AMATH 563: COMPUTATIONAL REPORT 5

NGHI NGUYEN

Applied Math Department, University of Washington, Seattle, WA
nghiiuu@uw.edu

1. Introduction

In this report, we designed, implemented, and tested a kernel-based algorithm for Generative Modeling; specifically, we solved a non-convex opitmization problem for the optimal transport operator that transports the points from a reference set with a normal distribution to a 2-D complex distribution. The 2-D distribution includes the "moons, "swissroll", and the "pinwheel" distribution. The optimization problem uses the Radial Basis Function (RBF), Laplacian, and the polynomial of various degrees kernels. The optimization problem minimizes over the Maximum Mean Discrepancy (MMD), a statistical divergence that measures the difference between two probability distribution. To quantify the convergence of the transported distribution to the target distribution, the norm of the gradient and the value of the MMD loss will be considered over number of iterations.

2. Theory

Define the reference space $\mathcal{Z} = \mathbb{R}^2$, the normal distribution $\eta = N(0, I) \in \mathbb{P}(\mathcal{Z})$ where N denotes the sample size, the image space $\mathcal{X} = \mathbb{R}^2$, and the target distribution $\nu \in \mathbb{P}(\mathcal{X})$. The goal Generative Modeling is to find an optimal transport map $T^* : \mathcal{Z} \to \mathcal{X}$ such that the distribution of the random variable $T(z)$, ie, $T^*_{\#}\eta \approx \nu$ for $z \in \mathcal{Z}$. The target distribution ν will be 2D complex distribution with the following shape: pinwheel, moons, and swissroll.

To quantify what we mean by an "optimal" transport and what the approximation sign $\approx$ measures, consider the following optimization problem where $Z \subset \mathcal{Z}$ and $X \subset \mathcal{X}$:

$$T^* = \underset{T \in \mathcal{T}}{\operatorname{argmin}} D(T(Z), X) + R(T) \tag{1}$$

where $Z = \{z_1, ..., z_N\}$ and $T(Z) = \{T(z_1), ..., T(z_N\}$. The optimization problem is equipped with a statistical divergence function D to measure the difference between the two probability distributions $T(Z)$ and X. The specific statistical divergence will be further discussed in the 3 section.

3. Methods

Before solving the optimization problem 1, we want to quantify the measure between the reference space $\mathbf{Z}$ and $\mathcal{X}$ before any transport mapping (Part 1). To do so, consider the following statistical divergence: **Maximum Mean Discrepancy (MMD)** for $X, X' \in \mathbb{P}(\mathcal{X})$

Date: June 9, 2025.

$$MMD(X, X') = \frac{1}{N^2}\sum_{i,j=1}^{N} G(x_i, x_j) + \frac{1}{(N')^2}\sum_{i,j=1}^{N'} G(x'_i, x'_j) - \frac{2}{NN'}\sum_{i=1}^{N}\sum_{j=1}^{N'} G(x_i, x'_j) \tag{2}$$

where $G : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a PDS kernel on the RKHS $\mathcal{G}$. The kernels we will be implemented in this report is the Radial Basis Function (RBF), Laplacian, and the polynomial degree. The kernels are defined as the following:

$$\begin{cases} \text{RBF}: & K(x,x') = \exp(-\frac{||x-x'||^2}{2l^2}) \\ \text{Laplacian}: & K(x,x') = \exp(-\frac{||x-x'||^2}{l}) \\ \text{Polynomial}: & K(x,x') = (bx^Tx' + c)^d \end{cases}$$

For simplicity, denote the parameters $\gamma_{RBF} = \frac{1}{2l^2}$ and $\gamma_L = \frac{1}{l}$ where l is the length-scale to the respective kernel. For part 1, we will use **Median Heuristic** to pick the length-scale l as the pairwise distance between $Z \subset \mathcal{Z}$ and $X \subset \mathcal{X}$. Implementing the MMD with the kernels above, the distance of the reference space and image space follows: Now that we have recorded the distance

Target	**RBF**	**Laplacian**	**Polynomial** (degree $d = 2$)
Swissroll	0.005956	0.010922	0.022059
Moons	0.003421	0.012184	0.387632
Pinwheel	0.002376	0.012630	0.000077

TABLE 1. MMD Distance between η and ν with different kernels

between the two distributions before the optimal transport map has been applied, we will focus on the optimization problem 1. The MMD statistical divergence used in part 1 will be minimize to get the following optimization problem for $n-$RKHS functions T_j's :

$$\begin{aligned} T^* =& \arg\min_{T\in\mathcal{T}} \; MMD_G^2(T_\#^N, \nu^N) + \frac{\lambda}{2}||T_j||_{\mathcal{T}}^2 \\ &\equiv \arg\min_{\underline{w}\in\mathbb{R}^n} \; MMD_G^2(Z+V, X) + \frac{\lambda}{2}\sum_{j=1}^{n} \underline{w}_j^T (K_j(Z,Z) + \sigma^2 I)^{-1}\underline{w}_j \\ &= \arg\min_{\underline{w}\in\mathbb{R}^n} f(\underline{w}) \end{aligned}$$

where λ is the regularization problem. The optimal transport could be written as $T_j^* = K_j(\cdot, Z)(K_j(Z,Z)+\sigma^2 I)^{-1}\underline{w}_j$ and the optimal coefficient $\underline{w}_j^*$ solves the optimization problem

$$\underline{w}_j^* = \arg\min_{\underline{w}\in\mathbb{R}^n} \; MMD_G^2(W, X) + \frac{\lambda}{2}\sum_{j=1}^{n} \underline{w}_j^T (K_j(Z,Z) + \sigma^2 I)^{-1}\underline{w}_j$$

.

We applied **L-BFGS**, a quasi-Newton algorithm, with the RBF kernel to solve the non-convex smooth optimization problem for T^*. A quasi-Newton method solves for the minimizer by approximating the Hessian for each step of the Newton's method, rather than directly computing the Hessian matrix due to time cost. At iteration k and current position $\underline{w}_k$, the next position is computed as

$$\underline{w}k + 1 = \underline{w} - \alpha_k d_k$$

where $\alpha_k = 1$ is the step size and $d_k = KH(x^k)\nabla f(\underline{w}_k)$ is the search direction. In the search direction, $H(\underline{w}^k) \approx (\nabla^2 f(\underline{w}_k))$ is the numerical approximation of the inverse of the Hessian matrix,

and $\nabla f(\underline{w}_k)$ is the gradient. To approximate the Hessian matrix, the L-BFGS requires the following computations: $\begin{cases} \textbf{Vector Step}: & s_k = x_{k+1} - x_k \\ \textbf{Gradient Change} & y_k = \nabla f(\underline{w}_{k+1}) - \nabla f(\underline{w}_k) \\ \textbf{Curvature Scale}: & \rho_k = \frac{1}{y_k^T s_k} \end{cases}$ for a two-loop recursion and the gradient is computed with jax.grad. The overall algorithm uses the JAX library and jax.opt.

Convergence Indicator: The L-BFGS algorithms ends either when the norm of the gradient vector at the updated position is less than a fixed tolerance ie. $||\nabla f(\underline{w}_{k+1})|| \leq \textbf{tol} = 1e-5$ or when the iteration reaches a max interation of 100. Given the non-convexity of the problem, we graphed the gradient norm and the MMD loss values for each iteration to examine if T^* could be a saddle point, local/global maximum, or local/global minimum. The graphs can be further examine in 4.

We continued with the L-BFGS algorithm with the Laplacian kernel and the polynomial kernel for degree $d = 1,\ 2,\ 4$ along with the RBF kernel. For each kernel, their respective parameters will be hand-tuned for $N = 5000$ sample points.

4. Results

1 shows the gradient norms and the MMD loss for each iteration (up until the max iteration 100) using the L-BFGS and RBF. From visual aspect, we can see that of the gradient norm decreases near zero as the iteration increases. Given that the optimization is non-convex, we could not assume that the solution of the optimization problem T^* is a minimizer, saddle point, or maximizer. However, the MMD values decrease rapidly in the first 20 iterations and stabilize near 0 after, indicating the distance between the two distributions is close to zero (a minimum value since statistical divergence are always non-negative) and thus we can diagnose convergence.

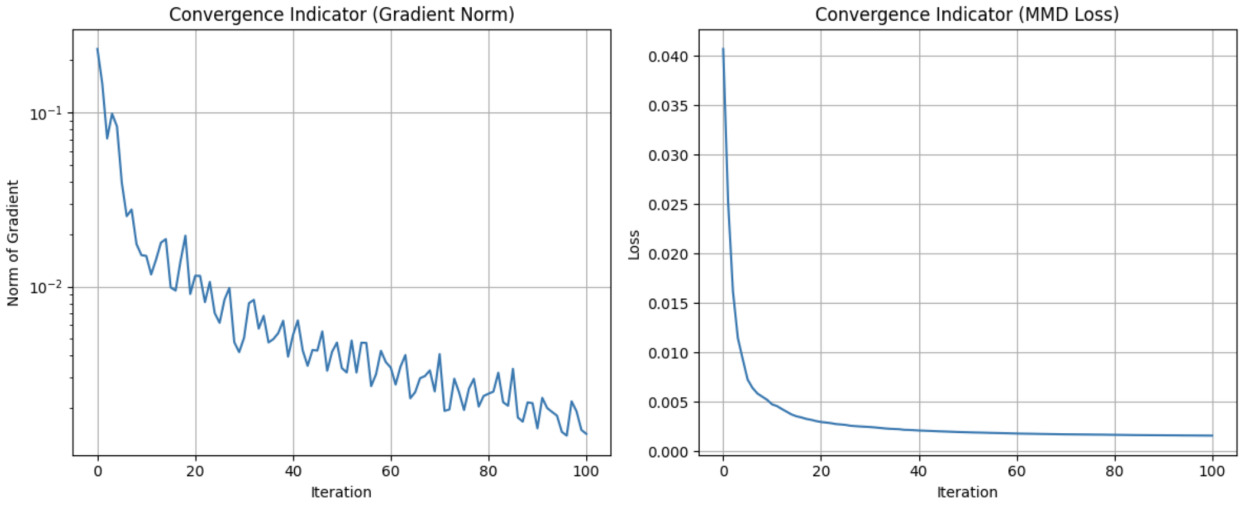

Figure 1. A plot of the gradient norm of the minimized problem and the MMD values over 100 iterations

2, 3, and 4 show the heatmap of the transported random variables in the normalized distribution to the 2D-complex distribution (pinwheel, swissroll, and moons) with the RBF kernel for 5000 points. The values of the parameter chosen are $\gamma_{RBF} = 100$ (corresponding with the lengthscale $l \approx 0.0707$) regularization parameter $\lambda = 1e-4$. Comparing the transported samples to the target

samples, we see that the transported samples have the same structural and color gradient as the target samples. This is consistent to the rapid decay of the MMD loss, emphasizing the transported samples converges to the target sample. 5, 6, and 7 show the heatmap of the transported samples with the Laplacian kernel; the values of the parameter chosen are $\gamma_L = 100$ (corresponding with the lengthscale $l = .1$) regularization parameter $\lambda = 1e - 4$.Similarly, we see that the transported samples have the same curvature and color schemes as the target heatmap, showing the convergence $T^*_{\#}\eta$ to ν. Thus, both the RBF kernel and the Laplacian kernel yields good optimal transport that maps for normalized distribution to 2D complex distribution. The performance of the RBF and Laplacian kernels is due to the Euclidean distance $||x - x||^2$; as the distance of the point increase $||x - x||^2 \to \infty$, both kernels $K(x, x') \to 0$, whereas he distance of the point decrease $||x - x||^2 \to 0$, both kernels $K(x, x') \to 1$. This means both kernels adapt to the distance of the samples and allowing for meaningful structures, which is crucial for complex distribution. On the contrary, the polynomial kernel does not perform as well as seen in 8,9, and 10 where the coefficient $b = .01$. Note that the plots for degree $d = 1,\ 2,\ 4$ are all the same, hence we only imported the plots for degree $d = 1$. For all target distributions (pinwheel, swissroll, and moons), the transport map acts as an identity map and the transported samples showed no visual change to the normalized samples. Again, contrary to the two kernels above, the polynomial kernel does not take in the Euclidean distance $||x - x||^2$, which does not allow the kernel to be sensitive to the structure and geometry of the distribution.

Variation of parameters for the polynomial kernel. In an attempt to improve the performance the polynomial degree, we picked the coefficient parameter to be a higher number $b = 75$ for degree $d = 1$. To shorten the running time, we picked $N = 500$ sample points as $N = 5000$ sample points took more than 50 minutes to run the pinwheel distribution. As see in 11, 12, 13, the transport map did not act like an identity map like for $b = .01$. The target map diagonally stretchy the normal distribution sample set but the color scheme and structure of the target sample set were not achieved. This once again shows that the polynomial degree performs poorly in generative modeling for 2D-complex target distribution.

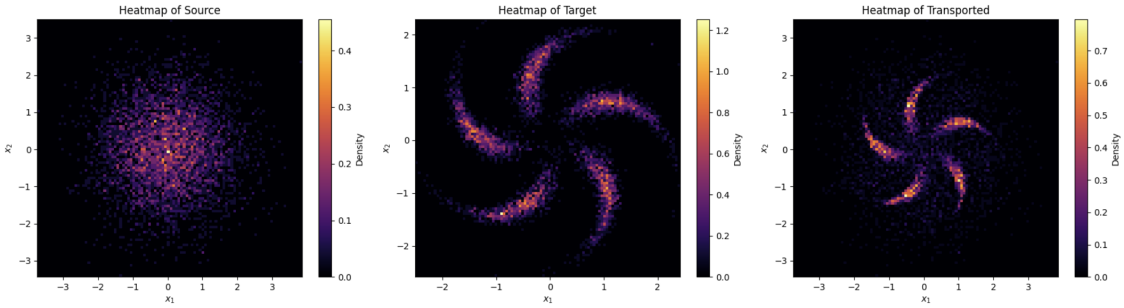

FIGURE 2. The RBF kernel with $\gamma_{RBF} = 100$ (corresponding with the lengthscale $l \approx 0.0707$) regularization parameter $\lambda = 1e-4$. The most left plot is the normal distributed samples, the middle is the target Pinwheel distribution, and the left plot is the transported sample.

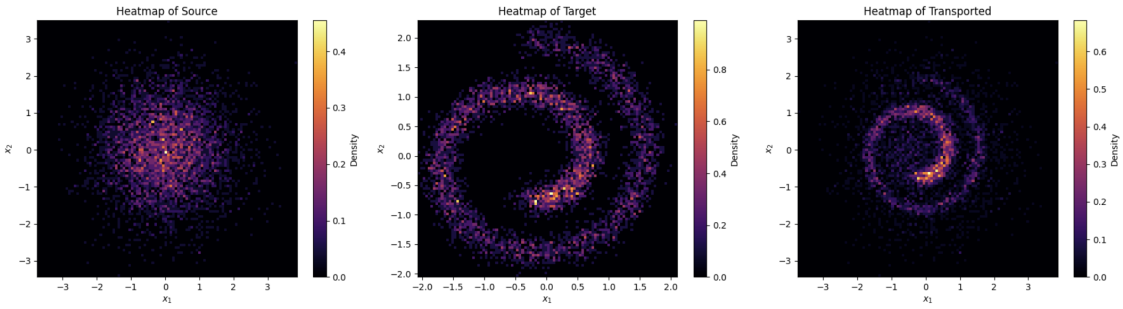

FIGURE 3. The RBF kernel with $\gamma_{RBF} = 100$ (corresponding with the lengthscale $l \approx 0.0707$) regularization parameter $\lambda = 1e-4$. The most left plot is the normal distributed samples, the middle is the target Swissroll distribution, and the left plot is the transported sample.

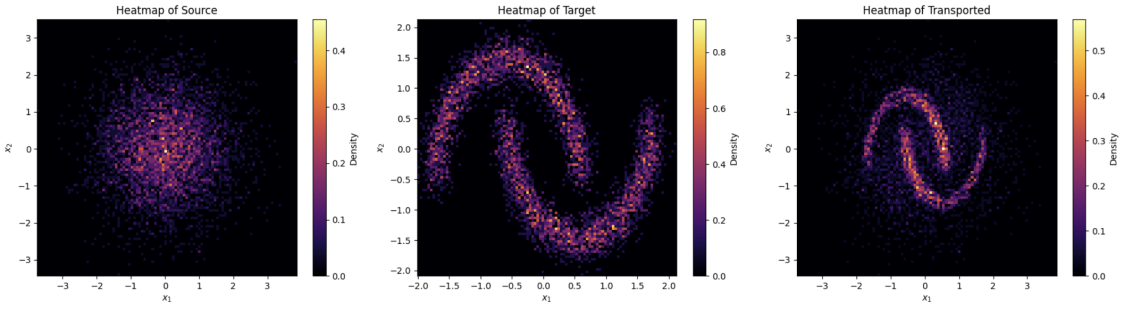

FIGURE 4. The RBF kernel with $\gamma_{RBF} = 100$ (corresponding with the lengthscale $l \approx 0.0707$) regularization parameter $\lambda = 1e-4$. The most left plot is the normal distributed samples, the middle is the target Moons distribution, and the left plot is the transported sample.

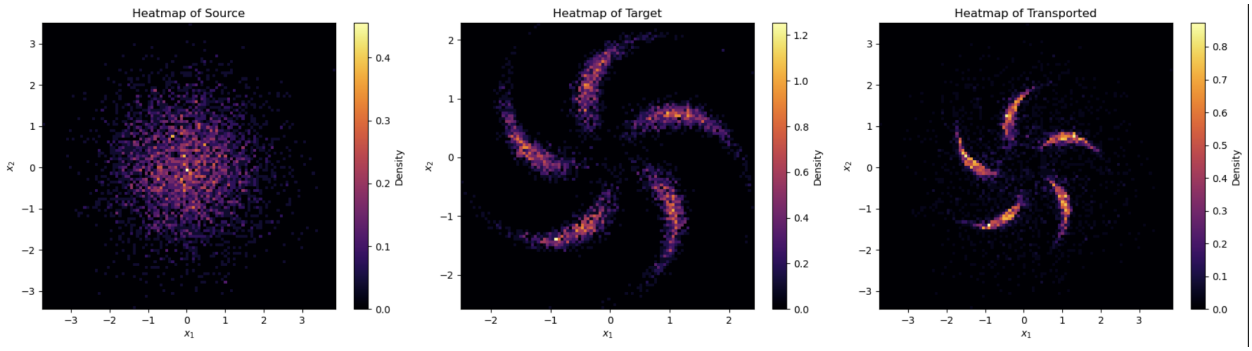

FIGURE 5. The Laplacian kernel with $\gamma_L = 100$ (corresponding with the lengthscale $l = .01$) regularization parameter $\lambda = 1e - 4$. The most left plot is the normal distributed samples, the middle is the target Pinwheel distribution, and the left plot is the transported sample.

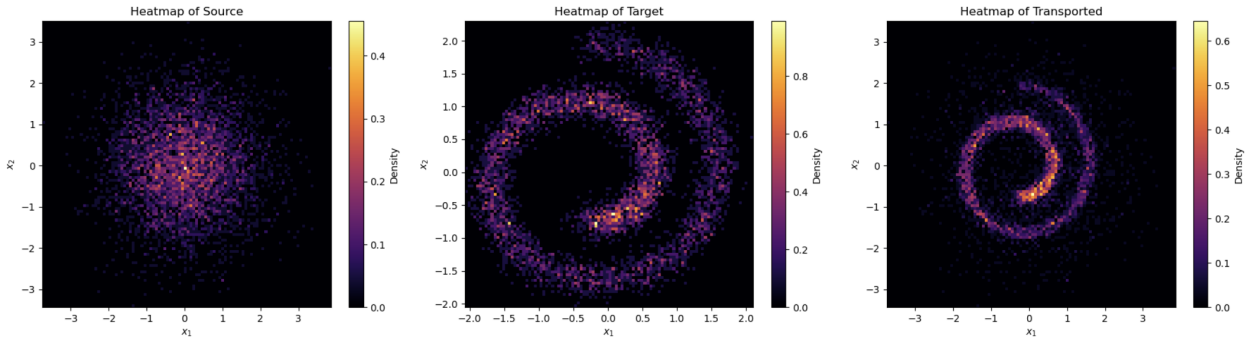

FIGURE 6. The Laplacian kernel with $\gamma_L = 100$ (corresponding with the lengthscale $l = .01$) regularization parameter $\lambda = 1e - 4$. The most left plot is the normal distributed samples, the middle is the target Swissroll distribution, and the left plot is the transported sample.

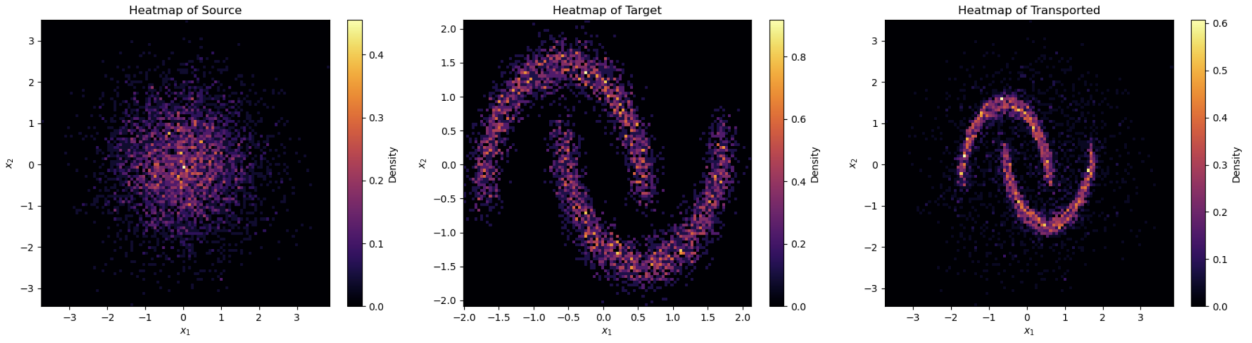

FIGURE 7. The Laplacian kernel with $\gamma_L = 100$ (corresponding with the lengthscale $l = .01$) regularization parameter $\lambda = 1e - 4$. The most left plot is the normal distributed samples, the middle is the target Moons distribution, and the left plot is the transported sample.

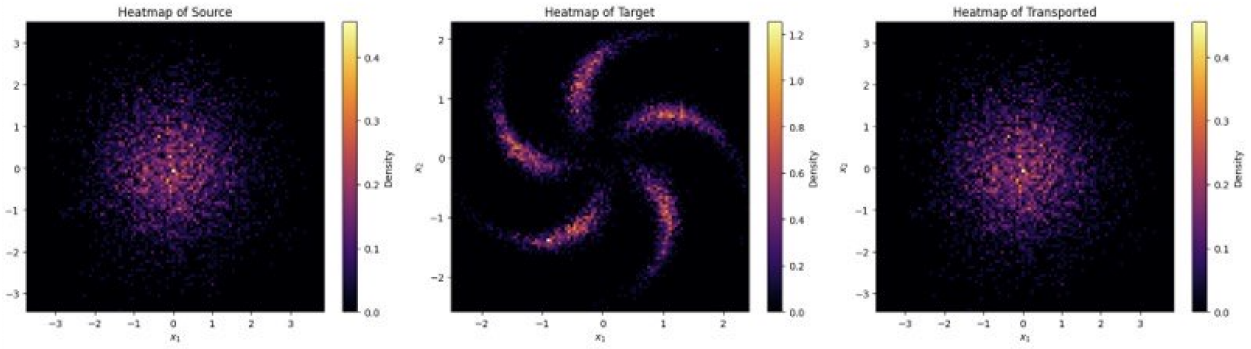

FIGURE 8. The Polynomial kernel with degree $d = 1$ (same as $d = 2, 4$) with the coefficient $b = .01$ and the regularization parameter $\lambda = 1e - 4$. The most left plot is the normal distributed samples, the middle is the target Pinwheel distribution, and the left plot is the transported sample.

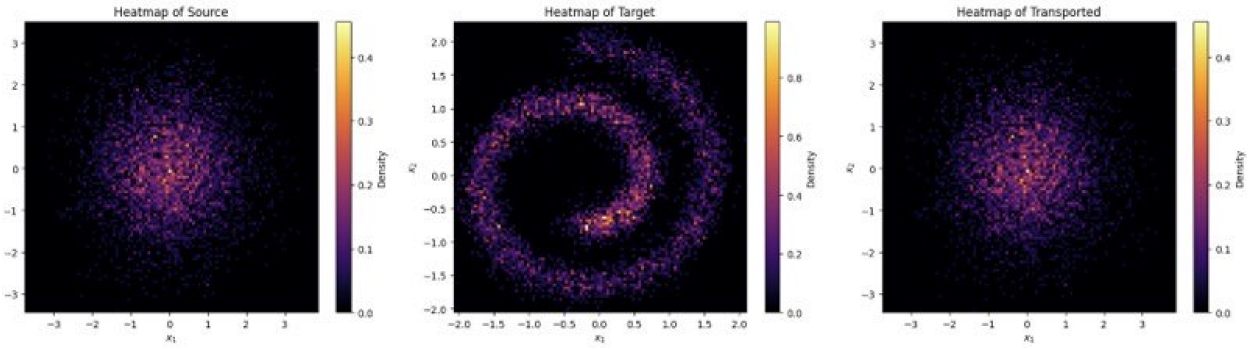

FIGURE 9. The Polynomial kernel with degree $d = 1$ (same as $d = 2, 4$) with the coefficient $b = .01$ and the regularization parameter $\lambda = 1e - 4$. The most left plot is the normal distributed samples, the middle is the target Swissroll distribution, and the left plot is the transported sample.

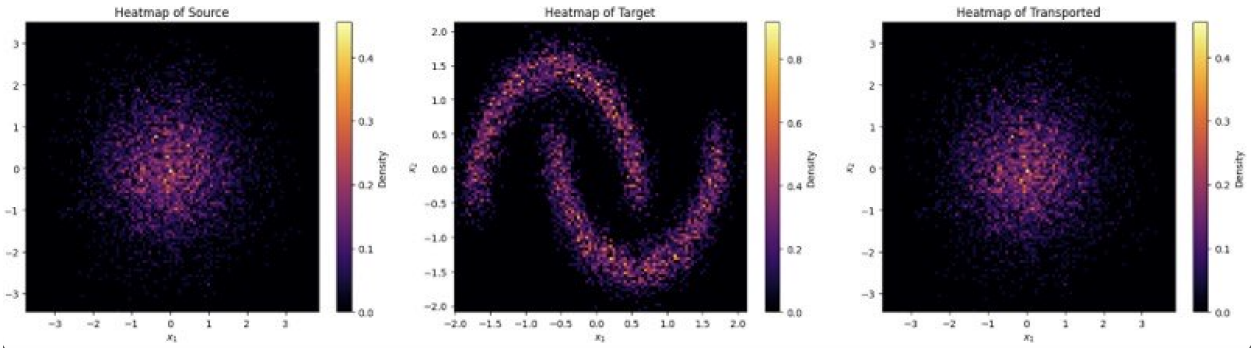

FIGURE 10. The Polynomial kernel with degree $d = 1$ (same as $d = 2, 4$) with the coefficient $b = .01$ and the regularization parameter $\lambda = 1e - 4$. The most left plot is the normal distributed samples, the middle is the target Moons distribution, and the left plot is the transported sample.

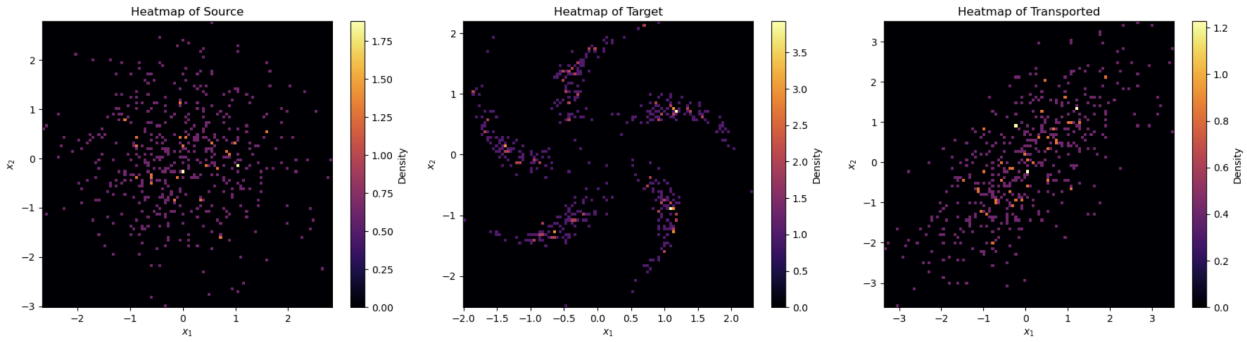

FIGURE 11. The Polynomial kernel with degree $d = 1$ with the coefficient $b = 75$ and the regularization parameter $\lambda = 1e - 4$. The most left plot is the normal distributed samples, the middle is the target Pinwheel distribution, and the left plot is the transported sample.

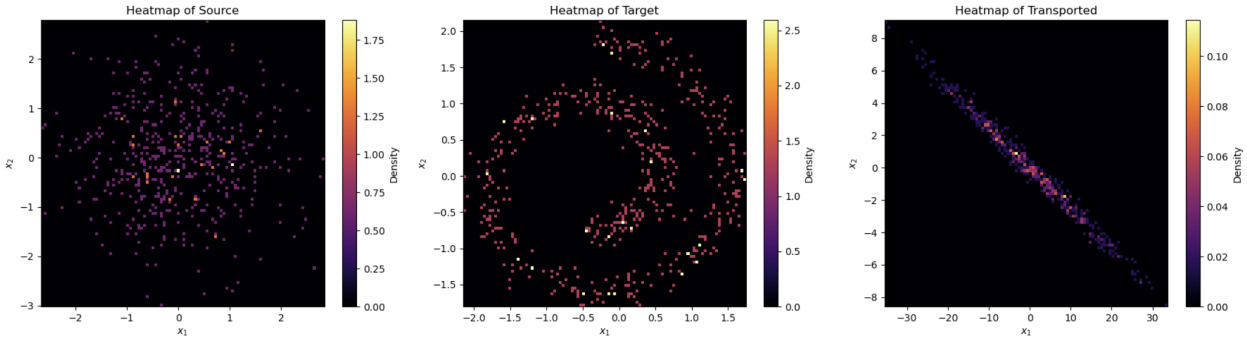

FIGURE 12. The Polynomial kernel with degree $d = 1$ with the coefficient $b = 75$ and the regularization parameter $\lambda = 1e - 4$. The most left plot is the normal distributed samples, the middle is the target Swissroll distribution, and the left plot is the transported sample.

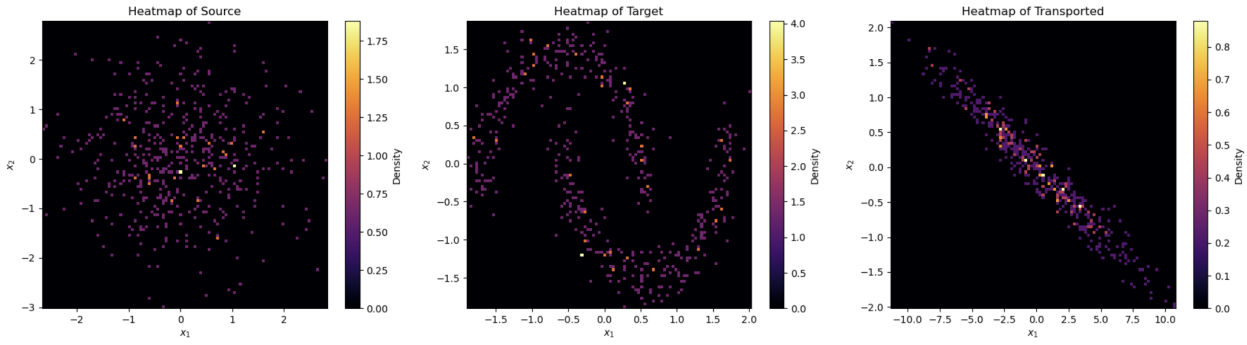

FIGURE 13. The Polynomial kernel with degree $d = 1$ with the coefficient $b = 75$ and the regularization parameter $\lambda = 1e - 4$. The most left plot is the normal distributed samples, the middle is the target Moons distribution, and the left plot is the transported sample.

5. Summary and Conclusions

In conclusion, the transport map living in the Reproducing Kernel Hilbert Spaces associated with the Radial Basis Function (RBF) and Laplacian kernels out-performed the polynomial basis with various degrees for transporting the set of normal distributed samples to the 2D-complex distribution. The Maximum Mean Discrepancy optimization problem's favor for the RBF and the Laplacian kernel could be due to how both kernels take into consideration the Euclidean distance of the points in a set, allowing for consideration and sensitivity of the structure and geometry of the 2D-complex distribution. Given the number of hyper-parameters for each kernel and the regularization parameter, we had to hand-tune each parameter in order to find the most suitable one. If this experiment continues, we hope to only stick with the RBF or the Laplacian kernel and use cross validation to pick the hyperparameter.

Acknowledgements

I would like to thank Emily, Howard, Johan, Christina, and Tomas for fruitful discussion.