UNIVERSITY OF CALIFORNIA,
IRVINE

# Does t-SNE See Curves? Theory and Practice

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematics

by

Kathryn Dover

Dissertation Committee:
Professor Roman Vershynin, Co-Chair
Assistant Professor Anna Ma, Co-Chair
Professor Mike Cranston

2023

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ALGORITHMS

# ACKNOWLEDGMENTS

I would like to first and foremost thank both of my advisors Roman Vershynin and Anna Ma for their continued guidance and support throughout my time in graduate school. Without them this work would not exist. I would also like to thank Qing Nie and Zixuan Cang for their collaboration on AVIDA and lending their expertise in the area of computational biology.

I would also like to acknowledge the National Science Foundation DMS-1954233 and DMS-2027299, the U.S Army 76649-CS, the NSF+Simons Research Collaborations on the Mathematical and Scientific Foundations of Deep Learning, ARCS Foundation, UC Irvine Math Department, UC Irvine School of Physical Sciences, and UC Irvine Graduate Division for funding and assistance.

I would also like to add a special thank you to both Professor Alessandra Pantano and my committee member Professor Mike Cranston for their support.

Thank you to my fellow grad students, you got me through classes, exams and kept me sane during the pandemic.

Finally, I would like to thank my family and friends for their continuous support during this long journey. A special thanks to Andrew for his support and encouragement.

# VITA

## Kathryn Dover

### EDUCATION

**Doctor of Philosophy in Mathematics** **2023**
University of California, Irvine *Irvine, California*

**Master of Science in Mathematics** **2019**
University of California, Irvine *Irvine, California*

**Bachelor of Science in Mathematics** **2017**
Harvey Mudd College *Claremont, California*

# ABSTRACT OF THE DISSERTATION

Does t-SNE See Curves? Theory and Practice

By

Kathryn Dover

Doctor of Philosophy in Mathematics

University of California, Irvine, 2023

Professor Roman Vershynin, Co-Chair
Assistant Professor Anna Ma, Co-Chair

Data visualization is the process of taking very high dimensional data and representing it in two or three dimensions. The main goal of data visualization is to create a representation of the data in such a way that a human observer can gain insight into the data's structure or patterns. There are many methods that produce visually appealing representations on a variety of datasets, but there is limited theory on why some of these methods work. The purpose of this dissertation is to identify quality metrics of good visualizations and use them to create new methods.

Chapter 1 is an introduction to t-stochastic neighbor embedding (t-SNE), one of the standard algorithms for data visualization. It is a non-convex method that represents the differences between data points as weighted probabilities in high and low dimensions and then minimizes the 'distance' between these two distributions using the Kullback-Leibler divergence. Despite its wide success, there is still very little mathematical understanding of the algorithm.

One of t-SNE's more interesting properties is its tendency to preserve local linear structure of data. For example, if t-SNE is given a dataset of multiple rings in high dimensions, its low dimensional representation will include multiple rings rather than multiple clusters. This preservation of fine local structure sets it apart from other methods. Chapter 1 also defines

and explores discrete curves, a new mathematical definition meant to represent these fine local structures both in high-dimensions and low-dimensions. Chapter 2 then rigorously proves that given a 1D structure in high dimensions, t-SNE will visualize that structure in its output.

This dissertation not only proves that t-SNE preserves this discrete curves in theory, but also demonstrates that knowledge can be applied successfully in practice, specifically for data integration of single-cell measurements. Chapter 3 introduces single-cell analysis, the study of human cells in the same population (liver cells, skin cells, etc.) that are genetically identical but behave differently. The differences between these cells can have important impacts on the health and function of the whole cell population. Single-cell analysis is the process of studying cell-to-cell variation within a certain cell population by looking at different properties of their genome, like gene expression and chromatin accessibility, which are referred to as single-cell measurements. Single-cell analysis has been applied in studying diseases, drug development, and in-depth analysis of stem cell differentiation.

One of the big challenges in single-cell analysis is processing the single-cell measurements. Due to technical limitations, it can be hard to obtain multiple types of measurements of the same cell. For example, for a small population of liver cells, a user may only have access to a dataset representing the gene expression of each cell and another dataset representing the chromatin accessibility. These datasets will not only be very high-dimensional and have local discrete structures, but they will not live in the same dimension since they represent different properties. Thus there is no direct way to identify corresponding features in both datasets since they represent different domains. Chapter 4 discusses AVIDA, an algorithm to process these datasets that produces a single dataset representing both single-cell measurement. AVIDA achieves this by using t-SNE and Optimal Transport methods to not only integrate these two datasets into the same domain, but to also generate a visualization that highlights the local underlying structures in the single-cell measurements.

# Chapter 1

# Introduction

The ability to visualize high-dimensional data has become an important part of data analysis for many different fields. Since visualizations are primarily used to interpret data via the shapes that they make, it is important that the geometry of the visualization truthfully conveys a particular quality of the data. Spectral methods such as Principal Component Analysis (PCA) [36], Locally Linear Embedding (LLE) [38], Isometric Feature Mapping (Isomap) [43] and Multidimensional Scaling (MDS) [44] are very popular methods that handle manifolds particularly well. In particular Isomap is guaranteed asymptotically to recover the true geometric structure of certain manifolds [6]. However, these methods do not necessarily handle non-manifold datasets very well and in practice real-life datasets rarely have a nice manifold structure.

Student-t Stochastic Neighbor Embedding (t-SNE) [45] was introduced in 2008 by van der Matten and Hinton to address this particular issue. t-SNE is a non-stochastic method designed to promote local distances in a dataset rather than global ones. An example of the power of t-SNE is shown in Figure 1.1. t-SNE is able to clearly and correctly cluster the data in the embedding while PCA is unable to get clear separation. This chapter gives a

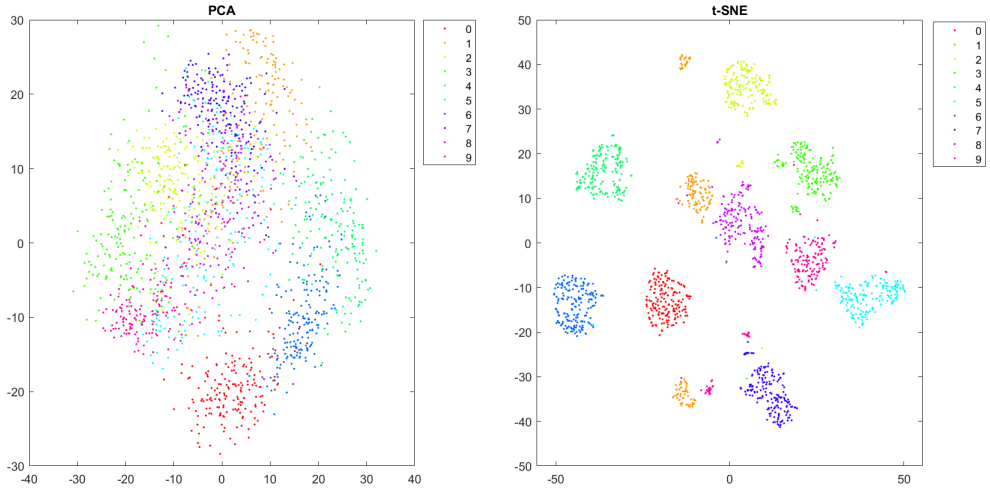


Figure 1.1: The MNIST dataset contains images of handwritten digits. These are the embeddings of MNIST made by PCA and t-SNE.

brief description of t-SNE and explains why we have decided to focus on it for this research.

## 1.1 What is t-SNE?

t-SNE takes a high-dimensional dataset $X = (x_i)_{i=1}^n$ in $\mathbb{R}^d$ and works to find a 2D embedding $Y = (y_i)_{i=1}^n$ in $\mathbb{R}^2$. t-SNE is able to find any lower dimensional embedding but here we focus on 2D embeddings for our study of visualization. The general idea behind t-SNE is to use joint probabilities to describe how "close" two points are in $X$ and find an embedding $Y$ where those probabilities are preserved, rather than preserving the Euclidean distance. First conditional probabilities are generated between any two points $x_i, x_j$ to be

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2)}, \tag{1.1}$$

where $\sigma_i$ is found using binary search such that for a given perplexity $\rho$, then $\sigma_i$ satisfies the equation

$$\rho = 2^{-\sum_{k \neq i} p_{k|i} \log_2(p_{k|i})}. \tag{1.2}$$

This perplexity value is chosen by the user and is often interpreted as the number of expected neighbors for a particular point and the induced $\sigma_i$ is the bandwidth of the Gaussian kernel such that most of the distribution's weight falls on the number of these expected neighbors. To create a joint probability between two points, the two conditional probabilities are averaged,

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}. \tag{1.3}$$

The probabilities between the points in the embedding are generated in a similar way but with the kernel from student t distribution:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq \ell}(1 + \|y_k - y_\ell\|^2)^{-1}}. \tag{1.4}$$

Let $P$ be the joint distribution of the points in $X$ and let $Q$ be the joint distribution of the points in $Y$. t-SNE minimizes the Kullback-Leibler (KL) divergence,

$$C(Y) = KL(P\|Q) = \sum_i \sum_j p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right), \tag{1.5}$$

between $P$ and $Q$ to find an embedding $Y$. The Kullback-Leibler divergence is way of measuring how "far" the distribution $P$ is from $Q$. By minimizing the divergence, t-SNE is finding an embedding $Y$ such that the induced probabilities $Q$ are most similar to $P$. t-SNE

uses gradient descent to minimize this cost function and the resulting gradient is given as

$$\frac{\delta C}{\delta y_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(1 + \|y_i - y_j\|^2)^{-1}(y_i - y_j). \tag{1.6}$$

This gradient is not the final version in the algorithm. t-SNE utilizes a phase known as early exaggeration. In the first 200 iterations of gradient descent, an exaggeration factor $\alpha > 1$ and step-size $h > 0$ are included in the gradient so that we are left with

$$\frac{\delta C}{\delta y_i} = 4h \sum_{j \neq i} (\alpha p_{ij} - q_{ij})(1 + \|y_i - y_j\|^2)^{-1}(y_i - y_j). \tag{1.7}$$

While this gradient looks complicated, the mechanics are fairly simple. Assume that two points $x_i$ and $x_j$ are close to each other in $X$. This would imply that $p_{ij}$ is very close to 1. However, if their corresponding representations $y_i$ and $y_j$ are far apart, then $q_{ij}$ is very close to zero. That means this gradient will move the points $y_i$ and $y_j$ closer together in the next iteration. The reverse would occur if the points were far apart in $X$ but the representations were too close together.

Additionally, since the choice of $\rho$ sets the bandwidth for the $p_{ij}$ values, t-SNE essentially

**Algorithm 1** t-SNE

**Input:** dataset $X = \{x_i\}_{i=1}^n$
*Parameters:* perplexity $\rho$, number of iterations $T$
**Output:** low-dimensional embedding $Y^{(T)} = \{y_i\}_{i=1}^n$
**begin**
  compute conditional probabilities $p_{j|i}$ with perplexity $\rho$ (Equation 1.1)
  set $p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n}$
  initialize $Y^{(0)} = \{y_i\}_{i=1}^n$ by sampling $\mathcal{N}(0, 10^{-4}I)$
  **for** $t = 1$ to $T$ **do**
    compute low-dimensional probabilities $q_{ij}$ (Equation 1.4)
    compute gradient $\frac{\partial C}{\partial Y}$ (Equation 1.6)
    set $Y^{(t)} = Y^{(t-1)} + \frac{\partial C}{\partial Y}$
  **end for**
**end**

just focuses on the local structures of a dataset. If two points are not in the same neighborhood, then their $p_{ij}$ values will be close to zero. As long as their representations are sufficiently far apart, their $q_{ij}$ values will be also very small and thus the gradient will have little movement.

The early exaggeration phase is a way to improve the optimization. In practice, $\alpha = 4$, so almost all of the $q_{ij}$'s are too small to model their corresponding $p_{ij}$'s. Thus, the optimization is encouraged to focus on modeling the large $p_{ij}$ values by fairly large $q_{ij}$ values. The effect is that points that are meant to be close together are highly attracted to each other and local structures are formed first before global structures are affected.

Since we will be discussing t-SNE's embeddings of high-dimensional datasets, it will be helpful to have some notation.

**Definition 1.1** (Visualization by t-SNE)**.** *Let $X = (x_i)_{i=1}^n$ in $\mathbb{R}^d$ be a subset of points in high dimensions. We denote the output of t-SNE given $X$ as an input as $Y = \text{TSNE}(X)$. $\text{TSNE}_e(X)$ is the output of t-SNE just after early-exaggeration.*

## 1.2 t-SNE Theory

While t-SNE is widely applied, there is limited theory about it. Linderman and Steinerberger [26] were able to show that t-SNE's early exaggeration stage is critical to cluster formation when the data is clusterable. Essentially by setting the $\alpha$ value to be sufficiently large, then $\alpha p_{ij}$ is much larger than $q_{ij}$ and thus points within the same neighborhood are pulled together and those who are not in the same neighborhood have very little gradient effect on each other. This allows t-SNE to focus on correctly structuring neighborhoods before attempting to orient the points in a correct global position.

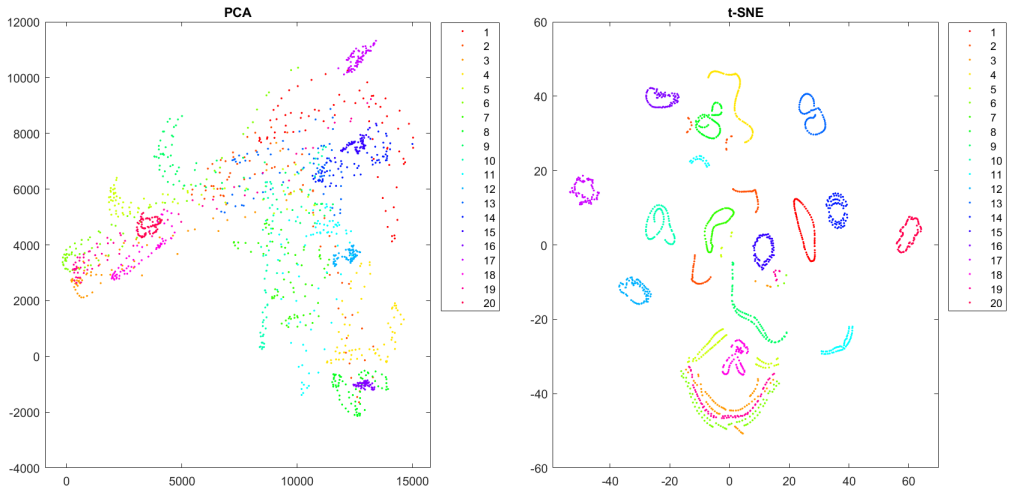


Figure 1.2: The COIL20 dataset contains images of objects at different orientations. These are the embeddings of COIL20 made by PCA and t-SNE.

Arora et al. [4] extended this result by proving that given sufficiently clusterable data, t-SNE's visualization is guaranteed to have visibly separate clusters. They were able to show that if a high-dimensional dataset had distinct clusters, then t-SNE's early exaggeration phase will correctly group points together by cluster. After leaving early exaggeration, the centers of the embedded clusters are pushed away from each other and with enough iterations become visibly distinct.

From [4], there have been additional insights into t-SNE's behavior. Zhang and Steinerberger [52] showed that the t-SNE's behavior in the limit can be analyzed with a mean-field model, specifically in the case of a single cluster. Very recently Cai and Ma [8] re-proved the separable clusters result in the context of Laplacian spectral clustering and illustrated the importance of early stopping as a form of implicit regularization.

All of these contributions focus on how t-SNE treats clusterable data. However, there have been many examples of real-life datasets that are not clusterable but t-SNE is still able to represent them well. The focus of our work is to investigate how t-SNE handles one-dimensional structures, such as curves. A good example of this is how t-SNE handles the

COIL20 dataset. This is a dataset that contains images of objects at different points in a rotation. Figure 1.2 shows that t-SNE not only clusters the points based on the object in the image, but orients them in a ring that corresponds to where the image is taken in the rotation. This is very interesting behavior that is not typically seen with methods like PCA. This work focuses on explaining why t-SNE preserves these local structures. Since we are considering the local preservation of these very particular structures, we have to define what properties the neighborhoods of these one-dimensional structures would have.

## 1.3 Discrete Curves

In order to discuss how t-SNE handles local 1D structures, it is important to clearly define them.

**Definition 1.2** ($\varepsilon$-Discrete Curve)**.** *Consider a finite sequence of points $(z_j)_{j=1}^k$ in $\mathbb{R}^d$. We say that this sequence of points is an $\varepsilon$-discrete curve if the following properties hold for $j, m$ such that $j - m, j, j + m \in [k]$:*

$$\frac{z_{j+m} + z_{j-m}}{2} = z_j + e_{j,m}, \tag{1.8}$$

*where $\|e_{j,m}\| \leq \varepsilon \|z_{j+m} - z_j\|$.*

This definition says that for any symmetrical triplet $\{z_{j-m}, z_j, z_{j+m}\}$, the middle point, $z_j$, must be close to the actual midpoint between $z_{j-m}$ and $z_{j+m}$. The use of $\varepsilon$ gives some room for error and allows for different structures of lines. A smaller $\varepsilon$ will only allow for discrete curves that are structured more like lines while a larger $\varepsilon$ will allow for a more curve-like structure. An example of this is shown in Figure 1.3. When $\varepsilon$ is set to be 0.001, the generated curve looks much more like a line. However, $\varepsilon$ set to be 0.01 allows for a less rigid structure.

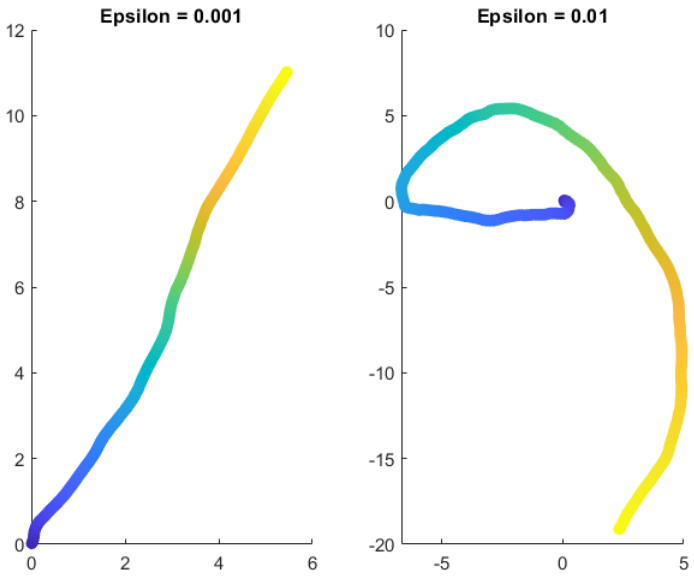


Figure 1.3: 500 randomly generated points to populate a $\varepsilon$-discrete curve where $\varepsilon$ is set to be 0.01 and 0.001.

This definition also allows us the flexibility to keep the property local. It would be too restrictive to expect an entire dataset have symmetrical triplets form a line; the only such dataset that would allow for that would be a straight line. Instead, the definition is meant to capture a local structure and require it of only points that are in each other's immediate neighborhood. Thus, this property can be used to describe parts of entire datasets, as shown in Figure 1.4. This high-dimensional dataset contains two large clusters connected by a small, densely populated line. Even though this dataset contains two very different kinds of structures, t-SNE is able to represent the local structures very well in the embedding.

Since this definition is applied locally, it is important to make some requirements of how the data is structured globally. We do not wish to consider the situation where a high-dimensional dataset or visualization has an overlap between neighborhoods.

**Definition 1.3** (No $a, b$-Critical Overlaps)**.** *Let $X = (x_i)_{i=1}^n$ in $\mathbb{R}^d$ and let $Y = \text{TSNE}(X)$. We say there are no $a, b$-critical overlaps in $Y$ if for any pair of indices $i, j$ such that $|i-j| > a$, $\|y_i - y_j\| \geq b$.*

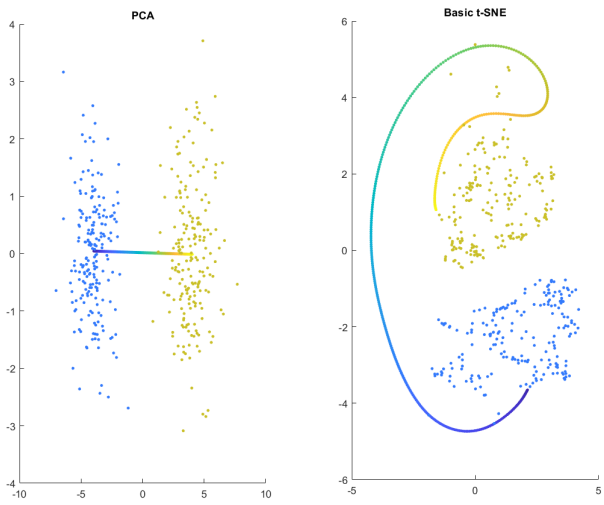


Figure 1.4: PCA and t-SNE embedding of two clouds connected by a line.

This definition encompasses the situation where the visualizations of two different neighborhoods are too close in low dimensions. An example of where this might occur is if the visualization has loops, as seen in Figure 1.5. In order to maintain the structure of the discrete curve, we require no such critical overlaps exist.

## 1.4 Main Results

In order to show that t-SNE is able to visualize an $\varepsilon-$discrete curve, we break the proof into two parts. We can show that t-SNE with early exaggeration induces a very particular structure on the gradient that encourages discrete curves to be formed.

**Theorem 1.1.** *Let $X = (x_i)_{i=1}^n$ in $\mathbb{R}^d$ be an $\varepsilon_x$-discrete curve. Then $Y = \mathrm{TSNE}_e(X)$ will also be an $\varepsilon_y$-discrete curve.*

For the formal version of this statement see Theorem 3.1. In the statement we make the distinction that while both $X$ and $Y$ are discrete curves, the $\varepsilon$ values that define them are not required to be the same. We also do not require $\varepsilon_x$ or $\varepsilon_y$ to be dependent on $n$, just that they

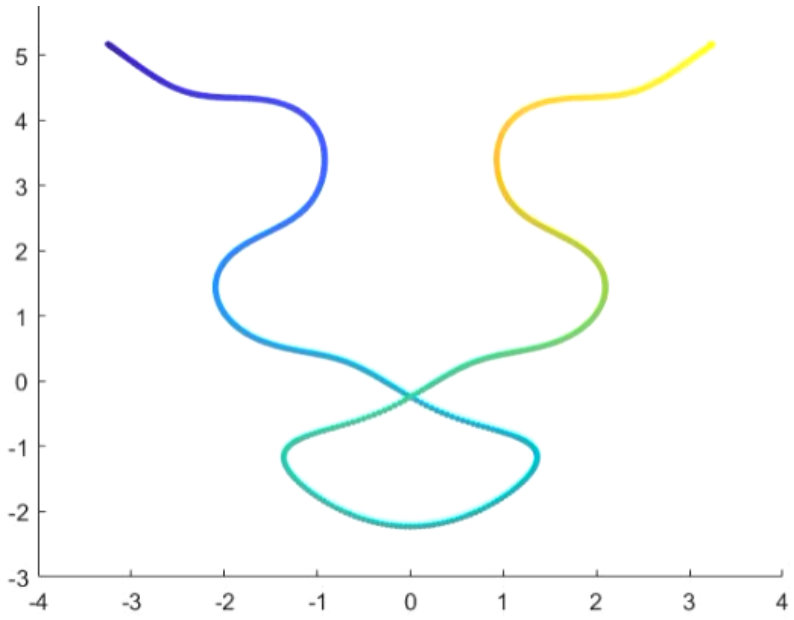


Figure 1.5: An example of a critical overlap in a representation of a curve.

are smaller than 1. We are able to show that given the particular properties of the discrete curves of $X$, the joint probabilities generated by t-SNE along with early exaggeration will essentially reduce t-SNE to a spectral method with a known solution. We then pair this result with the fact that a discrete curve in the visualization is a local minimum for the KL loss function.

**Theorem 1.2.** *Let $X = (x_i)_{i=1}^n$ in $\mathbb{R}^d$ be an $\varepsilon_x$-discrete curve and let $Y_e = \mathrm{TSNE}_e(X)$. If $Y_e$ has no $a, b$-critical overlaps, then $Y = \mathrm{TSNE}(X)$ is an $\varepsilon_y$-discrete curve.*

A formal statement and proof is given by Theorem 4.2. We show that given the assumptions about $X$, the joint probabilities generated by t-SNE induce local solutions that preserve the discrete curve structure. However, we have to assume that there are no overlaps between different neighborhoods as those will impact the iterations of t-SNE's gradient.

# Chapter 2

# Discrete Curves and t-SNE

Now that we have a definition for a discrete curve given in Def 1.2, we can use that to evaluate how t-SNE is impacted and treats such structures. We need some limitations on the density of the curve in order to do this. If a discrete curve is populated by $n$ points, we would expect the density of the curve to increase as $n$ increases rather than the distance of the curve increasing. We define density below.

**Definition 2.1** ($\lambda$-Dense Curve)**.** *Consider a finite sequence of points $(z_j)_{j=1}^k$ in $\mathbb{R}^d$. We say that this sequence of points is an $\lambda$-dense curve if the following properties hold for all $j, j+1 \in [k]$:*

$$\|z_j - z_{j+1}\| \leq \lambda. \tag{2.1}$$

In this section we quantify how distances scale in a $\varepsilon$-discrete curve and use that information to find properties of t-SNE's selection of $\sigma_i$ and gradient. The rest of this section is outlined as follows. Lemma 2.1 finds an upper bound on $\sigma_i$ when the underlying dataset is a $\lambda$-dense,

$\varepsilon$-discrete curve. Lemmas 2.2 and 2.3 estimate distances between points that satisfy the $\varepsilon$-discrete curve definition. These estimations are used in Lemma 2.4 to prove that differences in the gradient for symmetric pairs of points in a neighborhood are symmetric. In Lemma 2.5 we find upper bounds for t-SNE's gradient when the data is a discrete curve.

## 2.1 Discrete Curve and t-SNE Bandwidth

We will assume that our input data $X$ is a $\lambda_x$-dense, $\varepsilon_x$-discrete curve and $\rho$ is the perplexity chosen by the user and fixed. t-SNE uses binary search to find the appropriate Gaussian bandwidth with (1.2). Given that we assume $\lambda_x < 1/n$, we can find an upper bound on $\sigma_i$.

**Lemma 2.1.** *Let $X = (x_i)_{i=1}^n$ in $\mathbb{R}^d$ be a $\lambda_x$-dense, $\varepsilon_x$-discrete curve where $\lambda_x < 1/n$. Let $p_{j|i}$ be given as* (1.1) *and let the perplexity $\rho$ be a sufficiently small absolute constant. Then $\sigma_i \leq \frac{1}{\sqrt{\log(n)}}$.*

*Proof.* t-SNE uses binary search to find a $\sigma_i$ such that (1.2) holds.

$$\rho = 2^{-\sum_{j \neq i} p_{j|i} \log(p_{j|i})}$$

Rearranging (1.2) and applying Definition (1.1) gives us the following.

$$\begin{aligned}
\log_2(\rho) &= -\sum_{j\neq i} p_{j|i}\log(p_{j|i}) \\
&= -\left[\sum_{j\neq i} p_{j|i}\log\left(\frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{C_{P_i}}\right)\right] && (2.2)\\
&= -\left[\sum_{j\neq i} p_{j|i}\log(\exp(-\|x_i - x_j\|^2/2\sigma_i^2)) - \sum_{j\neq i} p_{j|i}\log(C_{P_i})\right] \\
&= \sum_{j\neq i} p_{j|i}\frac{\|x_i - x_j\|^2}{2\sigma_i^2} + \log(C_{P_i}) && (2.3)\\
&\geq \log(C_{P_i}) && (2.4)
\end{aligned}$$

(2.2) follows from applying the definition of $p_{j|i}$, (2.3) follows from the fact that the sum of all $p_{j|i}$ equals 1, and the last inequality (2.4) comes from the fact the first sum in (2.3) contains all positive terms. Now we wish to find a bound on the normalization term $C_{P_i}$. Because $(x_i)_{i=1}^n$ is a $\lambda_x$-dense, $\varepsilon_x$-discrete curve, then the largest pairwise distance between any two points will be at most $n\lambda_x$. Using this to bound $C_{P_i}$ we have

$$C_{P_i} = \sum_{j\neq i}\exp(-\|x_i - x_j\|^2/2\sigma_i^2) \geq \sum_{j\neq i}\exp(-n^2\lambda_x^2/2\sigma_i^2) = n\exp(-n^2\lambda_x^2/2\sigma_i^2).$$

Using the bound in (2.4) we have

$$\begin{aligned}
n\exp(-n^2\lambda_x^2/2\sigma_i^2) \leq C_{P_i} &\leq \rho, \\
\sigma_i &\leq \sqrt{\frac{n^2\lambda_x^2}{2\log(n/\rho)}}, \\
&\leq \frac{1}{\sqrt{\log(n)}}, && (2.5)
\end{aligned}$$

where (2.5) follows from the assumption that $\lambda_x < 1/n$. □

Even though we have an upper bound for $\sigma_i$, for simplicity of calculation we will assume

that $2\sigma_i = \rho^2/n^2 \leq \frac{1}{\sqrt{\log(n)}}$. This still satisfies the upper bound we found above and it is reasonable to assume that the bandwidth will be proportional to the density times the expected neighborhood size. There is also other instances in the theoretical work done about t-SNE where $\sigma_i$ was assumed to make the calculations simpler, see [4] and [26]. We can now use these properties about the $p_{ij}$ values to make insights about the overall structure of the matrix, $P$, of $p_{ij}$ values.

## 2.2 Linearity of Distances in Discrete Curves

Before we can show that t-SNE's gradient preserves the structures of discrete curves, we need some important information about the pairwise distances in discrete neighborhoods. First, we show that the pairwise distances between immediate neighbors in an $\varepsilon$-discrete curve are proportional to each other.

**Lemma 2.2.** *Assume that $(z_\ell)_{\ell=1}^k$ is an $\varepsilon$-discrete curve where $k^2\varepsilon < 1$, then for every $m < i < i + j \in [k]$,*

$$\|z_m - z_i\| \leq \left[\frac{|i-m|(1+\varepsilon)(1+|i-m|^2\varepsilon)}{j(1-j^2\varepsilon)}\right] \|z_{i+j} - z_i\|.$$

*Proof.* For simplicity, we first calculate the distance between two points, $z_m$ and $z_1$, and let $\|v\| = \|z_2 - z_1\|$. From the definition of a discrete curve, we have that

$$\|z_3 - z_2\| \leq \|v\| + |\delta_1|,$$

where $|\delta_1| \leq \varepsilon\|v\|$. Similarly, we can estimate the norm of $z_4 - z_3$ to get

$$\begin{aligned}\|z_4 - z_3\| &\leq \|z_3 - z_2\| + |\delta_2| \\ &\leq \|z_3 - z_2\| + \varepsilon\|z_3 - z_2\| \\ &\leq (1+\varepsilon)\|z_3 - z_2\| \\ &\leq (1+\varepsilon)^2\|v\|.\end{aligned}$$

We can continue this process iteratively to get that for any $\ell \leq k$,

$$\begin{aligned}\|z_\ell - z_{\ell-1}\| &= \|v\| + |\delta_{\ell-2}| \\ &\leq (1+\varepsilon)^{\ell-2}\|v\|\end{aligned}$$

where $|\delta_{\ell-2}| \leq \varepsilon(1+\varepsilon)^{\ell-2}\|v\| = ((1+\varepsilon)^j - 1)\|v\|$. If we now consider the entire distance $\|z_m - z_1\|$, then accounting for all of the errors, we get the following:

$$\begin{aligned}\|z_m - z_1\| &\leq (m-1)\|v\| + \sum_{\ell=1}^{m-1}\sum_{j=1}^{\ell}\delta_j \\ &\leq (m-1)\|v\| + \|v\|\sum_{\ell=1}^{m-1}\sum_{j=1}^{\ell} j\varepsilon + \varepsilon^2 \qquad (2.6) \\ &\leq (m-1)\|v\| + \varepsilon\|v\|\sum_{\ell=1}^{m-1}\sum_{j=1}^{\ell} j + \varepsilon \\ &\leq (m-1)\|v\| + (m-1)\varepsilon\|v\|\left[\frac{(m-2)(2m-3)}{12} + \frac{(2\varepsilon+1)(m-2)}{4}\right] \qquad (2.7) \\ &\leq (m-1)\|v\|\left[1 + m^2\varepsilon\right] \qquad (2.8)\end{aligned}$$

Line (2.6) comes from substituting our first-order approximation of $(1+\varepsilon)^j - 1 \leq j\varepsilon + \varepsilon^2$., Line (2.7) comes from evaluating the sum and factoring common terms, and Line (2.8) results from the fact $\frac{(m-2)(2m-3)}{12} + \frac{(2\varepsilon+1)(m-2)}{4}$ is bounded above by $m^2$. Additionally, since each of

the $\delta_j$ are added and could be negative, we have that a lower bound to be

$$\|z_m - z_1\| \geq (m-1)\|v\|(1-m^2\varepsilon).$$

Now we can do this same process for any $i, m, i+j$ in the neighborhood. Assume for simplicity that $m < i < i+j$. Then by following the steps in the above process, we have that

$$\|z_m - z_i\| \leq |i-m|\|z_i - z_{i-1}\|(1+|i-m|^2\varepsilon), \tag{2.9}$$

$$\|z_{i+j} - z_i\| \geq j\|z_i - z_{i+1}\|(1-j^2\varepsilon) \tag{2.10}$$

However, we know from the definition of a discrete curve that $\|z_i - z_{i-1}\| \leq (1+\varepsilon)\|z_{i+1} - z_i\|$. Substituting this into Line (2.9) we have

$$\begin{aligned}\|z_m - z_i\| &\leq |i-m|(1+\varepsilon)\|z_i - z_{i+1}\|\|(1+|i-m|^2\varepsilon), \\ &\leq \left[\frac{|i-m|(1+\varepsilon)(1+|i-m|^2\varepsilon)}{j(1-j^2\varepsilon)}\right]\|z_{i+j} - z_i\|.\end{aligned}$$

□

Here the requirement that $k^2\varepsilon < 1$ means that as a discrete curve with more curvature ($\varepsilon$ closer to 1), then the property above only holds for less and less $k$. This makes intuitive sense since we would not expect pairwise distances to scale linearly for points far away from each other if the discrete curve had extreme curvature.

Now we can show that our definition of a dense, discrete curve implies that the pairwise distances between points on the discrete curve break down nearly linearly. This property will be important later when we estimate the effect of t-SNE's gradient.

**Lemma 2.3.** *Assume that $(z_\ell)_{\ell=1}^k$ is a $\lambda$-dense, $\varepsilon$-discrete curve, then for every $c > b > a \in$*

$[k]$,

$$\|z_c - z_a\| = \|z_c - z_b\| + \|z_b - z_a\| + \delta_{c,b,a},$$

*such that*

$$\delta_{c,b,a} \leq [2(c-a)^2 + (b-a)^2]\lambda\varepsilon.$$

*Proof.* Let us take $z_c - z_a$ and expand it by adding and subtracting $z_\ell$ for $a+1 \leq \ell \leq c-1$:

$$z_c - z_a = (z_c - z_{c-1}) + (z_{c-1} - z_{c-2}) + \cdots + (z_{a+2} - z_{a+1}) + (z_{a+1} - z_a). \tag{2.11}$$

From the definition of an $\varepsilon$-discrete curve we have that $z_{a+2} - z_{a+1} = z_{a+1} - z_a + e_{a+1,1}$ where $\|e_{a+1,1}\| \leq \varepsilon\|z_{a+1} - z_a\|$. We essentially 'flip' the vectors with a cost of an error term. We can do this sequentially for each pairwise difference in (2.11). For example, with $z_{a+3} - z_{a+2}$, we can flip twice to get

$$\begin{aligned} z_{a+3} - z_{a+2} &= z_{a+2} - z_{a+1} + e_{a+2,1}, \\ &= z_{a+1} - z_a + e_{a+1,1} + e_{a+2,1}. \end{aligned}$$

Continuing this for each pairwise distance, we would get $z_c - z_{c-1} = z_{a+1} - z_a + \sum_{\ell=a+1}^{c-1} e_{\ell,1}$. Putting this into (2.11) we get

$$\begin{aligned} z_c - z_a &= (z_{a+1} - z_a) + \sum_{\ell=a+1}^{c-1} e_{\ell,1} + \cdots + (z_{a+1} - z_a) + e_{a+1,1} + (z_{a+1} - z_a), \\ &= (c-a)(z_{a+1} - z_a) + \sum_{h=a+1}^{c-1} \sum_{\ell=a+1}^{h} e_{\ell,1}. \end{aligned}$$

From the same process above, we have that

$$z_b - z_a = (b-a)(z_{a+1} - z_a) + \sum_{h=a+1}^{b-1} \sum_{\ell=a+1}^{h} e_{\ell,1}, \tag{2.12}$$

and

$$z_c - z_b = (c-b)(z_{b+1} - z_b) + \sum_{h=b+1}^{c-1} \sum_{\ell=b+1}^{h} e_{\ell,1}. \tag{2.13}$$

To compare $z_c - z_a$, $z_b - z_a$, and $z_c - z_b$, we want all the distances to be in terms of $z_{a+1} - z_a$. So looking at Line (2.13), we expand $z_{b+1} - z_b$ to get

$$z_c - z_b = (c-b)(z_{a+1} - z_a) + (c-b) \sum_{\ell=a+1}^{b} e_{\ell,1} + \sum_{h=b+1}^{c-1} \sum_{\ell=b+1}^{h} e_{\ell,1}.$$

From here, we can get a bound on the distances. By the reverse triangle inequality, we have

$$\left| \|z_c - z_a\| - \|(c-a)(z_{a+1} - z_a)\| \right| \leq \sum_{h=a+1}^{c-1} \sum_{\ell=a+1}^{h} \|e_{\ell,1}\| . \tag{2.14}$$

Pairing Line (2.14) with the triangle inequality gives us

$$\|z_c - z_a\| = (c-a)\|z_{a+1} - z_a\| + \delta_{c,a}, \tag{2.15}$$

where $\delta_{c,a} \leq \sum_{h=a+1}^{c-1} \sum_{\ell=a+1}^{h} \|e_{\ell,1}\|$. We can apply the same process to Lines (2.12) and (2.13) to get

$$\|z_b - z_a\| = (b-a)\|z_{a+1} - z_a\| + \delta_{b,a}, \tag{2.16}$$

$$\|z_c - z_b\| = (c-b)\|z_{a+1} - z_a\| + \delta_{c,b}, \tag{2.17}$$

where

$$|\delta_{b,a}| \leq \sum_{h=a+1}^{b-1} \sum_{\ell=a+1}^{h} \|e_{\ell,1}\|,$$
$$|\delta_{c,b}| \leq |c-b| \sum_{\ell=a+1}^{b} \|e_{\ell,1}\| + \sum_{h=b+1}^{c-1} \sum_{\ell=b+1}^{h} \|e_{\ell,1}\|.$$

Combining Lines (2.15), (2.16) and (2.17) gives us

$$\|z_c - z_a\| = \|z_c - z_b\| + \|z_b - z_a\| + \delta_{c,b,a},$$

where

$$\begin{aligned}|\delta_{c,b,a}| &\leq \sum_{h=a+1}^{c-1} \sum_{\ell=a+1}^{h} \|e_{\ell,1}\| + \sum_{h=a+1}^{b-1} \sum_{\ell=a+1}^{h} \|e_{\ell,1}\| + |c-b| \sum_{\ell=a+1}^{b} \|e_{\ell,1}\| + \sum_{h=b+1}^{c-1} \sum_{\ell=b+1}^{h} \|e_{\ell,1}\| \\ &= 2 \sum_{h=a+1}^{c-1} \sum_{\ell=a+1}^{h} \|e_{\ell,1}\| + \sum_{\ell=a+1}^{b} \|e_{\ell,1}\|.\end{aligned}$$

However, since each this is an $\varepsilon$-discrete, $\lambda$-dense curve, then each $\|e_{\ell,1}\| \leq \lambda\varepsilon$. Therefore we have that

$$\|\delta_{c,b,a}\| \leq [2(c-a)^2 + (b-a)^2]\lambda\varepsilon. \qquad \square$$

The above results shows that distances break down nearly linearly and we should expect the error to scale not only with the curvature of the discrete curve, but also with its density. We can see that the error also holds when a discrete curve is a perfect line, since when $\varepsilon = 0$, then this result says the error is 0, which is what we expect.

Now that we have this property about the distances in a discrete curve, we can move on to estimating the effect t-SNE's gradient has on discrete curves.

## 2.3 Gradient Upper Bounds of Dense, Discrete Curves

We can show that given a dense, discrete curve and a symmetric triplet in the curve, t-SNE's gradient will be nearly symmetric about the center of the triplet. To show this, we will have to assume that the neighborhood size $\rho$ stays sufficiently smaller than the size of the dataset $n$. We will also need to assume that in the visualization, any points that are not within $10\rho$ nearest neighbors will be sufficiently far away.

Since t-SNE updates the location of each point simultaneously, the iterations can be naturally written as a matrix equation. Let $Y \in \mathbb{R}^{n\times 2}$ be a matrix who rows are the points $(y_i)_{i=1}^n$. The joint probabilities $p_{ij}$ and $q_{ij}$ can be represented in $n \times n$ matrices where $P_{ij} = p_{ij}$ and $Q_{ij} = q_{ij}$. Using this notation, we can define an iterative matrix $M \in \mathbb{R}^{n\times n}$ where

$$M_{ij}^t = \begin{cases} 1 - h\sum_{k\neq i}(p_{ik} - q_{ik})(1 + \|y_{i,:}^t - y_{k,:}^t\|^2)^{-1} & i = j \\ h(p_{ij} - q_{ij})(1 + \|y_{i,:}^t - y_{j,:}^t\|^2)^{-1} & i \neq j \end{cases} \tag{2.18}$$

We can use $M$ to rewrite equation 1.7 to be the following

$$Y^{t+1} = M^t Y^t. \tag{2.19}$$

By writing the iteration updates in terms of the matrix $M$, we can study how t-SNE's visualizations change along with $M$. Below, we show how discrete curves impact the structure of $M$ and subsequently changes the output t-SNE creates.

**Lemma 2.4** (Gradient Symmetry)**.** *Let $(x_i)_{i=1}^n \subset \mathbb{R}^d$ be a $\lambda_x$-dense, $\varepsilon_x$-discrete curve and assume that for some $t > 0$, there is a visualization $(y_i^t)_{i=1}^n$ that is also a $\lambda_y$-dense, $\varepsilon_y$-discrete curve. Fixing $i$, assume $j \leq \rho$, then*

$$M_{i+j,k}^t - M_{i,k}^t = M_{i,k}^t - M_{i-j,k}^t + \Delta_{i,j,k}^t,$$

*such that* $|\Delta^t_{i,j,k}| \leq (j+\rho^2\varepsilon_x)^2\lambda_x^2+(j+\rho^2\varepsilon_y)^2\lambda_y^2+2\rho^3\lambda_y^2\varepsilon_y$ *for all* $k \in \mathcal{N}$ *where* $k \neq i-j, i, i+j$.

*Proof.* To simplify the notation in this proof, we adopt the following:

$$d^{(x)}_{f,g} = \|x_f - x_g\| \quad \text{and} \quad d^{(y)}_{f,g} = \|y^t_f - y^t_g\|.$$

Assume that $k < i - j$. For $v \in \{x, y\}$, since the neighborhoods of $x_i$ and $y^t_i$ are $\lambda_{(v)}$-dense, $\varepsilon_{(v)}$-discrete curves, we know there exists $\hat{\delta}^{(v)}_{i,j}$ such that

$$d^{(v)}_{i-j,i} = d^{(v)}_{i+j,i} + \hat{\delta}^{(v)}_{i,j}, \tag{2.20}$$

where $\hat{\delta}^{(v)}_{i,j} \leq j\lambda_{(v)}\varepsilon_{(v)}$. Applying Lemma 2.3 and line (2.20) gives

$$\begin{aligned} d^{(v)}_{i-j,k} &= d^{(v)}_{i,k} - d^{(v)}_{i-j,i} + \delta^{(v)}_{i-j,k,i}, \\ &= d^{(v)}_{i,k} - d^{(v)}_{i+j,i} + \tilde{\delta}^{(v)}_{i-j,k,i}, \end{aligned} \tag{2.21}$$

where $\delta^{(v)}_{i-j,k,i}$ is the error given in Lemma 2.3. Here $\tilde{\delta}^{(v)}_{i-j,k,i} \leq \delta^{(v)}_{i-j,k,i} + \hat{\delta}^{(v)}_{i,j}$ and thus $\tilde{\delta}^{(v)}_{i-j,k,i} \leq \lambda_{(v)}\varepsilon_{(v)}C_{i-j,k,i}$ where

$$C_{i-j,k,i} = [j + 2(i-k)^2 + (|i-j| - k)^2].$$

However, we know that since $i, k, i+j$ and $i-j$ are in the same neighborhood, then $C_{i-j,k,i} \leq 3\rho^2$. Thus $\tilde{\delta}^{(v)}_{i-j,k,i} \leq \rho^2\lambda_{(v)}\varepsilon_{(v)}$. Now we can use these distance equations to estimate the differences in gradient values generated by t-SNE.

Recall the iterative update generated by t-SNE's gradient in (2.18). Since we will use this approximation when t-SNE is not in early exaggeration, we fixed the exaggeration factor $\alpha = 1$ in (1.7). Our ultimate goal is to estimate how close the differences $M^t_{i+j,k} - M^t_{i,k}$

and $M^t_{i,k} - M^t_{i-j,k}$ are, and we will do so by taking a Taylor expansion of the function that represents the entries in t-SNE's gradient:

$$M(d^{(x)}, d^{(y)}) = h\left(\frac{\exp(-(d^{(x)})^2/2\sigma^2)}{nC_{P_i}} - \frac{(1+(d^{(y)})^2)^{-1}}{C_Q}\right)(1+(d^{(y)})^2)^{-1}.$$

Note that when $i \neq j$, then $M^t_{ij} = M(d^{(x)}_{ij}, d^{(y)}_{ij})$. Because $d^{(x)}_{ij}$ and $d^{(y)}_{ij}$ contribute such small amounts to $C_{P_i}$ and $C_Q$, we will treat them as constants in our expansion. Now we can use our distance equations from (2.21) and Lemma 2.3 to get that

$$\begin{aligned}
M^t_{i+j,k} &= M(d^{(x)}_{i+j,i} + d^{(x)}_{i,k} + \delta^{(x)}_{i+j,k,i}, d^{(y)}_{i+j,i} + d^{(y)}_{i,k} + \delta^{(y)}_{i+j,k,i}),\\
M^t_{i-j,k} &= M(d^{(x)}_{i,k} - d^{(x)}_{i+j,i} + \tilde{\delta}^{(x)}_{i-j,k,i}, d^{(y)}_{i,k} - d^{(y)}_{i+j,i} + \tilde{\delta}^{(y)}_{i-j,k,i}).
\end{aligned}$$

Now taking a first order Taylor expansion of $M$ about $(d^{(x)}_{i,k}, d^{(y)}_{i,k})$ and evaluating at $(d^{(x)}_{i+j,k}, d^{(y)}_{i+j,k})$ and $(d^{(x)}_{i-j,k}, d^{(y)}_{i-j,k})$:

$$M^t_{i+j,k} = M^t_{i,k} + (d^{(x)}_{i+j,i} + \delta^{(x)}_{i+j,k,i})M_x(d^{(x)}_{i,k}, d^{(y)}_{i,k}) + (d^{(y)}_{i+j,i} + \delta^{(y)}_{i+j,k,i})M_y(d^{(x)}_{i,k}, d^{(y)}_{i,k}) + \Delta_+, \tag{2.22}$$

$$M^t_{i-j,k} = M^t_{i,k} + (-d^{(x)}_{i+j,i} + \tilde{\delta}^{(x)}_{i-j,k,i})M_x(d^{(x)}_{i,k}, d^{(y)}_{i,k}) + (-d^{(y)}_{i+j,i} + \tilde{\delta}^{(y)}_{i-j,k,i})M_y(d^{(x)}_{i,k}, d^{(y)}_{i,k}) + \Delta_-. \tag{2.23}$$

Here $M_x$ and $M_y$ denote the partial derivatives of $M(\cdot,\cdot)$ with respect to $d^{(x)}_{\cdot,\cdot}$ and $d^{(y)}_{\cdot,\cdot}$ respectively. $\Delta_+$ and $\Delta_-$ denote the errors given by the expansion, which in this case will be bounded by $\left((d^{(x)}_{i+j,i} + \delta^{(x)}_{i+j,k,i})^2 + (d^{(y)}_{i+j,i} + \delta^{(y)}_{i+j,k,i})^2\right)$. If we substitute (2.23) into (2.22) and rearrange terms, we have that

$$\begin{aligned}
M^t_{i+j,k} - M^t_{i,k} &= M^t_{i,k} - M^t_{i-j,k}\\
&\quad + (\delta^{(x)}_{i+j,k} + \tilde{\delta}^{(x)}_{i-j,k})M_x(d^{(x)}_{i,k}, d^{(y)}_{i,k}) + (\delta^{(y)}_{i+j,k} + \tilde{\delta}^{(y)}_{i-j,k})M_y(d^{(x)}_{i,k}, d^{(y)}_{i,k})\\
&\quad + \Delta_+ + \Delta_-.
\end{aligned} \tag{2.24}$$

We will also need to estimate $M_x$ and $M_y$. A quick derivation gives us

$$M_x(d^{(x)}, d^{(y)}) = -2hd^{(x)} \exp(-(d^{(x)})^2)(1 + (d^{(y)})^2)^{-1},$$
$$M_y(d^{(x)}, d^{(y)}) = -2hd^{(y)}(1 + (d^{(y)})^2)^{-2} \exp(-(d^{(x)})^2) + 4hd^{(y)}(1 + (d^{(y)})^2)^{-3}.$$

Therefore we have that

$$M_x(d^{(x)}_{i,k}, d^{(y)}_{i,k}) \leq 0, \quad M_y(d^{(x)}_{i,k}, d^{(y)}_{i,k}) \leq d^{(y)}_{i,k}. \tag{2.25}$$

Recall that in (2.24), $(\delta^{(v)}_{i+j,k,i} + \tilde{\delta}^{(v)}_{i-j,k,i}) \leq 2\rho^2\lambda_{(v)}\varepsilon_{(v)}$ for $v \in \{x, y\}$. Using this and substituting in (2.25) into (2.24), we have that

$$(\delta^{(x)}_{i+j,k,i} + \tilde{\delta}^{(x)}_{i-j,k,i})M_x(d^{(x)}_{i,k}, d^{(y)}_{i,k}) + (\delta^{(y)}_{i+j,k,i} + \tilde{\delta}^{(y)}_{i-j,k,i})M_y(d^{(x)}_{i,k}, d^{(y)}_{i,k}) \leq 2\rho^3\lambda^2_y\varepsilon_y. \tag{2.26}$$

Finally, we can find an upper bound for $\Delta_- + \Delta_+$. From the Taylor expansion we know $\Delta_- + \Delta_+ \leq (d^{(x)}_{i+j,i} + \delta^{(x)}_{i,k})^2 + (d^{(y)}_{i+j,i} + \delta^{(y)}_{i,k})^2$. Since both $X$ and $Y$ are $\lambda_{(v)}$-dense curves, then each pairwise step is at most $\lambda_{(v)}$, making both $d^{(\cdot)}_{i+j,i}$ and $d^{(\cdot)}_{i-j,i}$ at most $j\lambda_{(v)}$. From Lemma 2.3, we know $\delta^{(\cdot)}_{i,k} \leq \rho^2\lambda_{(v)}\varepsilon_{(v)}$. Thus we have that $\Delta_+ + \Delta_- \leq \sum_{v\in x,y} \lambda^2_{(v)}(j + \rho^2\varepsilon_{(v)})^2$. Putting both of these values into (2.24) implies that

$$M^t_{i+j,k} - M^t_{i,k} = M^t_{i,k} - M^t_{i-j,k} + \Delta^t_{i,j,k}, \tag{2.27}$$

where $|\Delta^t_{i,j,k}| \leq (j + \rho^2\varepsilon_x)^2\lambda^2_x + (j + \rho^2\varepsilon_y)^2\lambda^2_y + 2\rho^3\lambda^2_y\varepsilon_y$. If instead $i - j < k < i$ then using our dense discrete curve requirement and Lemma 2.3, (2.21) will instead become,

$$d^{(v)}_{i-j,k} = d^{(v)}_{i-j,i} - d^{(v)}_{i,k} + \delta^{(v)}_{i-j,k,i},$$

and $d^{(v)}_{i+j,k}$ will be

$$d^{(v)}_{i+j,k} = d^{(v)}_{i-j,i} + d^{(v)}_{i,k} + \delta^{(v)}_{i+j,k,i} + \delta^{(v)}_{i+j,i-j,i}.$$

Then we can instead take the Taylor Expansion about $d^{(v)}_{i-j,k}$ and get the same order of approximation since the pairwise distances in the discrete curve are bounded by $\lambda_{(v)}$. □

This lemma essentially shows that for any symmetric triplet in a neighborhood, any other point in the neighborhood will have the same impact on the gradient value. This result will help show the effects the structure of a discrete curve has on t-SNE's gradient.

In addition to estimating the symmetry between gradient entries, it will be helpful to find some basic upper bounds on the gradient entries depending on the location of different datapoints. Below we estimate what the gradient will be depending on if a point is inside or outside a neighborhood.

**Lemma 2.5** (Gradient Upper Bounds)**.** *Assume that for $(x_i)_{i=1}^n \subset \mathbb{R}^d$ is a $\lambda_x$-dense, $\varepsilon_x$-discrete curve and its visualization $(y_i)_{i=1}^n$ is a $\lambda_y$-dense, $\varepsilon_y$-discrete curve where $\lambda_x, \lambda_y < 1/n$. If $2\sigma_i^2 = \rho^2/n^2$, then the following bounds hold.*

- *When $|i-j| \leq 10\rho$, then $|M^t_{i,j}| \leq \frac{4he}{\rho n}$.*
- *When $|i-j| > 10\rho$ and $\|x_i - x_j\|^2 \geq \rho^2 \log(n)/n$, $\|y_i^t - y_j^t\|^2 \geq \sqrt{16h/\varepsilon_y}$, then $|M^t_{i,j}| \leq \frac{\varepsilon_y}{4n^2}$*

*Proof.* As a reminder, $p_{ij}$ and $q_{ij}^t$ values are the following.

$$
\begin{aligned}
p_{j|i} &= \frac{\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/(2\sigma_i^2))} \\
p_{ij} &= \frac{p_{i|j} + p_{j|i}}{2n} \\
q_{ij}^t &= \frac{(1+\|y_i^t - y_j^t\|^2)^{-1}}{\sum_{k\neq \ell}(1+\|y_k^t - y_\ell^t\|^2)^{-1}}
\end{aligned}
$$

We can find an upper bound for $p_{j|i}$ by putting a lower bound on the normalization term.

$$
\begin{aligned}
p_{j|i} &= \frac{\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))}{\sum_{k\neq i}\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))} \\
&\leq \frac{\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))}{\rho\exp(-(\rho^2/n^2)/(2\sigma_i^2))} && (2.28) \\
&\leq \frac{\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))}{\rho/e} && (2.29)
\end{aligned}
$$

(2.28) follows from the assumption that the normalization term will contain at least the $\rho$ nearest neighbors and will be its smallest value when the distance is largest, which in this case is $\rho^2\lambda_x^2 = \rho^2/n^2$. (2.29) follows from the assumption that $2\sigma_i^2 = \rho^2/n^2$. A lower bound for $p_{i|j}$ can be found by upper bounding the normalization term.

$$
p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))}{\sum_{k\neq i}\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))} \geq \frac{\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))}{n}
$$

The upper bound on the normalization term comes from the fact that $\exp(-\|x_i - x_j\|^2/(2\sigma_i^2)) \leq 1$. Putting these bounds with the definition of $p_{ij}$ we have that

$$
\begin{aligned}
p_{ij} &\leq \frac{\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))}{n\rho/e}, && (2.30) \\
p_{ij} &\geq \frac{\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))}{n^2}. && (2.31)
\end{aligned}
$$

We can also find upper and lower bounds for $q_{ij}$. Since $(y_i)_{i=1}^n$ is a $\lambda_y$-dense, $\varepsilon_y$-discrete curve where $\lambda_y < 1/n$, then the maximum distance is at most 1 and the minimum pairwise

distance is 0. Thus we have

$$q_{ij} \leq \frac{(1+\|y_i^t - y_j^t\|^2)^{-1}}{\sum_{k\neq\ell}(1+\|y_k^t - y_\ell^t\|^2)^{-1}} \leq \frac{2(1+\|y_i^t - y_j^t\|^2)^{-1}}{n^2}, \tag{2.32}$$

$$q_{ij} \geq \frac{(1+\|y_i^t - y_j^t\|^2)^{-1}}{\sum_{k\neq\ell}(1+\|y_k^t - y_\ell^t\|^2)^{-1}} \geq \frac{(1+\|y_i^t - y_j^t\|^{-1})}{n^2}. \tag{2.33}$$

Let $|i-j| \leq 10\rho$ and assume that $p_{ij} \geq q_{ij}$. Then we have that

$$\begin{aligned} |p_{ij} - q_{ij}| &\leq |p_{ij}|\,, \\ &\leq \frac{\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))}{n\rho/e}, \\ &\leq \frac{e}{\rho n}. \end{aligned} \tag{2.34}$$

If we instead assume that $p_{ij} \leq q_{ij}$, then we have that

$$\begin{aligned} |p_{ij} - q_{ij}| &\leq \left|\frac{\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))}{n^2} - \frac{2(1+\|y_i^t - y_j^t\|^2)^{-1}}{n^2}\right|, \\ &\leq \frac{1}{n^2}|\exp(-1/\rho^2) - 2|, && (2.35) \\ &\leq \frac{1}{n^2} && (2.36) \end{aligned}$$

where (2.35) follows from the assumption that $\|x_i - x_j\|^2 \leq 1$ and $\|y_i^t - y_j^t\|^2 \geq 0$. Looking at both bounds in (2.34) and (2.36) we see that when $|i-j| \leq 10\rho$, $|p-q| \leq \frac{e}{\rho n}$.

Now consider when $|i-j| > 10\rho$. From our assumptions that means $\|x_i - x_j\|^2 \geq \frac{\rho^2 \log(n)}{n^2}$. If $p_{ij} \geq q_{ij}$, then

$$\begin{aligned} |p_{ij} - q_{ij}| &\leq \left|\frac{\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))}{n\rho/e} - \frac{(1+\|y_i^t - y_j^t\|^2)^{-1}}{n^2}\right|, \\ &\leq \frac{1}{n}\left|\frac{1/n}{\rho/e} - \frac{1}{2n}\right|, && (2.37) \\ &\leq \frac{e}{\rho n^2}, && (2.38) \end{aligned}$$

where (2.37) follows from the assumption that $\|x_i - x_j\|^2 \geq \frac{\rho^2 \log(n)}{n^2}$ and that $2\sigma_i^2 = \rho^2/n^2$. If instead $p_{ij} \leq q_{ij}$, then

$$\begin{aligned}
|p_{ij} - q_{ij}| &\leq \left| \frac{\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))}{n^2} - \frac{2(1 + \|y_i^t - y_j^t\|^2)^{-1}}{n^2} \right|, \\
&\leq \frac{1}{n^2} |0 - (1 + \|y_i^t - y_j^t\|^2)^{-1}|, \qquad (2.39) \\
&\leq \frac{\sqrt{\varepsilon_y}}{\sqrt{16h}n^2}, \qquad (2.40)
\end{aligned}$$

where (2.39) follows as $\exp(-x^2) \geq 0$ and (2.40) follows from the assumption that $\|y_i^t - y_j^t\|^2 \geq \sqrt{16h/\varepsilon_y}$. Thus at worst $|p_{ij} - q_{ij}|$ is $\frac{\sqrt{\varepsilon_y}}{\sqrt{16h}n^2}$ when $|i - j| > 10\rho$.

We can use these bounds to now upper bound the actual gradient values. For $|i - j| \leq 10\rho$, we can find an upper bound on the gradient value $M_{ij}^t$ given by (2.18).

$$\begin{aligned}
M_{ij}^t &= 4h|p_{ij} - q_{ij}|(1 + \|y_i^t - y_j^t\|^2)^{-1} \\
&\leq \frac{4he}{\rho n}(1 + \|y_i^t - y_j^t\|^2)^{-1} \\
&\leq \frac{4he}{\rho n} \qquad (2.41)
\end{aligned}$$

(2.41) follows as $\|y_i^t - y_j^t\|^2 \geq 0$. If instead $|i - j| \geq 10\rho$, we have that

$$\begin{aligned}
M_{ij}^t &= 4h|p_{ij} - q_{ij}|(1 + \|y_i^t - y_j^t\|^2)^{-1}, \\
&\leq \frac{4h\sqrt{\varepsilon_y}}{\sqrt{16h}n^2}(1 + \|y_i^t - y_j^t\|^2)^{-1}, \\
&\leq \frac{\varepsilon_y}{4n^2}, \qquad (2.42)
\end{aligned}$$

where (2.42) follows from the assumption that $\|y_i^t - y_j^t\|^2 \geq \sqrt{16h/\varepsilon_y}$. □

These lower bound requirements on the distances between points not in the same neighborhood correspond to our assumption that our discrete curves shouldn't have $a, b$-critical

overlaps. We are excluding curves that overlap themselves or have super tight coils.

# Chapter 3

# Early Exaggeration Promotes Discrete Curve Formation

Many non-convex optimization algorithms have a specific initialization that better optimize their solution. Similarly, t-SNE employs a period of *early exaggeration* that makes t-SNE's algorithm more advantageous. During this period t-SNE artificially enlarges the $p_{ij}$ values to promote attraction between similar points. In the previous section this was notated by $\alpha p_{ij}$ in the cost function, where in practice $\alpha = 4$ and is changed to $\alpha = 1$ after the first 200 iterations.

There has been some investigation into this part of the algorithm. Linderman and Steinerberger were able to show that given clusterable high-dimensional data, then t-SNE's early exaggeration phase promotes cluster formation when using appropriate choices for the step size $h$ and early exaggeration factor $\alpha$. [26]. This occurs because t-SNE initializes the visualization $Y$ in a very tiny square which forces the $q_{ij}$ values to be uniformly small. Pairing this with the larger $\alpha p_{ij}$ values, the gradient matrix $M$ develops a particular structure such that points in high dimensions move closer in the low dimensional representation.

Here we employ a similar analysis for early exaggeration. We consider high-dimensional data $X = (x_i)_{i=1}^n$ that is a $\lambda_x$-dense, $\varepsilon_x$-discrete curve and show that this property induces a certain structure on the $p_{ij}$ values. We then pair this with the early exaggeration factor $\alpha$ and show that the local solution for these neighborhoods are discrete curves and t-SNE visualizes them as such.

## 3.1 Induced Toeplitz Matrix

The high-dimensional dataset is a $\lambda$-dense, $\varepsilon$-discrete curve and will induce a particular structure in the $p_{ij}$ values. Lemma 3.2 shows that $p_{j|i} \approx \frac{\exp(-|i-j|^2\lambda^2/2\sigma_i^2)}{C_{P_i}}$. From here we can define a new variable $a$ to be $a = \exp(-\lambda^2/2\sigma_i^2)$ and thus $p_{j|i} \approx a^{|i-j|^2}$. Since we expect the neighborhoods of lines to have the same relative density of points we can assume that the induced $\sigma_i$ to be the same for a single neighborhood. This implies that $p_{j|i} \sim p_{i|j}$ for simplicity, we will denote $p_{ij} \sim p_{j|i}$. Using our new notation we have that the $n \times n$ matrix $P$ consisting of all $p_{ij}$ values can be written as

$$P = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1n-1} & p_{1n} \\ p_{21} & p_{22} & \dots & p_{2n-1} & p_{2n} \\ \vdots & \vdots & \dots & \vdots & \vdots \\ p_{n1} & p_{n2} & \dots & p_{nn-1} & p_{nn} \end{bmatrix} = \begin{bmatrix} 1 & a^2 & a^4 & \dots & a^{(n-1)^2} \\ a^2 & 1 & a^2 & \dots & a^{(n-2)^2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ a^{(n-1)^2} & a^{(n-2)^2} & a^{(n-3)^2} & \dots & 1 \end{bmatrix}.$$

Writing $P$ where the entries are powers of $a$ reveals a familiar structure. This is a Toeplitz matrix where the entry $a_k = \exp(-k^2\lambda^2/2\sigma^2)$. It is well know that $P$ is generated by the function $f(x)$:

$$f(x) = \sum_{k=-\infty}^{k=\infty} \exp(-k^2/a)\exp(ikx).$$

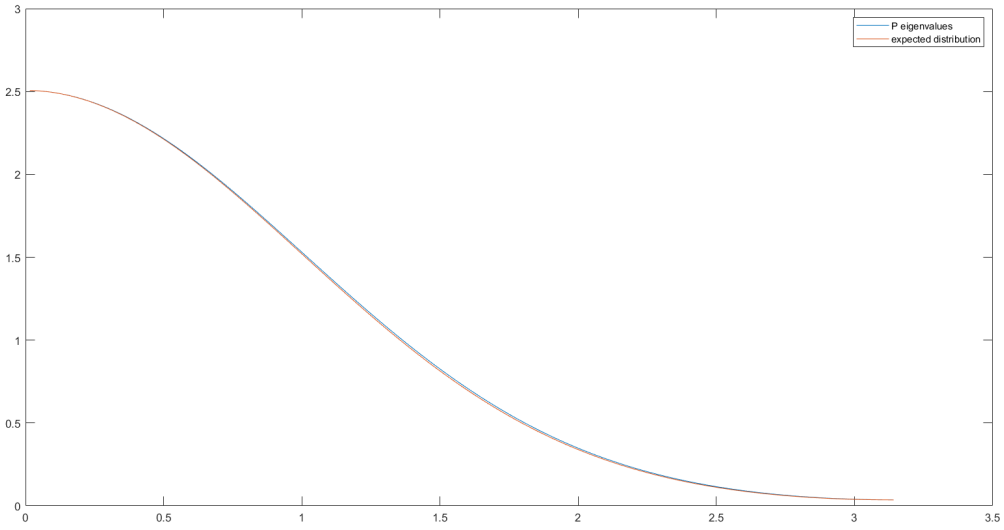


Figure 3.1: The eigenvalues of the $P$ matrix created by t-SNE compared with the eigenvalues expected from a Toeplitz matrix with appropriate $a$ values.

However this is exactly one of the Jacobi theta function so the solution to this equation is

$$f(x) = 1 + 2\sum_{\ell=1}^{\infty} \exp(-\ell^2/\sigma)\cos(\ell x) = \vartheta_3(x/2, \exp(-1/a)).$$

The Jacobi theta functions are well studied and a comprehensive review can be found here [5]. From the Szego Limit Theorem, in the limit the distribution of the eigenvalues look like the distribution of $f(x)$. For our particular $f(x)$ this implies that our $k$th eigenvalue $\lambda_k^{(n)}$ will equal $f(k\pi/n)$, or

$$f(k\pi/n) = 1 + 2\sum_{\ell=1}^{\infty} \exp(-\ell^2/\sigma^2)\cos(\ell k\pi/n).$$

In practice we can check that this is true. Using a dataset of 200 points on a dense line, we compared the eigenvalues of the $P$ matrix created by t-SNE and compared them to the eigenvalues of a Toeplitz matrix with the structure above with the comparable $a$ value. In Figure 3.1 we can see that there is perfect overlap, confirming our analysis. Now we can show rigorously that this Toeplitz matrix is very close to a matrix populated by the true $p_{ij}$

values.

**Lemma 3.1.** *Let $X = (x_i)_{i=1}^n$ be in $\mathbb{R}^d$ be a $\lambda_x$-dense, $\varepsilon_x$-discrete curve. Assume that for a fixed $i$, if $|i-j| \leq 5\rho$, then $\left| \|x_i - x_j\|^2 - |i-j|^2\lambda_x^2 \right| \leq \frac{\lambda_x}{\sqrt{n}}$. Then the following bound holds:*

$$\left| \exp(-\|x_i - x_j\|^2/2\sigma_i^2) - \exp(-|i-j|^2\lambda_x^2/2\sigma_i^2) \right| \leq \hat{\delta}_{ij}, \tag{3.1}$$

*where*

$$\hat{\delta}_{ij} \leq \begin{cases} 625(\varepsilon_x+1)^2\rho^4\lambda_x^4 & |i-j| \leq 5\rho \\ \frac{\lambda_x^2}{n} & |i-j| > 5\rho \end{cases}. \tag{3.2}$$

*Proof.* If $|i-j| \leq 5\rho$, then Lemma 2.3 tells us that

$$\|x_i - x_j\| = |i-j|\|x_i - x_{i+1}\| + \delta_{ij}, \tag{3.3}$$

where $|\delta_{ij}| \leq (i-j)^2\lambda_x\varepsilon_x$. Squaring (3.3) implies that

$$\|x_i - x_j\|^2 = |i-j|^2\|x_i - x_{i+1}\|^2 + \bar{\delta}_{ij}, \tag{3.4}$$

where $|\bar{\delta_{ij}}| \leq (i-j)^2\lambda_x^2\varepsilon_x + (i-j)^4\lambda_x^2\varepsilon_x^2$. However, since $|i-j| \leq 5\rho$ and $\lambda_x, \varepsilon_x < 1$ and $\|x_i - x_{i+1}\|^2 \leq \lambda_x^2$, then

$$\left| \|x_i - x_j\|^2 - |i-j|^2\lambda_x^2 \right| \leq \rho^2\lambda_x^2 + 25\rho^2\lambda_x^2\varepsilon_x \tag{3.5}$$

$$\leq 25(\varepsilon_x+1)\rho^2\lambda_x^2 \tag{3.6}$$

Now to find the difference in (3.1), we take a second order Taylor series about $\exp(-x/2\sigma_i^2)$

centered at $x = |i-j|^2\|x_i - x_{i+1}\|^2$. Evaluating the series at $\|x_i - x_j\|^2/2\sigma_i^2$ gives us,

$$\exp(-\|x_i-x_j\|^2/2\sigma_i^2) = \exp(-|i-j|^2\|x_i-x_{i+1}\|^2/2\sigma_i^2)+(\tilde{\delta}_{ij})\frac{\partial}{\partial_x}(\exp(-|i-j|^2\|x_i-x_{i+1}\|^2/2\sigma_i^2))+E_{ij} \tag{3.7}$$

where $|E_{ij}| \leq \bar{\delta}_{ij}^2 = 625(\varepsilon_x+1)^2\rho^4\lambda_x^4$. Given the fact the first derivative in (3.7) is negative, then we have that

$$\exp(-\|x_i - x_j\|^2/2\sigma_i^2) - \exp(-|i-j|^2\|x_i - x_{i+1}\|^2/2\sigma_i^2) \leq 625(\varepsilon_x+1)^2\rho^4\lambda_x^4. \tag{3.8}$$

If we instead consider $|i-j| > 5\rho$, we can use the assumption that $\left|\|x_i-x_j\|^2-|i-j|^2\lambda_x^2\right| \leq \frac{\lambda_x}{\sqrt{n}}$ to do a similar Taylor expansion to get

$$\exp(-\|x_i - x_j\|^2/2\sigma_i^2) - \exp(-|i-j|^2\|x_i - x_{i+1}\|^2/2\sigma_i^2) \leq \lambda_x^2/n. \tag{3.9}$$

□

We use the estimation in Lemma 3.1 to prove that the matrix $P$ population by $p_{j|i}$ values is close to the Toeplitz matrix descriped above.

**Lemma 3.2.** *Let $X = (x_i)_{i=1}^n$ be in $\mathbb{R}^d$ be a $\lambda_x$-dense, $\varepsilon_x$-discrete curve. Assume that for a fixed $i$, if $|i-j| > 5\rho$, then $\left|\|x_i - x_j\|^2 - |i-j|^2\lambda_x^2\right| \leq \frac{\lambda_x}{\sqrt{n}}$. If $P$ is the $n \times n$ matrix populated by $p_{j|i}$ values defined by t-SNE in (1.1) and $T_P$ is the $n \times n$ Toeplitz matrix such that $T_{P_{ij}} = \exp(-|i-j|^2\lambda_x^2/2\sigma_i^2)$, then*

$$\|P - T_P\|_F^2 \leq \frac{1}{n}, \tag{3.10}$$

*if $\lambda_x < 1/n$ and $\rho$ is constant.*

*Proof.* Using the difference estimation in Lemma 3.1 we can find the upper bound.

$$\|P - T_P\|_F^2 = \sum_i \sum_{j \neq i} |p_{j|i} - T_{P_{ij}}|^2 \tag{3.11}$$

$$= \sum_i \left( \sum_{j, |i-j| \leq 5\rho} |p_{j|i} - T_{P_{ij}}|^2 + \sum_{j, |i-j| > 5\rho} |p_{j|i} - T_{P_{ij}}|^2 \right) \tag{3.12}$$

$$\leq \sum_i \left( 5\rho(625(\varepsilon_x + 1)^2 \rho^4 \lambda_x^4) + (n - 5\rho)\frac{\lambda_x^2}{n} \right) \tag{3.13}$$

$$\leq 3125n(\varepsilon_x + 1)^2 \rho^5 \varepsilon_x^4 + n(n - 5\rho)\frac{\lambda_x^2}{n} \tag{3.14}$$

$$\leq \frac{\rho^5}{n^3} + \frac{1}{n}. \tag{3.15}$$

□

## 3.2 Early Exaggeration Structure

Using the definitions of $P$ and $M$ we can quantify the behavior of t-SNE during the early exaggeration phase. Assume that $X$ has satisfied the linear conditions from above. Then the updating iterative matrix $M^t$ from (1.7) can be written as $M^t = \hat{P} + \mathcal{E}_P^t + \mathcal{E}^t$ where

$$\hat{P}_{ij} = \begin{cases} 1 - \sum_{k \neq i} 4h\alpha p_{ik} & i = j \\ 4h\alpha p_{ij} & i \neq j \end{cases} \tag{3.16}$$

$$\mathcal{E}_{p_{ij}}^t = \begin{cases} \sum_{k \neq i} 4\alpha h p_{ik}(1 - (1 + ||y_i^t - y_k^t||^2)^{-1}) & i = j \\ 4h\alpha p_{ij}((1 + ||y_i^t - y_j^t||^2)^{-1} - 1) & i \neq j \end{cases} \tag{3.17}$$

$$\mathcal{E}_{ij}^t = \begin{cases} \sum_{k \neq i} 4hq_{ik}(1 + ||y_i^t - y_k^t||^2)^{-1} & i = j \\ -4hq_{ij}(1 + ||y_i^t - y_j^t||^2)^{-1} & i \neq j \end{cases}. \tag{3.18}$$

We can show that given the assumptions from the early exaggeration phase, $\hat{\mathcal{E}}^t = \mathcal{E}_P^t + \mathcal{E}^t$ adds only a small error each iteration.

**Lemma 3.3.** *Let $\mu > 0$ and let $P$ be an $n \times n$ matrix populated by the $p_{ij}$ values defined by t-SNE in Equation 1.3. Assume that for a particular iteration $t$, $(y_i^t)_{i=1}^n \subset \mathbb{R}^2$ such that $(1 + ||y_i^t - y_j^t||^2)^{-1} \geq 1 - \mu$. Then*

$$||\mathcal{E}_P^t||_F^2 + ||\mathcal{E}^t||_F^2 \leq \frac{32h^2}{n(1-\mu)^2} + 32\alpha^2 h^2 \mu^2.$$

*Proof.* Let us first find a bound for $\|\mathcal{E}_P^t\|_F^2$. For every $i \neq j$, we have that

$$\begin{aligned} |\mathcal{E}_{p_{ij}}^t|^2 &= |4\alpha h p_{ij}(1 - (1 + \|y_i^t - y_j^t\|^2)^{-1})|^2 \\ &\leq 16\alpha^2 h^2 p_{ij}^2 \mu^2 \qquad (3.19) \\ &\leq 16\alpha^2 h^2 p_{ij}^2 \mu^2. \qquad (3.20) \end{aligned}$$

Line 3.19 follows from the assumption $(1 + \|y_i^t - y_j^t\|^2)^{-1} \geq 1 - \mu$, which implies that $1 - (1 + \|y_i^t - y_j^t\|^2)^{-1} \leq \mu$. Line 3.20 follows from the fact that all $p_{ij}$ values are joint probabilities and are less or equal to than 1. When we have that $i = j$, the bound comes

out to be

$$
\begin{aligned}
|\mathcal{E}^t_{p_{ii}}|^2 &= \left| \sum_{k \neq i} 4\alpha h p_{ik} (1 - (1 + \|y^t_i - y^t_k\|^2)^{-1}) \right|^2 \\
&\leq \left( \sum_{k \neq i} (4\alpha h p_{ik})^2 \right) \left( \sum_{k \neq i} (1 - (1 + \|y^t_i - y^t_k\|^2)^{-1})^2 \right) && (3.21) \\
&= 16\alpha^2 h^2 \left( \sum_{k \neq i} p^2_{ik} \right) \left( \sum_{k \neq i} (1 - (1 + \|y^t_i - y^t_k\|^2)^{-1})^2 \right) \\
&\leq 16\alpha^2 h^2 n \mu^2 \sum_{k \neq i} p^2_{ik} && (3.22) \\
&\leq \frac{16\alpha^2 h^2 \mu^2}{n}. && (3.23)
\end{aligned}
$$

The inequality in Line 3.21 follows from Cauchy-Schwarz and the inequality in Line 3.22 follows from the assumption that $(1 + \|y^t_i - y^t_k\|^2)^{-1} \geq 1 - \mu$. For the inequality in Line 3.23, notice that since $\sum_{k \neq i} p_{ik} = \frac{1}{n}$, then $\left( \sum_{k \neq i} p_{ik} \right)^2 = \frac{1}{4n^2}$, so $\sum_{k \neq i} p^2_{ik} \leq \frac{1}{n^2}$ since every $p_{ik} \geq 0$. Using the inequalities in Lines 3.20 and 3.23, we have that

$$
\begin{aligned}
\|\mathcal{E}^t_P\|^2_F &= \sum_i \sum_{j \neq i} |\mathcal{E}^t_{p_{ij}}|^2 + \sum_i |\mathcal{E}^t_{p_{ii}}|^2 \\
&\leq 16\alpha^2 h^2 \mu^2 \sum_i \sum_{j \neq i} p^2_{ij} + \sum_i \frac{16\alpha^2 h^2 \mu^2}{n} \\
&= 32\alpha^2 h^2 \mu^2. && (3.24)
\end{aligned}
$$

Line 3.24 follows from the fact that $p_{ij}$ values are probabilities so $\sum_i \sum_{j \neq i} p_{ij} = 1$, implying that $\sum_i \sum_{j \neq i} p^2_{ij} \leq 1$ since each $p_{ij} \geq 0$.

Now let us find a bound for $\|\mathcal{E}^t\|_F^2$. For $i \neq j$ we have that

$$\begin{aligned}
|\mathcal{E}_{ij}^t|^2 &= |4hq_{ij}^t(1+\|y_i^t - y_j^t\|^2)^{-1}|^2 \\
&\leq 16h^2 q_{ij}^2 && (3.25)\\
&\leq \frac{16h^2}{n^4(1-\mu)^2} && (3.26)
\end{aligned}$$

Line 3.25 follows from the fact that $(1+\|y_i^t - y_j^t\|^2)^{-1} \leq 1$ for all $i \neq j$. Line 3.26 comes from the assumption that $(1+\|y_i^t - y_j^t\|^2)^{-1} \geq 1-\mu$. From the definition of $q_{ij}$, we know that

$$\begin{aligned}
q_{ij} &= \frac{(1+\|y_i^t - y_j^t\|^2)^{-1}}{\sum_{k\neq\ell}(1+\|y_k^t - y_\ell^t\|^2)^{-1}} \\
&\leq \frac{(1+\|y_i^t - y_j^t\|^2)^{-1}}{n^2(1-\mu)} && (3.27)\\
&\leq \frac{1}{n^2(1-\mu)} && (3.28)
\end{aligned}$$

The inequality in Line 3.27 follows from the assumption that $(1+\|y_i^t - y_j^t\|^2)^{-1} \geq 1-\mu$ and the inequality in Line 3.28 comes from the fact that $(1+\|y_i^t - y_j^t\|^2)^{-1} \leq 1$. We can plug the inequality from Line 3.28 into Line 3.25 to get the inequality in Line 3.26.

Now we can consider a bound on $|\mathcal{E}_{ii}^t|$. Similarly to above, we get that

$$\begin{aligned}
|\mathcal{E}_{ii}^t|^2 &= \left|\sum_{k\neq i} 4hq_{ik}(1+\|y_i^t - y_k^t\|)^{-1}\right|^2 \\
&\leq \left(\sum_{k\neq i} 16h^2 q_{ik}^2\right)\left(\sum_{k\neq i}(1+\|y_i^t - y_k^t\|^2)^{-2}\right) \\
&\leq 16h^2 n \sum_{k\neq i} q_{ik}^2 \\
&\leq \frac{16h^2}{n^2(1-\mu)^2}.
\end{aligned}$$

Again we use Cauchy-Schwarz and the bound on $q_{ik}$ from Line 3.26 to get this upper bound. Putting this together we have that

$$\begin{aligned}
\|\mathcal{E}^t\|_F^2 &= \sum_i \sum_{j \neq i} |\mathcal{E}_{ij}^t|^2 + \sum_i |\mathcal{E}_{ii}^t|^2 \\
&\leq \sum_i \sum_{j \neq i} \frac{16h^2}{n^4(1-\mu)^2} + \sum_i \frac{16h^2}{n^2(1-\mu)^2} \\
&\leq \frac{16h^2}{n^2(1-\mu)^2} + \frac{16h^2}{n(1-\mu)^2} \\
&\leq \frac{32h^2}{n(1-\mu)^2}
\end{aligned} \tag{3.29}$$

Adding the bounds from Lines 3.24 and 3.29 gives the final result.

□

Lemma 3.3 shows that the error induced by the matrices $\mathcal{E}_P^t, \mathcal{E}^t$ is dependent on the parameters $h, \alpha$ and $\mu$. So, given the particular parameters set by early exaggeration, we get the following theorem.

**Theorem 3.1.** *Let $X = (x_i)_{i=1}^n \subset \mathbb{R}^d$ be a $\lambda_x$-dense, $\varepsilon_x$-discrete curve and let $Y^t = \{y_i^t\}_{i=1}^n \subset \mathbb{R}^2$ be the the $t$-th iteration of the visualization of $X$ that t-SNE generates. Assume that t-SNE uses early exaggeration and $h$ is constant and $\alpha = 4$, and $(1 + \|y_i^t - y_j^t\|^2)^{-1} \geq 1 - (1/\sqrt{n})$. Then*

$$\|M^t\|_F^2 \leq \|\hat{P}\|_F^2 + \delta$$

*where $\delta \leq 1/n$.*

*Proof.* From our assumptions about $\alpha$ and $h$ the solution follows directly from Lemma 3.3.

□

Since in early exaggeration the gradient matrix $M$ is very close to the matrix $P$, and $P$ is very nearly a Toeplitz matrix, then t-SNE is essentially a spectral method and will locally preserve the discrete curve structure. However, since the $p_{ij}$ values are essentially 0 anytime $j \neq i$, then we can only expect locally for the solution to be ordered correctly, it does not necessarily prevent overlaps from happening. However, in the next section we are able to show that if there are no such overlaps, t-SNE will continue to preserve this structure even after early exaggeration.

# Chapter 4

# t-SNE Preserves Discrete Curves After Early Exaggeration

Now that we have shown that the early exaggeration phase of t-SNE generates discrete curves in the visualizations, it's important to show that it will preserve these structures beyond early exaggeration. Lemma 4.1 shows that the pairwise distance between points in a neighborhood in a given iteration can be bounded above by the same pairwise distance in the next iteration. Corollary 4.1 shows that density is preserved between iterations. Finally, we show in Lemma 4.2 that if points are a discrete curve in a given iteration, t-SNE preserves this property in the next iteration.

## 4.1 t-SNE Iterates Preserve Discrete Curves

Now that we have estimates for both distances and gradients from dense discrete curves, we can show that t-SNE will preserve the dense, discrete curve structure in its visualizations. We begin by finding a relationship between pairwise distances in sequential iterations.

**Lemma 4.1** (Comparing Distances Between Iterations)**.** *Assume we have high-dimensional dataset $(x_i)_{i=1}^n$ that is a $\lambda_x$-dense $\varepsilon_x$-discrete curve and that for some $t > 0$, the visualization $(y_i^t)_{i=1}^n$ of $(x_i)_{i=1}^n$ is also a $\lambda_y$-dense $\varepsilon_y$-discrete curve. If the following hold:*

- *The perplexity $\rho$, and the step size of t-SNE's gradient, $h$, satisfy $\frac{10he}{n} + 10(h+1)\rho^3(\lambda_x^2 + \lambda_y^2) < \frac{\varepsilon_y}{2}$,*
- *For $k$ such that $|i-k| > 10\rho$, we have $\|y_k^t - y_i^t\| > \sqrt[4]{16h/\varepsilon_y}$,*

*then for iteration $t+1$, for a fixed $i$ and $j \leq \rho$ we have that*

$$\|y_{i+j}^t - y_i^t\| \leq \frac{1}{1-\varepsilon_y}\|y_{i+j}^{t+1} - y_i^{t+1}\|.$$

*Proof.* By definition, we have that

$$y_{i+j}^{t+1} = y_{i+j}^t - \sum_{k \neq i+j} M_{i+j,k}^t (y_{i+j}^t - y_k^t),$$
$$y_i^{t+1} = y_i^t - \sum_{k \neq i} M_{i,k}^t (y_i^t - y_k^t).$$

Rearranging and combining these lines gives us

$$y_{i+j}^t - y_i^t = \underbrace{(y_{i+j}^{t+1} - y_i^{t+1})}_{A} - \underbrace{\left( \sum_{|i-k| \leq 10\rho, k \neq i+j} M_{i+j,k}^t (y_{i+j}^t - y_k^t) - \sum_{|i-k| \leq 10\rho, k \neq i} M_{i,k}^t (y_i^t - y_k^t) \right)}_{B}$$
$$- \underbrace{\left( \sum_{|i-k| > 10\rho} M_{i+j,k}^t (y_{i+j}^t - y_k^t) - M_{i,k}^t (y_i^t - y_k^t) \right)}_{C}.$$

We can combine and simplify the two summations in $B$ by using Lemma 2.4. Since these are $\lambda_x$,$\lambda_y$-dense $\varepsilon_x$,$\varepsilon_y$-discrete curves, know that for any $k$ where $|i-k| \leq 10\rho$, $M_{i+j,k}^t =$

$M^t_{i,k} + \delta_{i+j,i}$ such that $|\delta_{i+j,i}| < \rho^2(\lambda_x^2 + \lambda_y^2)$. Therefore we have two summations become the following:

$$\begin{aligned} B &= 2M^t_{i+j,i}(y^t_{i+j} - y^t_i) + \sum_{|i-k|\leq 10\rho, k\neq i+j,i} M^t_{i+j,k}(y^t_{i+j} - y^t_k) - M^t_{i,k}(y^t_i - y^t_k), \\ &= 2M^t_{i+j,i}(y^t_{i+j} - y^t_i) + \sum_{|i-k|\leq 10\rho, k\neq i+j,i} M^t_{i+j,k}(y^t_{i+j} - y^t_i) + \delta_{i+j,i}(y^t_i - y^t_k), \\ &= \left(2M^t_{i+j,i} + \sum_{|i-k|\leq 10\rho, k\neq i+j,i} M^t_{i+j,k}\right)(y^t_{i+j} - y^t_i) + \sum_{|i-k|\leq 10\rho, k\neq i+j,i} \delta_{i+j,i}(y^t_i - y^t_k). \end{aligned} \tag{4.1}$$

Now we want to find upper bounds on the norms of $B$ and $C$. Applying the triangle inequality to Line 4.1, we have that

$$\begin{aligned} \|B\| &\leq \left|2M^t_{i+j,i} + \sum_{k\neq i+j,i} M^t_{i+j,k}\right| \|y^t_{i+j} - y^t_i\| + \sum_{k\neq i+j,i} |\delta_{i+j,i}| \|y^t_i - y^t_k\| \\ &\leq \frac{40he}{n}\|y^t_{i+j} - y^t_i\| + 10\rho^3(\lambda_x^2 + \lambda_y^2)\|y^t_{i+t} - y^t_i\| \end{aligned} \tag{4.2}$$

Line 4.2 comes from two different estimations. For the first term we use the bound given by Lemma 2.5. For the second term we use the estimation from Lemma 2.2 and use the fact $|i-k|$ is at most $10\rho$.

Now we want to find a bound on the norm of $C$. Since these are from points outside of the $10\rho$ nearest neighbors, we have a different bound for $|M_{\cdot,k}|$. Instead, since we have lower bounds on the pairwise distances, Lemma 2.5 gives us $|M_{\cdot,k}| \leq \frac{\varepsilon_y}{4n^2}$. Thus we have that an

upper bound on the norm of $C$ to be

$$\begin{aligned}\|C\| &\leq 2 \sum_{|i-k|>10\rho} |M^t_{i+j,k}| \|y^t_{i+j} - y^t_k\| \\ &\leq 2(n-10\rho)\frac{\varepsilon_y}{4n^2}\frac{|n-10\rho|}{j}\|y^t_{i+j} - y^t_i\| \\ &\leq \frac{\varepsilon_y}{2}\|y^t_{i+j} - y^t_i\|. \end{aligned} \tag{4.3}$$

Combining the results of Line 4.2 and 4.3, we have the entire bound of $\|y^t_{i+j} - y^t_i\|$ to be the following.

$$\begin{aligned}\|y^t_{i+j} - y^t_i\| &\leq \|y^{t+1}_{i+j} - y^{t+1}_i\| + \left[\frac{40he}{n} + 10(h+1)\rho^3(\lambda_x^2 + \lambda_y^2) + \frac{\varepsilon_y}{2}\right]\|y^t_{i+j} - y^t_i\| \\ &\leq \|y^{t+1}_{i+j} - y^{t+1}_i\| + \left[\frac{\varepsilon}{2} + \frac{\varepsilon}{2}\right]\|y^t_{i+j} - y^t_i\| \end{aligned} \tag{4.4}$$

The bound in Line 4.4 comes from our assumption that $\frac{40he}{n} + 10(h+1)\rho^3(\lambda_x^2 + \lambda_y^2) < \frac{\varepsilon_y}{2}$. Therefore we can conclude that

$$\|y^t_{i+j} - y^t_i\| \leq \frac{1}{1-\varepsilon_y}\|y^{t+1}_{i+j} - y^{t+1}_i\|.$$

□

We can also use the proof in Lemma 4.1 to show that density is preserved after an iteration of t-SNE.

**Corollary 4.1.** *Assume we have a high-dimensional dataset $(x_i)_{i=1}^n$ that is a $\lambda_x$-dense, $\varepsilon_x$-discrete curve and that for some $t > 0$, the visualization $(y_i^t)_{i=1}^n$ of $(x_i)_{i=1}^n$ is also a $\lambda_y$-dense, $\varepsilon_y$-discrete curve. If the following hold:*

- *The perplexity $\rho$, and the step size of t-SNE's gradient, $h$, satisfy*
  $\frac{40he}{n} + 10(h+1)\rho^3(\lambda_x^2 + \lambda_y^2) < \frac{\varepsilon_y}{2}$,

- *For $k$ such that $|i-k| > 10\rho$, we have $\|y_k^t - y_i^t\| > \sqrt[4]{16h/\varepsilon_y}$,*

*then for iteration $t+1$, we have that*

$$\|y_{i+1}^{t+1} - y_i^{t+1}\| \leq (1+\varepsilon_y)\lambda_y.$$

*Proof.* By a similar estimation in Lemma 4.1, we have that

$$y_{i+1}^{t+1} - y_i^{t+1} = \underbrace{(y_{i+1}^t - y_i^t)}_{A} + \underbrace{\left( \sum_{|i-k|\leq 10\rho, k\neq i+j} M_{i+j,k}^t (y_{i+j}^t - y_k^t) - \sum_{|i-k|\leq 10\rho, k\neq i} M_{i,k}^t (y_i^t - y_k^t) \right)}_{B}$$
$$+ \underbrace{\left( \sum_{|i-k|>10\rho} M_{i+j,k}^t (y_{i+j}^t - y_k^t) - M_{i,k}^t (y_i^t - y_k^t) \right)}_{C}.$$

However, the bounds on $A$, $B$ and $C$ are still all applicable in this case. So by following the same proof (4.4) now becomes

$$\begin{aligned} \|y_{i+1}^{t+1} - y_i^{t+1}\| &\leq \|y_{i+1}^t - y_i^t\| + \varepsilon_y \|y_{i+1}^t - y_i^t\|, \\ &\leq (1+\varepsilon_y)\lambda_y, \end{aligned} \tag{4.5}$$

where (4.5) comes from the density assumption on immediate pairwise distances. □

Now that we have a relationship between the pairwise distances in the previous iteration and current iteration, we can show $\lambda$-dense, $\epsilon$-discrete curves will retain their structure in the next iteration of t-SNE. What this bound tells us is that as there is less curvature in the dataset, the upper bound will be tighter, so there is less room for expansion each iteration. However, for a discrete curve with a lot of curvature, there is more potential for the visualization to

expand.

**Lemma 4.2.** *Assume we have high-dimensional dataset $(x_i)_{i=1}^n$ that is a $\lambda_x$-dense $\varepsilon_x$-discrete curve and that for some $t > 0$, the visualization $(y_i^t)_{i=1}^n$ of $(x_i)_{i=1}^n$ is also a $\lambda_y$-dense $\varepsilon_y$-discrete curve. If the following hold:*

- *The perplexity $\rho$, and the step size of t-SNE's gradient, $h$, satisfy $\frac{40he}{n} + 10(h+1)\rho^3(\lambda_x^2 + \lambda_y^2) < \frac{\varepsilon_y}{2}$,*
- *For all $i, k$ where $|i-k| > 10\rho$, $\|x_i - x_k\| \geq \rho^2 \log(n)/n^2$ and $\|y_i^t - y_k^t\| \geq \sqrt[4]{3h/\varepsilon_y}$, where $h$ is the step size in t-SNE's gradient.*

*then for iteration $t+1$, for any $i, i-j, i+j \in [\rho]$ we have that*

$$\|y_{i+j}^{t+1} + y_{i-j}^{t+1} - 2y_i^{t+1}\| \leq \left[2\varepsilon_y + (1+10\rho)\rho^3(\lambda_x^2 + \lambda_y^2) + \frac{40he\varepsilon_y}{n}\right] \|y_{i+j}^t - y_i^t\|.$$

*Proof.* Let $\mathcal{N}$ be the $10\rho$ nearest neighbors of $i$. From the construction of t-SNE's gradient we have that

$$\begin{aligned}
y_{i+j}^{t+1} + y_{i-j}^{t+1} - 2y_i^{t+1} = &\underbrace{(y_{i+t}^t + y_{i-j}^t - 2y_i^t)}_{A} \\
&- \underbrace{(3M_{i+j,i}^t(y_{i+j}^t - y_i^t) + 3M_{i-j,i}^t(y_{i-j}^t - y_i^t))}_{B_1} \\
&- \underbrace{\left[\sum_{|i-k|\leq 10\rho, k\neq i,i+j,i-j} M_{i+j,k}^t(y_{i+j}^t - y_k^t) + M_{i-j,k}^t(y_{i-j}^t - y_k^t) - 2M_{i,k}^t(y_i^t - y_k^t)\right]}_{B_2} \\
&- \underbrace{\left[\sum_{|i-k|>10\rho} M_{i+j,k}^t(y_{i+j}^t - y_k^t) + M_{i-j,k}^t(y_{i-j}^t - y_k^t) - 2M_{i,k}^t(y_i^t - y_k^t)\right]}_{C}
\end{aligned}$$

We want to find bounds on the norms of these terms. Since we assume the previous iteration is a $\varepsilon_y$-discrete curve, then the norm of $A$ has the bound

$$\|A\| \leq \varepsilon_y \|y^t_{i+j} - y^t_i\|. \tag{4.6}$$

We can simplify the terms in $B_1$ by using the substitution $y^t_{i+j} - y^t_i = y^t_i - y^t_{i-j} + e^t_{ij}$ where $\|e^t_{ij}\| \leq \varepsilon_y \|y^t_{i+j} - y^t_i\|$. Substituting this in gives us

$$B_1 = -3\left((M^t_{i+j,i} - M^t_{i-j,i})(y^t_{i+j} - y^t_i) + 3M^t_{i-j,i}e^t_{ij}\right). \tag{4.7}$$

Before we take the norm of $B_1$, we consider the terms in $B_2$. Lemma 2.4 gives us a relationship between $M^t_{i+j,k}, M^t_{i-j,k}$ and $M^t_{i,k}$ so substituting that into $B_2$ we have

$$B_2 = \sum_{|i-k|\leq 10\rho, k\neq i+j, i-j, i} M^t_{i+j,k}(y^t_{i+j} - y^t_i) + M^t_{i-j,k}(y^t_{i-j} - y^t_i) + \sum_{|i-k|\leq 10\rho, k\neq i+j, i-j, i} \delta^t_{i+j,i,k}(y^t_i - y^t_k) \tag{4.8}$$

where $|\delta^t_{i+j,i,k}| \leq \rho^2(\lambda_x^2 + \lambda_y^2)$. Additionally, since we assume the previous iteration is a $\varepsilon$-discrete curve, we know that $y^t_{i-j} - y^t_i = y^t_i - y^t_{i+j} + e^t_{ij}$ where $\|e^t_{ij}\| \leq \varepsilon_y \|y^t_{i+j} - y^t_i\|$. Substituting this into Line 4.8 we have

$$B_2 = \sum_{|i-k|\leq 10\rho, k\neq i+j, i-j, i} (M^t_{i+j,k} - M^t_{i-j,k})(y^t_{i+j} - y^t_i) + M^t_{i-j,k}(e^t_{ij}) + \sum_{|i-k|\leq 10\rho, k\neq i, i+j, i-j} \delta^t_{i+j,i,k}(y^t_i - y^t_k). \tag{4.9}$$

However, the terms in the first sum in Line 4.10 are the same in those in Line 4.7 so we can combine them to get

$$B_1 + B_2 = \sum_{|i-k|\leq 10\rho} (M^t_{i+j,k} - M^t_{i-j,k})(y^t_{i+j} - y^t_i) + M^t_{i-j,k}(e^t_{ij}) + \sum_{|i-k|\leq 10\rho, k\neq i, i+j, i-j} \delta^t_{i+j,i,k}(y^t_i - y^t_k). \tag{4.10}$$

By a similar argument in Lemma 2.4, we know that $|M^t_{i+j,k} - M^t_{i-j,k}| \leq \rho^2(\lambda_x^2 + \lambda_y^2)$ and by Lemma 2.5 we know that $|M^t_{i-j,k}| \leq \frac{4he}{\rho n}$. So taking the norm of Line 4.10 we have

$$\begin{aligned}\|[B_1 + B_2\| \leq 10\rho^3(\lambda_x^2 + \lambda_y^2)\|y^t_{i+j} - y^t_i\| + \frac{40he\varepsilon_y}{n}\|y^t_{i+j} - y^t_i\| \\ + (10\rho - 3)\rho^2(\lambda_x^2 + \lambda_y^2)\frac{|i-k|}{j}\|y^t_{i+j} - y^t_i\|. \end{aligned} \tag{4.11}$$

The last estimation in Line 4.11 comes from the results in Lemma 2.2 to estimate $\|y^t_i - y^t_k\|$. Since $|i-k| < 10\rho$ and $j \geq 1$, then the upper bound in Line 4.11 becomes

$$\|B_1 + B_2\| \leq \left[(1 + 10\rho)\rho^3(\lambda_x^2 + \lambda_y^2) + \frac{40he\varepsilon_y}{n}\right]\|y^t_{i+j} - y^t_i\|. \tag{4.12}$$

Finally, we find a bound for the third part of the gradient, C. Here, we can use the assumption that for any $|i-k| > 10\rho$, $\|y^t_{i+j} - y^t_k\|$, $\|y^t_{i-j} - y^t_k\|$ and $\|y^t_i - y^t_k\|$ are all at least $d = \sqrt[4]{3h/\varepsilon_y}$. From Lemma 2.5 that means $|M^t_{\nu,k}| \leq \frac{h(1+d^2)^{-2}}{n^2} \leq \frac{\varepsilon_y}{3n^2}$ where $\nu \in \{i+j, i, i-j\}$. Applying this bound along with the results of Lemma 2.2 to the terms in $C$ to get the bound:

$$\|C\| \leq \frac{3(n-10\rho)\varepsilon_y}{3n^2}\frac{(n-10\rho)}{j}\|y^t_{i+j} - y^t_i\| \leq \varepsilon_y\|y^t_{i+j} - y^t_i\|. \tag{4.13}$$

Now we can apply the bounds from Lines 4.6, 4.12 and 4.13 to get

$$\begin{aligned}\|y^{t+1}_{i+j} + y^{t+1}_{i-j} - 2y^{t+1}_i\| &\leq \left[2\varepsilon_y + (1 + 10\rho)\rho^3(\lambda_x^2 + \lambda_y^2) + \frac{40he\varepsilon_y}{n}\right]\|y^t_{i+j} - y^t_i\|, \\ &\leq \frac{1}{1-\varepsilon_y}\left[2\varepsilon_y + (1 + 10\rho)\rho^3(\lambda_x^2 + \lambda_y^2) + \frac{40he\varepsilon_y}{n}\right]\|y^{t+1}_{i+j} - y^{t+1}_i\|, \end{aligned} \tag{4.14}$$

where (4.14) follows from the bound given in Lemma 4.1. □

Now that we have an upper bound on the error vector in the next iteration, we can bring this all together to categorize how t-SNE preserves discrete curves in each iteration.

**Theorem 4.1.** *Assume a high-dimensional dataset $(x_i)_{i=1}^n$ is a $\lambda_x$-dense, $\varepsilon_x$-discrete curve and that for some $t > 0$, the visualization $(y_i)_{i=1}^n$ of $(x_i)_{i=1}^n$ is also a $\lambda_y^t$-dense, $\varepsilon_y^t$-discrete curve. If the following hold:*

- *$\lambda_x^t$ and $\lambda_y^t$ are both less than $\frac{1}{n}$,*
- *The size of the neighborhood $\rho$, and the step size of t-SNE's gradient, $h$, satisfy $\frac{40he}{n} + 10(h+1)\rho^3(\lambda_x^2 + \lambda_y^2) < \frac{\varepsilon_y}{2}$,*
- *For all $i, k$ where $|i-k| > 10\rho$, $\|x_i - x_k\| \geq \rho^2 \log(n)/n$ and $\|y_i^t - y_k^t\| \geq \sqrt[4]{16h/\varepsilon_y}$*

*then $(y_i^{t+1})_{i=1}^n$ is a $\lambda_y^{t+1}$-dense, $\varepsilon_y^{t+1}$-discrete curve where*

$$\lambda_y^{t+1} = (1 + \varepsilon_y^t)\lambda_y^t, \tag{4.15}$$

$$\varepsilon_y^{t+1} = \frac{1}{1 - \varepsilon_y^t}\left[2\varepsilon_y^t + \frac{(1+10\rho)\rho^3}{n^2} + \frac{40he\varepsilon_y^t}{n}\right]. \tag{4.16}$$

*Proof.* The proof follows directly from Corollary 4.1 and Lemma 4.2. □

## 4.2 Discussion

Looking at the results in Theorem 4.1, we can see that as $n \to \infty$, then we have that

$$\varepsilon_y^{t+1} = \frac{2\varepsilon_y^t}{1 - \varepsilon_y^t}.$$

This implies that a a discrete curve becomes more well-populated, there will be less of an expansion each iteration and we can expect t-SNE to preserve the structure for more iterations. Additionally, we notice that if the discrete curve given is a perfect line ($\varepsilon_y = 0$)

then there will be no additive error for both the density and new $\varepsilon_y$. Theorem 4.1 requires that $\varepsilon_y^t < 1$, so it will no longer necessarily hold once $\varepsilon_y^t \geq 1$. We can estimate when this happens in the following lemma.

**Lemma 4.3.** *Assume that $\varepsilon_y^t$ and $\varepsilon_y^{t+1}$ are as described in Theorem 4.1. If $t \geq \log_2(1/\varepsilon_y^0) - 1$, then $\varepsilon_y^{t+1} \geq 1$.*

*Proof.* Since each $\varepsilon_y^k < 1$ for all $k \leq t$, then we have

$$\begin{aligned} \varepsilon_y^{t+1} &= \frac{2\varepsilon_y^t}{1-\varepsilon_y^t}, \\ &\geq 2\varepsilon_y^t, \\ &\geq 2^{t-1}\varepsilon_y^0, \\ &\geq 1, \end{aligned} \tag{4.17}$$

where the last line follows from the assumption that $t \geq \log_2(1/\varepsilon_y^t) - 1$. □

This Lemma quantifies the intuition that as the curvature, $\varepsilon_y$ gets smaller then the longer t-SNE's iterations will definitely preserve the structure.

While Theorem 4.1 shows that t-SNE will preserve the discrete curve structure over a certain number of iterations, there are some significant assumptions that needed to be made. First, both the high-dimensional data and visualization couldn't have global structures that are too dense, i.e. no critical overlaps and no tight coils. Because of t-SNE's random initialization, we cannot guarantee that after early exaggeration the visualization is not too tight of a coil nor has any overlaps. In fact, since t-SNE initializes in such a small box and globally doesn't expand much during early exaggeration, it is reasonable to assume there is more likely to be significant overlap as $n$ grows large.

# Chapter 5

# AVIDA

Up until now, this work has focused on the theoretical aspects of t-SNE's treatment of discrete curves. There are many instances in real-life datasets where t-SNE's ability to preserve discrete curves is important those who are use t-SNE to visualize datasets whose structure they may not understand. Alternating method for Visualing and Integrating Data (AVIDA) is a data integration framework where the use of t-SNE is critical to preserving discrete curves when integrating datasets. This chapter gives an overview of the importance of data integration in single-cell analysis and how AVIDA utilizes data visualization techniques to better integrate datasets in low-dimensions, oftentimes preserving discrete curves.

## 5.1 Background Information

Databases are expanding not only in size but also with increasing complexity. In many applications, multiple measurements of a system are taken across different samples or in different feature spaces which produce *multimodal data* such as texts attached to images [25]. Multimodality allows a more comprehensive investigation of a system. Establishing connections

among the modalities is the foundation of coherent analysis. Recently, the emerging multimodal single-cell omics has become a powerful tool to analyze different aspects of a biological system at the same time [53]. Fusing multimodal single-cell data is especially challenging when there is no direct correspondence between the measurements and the samples.

Single-cell RNA sequencing (scRNA-seq) is a recent technology that measures RNA abundance at transcriptomics level with single-cell resolution [42]. The maturation of the technology allows analysis with scRNA-seq assays across many samples that, for example, represent different ages or healthy and diseased individuals [41, 48]. On the other hand, the emerging single-cell assays provide a more comprehensive examination of a system, such as single-cell ATAC-seq (scATAC-seq) [7] that measures chromatin accessibility and single-cell Hi-C [33] that explores chromosome architecture.

Integrating the various single-cell assays across different samples provides a comprehensive characterization of a biological system. Many computational methods have been developed to integrate the same single-cell assays of multiple samples [21, 40, 49] or different single-cell assays [19, 20]. In the integration of multiple single-cell omics assays, most current methods rely on the known correspondence between features, for example by mapping chromatin loci to genes and assuming the similarity between the samples. The multi-omics integration becomes a harder problem when no prior correspondence is assumed, for example, a gene actually corresponds to multiple loci and accessible loci do not directly indicate gene expression. This leads to a general problem of integrating datasets without known correspondence between features.

When no feature correspondence is given, the structures of the individual datasets can be exploited and matched to integrate the datasets. For example, canonical correlation analysis examines covariances between the datasets but is limited to deriving linear correspondence between the features. When the datasets are represented as graphs with edges annotating pairs of similar data points within each dataset, the integration problem can be addressed

using various graph alignment methods [35, 51]. Among the graph alignment methods, *Gromov-Wasserstein optimal transport (GW-OT)* can align graphs based only on the graph structures [30]. It finds a coupling of the distributions representing the graphs that best preserves the intra-dataset distances between the nodes.

Optimal transport (OT) compares and finds connections between measures. It seeks the coupling between distributions with the minimum total coupling cost based on predefined costs between locations [24, 31, 47]. OT has been a versatile tool widely used in practical problems, such as generative deep learning models [3], domain adaptation [12], and image sciences [17]. It has been used to find correspondence between data points in single-cell gene expression data with common features [9,34,39]. The aforementioned GW-OT has been used in this field to exploit the structural information within individual datasets. SpaOTsc [9] uses fused Wasserstein-Gromov-Wasserstein optimal transport to improve the integration of spatial data and scRNA-seq data with few shared genes by matching the spatial structure and the structure in scRNA-seq data based on gene expression similarity. SCOT [13] uses Gromov-Wasserstein optimal transport to align scRNA-seq and scATAC-seq data by matching the structures represented by intra-dataset similarity among cells. Pamona [10] uses partial Gromov-Wasserstein optimal transport to partially align scRNA-seq and scATAC-seq data to address the partially overlapping cell populations among different samples.

In addition to studying shared structures revealed by the overlapping part of integrated data, it is equivalently important to examine the structures of non-overlapping parts which may depict a biological process uniquely captured by a certain assay [50]. Since most integration methods depend on similarities between samples, the dissimilar parts are often overlooked. Efforts have been made to keep the variation among samples examined with the same single-cell assay [50].

In the analysis of high-dimensional multimodal datasets, another crucial step is *dimensionality reduction*. Dimensionality reduction is the process of taking high-dimensional data and

finding a representation in lower dimensions that is still meaningful. It has many important applications because dimensionality reduction helps address the curse of dimensionality and other challenges that come with working with high-dimensional data [23]. Principal Component Analysis (PCA) [36] is the most traditional linear technique used in dimensionality reduction but there are many popular non-linear techniques, such as Local Linear Embedding [38], Isomap [43], UMAP [29], and t-SNE [45].

t-SNE is a popular dimensionality reduction and visualization technique that was introduced in 2008 by van der Matten and Hinton [45]. It has been applied to a variety of high dimensional data, including deep learning [28], physics [46], and medicine [1]. Given a high dimensional dataset, t-SNE outputs a low dimensional representation. t-SNE works by making pairwise affinities between points in high dimensions and pairwise affinities between points in low dimensions. It then uses gradient descent to find the set of points (in low dimensions) that minimize the KL divergence between the two sets of joint probabilities.

In the analysis of multimodal single-cell data, the dimensionality reduction and the integration steps are often performed separately or sequentially, including the existing methods that integrate datasets without known feature correspondences [10, 13]. However, these two steps are closely related in that they both preserve the structures from high dimension to low dimension or from the original spaces to the joint space. The benefit of combining these two steps has been shown in many recent works. For example, MultiMAP performs dimensionality reduction and integration utilizing both shared and non-shared features between datasets [22]. As another example, j-SNE learns a joint representation in low dimensions without shared features across multiple data sets with one-to-one correspondences [14]. In this work, we present a workflow called AVIDA (Alternating Method for Visualizing and Integrating Data), that integrates 2D representations of high dimensional data sets by alternating between dimension reduction and alignment. AVIDA operates without knowledge or the necessity of shared features or one-to-one correspondences across data sets. To demonstrate

this workflow, we use t-SNE for the dimension reduction module and Gromov-Wasserstein optimal transport for the integration module. Different choices for the dimension reduction module and alignment module can be utilized in this framework, depending on the application. We also include a small set of additional experiments in A.1, which utilize UMAP in the dimension reduction step instead of t-SNE to further demonstrate AVIDA's flexibility as a framework. In four synthetic datasets and two real biological datasets with ground truth, we show that AVIDA better preserves the structures of the individual datasets while achieving comparable integration quality compared to other methods.

## 5.2 Results

### 5.2.1 Overview of AVIDA

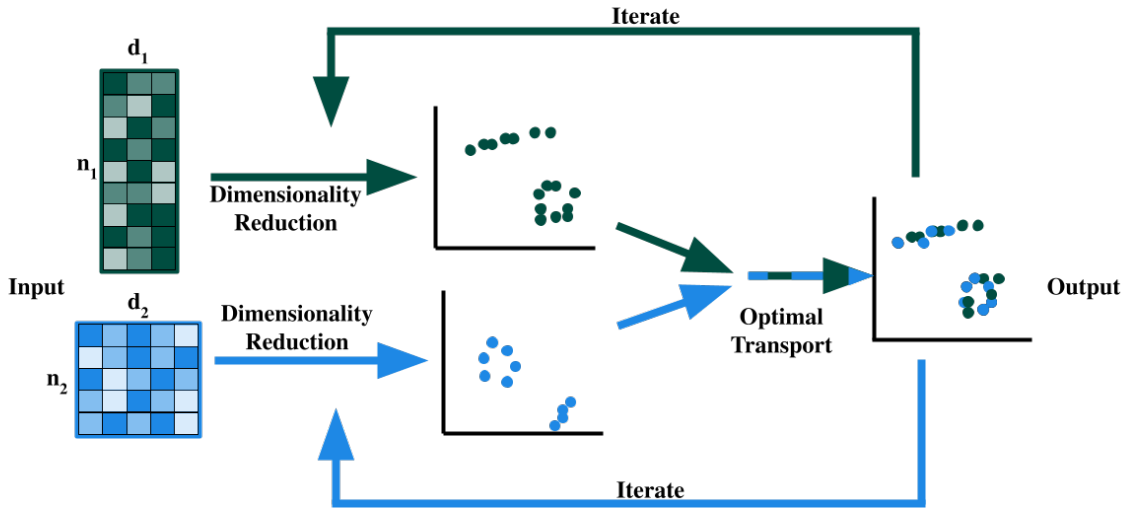


Figure 5.1: A visual schematic of AVIDA.

The proposed method is called the *alternating method for visualizing and integrating data*, or AVIDA. AVIDA alternates between improving the low dimensional representation through

a dimensionality reduction technique and the alignment of data points in low dimensions across different datasets. The purpose of alternating between dimensionality reduction and alignment is to find a balance between a good representation while still accurately aligning the datasets. We denote AVIDA as a function, taking as input the datasets $X_1, \ldots, X_k$, and is parameterized by choice of dimensionality reduction and alignment techniques: AVIDA$(X_1, X_2, ..., X_k; \text{DR}, \text{ALIGN})$. A simplified schematic of the method is shown in Figure 5.1. As shown in Figure 5.1, AVIDA can take as input two datasets and organizes the data as a pairwise distance matrix. Next, dimensionality reduction using the given pairwise distance matrix is performed on both datasets independently. An alignment method is used to "align" the datasets in the lower dimensional space and using the aligned data points, a new pairwise distance matrix is formed for each dataset, and the process iterates. This framework is flexible in its choice of dimensionality reduction technique (in fact, different dimension reduction algorithms can be used on different datasets if one so chooses) and alignment method.

Suppose one is given two datasets $X^{(1)}$ and $X^{(2)}$ and the goal is to create a joint representation of the datasets in a common lower dimensional space. Using some technique DR for dimensionality reduction (e.g., PCA, t-SNE, Random Forests, etc.) and GW-OT for alignment, we can formulate the objective function for AVIDA as AVIDA$(X_1, X_2; \text{DR}, \text{GW})$. The GW-OT objective is defined with respect to the low dimensional representation of points:

$$\text{GW}(Y^{(1)}, Y^{(2)}) = \sum_{i,j,i',j'} L_{i,j,i',j'} \mathbf{T}_{i,i'} \mathbf{T}_{j,j'} - \epsilon(H(\mathbf{T})), \tag{5.1}$$

where $H(\mathbf{T}) = \sum_{i,j} T_{ij} \log(T_{ij})$ is the Entropic regularization term and
$L_{i,j,i',j'} = \|d(y_i^{(1)}, y_j^{(1)}) - d(y_{i'}^{(2)}, y_{j'}^{(2)})\|^2$ with a chosen distance metric $d(\cdot, \cdot)$. This objective is minimized by using the projected gradient descent method with KL metric-based projections [37], $T \leftarrow \text{Proj}^{\text{KL}}_{U(\mathbf{a},\mathbf{b})}(T \odot e^{-\tau(L \otimes T + \epsilon \log(T))})$ where
$U(\mathbf{a}, \mathbf{b}) = \{T \in \mathbb{R}_+^{n_1 \times n_2} : T\mathbb{1} = \mathbf{a}, T^T\mathbb{1} = \mathbf{b}\}$ and $\tau$ is the step size. The implementation in

Python Optimal Transport [18] package is used. The representation for $Y^{(1)}$ will subsequently be mapped to $Y^{(2)}$ using the mapping found by minimizing (5.1) with respect to $T$, i.e., by setting $Y^{(1)} = TY^{(2)}$. Our combined loss function can be represented as

$$\begin{aligned}&\text{AVIDA}(X^{(1)}, X^{(2)}; \text{DR}, \text{GW}) \\ &= \min_{Y^{(1)}, Y^{(2)}} \text{DR}(X^{(1)}, Y^{(1)}) + \text{DR}(X^{(2)}, Y^{(2)}) + \text{GW}(Y^{(1)}, Y^{(2)}),\end{aligned} \tag{5.2}$$

where $\text{DR}(X^{(i)}, Y^{(i)})$ represents the objective loss associated with the dimensionality reduction technique DR. For example, if t-SNE is used for the DR step, the objective can be represented as the KL loss between probability distributions on the points in high and low dimensions. See 5.4 for more details.

### 5.2.2 AVIDA accurately reproduces the intra-dataset structures in integration of synthetic data

We compared $\text{AVIDA}(X_1, X_2; \text{TSNE}, \text{GW})$ to both Pamona and SCOT across four simulated datasets and two real-world single-cell multi-omics datasets. We chose Pamona and SCOT as a comparison because they are both advanced integration methods that do not require common features or one-to-one correspondence across data sets. To have a fair comparison with SCOT and Pamona, for these experiments we had SCOT and Pamona perform their alignment and then used t-SNE to visualize their low dimensional representations rather than UMAP or PCA. This way we are not comparing different kinds of visualization techniques to each other. To see how these methods would perform using UMAP instead of t-SNE, see Appendix A. Table 5.1 contains the performance metrics for $\text{AVIDA}(X_1, X_2; \text{TSNE}, \text{GW})$, SCOT and Pamona on both the simulated and real-life datasets. We used five different metrics to assess the performance of these methods: the fraction of samples closer than the true match (FOSCTTM), alignment, integration, accuracy, and representation loss. The

accuracy metric is only included on the datasets where the ground truth is known and an empty cell in the table implies the dataset did not meet that requirement. Details on the metrics are included in Section 5.4.2.

| Dataset | Method | FOSCTTM | Integration | Accuracy | Alignment | Representation Loss |
|---|---|---|---|---|---|---|
| **Bifurcated Tree** | AVIDA | 0.1202 | 1.0820 | **4.3863** | **0.5157** | **0.3275** |
| | Pamona | **0.1108** | **0.2933** | 7.6098 | 0.9897 | 1.0969 |
| | SCOT | 0.2103 | 1.0016 | 12.2095 | 0.75 | 2.1466 |
| **Circular Frustrum** | AVIDA | 0.1187 | 0.9699 | 2.9377 | **0.4267** | **0.3955** |
| | Pamona | **0.0186** | **0.2532** | **1.2577** | 0.9363 | 0.8377 |
| | SCOT | 0.0515 | 1.0032 | 4.3857 | 0.9727 | 1.7083 |
| **Dumbbell** | AVIDA | 0.5228 | 0.5568 | 25.1281 | 0.6385 | **0.1220** |
| | Pamona | 0.5055 | **0.3679** | 32.1714 | 0.7785 | 0.6176 |
| | SCOT | **0.4754** | 2.565 | **11.2244** | **0.2070** | 3.6008 |
| **Distant Rings** | AVIDA | 0.3138 | 0.6847 | 5.3429 | **0.639** | **0.1916** |
| | Pamona | 0.2580 | 1.2407 | 1.0 | 0.993 | 1.1784 |
| | SCOT | **0.0056** | **0.0791** | **0.2759** | 0.9125 | 0.9261 |
| **sc-GEM** | AVIDA | 0.2070 | 0.4700 | **2.4996** | 0.8994 | **0.4879** |
| | Pamona | 0.2108 | **0.3567** | 10.894 | 0.7237 | 1.4298 |
| | SCOT | **0.1818** | 2.3164 | 6.9267 | **0.5616** | 0.8790 |
| **scNMT-seq** | AVIDA | 0.2745 | 0.3631 | 4.5787 | **0.6619** | **1.0489** |
| | Pamona | 0.3889 | **0.2446** | **0.7032** | 0.9746 | 4.2435 |
| | SCOT | 0.**2675** | 2.4333 | 28.6287 | 0.7522 | 1.1979 |

Table 5.1: Metrics for AVIDA($X_1, X_2$; TSNE, GW) (labeled as AVIDA above), Pamona and SCOT experiments.

Our four simulated datasets include a bifurcated tree, a circular frustum (from [27]), a dumbbell, and distant rings. The dumbbell and distant rings datasets are introduced in order to highlight the difference between AVIDA and SCOT and Pamona. The dumbbell dataset consists of two rings that are connected by a line. We consider the following split of the dumbbell data set: $X_1$ contains data points from the two rings and a subset of the points along the line connecting the two rings. Then dataset $X_2$ contains all the points along the line connecting the two rings. Thus, the dumbbell dataset allows us to investigate the performance of AVIDA when there is only a partial direct correspondence between data sets.

We also introduce the distant rings dataset. The rings dataset consists of two rings that are far apart from each other in high dimensions. We set the sizes of their radii to be much smaller than the distance between the centers of the rings. Then, the datasets $X_1$ and $X_2$ are

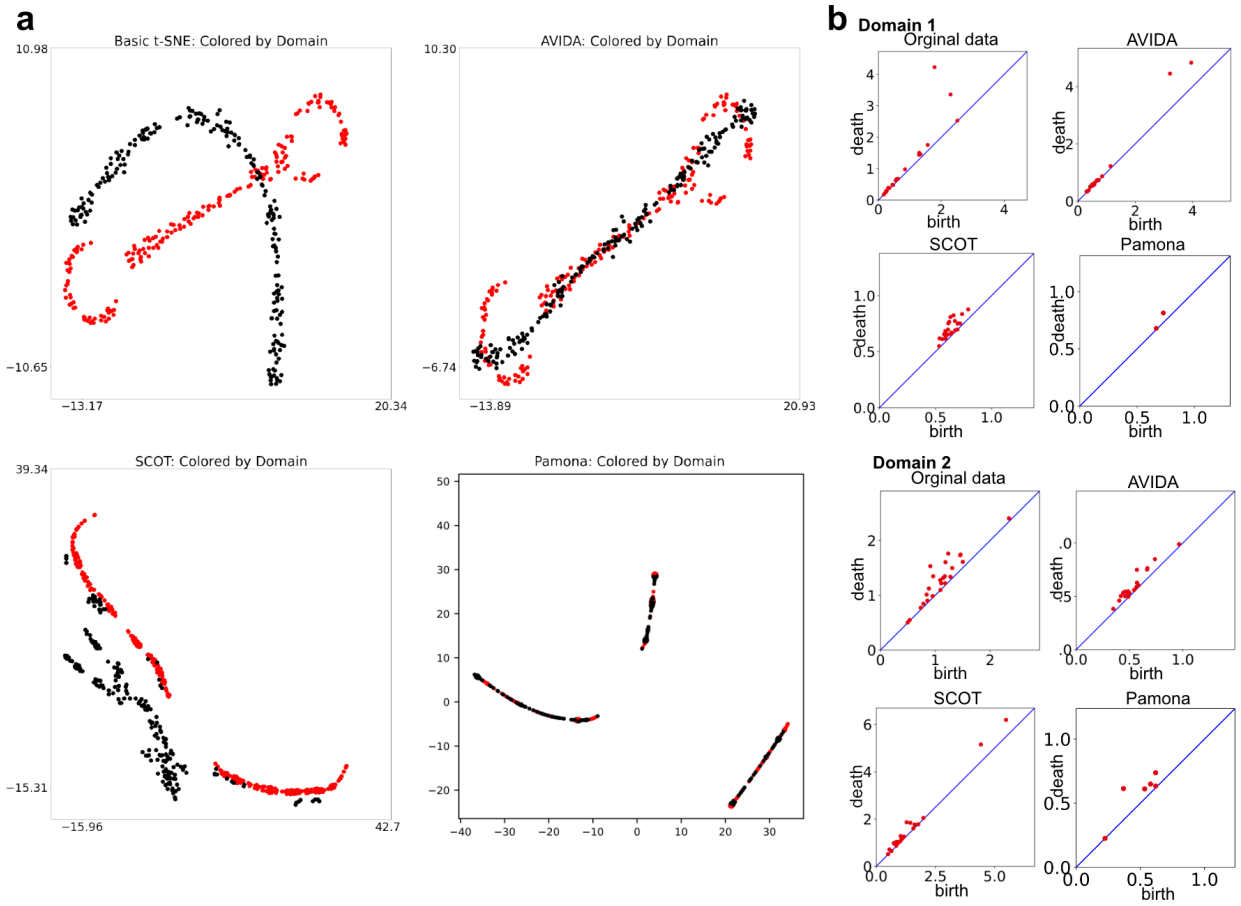


Figure 5.2: (a) Pamona, AVIDA, SCOT, and t-SNE representation of the dumbbell dataset. (b) The $H_1$ persistence diagrams of Vietoris-Rips filtration with Euclidean distance of the original data, and AVIDA and SCOT embeddings. The birth and death values are the scales at which topological features appear and disappear. A point farther away from the diagonal (blue line) represents a significant 1-dimensional loop. "Domain 1" and "Domain 2" correspond to the points colored red and black respectively in (a).

generated such that they both contain the entirety of the two rings dataset, i.e. $X_1 = X_2$. This is done so that there is a direct correspondence between points in $X_1$ and $X_2$. Thus, the rings dataset allows us to investigate the performance of AVIDA when there is a full direct correspondence between data sets. In addition, the difference in scale of the distances in the rings dataset allows us to highlight the advantage of using AVIDA rather than other forms of alignment.

The specific parameters used to generate these datasets are given in Section 5.4. The eval-

uations of these methods on the various metrics are given by Table 5.1.

Looking at Figures 5.2 and 5.3, it is clear why we want to introduce these datasets. In Figure 5.2, AVIDA clearly preserved the local structure of both datasets while Pamona and SCOT highlight the linear structure found in both datasets. This is demonstrated by both visual inspection of the loop structures preserved by AVIDA, as shown in Figure 5.2(a) and Figure 5.3(a) and the persistence diagrams, as shown in Figure 5.2(b) and Figure 5.3(b). The persistence diagram is the result of persistent homology [16, 54] which grows a simplicial complex on a point cloud and tracks the scale at which the topological features appear (birth value) and disappear (death value). A topological feature with large persistence value (difference between birth and death values) is considered significant and we are interested in the one dimensional $H_1$ features that correspond to circles in data. Details of persistent homology are discussed in Section 4.2.2. AVIDA is the only method that is able to successfully integrate the two representations generated by t-SNE's representation. Figure 5.3 shows that Pamona's method collapses both rings to a single point, destroying the local structure of the data. SCOT is able to integrate the datasets while still preserving some linear structure but compared to t-SNE's actual 2D representation, AVIDA produces a 2D representation with the most accurate local structure. Since AVIDA allows t-SNE to construct the local structure of the line before mapping, that structure is preserved in the final representation.

However, if we were to look at the FOSCTTM and accuracy scores in Table 5.1 for Figure 5.2 and Figure 5.3, Pamona scores best because all the points are correctly mapped close together. The datasets illustrate our need for a representation metric since the traditional metrics do not penalize for poor representations in 2D. We use t-SNE's loss function as our representation loss since it is a popular dimensionality reduction technique, however, it could easily be replaced by a loss function from other methods (e.g. UMAP).

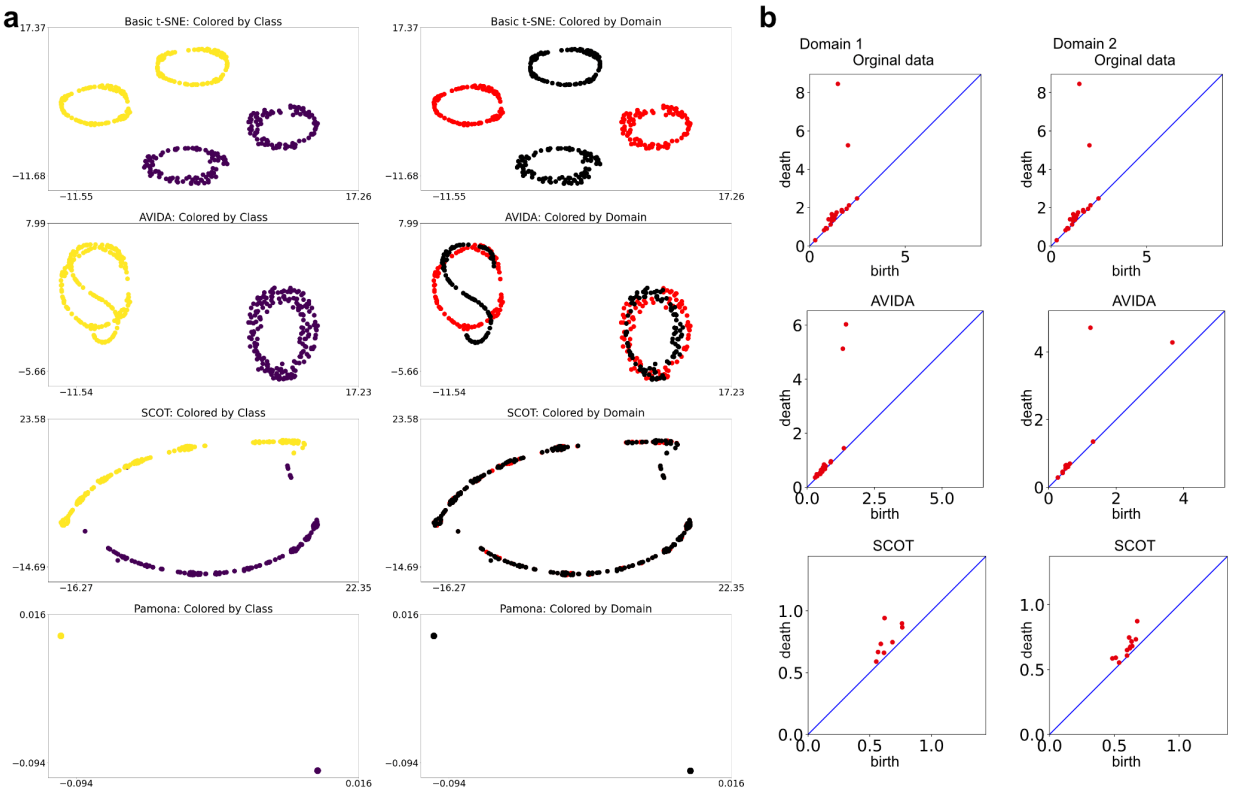


Figure 5.3: (a) t-SNE, AVIDA, SCOT and Pamona representation of the distant rings dataset. (b) The $H_1$ persistence diagrams of Vietoris-Rips filtration with Euclidean distance of the original data, and AVIDA and SCOT embeddings. The birth and death values are the scales at which topological features appear and disappear. A point farther away from the diagonal (blue line) represents a significant 1-dimensional loop. The $H_1$ diagrams of Pamona embeddings are empty. "Domain 1" and "Domain 2" correspond to the points colored red and black respectively in (a).

### 5.2.3 AVIDA achieves a balance between structure representation and multimodal dataset alignment

We also compare the outputs from two real-world single-cell multi-omics datasets. The first is sc-GEM, a dataset from [11] which contains both gene expression and DNA methylation at multiple loci on human somatic cell samples under coversion to induced pluripotent stem cells. The second is scNMT-seq, a dataset of chromatin accessibility, DNA methylation, and gene expression on mouse gastrulation samples collected at four different time states from [2]. The evaluations of AVIDA, SCOT, and Pamona on these datasets are also given

in Table 5.1. In Figure 5.4, we can see the different 2D representations for sc-GEM. The left

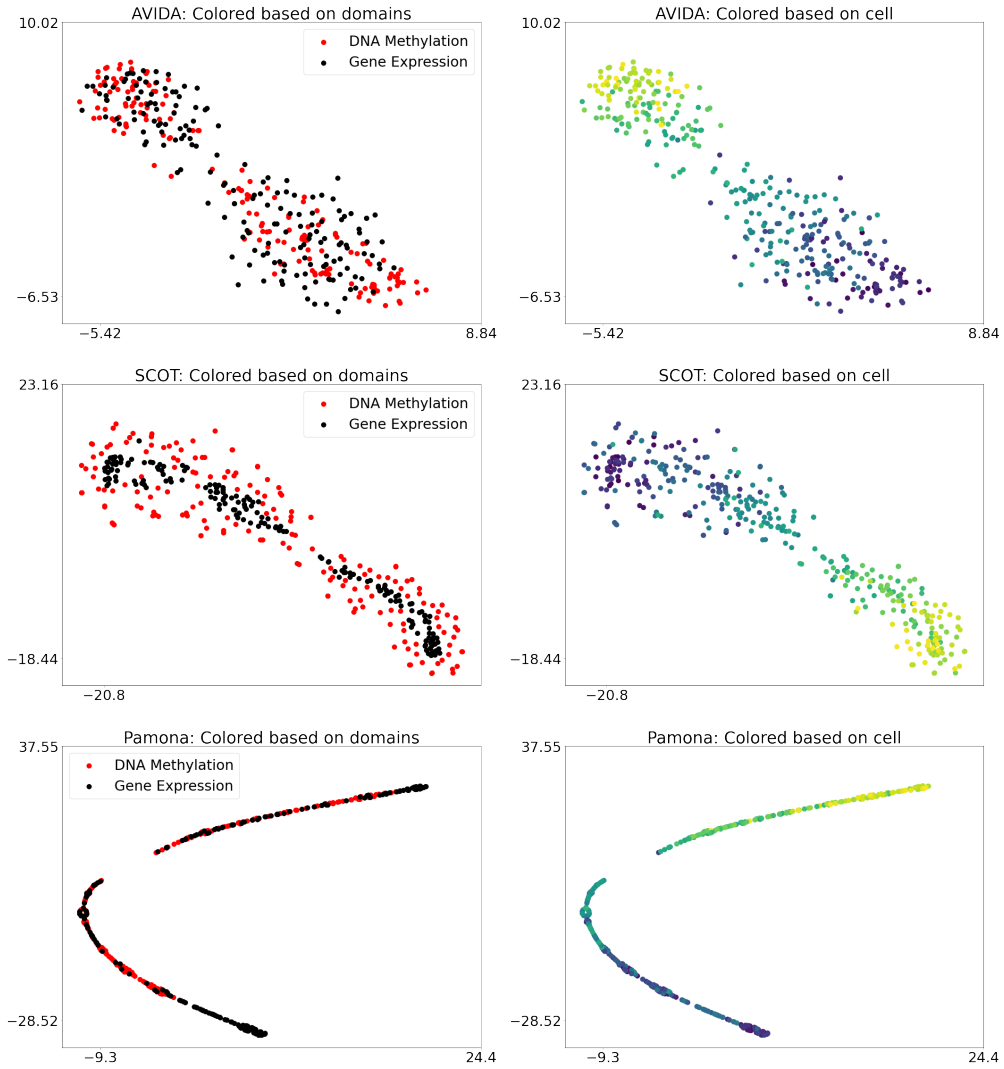


Figure 5.4: AVIDA, SCOT and Pamona representation of sc-GEM. The visualizations for each of the methods were made by t-SNE.

column of the figure shows the integration between the two datasets and the right column has the data points colored by cell. From these representations, we can see that AVIDA is able to fully integrate the two different datasets where there is some noticeable separation in the SCOT representation. Since this dataset contains the conversion from somatic cells to stem cells, we hope to see a gradient of colors from one end of the representation to the other which all methods are able to achieve. This is a good example of how AVIDA's performance on integration of real-life datasets is comparable to both SCOT and Pamona.

We can also confirm this observation in Table 5.1. AVIDA is able to achieve FOSCTTM and alignment scores that are comparable to SCOT and Pamona while simultaneously having the best representation loss. The same holds true for scNMT-seq as well. These examples illustrate that AVIDA is comparable to both Pamona and SCOT on real-life datasets while also performing well on the adversarial datasets: the dumbbell and distant rings datasets.

While we did not plot every dataset's low dimensional representation here, Figure 5.5 compares the FOSCTTM and representation losses for each 2D representation generated by SCOT, AVIDA, and Pamona. The shapes designate the dataset's low dimensional representation and the different colors represent the method that was used. We can see that across the different datasets, all three methods have comparable FOSCTTM scores, indicating that the integration of the datasets are similar. However, we can also see that AVIDA by far has the best representation loss, indicating a more accurate low dimensional representation.

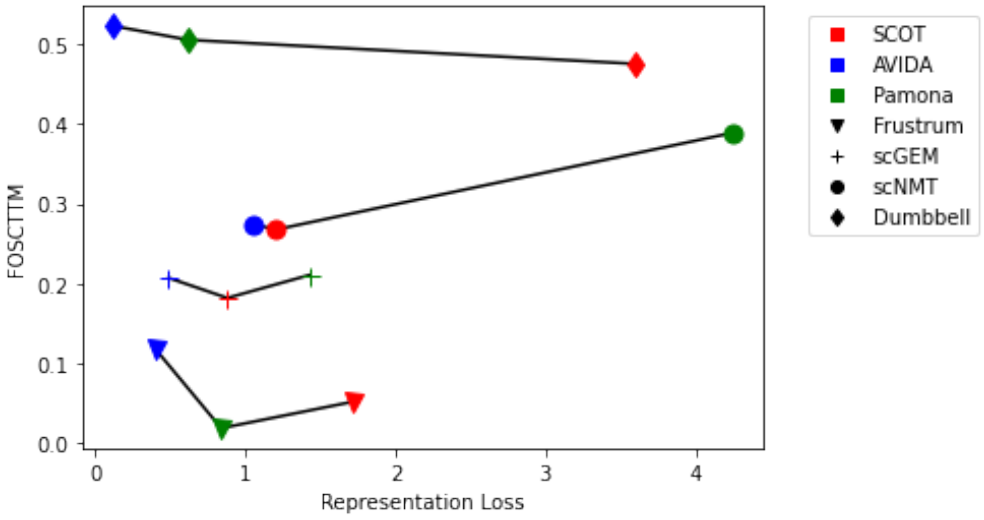


Figure 5.5: A comparison of methods using integration and 2D representation.

## 5.3 Discussion

Motivated by the similar fundamental assumptions in dimension reduction and data integration that they both try to preserve the structures of datasets, we developed an alternating method, AVIDA, which combines these two processes for joint 2D representation of datasets without shared features. Comparing with the methods that perform integration first and then dimension reduction, AVIDA better preserves the detailed structures of the datasets being integrated especially the structures present in only one of the datasets. This property allows the identification of mechanisms that can only be revealed with one of the technologies.

In this work, we demonstrate the method using t-SNE for dimension reduction and Gromov-Wasserstein optimal transport for data integration. In general, other dimension reduction methods and integration methods could be used. The representation loss used in the comparison can also be used as a control metric about how well the structures of individual datasets are preserved in the joint representation. This metric can be used to find a balance between integration and representation when other methods are used for the dimension reduction and integration modules. The comparison indicates that a method could do a perfect job in integration while missing structures presented in the individual datasets. It is thus important to also evaluate the quality of the structure representation of individual datasets when developing joint dimension reduction methods for high-dimensional multimodal datasets.

Despite the improvements on performing the two processes separately, the quality of the joint 2D representation still heavily depends on the performance of the specific dimension reduction method and integration method. While the quality of dimension reduction can be checked by comparing it to the structures present in the original high dimensional datasets, it is hard to evaluate the integration quality without ground truth. It is thus also important to further validate the result with prior knowledge or assess the robustness of the integration with, for example, subsampling.

Upon the joint representation of multimodal datasets, one major downstream task is to find the correspondence between the non-overlapping features across the datasets. A potential method for this is to track the contributions of original features to the common low dimensional representations and subsequently find the correspondence between them.

## 5.4 Methods

AVIDA is a framework that takes input data sets $\{X^{(\ell)}\}_{i=1}^{N}$ where the data sets $X^{(\ell)} \in \mathbb{R}^{n_\ell \times d_\ell}$ need not be in the same feature space. The output of AVIDA is a low dimensional representation of all data sets simultaneously in a single feature space. This is accomplished by alternating between dimensionality reduction and alignment. The AVIDA framework is presented in Algorithm 2. The choice of dimensionality reduction technique and alignment method is up to the user and can be chosen based on the use case. In Section 5.4.1, we present a detailed implementation of AVIDA using t-SNE for dimensionality reduction and GW-OT for alignment.

**Algorithm 2** AVIDA

**Input**:$N$ datasets $X^{(\ell)} = \{x_i^{(\ell)}\}_{i=1}^{n_\ell} \subset \mathbb{R}^{d_\ell}$, target dimension $d$, Dimensionality Reduction Method $DR(\cdot)$, Alignment Method $ALIGN(\cdot)$.
**Output**: Low-dimensional representations $Y^{(\ell)} = \{y_i^{(\ell)}\}_{i=1}^{n_\ell} \subset \mathbb{R}^{d}$.
Initialize $Y_0^{(\ell)}$ for $\ell \in [N]$ and set $t = 0$.
**do**
  Dimensionality reduction step:
    $\hat{Y}_t^{(\ell)} = DR(X^{(\ell)}, Y_t^{(\ell)})$ for $\ell \in [N]$. ▷ Input dataset $X^{(\ell)}$ and initialization $Y_t^{(\ell)}$
  Alignment step:
    $[Y_{t+1}^{(1)}, \cdots, Y_{t+1}^{(N)}] = ALIGN(\hat{Y}_t^{(1)}, \cdots, \hat{Y}_t^{(N)})$.
  Increment iteration count: $t = t + 1$.
**while** stopping criteria not satisfied
Return $Y^{(\ell)} = Y_t^{(\ell)}$ for $\ell \in [N]$.

### 5.4.1 AVIDA with t-SNE and GW-OT

In this section, we present our implementation of the AVIDA framework using t-SNE for dimensionality reduction and GW-OT for alignment, i.e., AVIDA($X_1, X_2$; TSNE, GW). For simplicity, we assume there are two input data sets $X^{(1)} = \{x_i^{(1)}\}_{i=1}^{n_1} \subset \mathbb{R}^{d_1}$ and $X^{(2)} = \{x_i^{(2)}\}_{i=1}^{n_2} \subset \mathbb{R}^{d_2}$ and that the low dimensional output feature space has dimension $d = 2$, i.e., $Y^{(1)} = \{y_i^{(1)}\}_{i=1}^{n_1} \subset \mathbb{R}^2$ and $Y^{(2)} = \{y_i^{(2)}\}_{i=1}^{n_2} \subset \mathbb{R}^2$.

In the dimensionality reduction step, t-SNE generates pairwise affinity values $\{p_{ij}^{(\ell)}\}$ for each of the dataset $X^{(\ell)}$, as given by

$$p_{j|i}^{(\ell)} = \frac{\exp(-\|x_i^{(\ell)} - x_j^{(\ell)}\|^2/2\sigma_i^{(\ell)})}{\sum_{k\neq i}\exp(-\|x_k^{(\ell)} - x_i^{(\ell)}\|^2/2\sigma_i^{(\ell)})} \tag{5.3}$$

$$p_{ij}^{(\ell)} = \frac{p_{j|i}^{(\ell)} + p_{i|j}^{(\ell)}}{2n_\ell}, \tag{5.4}$$

where the $\sigma_i^{(\ell)}$'s satisfy

$$\rho = 2^{-\sum_{j\neq i} p_{j|i}^{(\ell)} \log(p_{j|i}^{(\ell)})}, \tag{5.5}$$

for a perplexity value $\rho$ chosen by the user. To obtain $y_i^{(\ell)}$, t-SNE minimizes the Kullback-Leibler divergence between $\{p_{ij}^{(\ell)}\}_{j\neq i}$ and $\{q_{ij}^{(\ell)}\}_{j\neq i}$ using gradient descent. The target probabilities $q_{ij}^{(\ell)}$ are defined as:

$$q_{ij}^{(\ell)} = \frac{(1 + \|y_i^{(\ell)} - y_j^{(\ell)}\|^2)^{-1}}{\sum_{i',j'}(1 + \|y_{i'}^{(\ell)} - y_{j'}^{(\ell)}\|^2)^{-1}}. \tag{5.6}$$

To obtain $y_i^{(\ell)}$, t-SNE minimizes the Kullback-Leibler divergence between $\{p_{ij}^{(\ell)}\}_{j\neq i}$ and $\{q_{ij}^{(\ell)}\}_{j\neq i}$ using gradient descent:

$$KL(P_\ell||Q_\ell) = \sum_{i,j=1}^{n_\ell} p_{ij}^{(\ell)} \log\left(\frac{p_{ij}^{(\ell)}}{q_{ij}^{(\ell)}}\right), \tag{5.7}$$

The t-SNE method utilizes a "early exaggeration" phase to artificially highlights the attractions between points in similar neighborhoods, promoting clusters. This period is a very important tool that allows t-SNE to develop local structures in its representation. The early exaggeration phase occurs in the first 200 iterations of gradient descent in which $p_{ij}^{(\ell)}$ values are scaled by a factor of 4. It has been shown that the early exaggeration phase in t-SNE promotes clustering of similar points [26]. After the first 200 iterations, the $p_{ij}^{(\ell)}$ values are returned to their original value and t-SNE continues to perform gradient descent.

In the alignment step of AVIDA, GW-OT is used to align data points across data sets. Given the current low dimensional representations outputs from t-SNE, $Y^{(1)}$ and $Y^{(2)}$, the following optimization problem is solved to compute the transport matrix $\mathbf{T}$:

$$\begin{aligned}&\mathrm{GW}(Y^{(1)}, Y^{(2)})\\ &= \min_{\mathbf{T}} \sum_{i,j,i',j'} \|d(y_i^{(1)}, y_j^{(1)}) - d(y_{i'}^{(2)}, y_{j'}^{(2)})\|^2 \mathbf{T}_{i,i'}\mathbf{T}_{j,j'} - \epsilon(H(\mathbf{T})),\end{aligned} \tag{5.8}$$

where $H(\mathbf{T}) = \sum_{i,j} \mathbf{T}_{ij}\log(\mathbf{T}_{ij})$ is an Entropic regularization term and $d(\cdot,\cdot)$ is a chosen distance metric. The representation for $Y^{(1)}$ is mapped to $Y^{(2)}$ using the mapping found by minimizing (5.8), or by computing $Y^{(1)} = TY^{(2)}$. AVIDA$(X^{(1)}, X^{(2)}; \mathrm{TSNE}, \mathrm{GW})$ continues alternating between minimizing the KL loss in t-SNE and using optimal transport to align points until a stopping criteria is reached. In this implementation, we choose to limit the number of iterations to 1000 and perform alignment every 100 iterations after the early exaggeration phase (i.e., after the first 200 iterations) of t-SNE. The pseudo-code for AVIDA$(X^{(1)}, X^{(2)}; \mathrm{TSNE}, \mathrm{GW})$ is provided in Algorithm 3.

**Algorithm 3** AVIDA($X_1, X_2$; TSNE, GW)

**Input**: datasets $X^{(1)} = \{x_1^{(1)}, \ldots, x_{n_1}^{(1)}\}$, $X^{(2)} = \{x_1^{(2)}, \ldots, x_{n_2}^{(2)}\}$, perplexity $\rho$, and regularization parameter $\epsilon$
**Output**: low-dimensional representations: $Y_0^{(1)} = \{y_1^{(1)}, \ldots, y_{n_1}^{(1)}\}$, $Y_0^{(2)} = \{y_1^{(2)}, \ldots, y_{n_2}^{(2)}\}$
Compute pairwise affinities $p_{ij}^{(1)}$, $p_{ij}^{(2)}$ with perplexity $\rho$ (using Eq. (5.3) and Eq. (5.4))
Initialize solutions $Y_0^{(1)}, Y_0^{(2)}$ with points drawn i.i.d. from $\mathcal{N}(0,10^{-4}I)$
**while** $t < 1000$ **do**
  **if** $\mod(t, 100) \neq 0$ **then**
    **for** $\ell = 1, 2$ **do**
      Compute pairwise affinities $q_{ij}^{(\ell)}$ (using Eq. 5.6)
      Compute gradients $\Delta_t^{(\ell)} = \frac{\delta}{\delta Y_t^{(\ell)}} \text{TSNE}(X^{(\ell)}, Y_t^{(\ell)})$ (using Eq. 5.7)
      Set $Y_t^{(\ell)} = Y_t^{(\ell)} + \Delta_t^{(\ell)}$
    **end for**
  **else**
    Compute the GW-OT mapping, $\mathbf{T}$, between $Y_t^{(1)}$ and $Y_t^{(2)}$ (using Eq. 5.1)
    Set $Y_{(t+1)}^{(\ell)} = \mathbf{T}\, Y_t^{(\ell)}$
  **end if**
  $t \leftarrow t + 1$
**end while**

### 5.4.2 Metrics, parameters, hardware

The metrics used in Section 5.2 are described in detail in this section. For reproducibility, we also include the hardware settings under which these experiments were run and the user-selected parameters employed to obtain our numerical results.

#### Metrics

To compare AVIDA($X_1, X_2$; TSNE, GW), Pamona, and SCOT five different metrics are employed: fraction of samples closer than the true match (FOSCTTM), alignment, integration, accuracy, and representation loss. The FOSCTTM and alignment are metrics proposed in previous works. FOSCTTM was originally proposed by Liu et al. [27] and was used to validate the performance of SCOT. The alignment score was used in [10] to compare Pamona

and SCOT. In addition to the metrics used in previous works, we also introduce a few others to capture various aspects of the output representation. The additional metrics we measure are integration, accuracy, and representation loss. In this section, we define each and the conditions under which these metrics are meaningful. For notational simplicity, $D \in \mathbb{R}^{n_1 \times n_2}$ such that $D_{ij} = d(y_i^{(1)}, y_j^{(2)})$ denote the pairwise distance matrix between points in $Y^{(1)}$ and points in $Y^{(2)}$.

The FOSCTTM captures roughly the accuracy of the representation. FOSCTTM operates under the assumption that every point has a "true match" and that the "true matches" should be close together in the lower dimensional representation. This is formalized as follows. Assume, for simplicity, and $n_1 = n_2 = n$ and without loss of generality that the true match of $x_i^{(1)}$ is $x_i^{(2)}$ for all $i \in [n]$. The FOSCTTM is defined as:

$$\text{FOSCTTM} = \sum_{i=1}^{n} \frac{|\{j : D_{ij} < D_{ii}\}|}{n-1} + \sum_{j=1}^{n} \frac{|\{i : D_{ij} < D_{jj}\}|}{n-1}. \tag{5.9}$$

In other words, for each point $Y^{(1)}$, determine the fraction of the points $y_i^{(2)}$ that are closer to $y_i^{(1)}$ than $y_i^{(2)}$. Then, repeat the process for points in $Y^{(2)}$. Smaller values of FOSCTTM indicate better performance.

Under these same assumptions (that every point has a true match), we can also define an accuracy score. The idea is that points that are true matches should appear close together in the lower dimensional representation. This is measured by taking a simple trace of the matrix $D$:

$$\text{Accuracy} = \sum_{i=1}^{n} D_{ii} = tr(D)$$

The Alignment score used in this work was also used in [10]. The alignment score measures how well aligned the two datasets being integrated are in low dimensions. For the alignment score, we assume that each data set has class labels and that those class labels can be shared

across data sets. The points in each data set are split into "shared" and "dataset specific". "Shared" data points have representation in both $Y^{(1)}$ and $Y^{(2)}$ whereas "dataset specific" data points only appear in one of the datasets. The alignment score is computed as follows. Let $S^{(1)} \cup P^{(1)} = Y^{(1)}$ and $S^{(2)} \cup V^{(2)} = Y^{(2)}$ where sets $S^{(\ell)}$ denote the set of all points corresponding to "shared" data points and $V^{(\ell)}$ denote the set indices of all dataset specific points in $Y^{(\ell)}$. The alignment score is defined as:

$$\text{Alignment} = 1 - \frac{|\bar{x}_s - k/(\ell+1)|}{k - k/(\ell+1)},$$

where $\bar{x}_s$ is the average number of nearest neighbors that are shared points from the same dataset.

The aforementioned metrics have been utilized in previous works. We also propose to use the following for evaluating the representation of the low dimensional data. First, we employ a symmetrized Kullback-Leibler loss with a student t-distribution kernel to evaluate how well the output represents the high dimensional data in an integrated fashion. We refer to this as the Representation Loss:

$$\begin{aligned}\text{Representation Loss} = &\frac{1}{2}\left(\text{KL}(X^{(1)}\|Y^{(1)}) + \text{KL}(Y^{(1)}\|X^{(1)})\right) \\ &+ \frac{1}{2}\left(\text{KL}(X^{(2)}\|Y^{(2)}) + \text{KL}(Y^{(2)}\|X^{(2)})\right).\end{aligned}$$

The choice of this representation loss as a way to measure the quality of the representation in 2D is based on the fact that popular data dimensionality reduction techniques such as UMAP and t-SNE, both use a version of the KL loss. We recognize that there are other dimensionality reduction techniques, such as PCA or Laplacian Eigenmaps. However such techniques are spectral methods whose loss functions are evaluated by manifold-based metrics similar to FOSCTTM (5.9) and Integration (5.10). This representation loss is a way to measure the quality of the representation in cases of structures that are not best described

by the alignment of nearest neighbors, such as clusters or rings. Since t-SNE and UMAP are most adept at preserving these structures in low dimensions, it seems natural to modify their loss function as a way to measure the quality of the 2D representations.

Lastly, we want to evaluate how well integrated the two data sets are in low dimensions. We say that integration is the average, minimum distance between a data point in $Y_1$ and any data point in $Y_2$. The integration is defined as:

$$\text{Integration} = \frac{1}{n_1}\sum_{i=1}^{n_1} \min_j D_{ij} + \frac{1}{n_2}\sum_{j=1}^{n_2} \min_i D_{ij}. \tag{5.10}$$

**Persistent homology**

Persistent homology [16, 54] is used to evaluate the conservation of local geometries of the synthetic datasets. On a point cloud, a filtration of a simplicial complex $K$ such that $\emptyset = K^0 \subset K^1 \subset \cdots \subset K^m = K$ is constructed based on certain rules such as the Vietoris-Rips filtration, which we employ here. For each simplicial complex $K^i$, the rank of the $k$th homology group $H_k(K^i)$ represents the $k$th Betti number of $K^i$. For the examples here, we focus on the 1st homology group which represents the 1-dimensional holes in the data such as loops and rings. Along the filtration, the appearance and disappearance of these homology groups are tracked by computing the $p$-persistent $k$th homology group of $K^i$, $H_k^p(K^i)$ which records the homology classes of $K^i$ that persist at least until $K^{i+p}$. Each homology class is then represented by a pair of filtration values at which the class appears and disappears, usually called the birth and death values. These outputs of persistent homology can be visualized as persistence diagrams by taking the birth and death values as 2D coordinates. A more persistent homology class (with a large difference between death and birth values or equivalently farther away from the diagonal in the persistence diagram plots) is considered a significant feature. For the examples here, we are interested in the significant 1-dimensional loops which are captured as significant off-diagonal points in the $H_1$ persistence diagram.

We refer interested readers to [15] for complete details of persistent homology. Here, the package Dionysus 2 [32] was used for persistent homology computation with Vietoris-Rips filtration on Euclidean distance.

#### Parameters

The default perplexity value in most standard implementations of t-SNE is 30. However, depending on the dataset, the perplexity value may need to be adjusted. Table 5.2 shows the perplexity value choices for each experiment presented in Section 5.2. In addition to

| **Dataset** | **Bifurcated Tree** | **Circular Frustrum** | **Dumbbell** | **Distant Rings** | **sc-GEM** | **scNMT-seq** |
|---|---|---|---|---|---|---|
| **Perplexity Value** | 30 | 60 | 30 | 30 | 50 | 100 |

Table 5.2: Perplexity choices for each dataset.

perplexity, another important parameter is $\varepsilon$ in Equation 5.1. For all of our experiments, $\varepsilon$ was set to be $5 \times 10^{-3}$ but depending on the dataset could be adjusted.

#### Hardware

We ran the experiments on an Intel i7-10750H CPU (base frequency 2.60GHz) with 8GB memory.

### 5.4.3 Datasets

For our analysis, we introduced two synthetic datasets: the dumbbell dataset and distant rings dataset. The dumbbell dataset consists of two sub-datasets, $X^{(d,1)}, X^{(d,2)} \subset \mathbb{R}^2$ with

200 datapoints each. For all $0 \leq i \leq 200$,

$$X_{i,1}^{(d,1)} \sim 50U(0,1)$$
$$X_{i,2}^{(d,1)} \sim N(0,1)$$

where $U(0,1)$ is the uniform distribution and $N(0,1)$ is the normal distribution. This essentially constructs $X^{(d,1)}$ as a line in 2D with a little bit of noise. To construct the two rings in $X^{(d,2)}$, we consider $\theta \sim U(0,2\pi)$ and $r \sim N(3,0.5)$, then use it in our construction.

$$X_{i,1}^{(d,2)} \sim r\cos(\theta), \quad 1 \leq i \leq 50$$
$$X_{i,2}^{(d,2)} \sim r\sin(\theta), \quad 1 \leq i \leq 50$$
$$X_{i,1}^{(d,2)} \sim r\cos(\theta) + 14, \quad 50 < i \leq 100$$
$$X_{i,2}^{(d,2)} \sim r\sin(\theta), \quad 50 < i \leq 100$$

The first 50 points in $X^{(2)}$ are a slightly noisy circle centered at 0, where the next 50 points in the dataset are the same slightly noisy circle centered instead at 14. These two rings are then connected by a line.

$$X_{i,1}^{(d,2)} \sim U(3,10), \quad 100 < i \leq 200$$
$$X_{i,2}^{(d,2)} \sim N(0,0.2), \quad 100 < i \leq 200$$

This line is the last 100 points and also has small noise across one dimension.

The distant rings dataset also contains two subdatasets, $X^{(c,1)}, X^{(c,2)} \subset \mathbb{R}$. Again, we let $\theta \sim U(0,2\pi)$ and now we define $r_1 \sim N(5,1)$ and $r_2 \sim N(5,0.1)$ and define two different

rings.

$$X_{:,1}^{(c,1)} \sim r_1 \cos(\theta)$$
$$X_{:,2}^{(c,1)} \sim r_1 \sin(\theta)$$
$$X_{:,1}^{(c,2)} \sim r_2 \cos(\theta) + 100$$
$$X_{:,2}^{(c,2)} \sim r_2 \sin(\theta) + 100$$

Essentially for each dataset, we construct two rings where the distance between them dwarfs the radius of each ring. To make these two rings distinct, we constructed one ring to have much less noise than the other.

# Bibliography


[1] W. M. Abdelmoula, B. Balluff, S. Englert, J. Dijkstra, M. J. T. Reinders, A. Walch, L. A. McDonnell, and B. P. F. Lelieveldt. Data-driven identification of prognostic tumor subpopulations using spatially mapped t-sne of mass spectrometry imaging data. *Proceedings of the National Academy of Sciences*, 113(43):12244–12249, 2016.

[2] R. Argelaguet, S. J. Clark, H. Mohammed, L. C. Stapel, C. Krueger, C.-A. Kapourani, I. Imaz-Rosshandler, T. Lohoff, Y. Xiang, C. W. Hanna, et al. Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature*, 576(7787):487–491, 2019.

[3] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[4] S. Arora, W. Hu, and P. K. Kothari. An analysis of the t-sne algorithm for data visualization. In *Conference On Learning Theory*, pages 1455–1462. PMLR, 2018.

[5] R. Bellman. *A Brief Introduction to Theta Functions*. Holt, Rinehart and Winston, 1961.

[6] M. Bernstein, V. De Silva, J. C. Langford, and J. B. Tenenbaum. Graph approximations to geodesics on embedded manifolds. Technical report, Citeseer, 2000.

[7] J. D. Buenrostro, B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, 2015.

[8] T. T. Cai and R. Ma. Theoretical foundations of t-sne for visualizing high-dimensional clustered data. *arXiv preprint arXiv:2105.07536*, 2021.

[9] Z. Cang and Q. Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature communications*, 11(1):1–13, 2020.

[10] K. Cao, Y. Hong, and L. Wan. Manifold alignment for heterogeneous single-cell multi-omics data integration using pamona. *Bioinformatics*, 38(1):211–219, 2022.

[11] L. F. Cheow, E. T. Courtois, Y. Tan, R. Viswanathan, Q. Xing, R. Z. Tan, D. S. Tan, P. Robson, Y.-H. Loh, S. R. Quake, et al. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nature methods*, 13(10):833–836, 2016.

[12] N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014.

[13] P. Demetci, R. Santorella, B. Sandstede, W. S. Noble, and R. Singh. Scot: Single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology*, 29(1):3–18, 2022.

[14] V. H. Do and S. Canzar. A generalization of t-sne and umap to single-cell multimodal omics. *Genome Biology*, 22(1):1–9, 2021.

[15] H. Edelsbrunner and J. L. Harer. *Computational topology: an introduction*. American Mathematical Society, 2022.

[16] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. In *Proceedings 41st annual symposium on foundations of computer science*, pages 454–463. IEEE, 2000.

[17] S. Ferradans, N. Papadakis, G. Peyré, and J.-F. Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

[18] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[19] M. Forcato, O. Romano, and S. Bicciato. Computational methods for the integrative analysis of single-cell data. *Briefings in bioinformatics*, 22(3):bbaa042, 2021.

[20] Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck III, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

[21] B. Hie, B. Bryson, and B. Berger. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nature biotechnology*, 37(6):685–691, 2019.

[22] M. S. Jain, K. Polanski, C. D. Conde, X. Chen, J. Park, L. Mamanova, A. Knights, R. A. Botting, E. Stephenson, M. Haniffa, et al. Multimap: dimensionality reduction and integration of multimodal data. *Genome biology*, 22(1):1–26, 2021.

[23] L. Jimenez and D. Landgrebe. Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 28(1):39–54, 1998.

[24] L. Kantorovitch. On the translocation of masses. *C. R. (Doklady) Acad. Sci. URSS (N.S.)*, 37(199-201), 1942.

[25] D. Lahat, T. Adali, and C. Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

[26] G. C. Linderman and S. Steinerberger. Clustering with t-sne, provably. *SIAM Journal on Mathematics of Data Science*, 1(2):313–332, 2019.

[27] J. Liu, Y. Huang, R. Singh, J.-P. Vert, and W. S. Noble. Jointly embedding multiple single-cell omics measurements. In *Algorithms in bioinformatics:... International Workshop, WABI..., proceedings. WABI (Workshop)*, volume 143. NIH Public Access, 2019.

[28] B. Lorincz, A. Stan, and M. Giurgiu. An objective evaluation of the effects of recording conditions and speaker characteristics in multi-speaker deep neural speech synthesis. *ArXiv*, abs/2106.01812, 2021.

[29] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[30] F. Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

[31] G. Monge. Memoire sur la theorie des deblais et des remblais. *Histoire de l'Academie Royale des Sciences*, (666-704), 1781.

[32] Morozov, Dmitriy. Dionysus 2.

[33] T. Nagano, Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay, and P. Fraser. Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013.

[34] M. Nitzan, N. Karaiskos, N. Friedman, and N. Rajewsky. Gene expression cartography. *Nature*, 576(7785):132–137, 2019.

[35] K. O'Connor, B. Yi, K. McGoff, and A. B. Nobel. Graph optimal transport with transition couplings of random walks. *arXiv preprint arXiv:2106.07106*, 2021.

[36] K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[37] G. Peyré, M. Cuturi, and J. Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672. PMLR, 2016.

[38] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[39] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[40] T. Stuart and R. Satija. Integrative single-cell analysis. *Nature reviews genetics*, 20(5):257–272, 2019.

[41] M. J. Stubbington, O. Rozenblatt-Rosen, A. Regev, and S. A. Teichmann. Single-cell transcriptomics to explore the immune system in health and disease. *Science*, 358(6359):58–63, 2017.

[42] V. Svensson, R. Vento-Tormo, and S. A. Teichmann. Exponential scaling of single-cell rna-seq in the past decade. *Nature protocols*, 13(4):599–604, 2018.

[43] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[44] W. S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

[45] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[46] M. Verma, G. Matijevič, C. Denker, A. Diercke, E. Dineva, H. Balthasar, R. Kamlah, I. Kontogiannis, C. Kuckein, and P. S. Pal. Classification of high-resolution solar h$\alpha$ spectra using t-distributed stochastic neighbor embedding. *The Astrophysical Journal*, 907(1):54, 2021.

[47] C. Villani. *Topics in optimal transportation*. American Mathematical Society, Providence, RI, 2003.

[48] D. E. Wagner, C. Weinreb, Z. M. Collins, J. A. Briggs, S. G. Megason, and A. M. Klein. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 360(6392):981–987, 2018.

[49] J. D. Welch, V. Kozareva, A. Ferreira, C. Vanderburg, C. Martin, and E. Z. Macosko. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887, 2019.

[50] L. Zhang and Q. Nie. scmc learns biological variation through the alignment of multiple single-cell genomics datasets. *Genome biology*, 22(1):1–28, 2021.

[51] S. Zhang and H. Tong. Network alignment: recent advances and future directions. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3521–3522, 2020.

[52] Y. Zhang and S. Steinerberger. t-sne, forceful colorings and mean field limits. *arXiv preprint arXiv:2102.13009*, 2021.

[53] C. Zhu, S. Preissl, and B. Ren. Single-cell multimodal omics: the power of many. *Nature methods*, 17(1):11–14, 2020.

[54] A. Zomorodian and G. Carlsson. Computing persistent homology. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 347–356, 2004.

# Appendix A

# AVIDA and UMAP

## A.1 Using Alternate Dimensionality Reduction Techniques

We introduce AVIDA as a framework that allows for different methods for dimension reduction (or visualization) and alignment can be used depending on the dataset and applications. UMAP is another common dimensionality reduction technique utilized in computational biology. Here, we demonstrate AVIDA using UMAP for the dimension reduction module and GW-OT for the alignment module. The purpose of these brief experiments is to demonstrate AVIDA's viability as a framework. The experiments here essentially replicate a small subset of the experiments presented in the main section of our paper with the main difference being the utilization of UMAP for dimension reduction instead of t-SNE. To create 2D representations for SCOT and Pamona, we also used UMAP.

In Figure A.1, we apply AVIDA($X_1, X_2$; UMAP, GW) to the sc-GEM dataset, a dataset from [11] which contains both gene expression and DNA methylation at multiple loci on

human somatic cell samples under coversion to induced pluripotent stem cells. We can see comparing Figure A.1 (which uses UMAP for dimension reduction) with Figure 5.4 (which uses t-SNE for dimension reduction), using UMAP produces nearly the same clusters, but here we see a more distinct separation between the two point clouds, both for AVIDA and for Pamona. This shows that there may be datasets where another dimensionality reduction technique might be superior over other choices. However, the reverse can also be true.

In Figure A.2 we apply AVIDA($X_1, X_2$; UMAP, GW) to the rings data set described in Section 5.4.3 and see that using UMAP does not preserve the local structure as well as using t-SNE, as shown in Figure 5.3, for all three of the data integration methods. It is not surprising different dimensionality reduction techniques for the same dataaset will produce different representations and we encourage any users of AVIDA to incorporate the dimensionality reduction technique that works best on the dataset they are working with.

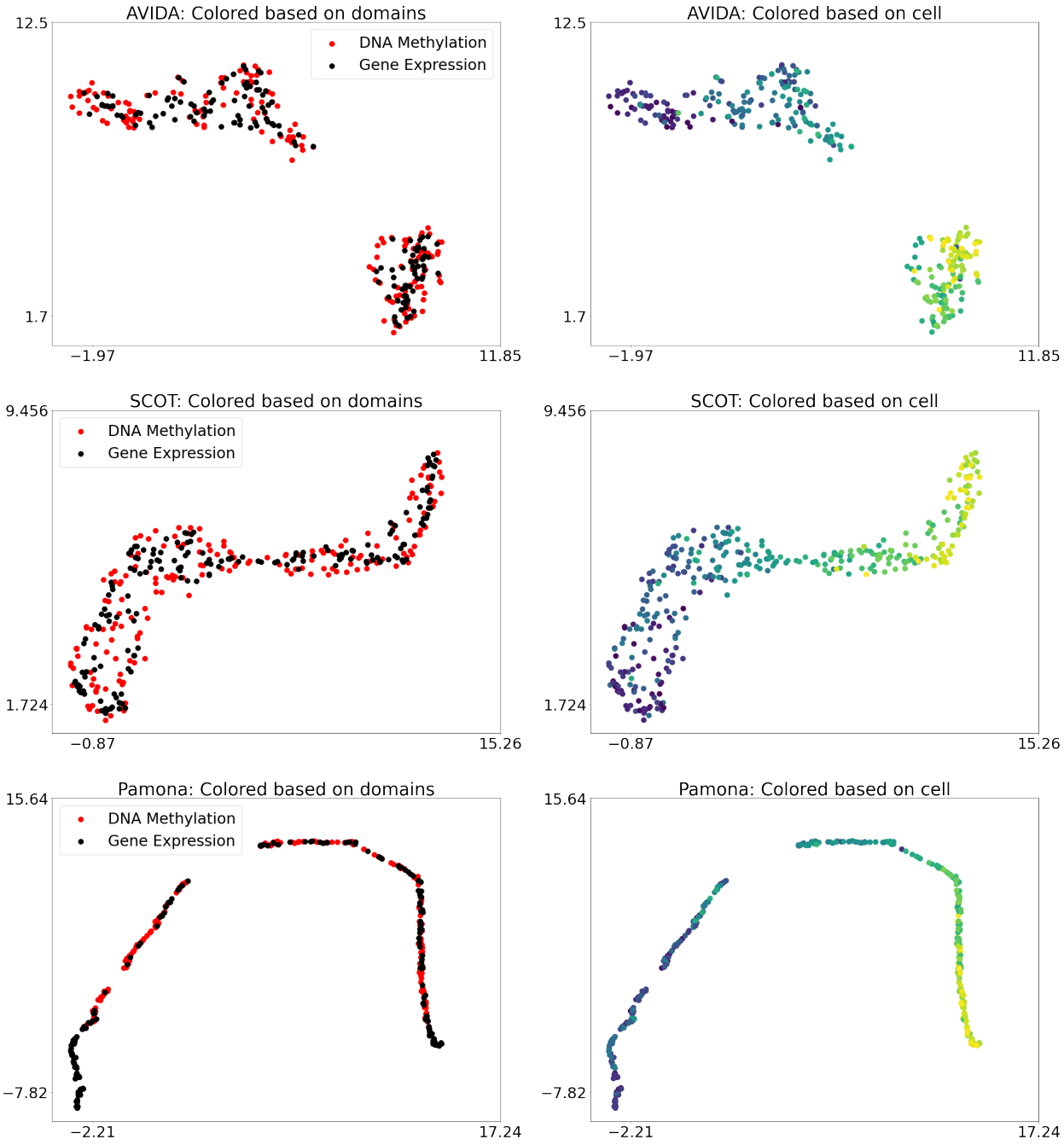


Figure A.1: AVIDA, SCOT, and Pamona representation of the scGEM dataset. In this experiment, UMAP was applied to SCOT and Pamona's output and UMAP's gradient was incorporated into AVIDA for the dimension reduction module.

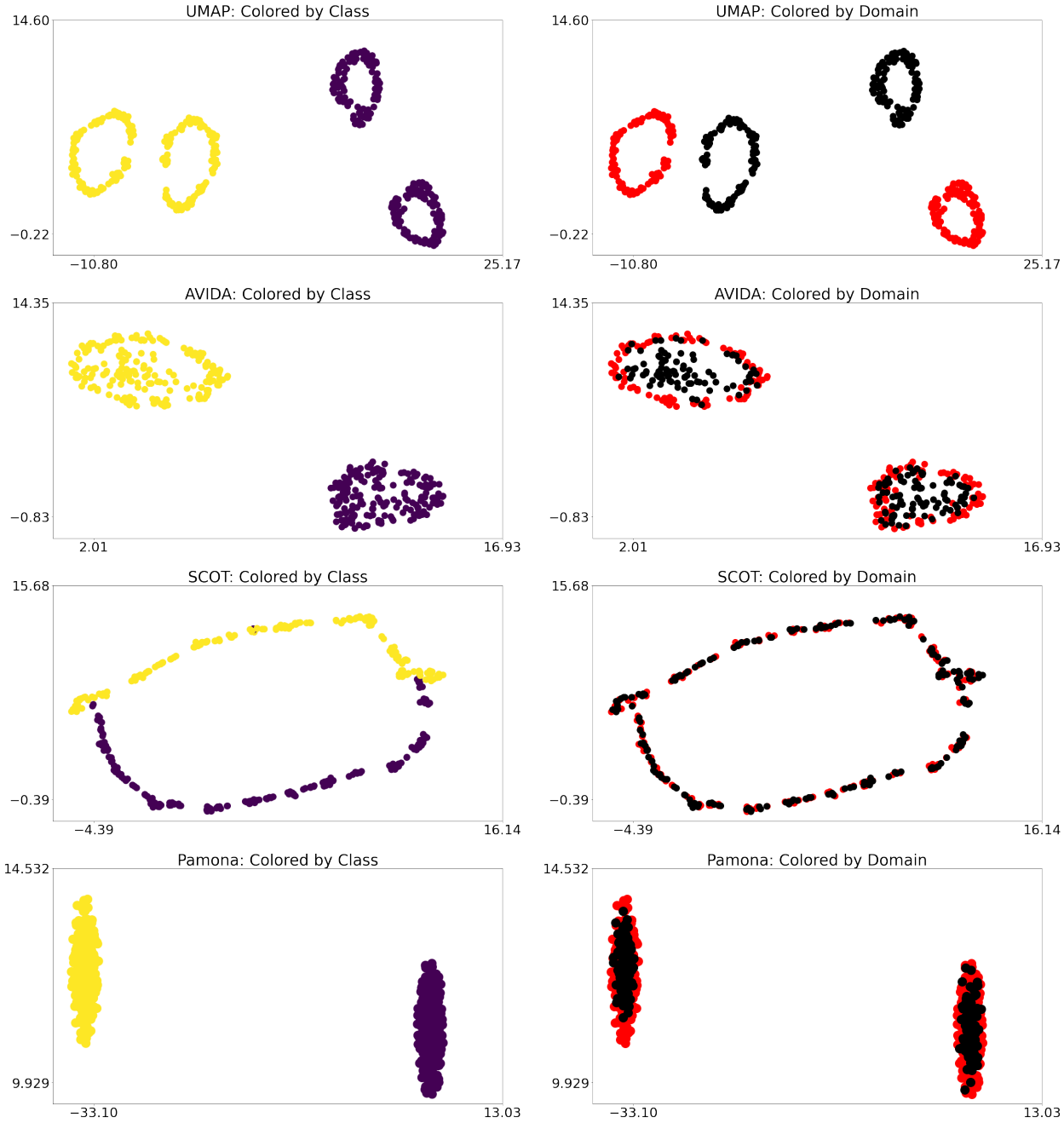


Figure A.2: UMAP, AVIDA, SCOT and Pamona representation of the distant rings dataset. In this experiment, UMAP was applied to SCOT and Pamona's output and UMAP's gradient was incorporated into AVIDA for the dimension reduction module.