transport maps often produce artifacts due to averaging, resulting in suboptimal or blurred solutions.

This report focuses on a recent method that integrates *score-based generative models* and *Langevin dynamics* into regularized OT to address these challenges. Score-based generative models, which rely on estimating the score (*i.e.* the gradient of the log-density) of target distributions, have proven effective in generating high-quality samples in applications such as image generation. By combining these models with the Sinkhorn coupling framework, the method bypasses the need for barycentric projection and instead generates samples directly from the conditional transport plan using Langevin dynamics. This approach eliminates averaging artifacts. The framework, introduced in [1], represents a novel contribution to the intersection of generative modeling and optimal transport.

1.1 Related Work

The proposed method builds mainly on the work of Seguy et al. [2], that introduced a large-scale stochastic dual approach for estimating Sinkhorn couplings, using neural networks to parametrize dual variables and efficiently optimize the OT problem. Their method also employed barycentric projection to derive transport maps, which has since been observed to induce undesirable artifacts. Meanwhile, Song and Ermon [3] demonstrated the effectiveness of Langevin dynamics for generative modeling by iteratively sampling from the score of a distribution. This approach, further enhanced by annealed Langevin sampling, has shown remarkable success in generating high-quality samples.

More recently, Daniels et al. [1] extended these ideas by integrating score-based generative modeling into the Sinkhorn framework, enabling direct sampling from the conditional coupling without requiring projection-based post-processing.

1.2 Contributions and Context

This report explores and evaluates the method proposed by Daniels et al. [1] in the context of toy 1D and 2D distributions. By approximating scores of the target distributions and combining them with Sinkhorn regularization, the method addresses the limitations of barycentric projection and improves the quality of generated samples. The contributions of this work are both theoretical and numerical:

- Theoretical results establish strong convexity and duality properties of the regularized OT problem, ensuring convergence of gradient-based optimization methods.
- Numerically, the integration of score-based generative models allows for accurate sampling of conditional distributions, bypassing the artifacts inherent to projection methods.

These contributions are analyzed within the broader framework of optimal transport, a topic studied in the Computational Optimal Transport course of the MVA program. The duality and regularization properties studied in this report draw heavily on concepts from the Convex Optimization course, where topics such as strong convexity/duality and smoothness were rigorously examined. The report bridges these theoretical insights with practical experimentation, validating the proposed method on simple distributions and investigating the impact of hyperparameters, including neural network hidden layers' width h, regularization strength λ, Langevin dynamics step size ϵ and number of sampling steps T, on quantitative and qualitative performance.

2 Sinkorn Conditional Neural Sampling

We will now explain the method, starting by introducing useful definitions and propositions from [1] that are necessary to understand the the experiments conducted. All sentences and formulas in Sub-sections 2.1 and 2.2, unless explicitly specified otherwise, come from their article.

2.1 Regularized Optimal Transport

Definition 2.1 (Regularized OT) Let $\sigma \in \mathcal{M}_+(\mathcal{X})$ and $\tau \in \mathcal{M}_+(\mathcal{Y})$ be probability measures supported on compact sets $\mathcal{X}, \mathcal{Y}$. Let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a convex, lower semi-continuous function representing cost of transporting a point $x \in \mathcal{X}$ to $y \in \mathcal{Y}$. The regularized optimal transport distance $\mathrm{OT}_\lambda(\sigma, \tau)$ is given by

$$\text{OT}_\lambda(\sigma,\tau) = \min_\pi \mathbb{E}_\pi[c(x,y)] + \lambda H(\pi)$$
$$\text{subject to } \pi_X = \sigma, \quad \pi_Y = \tau$$
$$\pi(x,y) \geq 0$$

where $H : \mathcal{M}_+(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ is a convex regularizer and $\lambda \geq 0$ is a regularization parameter.

We are mainly concerned with optimal transport of empirical distributions, where $\mathcal{X}$ and $\mathcal{Y}$ are finite and σ, τ are empirical probability vectors.

We refer to the objective $K_\lambda(\pi) = \mathbb{E}_\pi[c(x,y)] + \lambda H(\pi)$ as the *primal objective*, and we will use $J_\lambda(\varphi, \psi)$ to refer to the associated *dual objective*, with dual variables φ, ψ.

Two common regularizers are $H(\pi) = KL(\pi||\sigma \times \tau)$ and $H(\pi) = \chi^2(\pi||\sigma \times \tau)$, called entropy and l2 regularization respectively:

$$KL(\pi||\sigma \times \tau) = \mathbb{E}_\pi \left[log \left(\frac{d\pi(x,y)}{d\sigma(x)d\tau(y)} \right) \right], \quad \chi^2(\pi||\sigma \times \tau) = \mathbb{E}_{\sigma\times\tau} \left[\left(\frac{d\pi(x,y)}{d\sigma(x)d\tau(y)} \right)^2 \right]$$

where $\frac{d\pi(x,y)}{d\sigma(x)d\tau(y)}$ is the Radon-Nikodym derivative of π with respect to the product measure $\sigma \times \tau$.

Proposition 2.2 *In the empirical setting of Definition 2.1, the entropy regularized primal problem $K_\lambda(\pi)$ is λ-strongly convex in l_1 norm. The dual problem $J_\lambda(\varphi, \psi)$ is concave, unconstrained, and $\frac{1}{\lambda}$-strongly smooth in l_∞ norm. Additionally, these objectives witness strong duality: $\inf_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} K_\lambda(\pi) = \sup_{\varphi,\psi\in\mathbb{R}^{2d}} J_\lambda(\varphi,\psi)$ and the extrema of each objective are attained over their respective domains.*

Proposition 2.3. *In the setting of Proposition 2.2, the KL-regularized dual objective takes the form*

$$J_\lambda(\varphi,\psi) := \mathbb{E}_\sigma[\varphi(x)] + \mathbb{E}_\tau[\psi(y)] - \lambda \mathbb{E}_{\sigma\times\tau} \left[\frac{1}{e} \exp\left(\frac{1}{\lambda}(\varphi(x) + \psi(y) - c(x,y)) \right) \right].$$

The optimal solutions $\varphi^*, \psi^* = \arg\max_{\varphi,\psi\in\mathbb{R}^{2d}} J_\lambda(\varphi,\psi)$ and $\pi^* = \arg\min_{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} K_\lambda(\pi)$ satisfy

$$\pi^*(x,y) = \frac{1}{e} \exp\left(\frac{1}{\lambda}(\varphi^*(x) + \psi^*(y) - c(x,y)) \right) \sigma(x)\tau(y).$$

The solution $\pi^*(x,y)$ of the entropy regularized problem is called the Sinkhorn coupling between σ and τ in reference to Sinkhorn's Algorithm [5]. For arbitrary choices of regularization, we call $\pi^*(x,y)$ a Sinkhorn coupling.

From Proposition 2.2 and 2.3, the major requirements for the regularizer are: $H(\pi)$ (and hence $K_\lambda(\pi)$) must be strongly convex in l_1 norm and $H(\pi)$ must induce a nice analytic form of π^* in terms of φ^* and ψ^*.

We denote by D_f regularized problem, the Regularized OT problem with a regularizer H that belongs to the class of f-Divergences, which are statistical divergences of the form $D_f(p||q) = \mathbb{E}_q \left[f\left(\frac{p(x)}{q(x)} \right) \right]$ where $f : \mathbb{R} \to \mathbb{R}$ is convex, $f(1) = 0$, p and q are probability measures, and p is absolutely continuous with respect to q.

While Propositions 2.2 and 2.3 are already well-known to the literature (shown in [6] and [4]), the following proposition (2.4) is proven in [1].

Proposition 2.4 *Let $f(v) : \mathbb{R} \to \mathbb{R}$ be a differentiable α-strongly convex function with convex conjugate $f^*(v)$. Set $f^{*\prime}(v) = \partial_v f^*(v)$. Define the violation function $V(x,y;\varphi,\psi) = \varphi(x) + \psi(y) - c(x,y)$. Then,*

1. *The D_f regularized primal problem $K_\lambda(\pi)$ is $\lambda\alpha$-strongly convex in l_1 norm. With respect to dual variables $\varphi \in \mathbb{R}^{|\mathcal{X}|}$ and $\psi \in \mathbb{R}^{|\mathcal{Y}|}$, the dual problem $J_\lambda(\varphi, \psi)$ is concave, unconstrained, and $\frac{1}{\lambda\alpha}$-strongly smooth in l_∞ norm. Strong duality holds: $\forall \pi, \varphi, \psi, K_\lambda(\pi) \geq J_\lambda(\varphi, \psi)$, with equality for some triplet π^*, φ^*, ψ^*.*

2. *$J_\lambda(\varphi, \psi)$ takes the form $J_\lambda(\varphi, \psi) = \mathbb{E}_\sigma[\varphi(x)] + \mathbb{E}_\tau[\psi(y)] - \mathbb{E}_{\sigma\times\tau}[H^*(V(x, y; \varphi, \psi))]$, where $H_f^*(v) = \lambda f^*(\lambda^{-1}v)$.*

3. *The optimal solutions $(\pi^*, \varphi^*, \psi^*)$ satisfy*

$$\pi^*(x, y) = M_f(V(x, y; \varphi, \psi))\sigma(x)\tau(y)$$

where $M_f(x, y) = f^{*\prime}(\lambda^{-1}v)$

A similar method was proposed in [2], who then, with access $\pi(x, y)$, approximate an optimal transport map using a barycentric projection. This method is observed to induce averaging artifacts, which is why the authors of [1] propose a direct sampling strategy with a score-based generative model, using Langevin dynamics [3]. Their method eliminates averaging artifacts incurred by such barycentric projection.

2.2 Langevin Sampling and Score Based Generative Modeling

Thanks to Proposition 2.4, given optimal dual variables $\varphi^*(x), \psi^*(y)$, it is easy to evaluate the density of the corresponding optimal coupling. To sample from this coupling, we use Langevin Sampling [3]. The key quantity used in Langevin sampling of a generic probability measure $p(x)$ is its *score function*, given by $\nabla_x \log p(x)$ for $x \in \mathcal{X}$. The algorithm is an iterative Monte Carlo method which generates approximate samples $\tilde{x}_t$ by iterating the map

$$\tilde{x}_t = \tilde{x}_{t-1} + \epsilon\nabla_x \log p(\tilde{x}_{t-1}) + \sqrt{2\epsilon}z_t$$

where $\epsilon > 0$ is a **step size parameter** and where $z_t \sim \mathcal{N}(0, I)$ independently at each time step $t \geq 0$. When $\epsilon \to 0$ and $T \to \infty$, the samples $\tilde{x}_T$ converge weakly in distribution to $p(x)$. Using a method introduced by [3], the score can be estimated using a neural network $s_\vartheta(x)$ trained on samples of $p(x)$ that approximate $s_\vartheta(x) \approx \nabla_x \log p(x)$ for $x \in \mathcal{X}$. To generate samples, we iterate Langevin dynamics with this estimate instead of the true score.

To scale this method to high dimensional image datasets, [3] propose an iterative denoising method with an annealing scheme, which samples noised versions of $p(x)$ as the noise is gradually reduced. One first samples a noised distribution $p(x) * \mathcal{N}(0, \tau_1)$, at noise level τ_1. The noisy samples, which are presumed to lie near high density regions of $p(x)$, are used to initialize additional rounds of Langevin dynamics at diminishing noise levels $\tau_2 > ... > \tau_N > 0$. At the final round, Annealed Langevin Sampling outputs approximate samples according to the noiseless distribution.

2.3 Conditional Sampling of Regularized Optimal Transport Plans

[1]'s approach, named SCONES, is split into two parts: approximate the density of the optimal Sinkhorn coupling, and then sample optimal couplings using Langevin dynamics.

2.3.1 Sinkhorn Coupling Approximation

We will use the same notations as in Sub-section 2.1.
The goal is to approximate the optimal Sinkhorn coupling $\pi^*(x, y)$ that minimizes $K_\lambda(\pi)$ by applying the large-scale stochastic dual approach described in Sub-section 2.1 from [2], training the neural networks $\varphi_{\theta_1}, \psi_{\theta_2}$ that are the dual variables, to maximize $J_\lambda(\varphi_{\theta_1}, \psi_{\theta_2})$ with respect to $\theta = (\theta_1, \theta_2)$ using gradient descent. Once the optimal θ^* is obtained, we have the associated optimal transport plan given by Proposition 2.4: $\hat{\pi}(x, y) = M(V(x, y; \varphi_{\theta_1^*}, \psi_{\theta_2^*}))\sigma(x)\tau(y)$. This method is descibed in Algorithm 1.

Algorithm 1 Density Estimation.

Input: Step size γ, batch size m.
Input: Nets $\varphi_{\theta_1}, \psi_{\theta_2}$.
Input: Datasets σ, τ. Time steps $T > 0$.
Output: Trained $\varphi_{\theta_1^*}, \psi_{\theta_2^*}$.
  for $t = 1 \dots T$ **do**
    Sample $X_1, \dots, X_m \sim \sigma$, and $Y_1, \dots, Y_m \sim \tau$.
    Stochastic gradient update $\varphi_{\theta_1}, \psi_{\theta_2}$:
    $\Delta_1 \leftarrow \sum_{i,j=1}^{m} \nabla_{\theta_1}[\varphi_{\theta_1}(X_i) - H^*(V(X_i, Y_j))]$.
    $\Delta_2 \leftarrow \sum_{i,j=1}^{m} \nabla_{\theta_2}[\psi_{\theta_2}(X_i) - H^*(V(X_i, Y_j))]$.
    $\theta_1 \leftarrow \theta_1 + \gamma\Delta_1$.
    $\theta_2 \leftarrow \theta_1 + \gamma\Delta_2$.
  end for
  Output parameters $\{\theta_1, \theta_2\}$

After obtaining $\hat{\pi}$, we sample $\hat{\pi}_{Y|X=x}(y)$ using Langevin dynamics, using the score estimator for the conditional distribution:

$$\begin{aligned}\nabla_y \log \hat{\pi}_{Y|X=x}(y) &= \nabla_y[\log(M(V(x, y; \phi_{\theta^*}, \psi_{\theta^*}))\sigma(x)\tau(y)) - \log(\sigma(x))] \\ &\approx \nabla_y \log(M(V(x, y; \phi_{\theta^*}, \psi_{\theta^*}))) + s_\vartheta(y).\end{aligned}$$

which is approximated by directly differentiating $\log M(V(x, y))$ using standard automatic differentiation tools and adding the result to an unconditional score estimate $s_\vartheta(y)$. This method is described in Algorithm 2.

Algorithm 2 SCONES Sampling Procedure.

Input: Noise levels $\tau_1 > \ldots > \tau_N$.
Input: Dual vars. $\tilde{\phi}(x)$, $\tilde{\psi}(y)$. Source $x \in \mathcal{X}$.
Input: Time steps $T > 0$. Step size $\epsilon > 0$.
Output: Data sample $\tilde{y} \sim \pi_{Y|X=x}(y)$.
  Initialize $\tilde{y}_{1,0} \sim \mathcal{N}(0, I)$.
  for τ_i, $i = 1 \dots N$ **do**
    for $t = 1 \dots T$. **do**
      Sample $z \sim \mathcal{N}(0, \tau_i)$.
      Compute score update:
      $\Delta_s \leftarrow s_\vartheta(\tilde{y}_{i,t-1})$.
      $\Delta_\pi \leftarrow \nabla_y \log M(V(x, \tilde{y}_{i,t-1}; \tilde{\phi}, \tilde{\psi}))$.
      $\tilde{x}_{i,t} = \tilde{x}_{i,t-1} + (\epsilon/2)(\Delta_s + \Delta_\pi) + \sqrt{\epsilon}z$.
    end for
    Initialize $\tilde{x}_{i+1,0} = \tilde{x}_{i,T}$.
  end for
  Output sample $\tilde{x}_{N,T}$.

2.4 Theoretical Guarantees

2.4.1 Maximization of the Dual Objective

The goal of training φ_θ and ψ_θ is to approximate the continuous data distribution between the empirical coupling and the population coupling, up to optimization error and statistical estimation error. Under Assumption 4.1, one can prove the convergence of Algorithm 1 to the global maximizer of $J_\lambda(\varphi, \psi)$, and find a quantitative bound on the optimization error.

Assumption 4.1. [1] Let $f_\theta(x)$ be a neural network with parameters $\theta \in \Theta$, where Θ is a set of feasible weights, reachable by gradient descent. Fix a dataset $\{X_i\}_{i=1}^N$ and let $\mathcal{K}_\theta \in \mathbb{R}^{N \times N}$ be the Gram matrix of coordinates $[K_\theta]_{ij} = \langle \nabla_\theta f_\theta(X_i), \nabla_\theta f_\theta(X_j) \rangle$. Then $f_\theta(x)$ must satisfy:

1. There exists $R \gg 0$ so that $\Theta \subseteq B(0, R)$, where $B(0, R)$ is the Euclidean ball of radius R.

2. There exist $\rho_M > \rho_m > 0$ such that for $\theta \in \Theta$,

$$\rho_M \geq \lambda_{\max}(\mathcal{K}_\theta) \geq \lambda_{\min}(\mathcal{K}_\theta) \geq \rho_m > 0.$$

3. For $\theta \in \Theta$ and for all data points $\{X_i\}_{i=1}^N$, the Hessian matrix $D_\theta^2 f_\theta(X_i)$ is bounded in spectral norm: $||D_\theta^2 f_\theta(X_i)|| \leq \frac{\rho_M}{C_h}$, where $C_h \gg 0$ depends only on R, N and the regularization λ.

Fully-connected networks with smooth and Lipschitz-continuous activations satisfy Assumption 4.1 when the width of all layers is sufficiently large [7].

Theorem 4.2 shows that when φ, ψ are parametrized by neural networks satisfying assumption 4.1, gradient descent converges to a global maximizer of $J_\lambda(\varphi, \psi)$, with a bound on the number of iterations.

Theorem 4.2. Suppose $J_\lambda(\varphi, \psi)$ is $\frac{1}{s}$-strongly smooth in l_∞ norm. Let φ_θ, ψ_θ be neural networks satisfying Assumption 4.1 for the dataset $\{(x_i, y_i)\}_{i=1}^N$, $N = |\mathcal{X}| \cdot |\mathcal{Y}|$.
Then gradient descent of $J_\lambda(\varphi_\theta, \psi_\theta)$ with respect to θ at learning rate $\eta = \frac{\lambda}{2\rho_M}$ converges to an ϵ-approximate global maximizer of J_λ in at most $\left(\frac{2\kappa R^2}{s}\right)\epsilon^{-1}$ iterations, where $\kappa = \frac{\rho_M}{\rho_m}$.

Finally, Theorem 4.3 guarantees that approximately maximizing $J_\lambda(\varphi, \psi)$ is sufficient to produce a close approximation of the true empirical Sinkhorn coupling.

Theorem 4.3. Suppose $K_\lambda(\pi)$ is s-strongly convex in l_1 norm and let $\mathcal{L}(\varphi, \psi, \pi)$ be the Lagrangian of the regularized optimal transport problem. For $\hat{\varphi}, \hat{\psi}$ which are ϵ-approximate maximizers of $J_\lambda(\varphi, \psi)$, the pseudo-plan $\hat{\pi} = M_f(V(x, y; \hat{\varphi}, \hat{\psi}))\sigma(x)\tau(y)$ satisfies

$$|\hat{\pi} - \pi^\star|_1 \leq \sqrt{\frac{2\epsilon}{s}} \leq \frac{1}{s}\left|\nabla_{\hat{\pi}}\mathcal{L}(\hat{\varphi}, \hat{\psi}, \hat{\pi})\right|_1.$$

3 Experiments

In this section, we report various experiments' results, comparing the barycentric projection method (BP) [2] with SCONES [1]. We experiment on simple distributions (two dimensions). We use $H(\pi) = KL(\pi||\sigma \times \tau)$ as regularizer function. To compare the performance, we use

- Bures-Wasserstein Unexplained Variance Percentage [8] BW-UV $(\hat{\pi}, \pi^\lambda)$, where π^λ is the closed form solution given by [9] and where $\hat{\pi}$ is the joint empirical covariance of $k = 10000$ samples $(x, y) \sim \pi$ generated using either SCONES or Barycentric Projection.
- 2-Sliced Wasserstein Distance $\mathcal{SWD}_2$ from [10], from the POT library [11], defined as

$$\mathcal{SWD}_2(\mu, \nu) = \mathbb{E}_{\theta \sim \mathcal{U}(\mathbb{S}^{d-1})}[\mathcal{W}_2^2(\theta_\# \mu, \theta_\# \nu)]^{\frac{1}{2}}$$

  where $\theta_\# \mu$ stands for the pushforwards of the projection $X \in \mathbb{R}^d \mapsto \langle \theta, X \rangle$.

We quantify the influence of several parameters: width of the network h, sampling step size ϵ, regularization strength λ, and number of sampling steps T.

3.1 Gaussian to Gaussian

In this first experiment, we implement the method to learn the coupling between two Gaussian distributions, in 2D, for varying sampling step-sizes $\epsilon \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$, and varying hidden dimensions for the three layers of the fully connected networks that parametrize the dual variables: $h \in \{8, 16, 32, 64, 128, 256\}$.
We train SCONES according to Algorithm 1 and generate samples according to Algorithm 2. In place of a score estimate, we use the ground truth target score $\nabla_y \log \tau(y) = \Sigma_2^{-1}(y - \mu_2)$ and omit annealing.
We run each method three times for given ϵ, h and average the different BW-UVP obtained, that are shown in Figure 1. We fixed $\lambda = 4$.
For all ϵ, h the SCONES BW-UVP is much lower than the one of the BP. The lowest value for BP is

achieved with neural networks of 64-dimensional hidden layers, at 0.128 (same for all ϵ), whereas the lowest BW-UVP for SCONES is achieved with neural networks of 16-dimensional hidden layers, for $\epsilon = 0.05$, at 0.0176. The smaller networks ($h \in \{8, 16\}$) are performing worse than the bigger ones for BP, while SCONES seems to benefit from smaller widths, as shown in Table 1. It seems the greater the ϵ the lower the BW-UVP for SCONES, though the decrease is negligeable for $\epsilon > 0.001$.

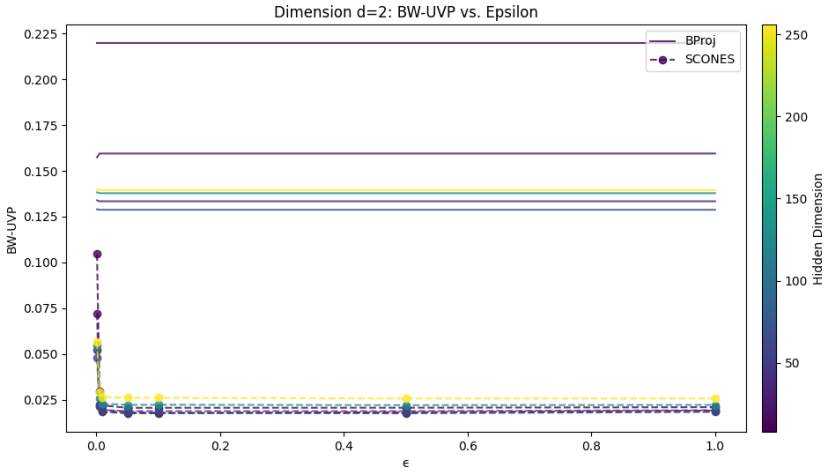

Figure 1: Evolution of the BW-UVP with different sampling step-sizes ϵ with SCONES, for different neural networks width, compared with barycentric projection.

($\epsilon = 0.05$)	$h = 16$	$h = 32$	$h = 64$	$h = 8$	$h = 128$	$h = 256$
BP	0.159489	0.133408	**0.128768**	0.219851	0.137805	0.139487
SCONES	**0.017613**	0.018263	0.018800	0.020580	0.022184	0.026093

Table 1: Comparison of mean BW-UVP values for $\epsilon = 0.05$ and different neural networks widths, ranked by SCONES mean from lowest to highest.

Fixing $\epsilon = 0.001$, we now experiment various $\lambda \in \{0.1, 0.5, 1, 2, 4, 8, 9, 10, 15, 20\}$. Results are shown in Figure 2. Both methods show good performance for $\lambda = 2$, and the BW-UVP starts increasing for $\lambda \geq 4$. Interestingly, the wider neural networks for BP ($h \in \{64, 128, 256\}$) show similar or better performance for low λ values (< 1) than all neural networks sizes for SCONES. Then, for $\lambda \geq 1$, every version of SCONES has a lower BW-UVP than BP, showing little difference between different hidden dimension values.

We also experiment various sampling steps T (see Algorithm 2), ranging from 1 to 2000, and track the time taken (Figure 3, right), the BW-UVP (Figure 3, left), and some samples drawn from the optimal coupling (Figure 10). We fix $\epsilon = 0.001$, $\lambda = 20$ and $h = 64$. Increasing the number of Langevin sampling steps T improves the BW-UVP metric, which is decreasing exponentially, but linearly increases the compute time, whereas BP is almost instantaneous (0.001 second on CPU). Figures 4 and 11 show results for $\lambda = 4$, that are quantitatively similar, though the BW-UVP for SCONES begins to stabilize around iteration 1500, while for $\lambda = 20$ it kept improving. While SCONES is much better quantitatively and qualitatively than BP for a high λ, the improvement is less evident for a smaller λ, as BP has a low BW-UVP and the samples look more similar to the target distribution.

While SCONES demonstrates strong performance overall, particularly at higher regularization strengths, its advantages are not absolute. For low regularization values, the barycentric projection method achieves comparable or even superior BW-UVP scores, especially when wider neural networks are used. This suggests that SCONES may not always justify its higher computational cost in settings where BP already performs well. The results also highlight that SCONES benefits significantly from careful tuning of sampling steps, which may limit its practicality in time-sensitive

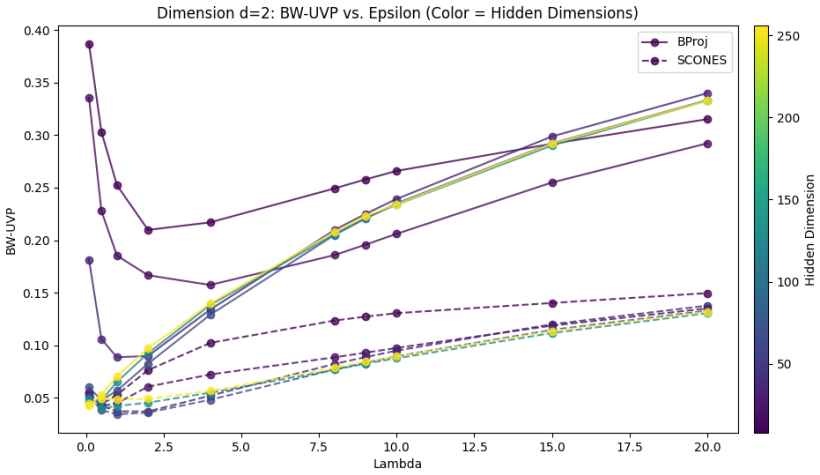

Figure 2: Evolution of the BW-UVP with different regularization strength λ with SCONES, for different neural networks width, compared with barycentric projection.

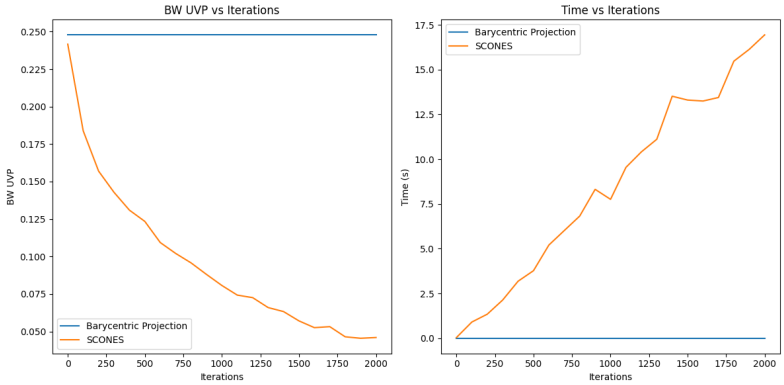

Figure 3: BW-UVP (left) and Time vs Langevin sampling iteration T for $\lambda = 20, \epsilon = 0.001, h = 64$, for $T \in \{1, 100, 1000, 2000\}$.

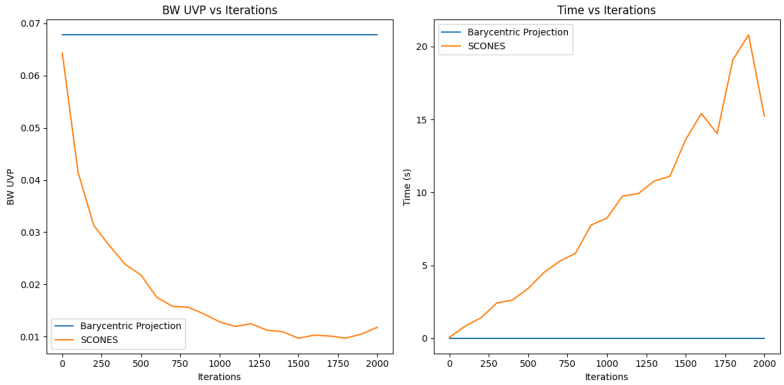

Figure 4: BW-UVP (left) and Time vs Langevin sampling iteration T for $\lambda = 4, \epsilon = 0.001, h = 64$, for $T \in \{1, 100, 1000, 2000\}$.

applications.

The main value of SCONES lies in its ability to excel at high regularization, where BP struggles, and in its flexibility to achieve higher precision in the BW-UVP metric as sampling steps increase. However, the diminishing returns in performance improvement with more sampling steps raise questions about its scalability for larger or more complex datasets. These findings suggest that while SCONES is a powerful tool, it may be most suitable for scenarios where precision in the transport metric is crucial and computational cost is less of a concern. BP, on the other hand, remains a strong and efficient alternative for simpler or less regularized problems.

3.2 Gaussian to Swiss Roll

For this experiment, we will test the method on a Gaussian 2D source distribution, with a 2D Swiss Roll target distribution. We will compare SCONES et BP using the 2-Sliced Wasserstein Distance $\mathcal{SWD}_2$ [10]. We will denote by $\mathcal{SWD}_2$ the distance from the source Gaussian distribution to the target Swiss-Roll distribution produced by SCONES or BP, instead of $\mathcal{SWD}_2(\mu, \hat{\nu}_{\text{SCONES}})$ and $\mathcal{SWD}_2(\mu, \hat{\nu}_{\text{BP}})$, when it is clear which method the distance refers to.

We experiment with lower sampling steps than for Gaussian to Gaussian, ranging from 1×10^{-6} to 1×10^{-5}, as $\epsilon > 1 \times 10^{-5}$ makes the output data distribution scale too much and often results in $\mathcal{SWD}_2 = \infty$, as shown in figure 6. We first experiment with neural networks with hidden dimension $h \in \{8, 16, 32, 64, 128, 256, 512\}$. We set $\epsilon \in \{1 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-5}, 5 \times 10^{-5}, 8 \times 10^{-5}\}$, and $\lambda = 2$. The SCONES $\mathcal{SWD}_2$ is equal to infinity for $\epsilon > 1 \times 10^{-5}$, and those values are hence not shown on the plot in Figure 5.

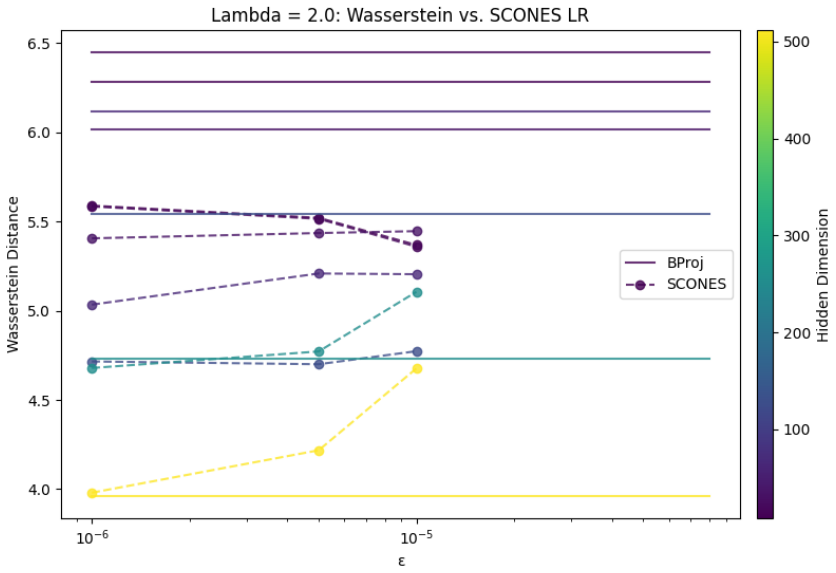

Figure 5: Evolution of the 2-Sliced Wasserstein Distance with different sampling step-sizes ϵ with SCONES, for different neural networks width, compared with barycentric projection. X-axis in log scale.

Clearly, the results depend on the neural networks' width, with $h = 512$-variant achieving a much lower distance. For that parametrization, SCONES achieves a comparable distance for $\epsilon = 1 \times 10^{-6}$, at 3.98 but not better that BP, achieving 3.96. These distances are quite high, indicating the methods do not perform really well, which was expected as the target distribution is more complex than the Gaussian one, and the neural networks are not large enough, as stated in Theorem 4.2, to verify Assumption 4.1, and the dual objective is not maximized. On the other hand, SCONES achieve

a lower $\mathcal{SWD}_2$ than BP with $h < 512$-neural networks variants, for low ϵ, the difference being significant for smaller models.

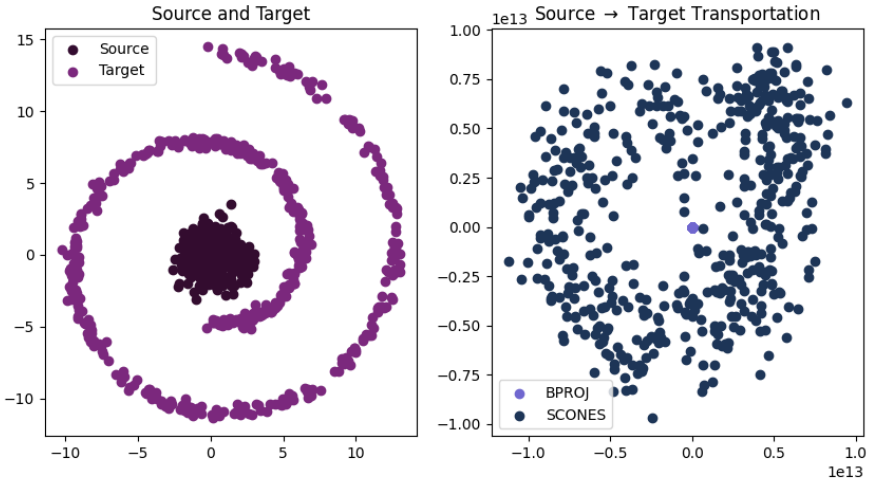

Figure 6: Transport failure for SCONES (right), with $\epsilon = 0.0001$. Axis are $\times 1 \cdot 10^{13}$ on the right hand side figure.

We will now experiment with greater neural networks' widths, $h \in \{512, 1024, 2048, 4096\}$, $\epsilon \in \{1 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-5}\}$. We fix $\lambda = 2$ and $T = 300$.

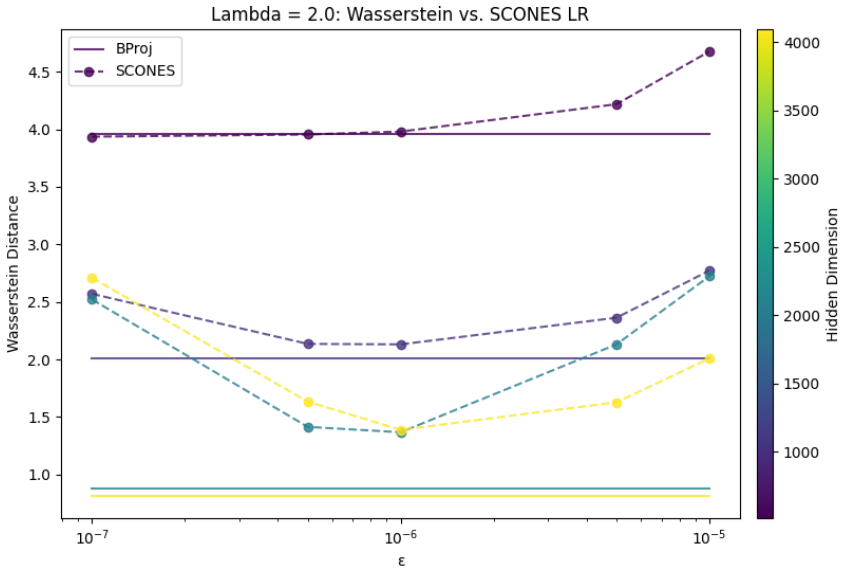

Figure 7: Evolution of the 2-Sliced Wasserstein Distance with different sampling step-sizes ϵ with SCONES, for different neural networks width, compared with barycentric projection. X-axis in log scale.

For this setting, BP's $\mathcal{SWD}_2$ is clearly much better than SCONES', while being much faster, no matter the step-size value. For smaller ϵ ($\leq 1 \times 10^{-6}$), $h = 2048$ achieves a better $\mathcal{SWD}_2$ than $h = 4096$ for SCONES, which suggest it might not be suitable to increase more the neural networks

width. For BP, $h = 2048$ and $h = 4096$ achieves almost the same performance, suggesting the same thing.

While SCONES fails against BP for $\lambda = 2$ and $T = 300$ using the $\mathcal{SWD}_2$ metric, it looks qualitively better than BP, and even more for higher values of T, as shown in Figure 8. However, despite this clear increase in quality, the metric does not follow what we visually see, even decreasing with T (Figure 9), indicating it might not be suitable for this specific distribution.

4 Conclusion

In this report, we investigated score-based generative neural networks for regularized optimal transport, focusing on the SCONES approach [1] as an alternative to the barycentric projection (BP) method [2]. Our experiments on low-dimensional Gaussian distributions and more intricate Swiss Roll targets reveal that SCONES successfully mitigates the averaging artifacts inherent in barycentric projection and can produce samples that more closely align with the true transport plan under appropriate parameter settings. In particular, for high regularization strengths, SCONES often outperforms BP in terms of both quantitative metrics, such as BW-UVP for Gaussian-to-Gaussian transport, and qualitative sample fidelity. Additionally, the flexibility afforded by Langevin sampling allows practitioners to trade computation for increasingly accurate sample approximations of the transport plan.

However, these advantages do not come without drawbacks. The method's reliance on iterative sampling introduces a significant computational overhead compared to barycentric projection, especially for large T in the Langevin dynamics. Moreover, the performance of SCONES depends heavily on tuning key hyperparameters, including the number of sampling steps T, the step size ϵ, and the size of the neural network. In particular, we observed that too large or too small values of ϵ may lead to suboptimal or unstable results (e.g., exploding sample variances), indicating a balance that must be struck in practice. The choice of the regularization parameter λ likewise plays a crucial role, as it influences both the Sinkhorn coupling's smoothness and the difficulty of sampling.

Looking ahead, one promising direction is to develop efficient heuristics or adaptive procedures for hyperparameter selection, potentially reducing the manual tuning required for SCONES. Another direction might be to explore better initialization strategies to further improve sample quality and stability. Finally, integrating more sophisticated sampling techniques may alleviate some of the computational bottlenecks observed here.

5 Connection with the Course

The methodology and theoretical foundations presented in this work have strong connections with the notions introduced in the *Computational Optimal Transport* course:

- **Entropy regularization and Sinkhorn's algorithm**: Regularized optimal transport is the topic covered by the paper, so these concepts are central.
- **Duality:** The course also is about the dual formulation of optimal transport problems, which is directly used in the paper.
- **Wasserstein distance**: I used a variant of the standard Wasserstein distance seen during the course to evaluate the method.
- **The numerical tour #2**: Playing with the Sinkhorn's algorithm and Iterative Bregman Projection algorithm was really helpful, and has helped me understand the barycentric projection method.

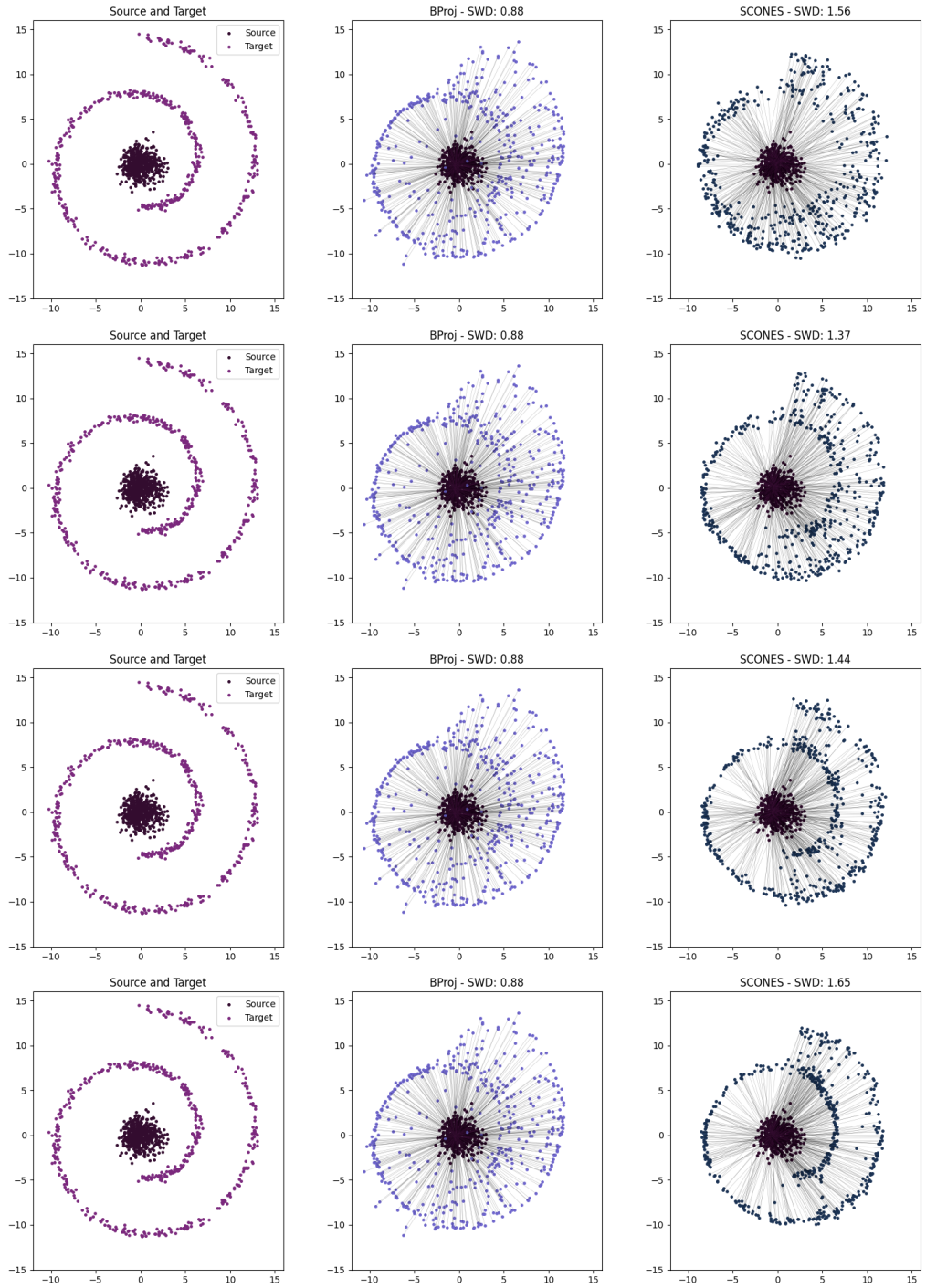

Figure 8: Visualization of predicted samples from a Gaussian 2D source distribution to a Swiss Roll 2D target distribution for $\lambda = 2,\ T = 100, 300, 600, 1000$ (top to bottom) $\epsilon = 1 \times 10^{-6}$

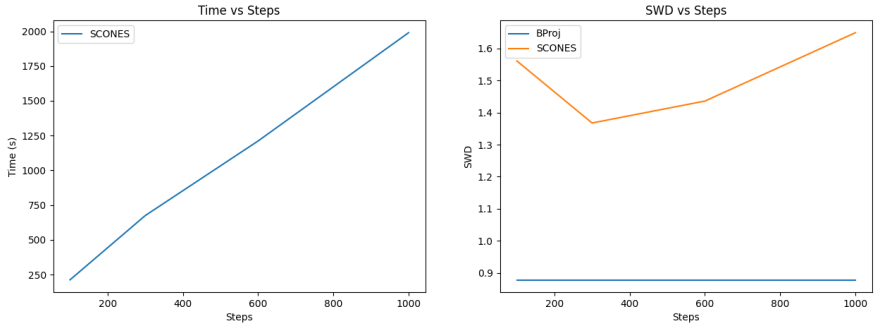

Figure 9: SCONES running time vs. T (left) and SCONES $\mathcal{SWD}_2$ vs. T (right)

6 References

References

[1] Max Daniels, Tyler Maunu, and Paul Hand. Score-based generative neural networks for large-scale optimal transport, 2022.

[2] Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation, 2018.

[3] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models, 2020.

[4] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013.

[5] Richard Sinkhorn. A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices. *The Annals of Mathematical Statistics*, 35(2):876 – 879, 1964.

[6] Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport, 2018.

[7] Yuesheng Xu and Haizhang Zhang. Uniform convergence of deep neural networks with lipschitz continuous activation functions and variable widths, 2023.

[8] Jiaojiao Fan, Amirhossein Taghvaei, and Yongxin Chen. Scalable computations of wasserstein barycenter via input convex neural networks, 2021.

[9] Hicham Janati, Boris Muzellec, Gabriel Peyré, and Marco Cuturi. Entropic optimal transport between unbalanced gaussian measures has a closed form, 2020.

[10] 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, April 2014.

[11] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

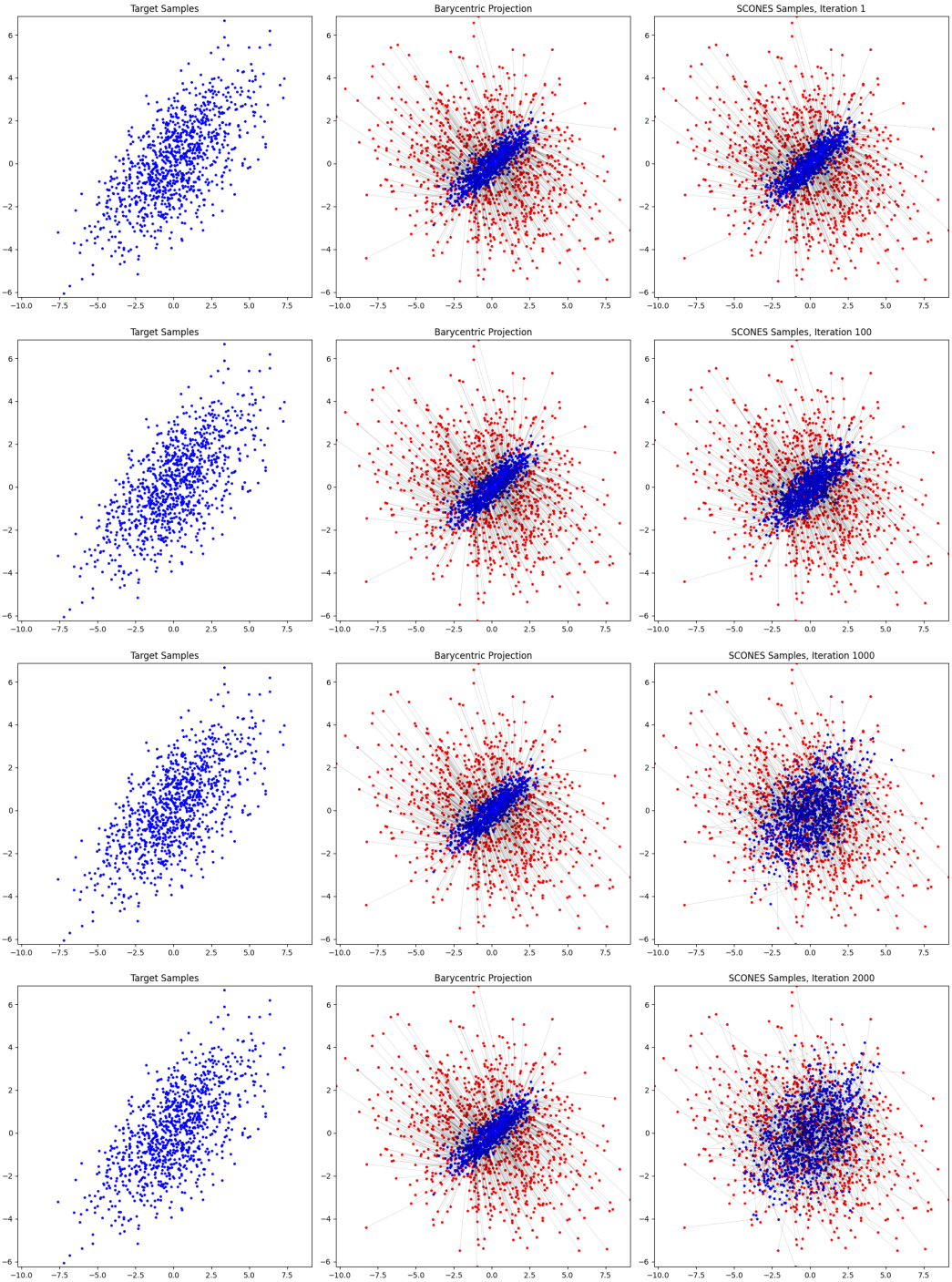

Figure 10: Visualization of predicted samples from a Gaussian 2D source distribution (Red) to a Gaussian 2D target distribution (Blue, left) for $\lambda = 20, \epsilon = 0.001, h = 64$, for $T \in \{1, 100, 1000, 2000\}$

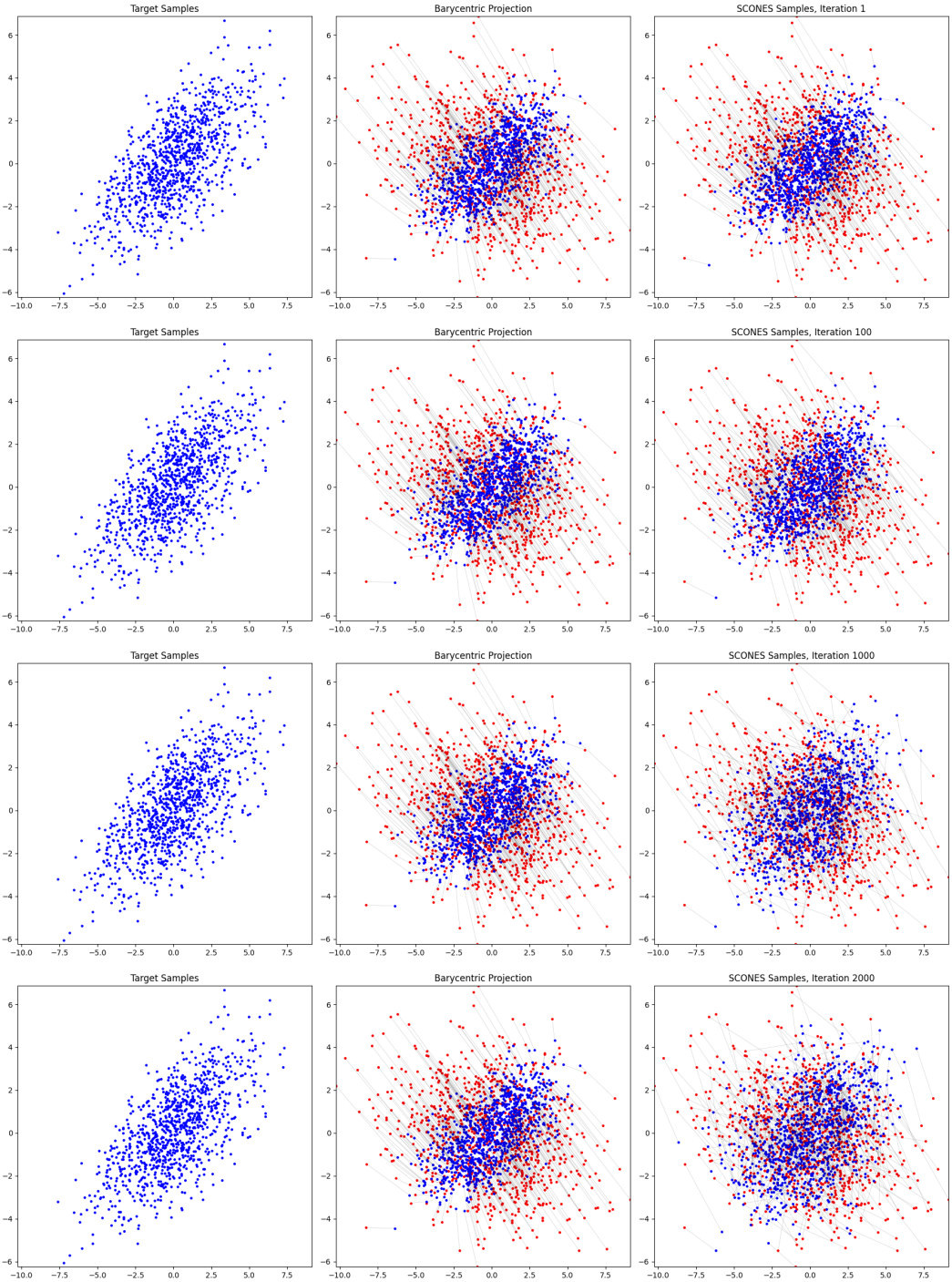

Figure 11: Visualization of predicted samples from a Gaussian 2D source distribution (Red) to a Gaussian 2D target distribution (Blue, left) for $\lambda = 4, \epsilon = 0.001, h = 64$, for $T \in \{1, 100, 1000, 2000\}$