t-SNE performance on cartoon data

T. Agarwal, S. Newton, P. Sasan, G. Tendolkar

1 Introduction

The 21st century is indeed the age of Big Data. Many datasets are so large that existing computing resources are insufficient to process them. While there have been rapid improvements on the hardware front, the software implementations are still trying to catch up. One of the challenges arising from these huge datasets is the sheer quantity of associated features they contain. Therefore, some tools are needed to understand the data and extract the essential features to make relevant predictions. Also, most large-scale datasets are unlabeled and an effective unsupervised machine learning technique is frequently needed.

Fortunately, there have been several successful approaches to deal with these challenges, and two of the most effective approaches have been dimensionality reduction and clustering. t-SNE, an approach introduced by (Maaten & Hinton, 2008), is a highly effective dimensionality reduction technique that also helps to visualize local clustering in the data because of its inherent design. Hence, t-SNE has been widely adapted as the go-to method to visualize high-dimensional data in 2 or 3 dimensions.

Despite its popularity, it is important to note that t-SNE is a heuristics-based algorithm. In fact, until very recently, there was no theoretical analysis of the algorithm at all. So, it warrants some theoretical analysis and that is the objective of this work. The paper is structured as follows.

First, the original algorithm is explained. Then the analysis of properties of t-SNE on clustered data are reviewed using approach from (Linderman & Steinerberger, 2017). Then we discuss (Arora, Hu, & Kothari, 2018) which extends the results from (Linderman & Steinerberger, 2017) to present a more rigorous analysis of t-SNE. Furthermore, the performance of t-SNE using different objective functions, in particular various f-divergences, is explored using the results from (Im, 2018). Finally, we try to prove theoretically, whether t-SNE preserves the structure of data which lies on a circle in the input space. We also tried to extend the analysis of (Linderman & Steinerberger, 2017) for the same type of input data.

2 Background

2.1 t-SNE: The Algorithm

The original work by (Maaten & Hinton, 2008) is summarized in this section. Consider a given dataset $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ such that $x_i \in \mathbb{R}^d$ for all $i \in [N]$. The algorithm aims to produce a 2 or 3 dimensional embedding of $\mathcal{X}$, say $\mathcal{Y} = \{y_1, y_2, ..., y_N\}$, such that $y_i \in \mathbb{R}^q$ for all $i \in [N]$ where usually $q \in \{2, 3\}$. This makes it easy to visualize the data and draw some important conclusions from it. The algorithm takes a probabilistic approach for the task. It converts the Euclidean distances between points, say x_i and x_j, into conditional probabilities $p_{j|i}$. These probabilities basically reflect similarity between these points, and therefore a Gaussian Kernel normalized to 1 is used:

$$p_{j|i} = \frac{exp(\frac{-||x_i - x_j||^2}{2\sigma_i^2})}{\sum_{j \in [n] \setminus \{i\}} exp(\frac{-||x_i - x_j||^2}{2\sigma_i^2})} \tag{1}$$

where σ_i controls the bandwidth in different regions of the high dimensional space. Assume σ_i induces a conditional Gaussian distribution P_i around the point x_i. Then, the suggested way of choosing σ_i, by the authors, is to perform a binary search such that the resulting P_i always has a fixed perplexity, $Perp(P_i)$ where

$$Perp(P_i) = 2^{H(P_i)}$$

with $H(P_i)$ being the Shannon-entropy of P_i in bits evaluated as

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}$$

This perplexity is defined by the user and can be roughly thought of as the number of neighbors each point x_i is expected to have. Therefore, one may expect a high perplexity value to favor the representation of the global structure in the embedding and a lower value to capture the local structure better. The usual values suggested for perplexity are from 5 to 50. Although, the authors state that the t-SNE algorithm is fairly robust to changes in perplexity values, practitioners of the algorithm disagree. In fact, the algorithm is especially sensitive to changes in this perplexity hyperparameter. Some supporting empirical findings can be found in the article (Wattenberg, Viégas, & Johnson, 2016).

In a similar way to (1), $q_{j|i}$ can be defined. The final target will be to match the conditional distributions P_i and Q_i (Q_i being the conditional distribution around the point y_i) for all $i \in [N]$, as closely as possible by minimizing a cost function. One obvious candidate for this task is the KL-divergence between the 2 distributions. Hence, the resulting cost function is:

$$C = \sum_i KL(P_i|Q_i) = \sum_i \sum_j p_{j|i} log(p_{j|i}|q_{j|i}) \tag{2}$$

Due to the asymmetry of the KL-divergence cost, this function, also called the SNE cost function (Hinton & Roweis, 2003), primarily focuses on preserving the local structure of the data in the low dimensional map. As mentioned earlier, this may be partly compensated by a high value of perplexity. The authors of (Maaten & Hinton, 2008) point out that this cost function is difficult to optimize numerically using Gradient-Descent, even after availability of a rich set of heuristics. They propose a different approach instead. The idea is to obtain pairwise joint distributions P and Q instead of conditional distributions P_i and Q_i by estimating some p_{ij} and q_{ij} and then minimizing the KL-divergence between P and Q, i.e.

$$C_J = KL(P|Q) = \sum_i \sum_j p_{ij} log(p_{ij}|q_{ij}) \tag{3}$$

The straightforward approach to do this would be to define p_{ij} and q_{ij} as follows:

$$p_{ij} = \frac{exp(\frac{-||x_i - x_j||^2}{2\sigma^2})}{\sum_{k,l \in [N], k \neq l} exp(\frac{-||x_l - x_s||^2}{2\sigma^2})} \tag{4}$$

$$q_{ij} = \frac{exp(-||y_i - y_j||^2)}{\sum_{k,l \in [N], k \neq l} exp(-||y_l - y_s||^2)} \tag{5}$$

Where q_{ij} values are given a constant $\sigma = (\sqrt{2})^{-1}$.

But, such a definition runs into problems. The first problem is with the definition of p_{ij}. For an outlier point x_i, all the values p_{ij} would be extremely small and so the location of corresponding y_i has negligible effect on the cost function. The problem is resolved by using a heuristic-based definition of p_{ij} such that $p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}$. This ensures $\sum_j p_{ij} > \frac{1}{2N}$ for all points x_i making the contribution of all x_i, $i \in [N]$, to be significant. Also, note that these p_{ij} form a valid probability distribution which sums to 1.

The second problem is overcrowding of points in the low-dimensional space. Simply put, the volume of the low-dimensional space that is available to accommodate moderately distant points x_i, for $i \in [N]$, is not be nearly as large enough as compared with the volume available to accommodate same points in the high-dimensional space. As a result, most of the points y_i, for $i \in [N]$, have a tendency to collapse into a very small region making it difficult to visualize clusters properly and draw useful conclusions. This problem is alleviated by using a Student t-distribution, which has a heavier tail compared to a Gaussian distribution, to model q_{ij}. The fundamental insight into this resolution is that moderately spaced points in the high-dimensional space have to be placed farther apart in the low-dimensional space to get a useful embedding. The heavier tail of the t-distribution

achieves that by assigning a higher value of the distance $||y_i - y_j||$ between any two points, for the same value of q_{ij}, as compared to the Gaussian distribution. The resulting modified definitions of p_{ij} and q_{ij} are as follows:

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N} \tag{6}$$

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k,l \in [N], k \neq l} (1 + ||y_l - y_s||^2)^{-1}} \tag{7}$$

Where the conditionals $p_{j|i}$ are obtained from (1).

The t-SNE algorithm then constitutes equations (6), (7) and (3). Another advantage of this approach is the simple form of the gradient update and the reduced computational cost to evaluate the density of a point under a t-distribution. The final gradient updates, as calculated in Appendix A of (Maaten & Hinton, 2008) are:

$$\frac{\partial C_J}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + ||y_i - y_j||^2)^{-1} \tag{8}$$

It was observed that the convergence rate decreased quickly as the number of points increased. To combat this (Maaten & Hinton, 2008) proposed a slight modification which they called the early exaggeration. For the first few steps instead of using p_{ij} as input they use αp_{ij} as input where $\alpha > 1$. It was observed experimentally that this helped the algorithm find the global structure of the dataset quickly and thus increased the convergence speed.

2.2 Analysis of the Algorithm

2.2.1 Analyzing the intra-cluster behavior

One of the most widely used applications of t-SNE is to observe clusters in the data. (Linderman & Steinerberger, 2017) proved that if clusters are present in the original data and if the points within the clusters are close to each other, then these points remain close to each other in the t-SNE embedding. For this, we assume there is a clustering function $\pi : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, k\}$ which assigns each data point to its corresponding cluster. The main result of the paper follows from the following:

Assumption 1 *(Linderman & Steinerberger, 2017)*
(i) If $\pi(i) = \pi(j)$ *, then* $p_{ij} \geq \frac{1}{10n|\pi^{-1}(\pi(i)|}$
(ii) $\frac{1}{100} \leq \alpha h$
(iii) $\sum_{\substack{i \neq j \\ same\ cluster}} p_{ij} \leq \frac{9}{10}$

The first assumption quantifies the acceptable distance of points within the same cluster. It says that points within the same cluster should be close to each other. The first assumption, however, is not something unique to clusters; there could be other geometrical objects satisfying the constraints.

Under these assumptions, they proved the following result using the tools of section 2.2.2.

Theorem 2.1 The diameter of the embedded cluster $\{y_j : 1 \leq j \leq n \ , \ \pi(j) = \pi(i)\}$ decays exponentially until its diameter satisfies, for some universal constant $c > 0$,

$$diam\{y_j : 1 \leq j \leq n \ , \ \pi(j) = \pi(i)\} \leq c \cdot h\left(\alpha \sum_{\substack{j \neq i \\ \text{other clusters}}} p_{ij} + \frac{1}{n}\right) \tag{9}$$

2.2.2 Discrete dynamical systems

Following the work of (Linderman & Steinerberger, 2017), define a discrete dynamical system by letting $z_1, \ldots, z_n \in \mathbb{R}^d$ be points such that $z_i(0) = 0$ and

$$z_i(t+1) = z_i(t) + \sum_{j=1}^{n} \alpha_{i,j,t}(z_j(t) - z_i(t)) + \epsilon_i(t) \tag{10}$$

If we do not put any constraints on α and ϵ, the system can move from any point to any other point. But imposing the following three conditions on them produces some useful results which for our problem. The 3 conditions are:

1. $|\alpha_{i,j,t}| \geq \delta > 0$
2. $\sum_{j=1}^{n} \alpha_{i,j,t} \leq 1$
3. $||\epsilon_i(t)|| \leq \epsilon$

Restricting α and ϵ with the above three constraints (Linderman & Steinerberger, 2017) prove the following results.

Lemma 2.2 (Stability of the convex hull). With the assumptions above, we have

$$conv\{z_1(t+1), z_2(t+1), \ldots, z_n(t+1)\} \subseteq conv\{z_1(t), z_2(t), \ldots, z_n(t)\} + B(0, \epsilon) \tag{11}$$

Lemma 2.3 (Contraction Inequality). If the diameter is large such that $diam\{z_1(t), z_2(t), \ldots, z_n(t)\} \geq \frac{10\epsilon}{n\delta}$, then

$$diam\{z_1(t+1), z_2(t+1), \ldots, z_n(t+1)\} \leq \left(1 - \frac{n\delta}{20}\right) diam\{z_1(t), z_2(t), \ldots, z_n(t)\} \tag{12}$$

If we can write the t-SNE update equation (8) in the form of a discrete dynamical system satisfying the above constraints, then the contraction inequality can prove theorem 2.1. Taking

- $\alpha_{i,j,t} = \alpha h p_{ij} q_{ij} Z$
- $\epsilon_i(t) = \sum_{\substack{j \neq i \\ \text{other clusters}}} \alpha h p_{ij} q_{ij} Z(y_j(t) - y_i(t)) - h \sum_{j \neq i} q_{ij}^2 Z(y_j(t) - y_i(t))$

we can show the t-SNE update is of the form the discrete dynamical system and (Linderman & Steinerberger, 2017) shows it satisfies the constraints and thus the theorem follows from the lemma.

(Linderman & Steinerberger, 2017) showed that if points within a fixed cluster are close to each other, then they will be close in the t-SNE embedding as well. However, this is not sufficient to guarantee a full visualization since it does not consider the distance between embedded clusters. These questions are addressed by (Arora et al., 2018). They also give a more formal definition of what it means to have a good visualization.

2.3 Extension to inter-cluster behavior

The previous section showed that if the data satisfies assumptions 1, then the diameter of embeddings of data within clusters decreases exponentially with t-SNE steps. But the proof does not guarantee a good global visualisation. For example, theorem 2.1 does not prove that clusters in embedded data corresponding to clusters in the input data are well separated. (Arora et al., 2018) extends the above theorem to show that if the original data is "well-separated" and "spherically clustered", then t-SNE with early exaggeration will output a "good visualisation".

Towards that goal, the paper first defines "well-separated", "spherically clustered" and "good visualisation", which are different, and more formal, from previous uses of such phrases. Then it reuses theorem 2.1, and proves a few more, to show that the diameters of embedded clusters are much smaller than the inter-cluster distances.

Let us say that for a given collection of points $\mathcal{X} = \{x_1, x_2, ..., x_n\} \in \mathbb{R}^d$, there exists a ground truth clustering described by k partitions $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_k$ of $[n]$. A visualisation is described by a 2-dimensional embedding $\mathcal{Y} = \{y_1, y_2, ..., y_n\} \in \mathbb{R}^2$, where each x_i corresponds to a y_i.

The paper first formalizes the definition of a "good visualisation".

Definition 1 *(Visible cluster) Let $\mathcal{Y}$ be a 2 dimensional embedding of $\mathcal{X}$ with ground truth clusters $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_k$. Given $\epsilon > 0$, a cluster $\mathcal{C}_l \in \mathcal{X}$ is said to be $(1-\epsilon)$ visible in $\mathcal{Y}$ if there exists partitions $\mathcal{P}, \mathcal{P}_{err} \in [n]$ such that:*
(i) $|(\mathcal{P} \setminus \mathcal{C}_l) \bigcup \mathcal{C}_l \setminus \mathcal{P}| < \epsilon |\mathcal{C}_l|$
(ii) $|\mathcal{P}_{err}| < \epsilon n$
(iii) for every $i, i' \in \mathcal{P}$ and $j \in [n] \setminus (\mathcal{P} \bigcup \mathcal{P}_{err})$, $\|y_i - y_{i'}\| \leq \frac{1}{2}\|y_i - y_j\|$
In such case, we say that $\mathcal{P}$ $(1-\epsilon)$-visualises $\mathcal{C}_i$ in $\mathcal{Y}$

Definition 2 *(Full visualisation) Given $\epsilon > 0$, we say that $\mathcal{Y}$ $(1-\epsilon)$-visualises $\mathcal{X}$ if there exists a partition $\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_k, \mathcal{P}_{err}$ of $[n]$ such that:*
(i) for each $i \in [k]$, $\mathcal{P}_i$ $(1-\epsilon)$-visualises $\mathcal{C}_i$ in $\mathcal{Y}$
(ii) $|\mathcal{P}_{err}| \leq \epsilon n$

Note that any 2-dimensional embedding that satisfies criteria for a full visualisation will be a good visualisation by most subjective human standards. But there could be other equally good or even better visualisations that fail to satisfy the full visualisation criteria. For example, points on two parallel lines in 2 dimensions would be a "good" lower dimensional representation for points on two parallel hyperplanes in higher dimension. But points on parallel lines in two dimension is not a full visualisation as defined in this paper. Nevertheless, the paper proves that t-SNE with early exagerration on data that is "well-separated" and "well-clustered", outputs a full visualisation.

Then (Arora et al., 2018) formalizes the definition of being well-separated and well-clustered.

Definition 3 *(Well-separated, spherical data) Let* $\mathcal{X} = \{x_1, x_2, ..., x_n\} \in \mathbb{R}^d$ *be clusterable data with* $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_k$ *defining individual clusters such that for each* $l \in [k], |\mathcal{C}_l| \geq 0.1(n/k)$*, we say that* $\mathcal{X}$ *is* γ*-well-separated and* γ*-spherical if for some* $b_1, b_2, ..., b_k > 0$*, we have:*
(i) (γ*-spherical) for any* $l \in [k]$ *and* $i, j \in \mathcal{C}_l (i \neq j)$*, we have* $\|x_i - x_j\|^2 \geq \frac{b_l}{1+\gamma}$*, and for any* $i \in \mathcal{C}_l$*, we have* $\left|\{j \in \mathcal{C}_l \setminus \{i\} : \|x_i - x_j\|^2 \leq b_l\}\right| \geq 0.5|\mathcal{C}_l|$
(ii) (γ*-well-separated) for any* $l, l' \in [k]$ $(l \neq l')$*,* $i \in \mathcal{C}_l$ *and* $j \in \mathcal{C}_{l'}$*, we have* $\|x_i - x_j\|^2 \geq (1 + \gamma \log n) \max\{b_l, b_{l'}\}$

Definition 3 says that to be γ-well-separated and γ-spherical means

- Cardinality of every cluster is lower bounded
- Distances between points of same cluster are concentrated around a single value
- Distances between points of different clusters are larger than intra-cluster distances of both clusters

The main theorem of the paper goes as follows,

Theorem 2.4 Let $\mathcal{X} = \{x_1, x_2, ..., x_n\} \in \mathbb{R}^d$ be γ-well-separated and γ-spherical, clusterable data with $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_k$ defining k individual clusters of size atleast $0.1(n/k)$, where $k << n^{1/5}$. Choose $\sigma_i^2 = \frac{\gamma}{4} \min_{j \in [n] \setminus \{i\}} \|x_i - x_j\|^2$ $(\forall i \in [n])$, $h = 1$ and any α satisfying $k^2 \sqrt{n} \log n << \alpha << n$. Let $\mathcal{Y}^T$ be the output of t-SNE after $T = \Theta(\frac{n \log n}{\alpha})$ iterations on input $\mathcal{X}$ with the above parameters. Then, with probability at least 0.99 over choice of initialisation, $\mathcal{Y}^T$ is a full visualisation of $\mathcal{X}$.

Note that this paper uses above defined σ_i as the standard deviations instead of using perplexity based calculations. α is the early exaggeration parameter and h is the learning rate.

2.3.1 Proof Outline

The proof first reuses the proofs in (Linderman & Steinerberger, 2017) to show that after $T = \Theta\left(\frac{\log 1/\epsilon}{\delta \eta}\right)$ iterations, the individual cluster diameters are upper bounded by $\mathcal{O}\left(\frac{\epsilon}{\delta \eta}\right)$. It then shows that after same $T = \Theta\left(\frac{\log 1/\epsilon}{\delta \eta}\right)$ iterations, the distance between centroids of embedding partitions

is lower bounded by $\Omega\left(\frac{1}{k^2\sqrt{n}}\right)$. To prove this, they use Berry-Esseen theorem to show that inter-centroid distances of clusters of a randomly initialized points is lower bounded and changes only a small amount in every gradient descent step of t-SNE. Moreover, for given ϵ, δ and η are parameters that are chosen such that $\frac{1}{k^2\sqrt{n}} >> \frac{\epsilon}{\delta\eta}$. Thus they show that inter-cluser distances are larger than intra-cluster distances for "well-clustered" and "well-separable" data.

2.4 Generalizing the objective function

Recall t-SNE minimizes KL-divergence between probability distributions modeling input data and output embedding. KL-divergence measures information loss from approximating the input distribution by the output distribution; however, there is no strong theoretical justification for using KL-divergence rather than some other measure of dissimiliarity between probability distributions.

It is reasonable to investigate f-divergences, of which KL-divergence is a particular instance. This is precisely what the paper (Im, 2018) does. The general form of an f-divergence is $D_f(P||Q)f(\frac{dP}{dQ})dQ$ for a probability distribution P absolutely continuous with respect to probability distribution Q. See (f divergence, 2019) for an overview of f-divergences. In our case with discrete P and Q, this reduces to form $\sum_{i\neq j} q_{ij} f(\frac{p_{ij}}{q_{ij}})$.

KL-divergence is thus an f-divergence with $f(x) = xlog(x)$. Common f-divergences are given in the following table (a larger table is available in (Im, 2018)):

Name	f	Objective
KL	$xlog(x)$	$D_{KL}(P\|\|Q) = \sum_{i\neq j} p_{ij} log(\frac{p_{ij}}{q_{ij}})$
Reverse KL	$-log(x)$	$D_{RKL}(P\|\|Q) = \sum_{i\neq j} q_{ij} log(\frac{q_{ij}}{p_{ij}})$
Hellinger Distance	$(\sqrt{x}-1)^2$	$D_{HL}(P\|\|Q) = \sum_{i\neq j} (\sqrt{p_{ij}} - \sqrt{q_{ij}})^2$

As the previous sections explain, t-SNE preserves clustering of sufficiently clustered data, and on toy data sets t-SNE appears to preserve low-dimensional structure. For instance, in (Wattenberg et al., 2016), t-SNE embeds a nice data set lying on a trefoil knot as a graph. Therefore, it is natural to ask what kinds of structure t-SNE preserves from high dimensional data to its low-dimensional embedding.

Unfortunately, the objective function for t-SNE, $\min_Y \sum_{i\neq j} p_{ij} log(\frac{p_{ij}}{q_{ij}})$ is highly non-convex and difficult to understand formally since every pairwise distance between data points contributes non-linearly to a term in the summation. Thus, as a first attempt at answering the above question, it is reasonable to consider a simplified model of t-SNE introduced in (Im, 2018).

To motivate the following simplified model of t-SNE, consider data lying on a non-self-intersecting curve. Each data point has two neighbors, one on either side. If t-SNE could perfectly preserve this local neighbor structure, we would expect it to output another curve, from which we could deduce that the data has strong 1-dimensional structure. At the other extreme, imagine data with three

obvious clusters which do not appear to have any structure internal to each cluster. If t-SNE could output three clusters, then it has succeeded in visualizing the structure of the data. Notice that we do not care whether the points in a cluster are mixed together haphazardly such that the local structure within each cluster is destroyed. These examples suggest that to understand t-SNE we should determine what happens to a small neighborhood of each data point.

From (Im, 2018), the simplified model considered is a binary neighborhood model. The idea is to simplify the t-SNE objective function by modifying the affinities p_{ij} and q_{ij} for each pair of data points such that p_{ij} is close to 1 if data point j lies in a small neighborhood of data point i and p_{ij} is close to 0 otherwise, and similarly for q_{ij}. Then the objective function weighs the contributions of nearby points much more heavily than far away points. While this obviously distorts the output of t-SNE, one hopes that it exaggerates what t-SNE is already doing, rather than introducing a new phenomenon.

Specifically, define $N_\epsilon(x_i) = \{x_j \mid p_{j|i} > \epsilon\}$ and $N_\epsilon(y_i) = \{y_j \mid q_{j|i} > \epsilon\}$, the points lying in a small neighborhood of x_i and y_i respectively for each i.

Fix $0 < \delta << 1$. For each i let

$$a_i \geq \frac{1-\delta}{|N_\epsilon(x_i)|}, \quad c_i \geq \frac{1-\delta}{|N_\epsilon(y_i)|}, \quad b_i \leq \frac{\delta}{n - |N_\epsilon(x_i)| - 1}, \quad d_i \leq \frac{\delta}{n - |N_\epsilon(y_i)| - 1}.$$

Finally, using these quantities, define

$$p_{ij} = \begin{cases} a_i & x_j \in N_\epsilon(x_i) \\ b_i & x_j \notin N_\epsilon(x_i) \end{cases}, \text{ and } q_{ij} = \begin{cases} c_i & y_j \in N_\epsilon(y_i) \\ d_i & y_j \notin N_\epsilon(y_i) \end{cases}.$$

The paper defines two important quantites: recall and precision. Define $recall(i) = \frac{|(N_\epsilon(x_i) \cap N_\epsilon(y_i))|}{|N_\epsilon(x_i)|}$ to be the ratio of neighbors of data point x_i that t-SNE preserves after embedding to total number of neighbors of x_i. Essentially, recall measures if local neighborhoods are preserved, without concern for whether non-neighbors become spurious neighbors upon embedding. Also define $precision(i) = \frac{|(N_\epsilon(x_i) \cap N_\epsilon(y_i))|}{|N_\epsilon(y_i)|}$ to be the ratio of neighbors of data point x_i that are also neighbors of y_i after embedding to the total number of neighbors of the embedded y_i. Precision roughly measures how distored local neighborhoods become by the addition of spurious neighbors in the embedding, without concern for whether neighbors are lost upon embedding.

Using this language, (Im, 2018) proves the following,

Proposition In the binary neighborhood model, for $0 < \delta << 1$,

1. $D_{KL}(P||Q) \propto \sum_i (1 - recall(i))$
2. $D_{RKL}(P||Q) \propto \sum_i (1 - precision(i))$

3. $D_{HL}(P||Q) \propto \sum_i \Big[(1 - recall(i)) \cdot \big(1 - O((\delta \cdot |N_\epsilon(x_i)|)^{\frac{1}{2}})\big) + (1 - precision(i)) \cdot \big(1 - O((\delta \cdot |N_\epsilon(y_i)|)^{\frac{1}{2}})\big) + precision(i) \cdot \big(\sqrt{\frac{|N_\epsilon(x_i)|}{|N_\epsilon(y_i)|}} - 1\big)^2\Big]$

Interpreting this theorem, one should not be surprised that KL-divergence tends to preserve clusters because high recall implies very few spurious neighbors after embedding. Interestingly, reverse KL-divergence favors high precision, which may preserve finer global structures in the data. Hellinger distance, as well as several other f-divergences considered in (Im, 2018), show an intermediate behavior that balances precision and recall. This diversity among f-divergences suggests that for exploratory data analysis, it may be desirable to run t-SNE using several f-divergences to see several perspectives of the data.

The binary neighborhood model suggests that the t-SNE optimal embedding of data lying on a circle may not be points lying on a circle; specifically, recall can be maximized, and hence KL-divergence minimized, by collapsing the entire dataset into a small neighborhood. However, for both empirical and intuitive reasons, it is expected that t-SNE preserves circular structure (see the Innovation section below). This reveals an obvious deficiency in the model, namely that it exaggerates the attractive nature of t-SNE with KL-divergence and underappreciates the repulsive contributions to the KL-divergence from data points being squished too close together in the t-SNE output. Despite this apparent drawback, given the tractibility of this model it is a promising starting point for proving behaviour of t-SNE and its variants using f-divergences other than KL-divergence.

3 Organization

Our group met roughly bi-weekly to discuss our progress and provide useful inputs to each other, especially regarding the innovation and class presentations. We spent the first month discussing the reference papers and deciding the structure and flow of our project. Each of us read all the reference papers but focused on one of them for our presentation. Before each presentation, the presenter gave a tutorial on his topic of focus to the other members and received critical feedback. The focus of each member is briefly described as follows.

Tushar focused on the fundamental understanding of the algorithm from the original papers (Hinton & Roweis, 2003) and (Maaten & Hinton, 2008). Hence, he was responsible for Introduction and the subsection 2.1. Prateek and Gaurav concentrated on the analysis of t-SNE on clustered data using approach from (Linderman & Steinerberger, 2017) and (Arora et al., 2018). Therefore, they worked on subsections 2.2 and 2.3 where they summarized their understanding. Scott explored different objective functions, in particular various f-divergences, following (Im, 2018). Hence, he wrote subsection 2.4 and tried to leverage symmetries in data to understand optimal theoretical outputs of t-SNE. The innovation and conclusion sections of the paper were a collective effort.

4 Innovation

A trivial yet important observation is that the t-SNE objective function depends only on the pairwise distances between high-dimensional data points and between embedded points. So, the output of t-SNE is invariant under ambient isometries of $\mathbb{R}^d$.

One idea is to use symmetry in a nice data set to prove that t-SNE preserves some structure. Then, to get a more general result allowing noisy data, one could try to show that under small perturbations of the data, the structure is still preserved. To tackle the problem of whether t-SNE preserves low-dimensional structure, we considered N data points lying equally spaced on a circle (i.e. as vertices of a regular N-gon).

Observe that the t-SNE objective function $D_{KL}(P||Q) = \sum_{i \neq j} p_{ij} log(\frac{p_{ij}}{q_{ij}})$ is preserved under the action of the dihedral group D_N on the labels of the input data. That is, for $\sigma \in D_N$, since $p_{\sigma(i)\sigma(j)} = p_{ij}$ by the symmetry of the input data,

$$\sigma.D_{KL}(P||Q) = \sigma.\sum_{i \neq j} p_{ij} log\left(\frac{p_{ij}}{q_{ij}}\right) := \sum_{i \neq j} p_{\sigma(i)\sigma(j)} log\left(\frac{p_{\sigma(i)\sigma(j)}}{q_{ij}}\right) = \sum_{i \neq j} p_{ij} log\left(\frac{p_{ij}}{q_{ij}}\right) = D_{KL}(P||Q).$$

So the optimal output, if unique, must also have D_N symmetry; that is, lie on a circle as the vertices of some regular N-gon. However, it is not clear that the t-SNE objective function has a unique global minimum. Interestingly, this argument does not depend on the use of KL-divergence, and holds under any f-divergence.

Another example data set with a lot of symmetry is the vertices of a regular N-simplex in $\mathbb{R}^d$ for $N \leq d$. The t-SNE objective function then has S_N symmetry, for S_N the symmetric group on N letters, since the distance between each pair of points is the same. But no set of points in $\mathbb{R}^2$ have such symmetry for $N > 3$. We were unable to determine what configuration minimizes the t-SNE objective function, but there should be as much symmetry as possible in an output that globally minimizes the objective function.

These simple questions need to be completely answered before tackling more general questions such as whether, given data lying on an embedded graph in high dimensions, t-SNE minimizes the crossing number of the embedded graph in $\mathbb{R}^2$. An arbitrary instance of this graph embedding problem will not have as much symmetry to leverage, so new techniques are needed to address it.

As an alternative approach, taking motivation from (Linderman & Steinerberger, 2017), the evolution of t-SNE embedding is explored using discrete-dynamical systems and input data point lying uniformly on a circle.

In our experiments using a simulator built by us (`https://codetendolkar.github.io/`), we observed a pattern that when we have equally spaced data points on a circle, in the first few iterations, the t-SNE embedding forms a disc where the points which were close to each other in the input get

(a) 20 Iterations **(b)** 100 Iterations **(c)** 200 Iterations

Figure 1: T-SNE visualization during optimization

closer in the embedding and the points which were far move farther apart. After that, it takes a number of iterations for the points to take a shape of a circle.

Taking motivation from (Linderman & Steinerberger, 2017), we tried to follow the evolution of t-SNE embedding using discrete-dynamical systems for the input data points as above. Under certain conditions on the distance between neighbors and perplexity, we can show mathematically that the points would follow the trend observed in the simulations for the first few iterations. Tracking the evolution beyond the first few iterations become very difficult and thus we were not able to show if the optimal t-SNE embedding would necessarily output a circle using this technique. Unfortunately since we were not able track t-SNE embedding evolution completely and unable to give any guarantees on the final t-SNE result for points on a circle we have omitted this insignificant result for brevity.

5 Conclusion

t-SNE remains a state-of-the-art approach to dimensionality reduction and data visualization, despite the continued lack of theoretical underpinnings. In this paper, we reviewed the current literature on t-SNE theory. The strongest result has shown that t-SNE preserves clusters under suitable conditions. Rigorous mathematical formalization of a heuristic based algorithm was one of the key take-aways from this project. Specifically, the assumptions defining a good visualization is an important first step towards characterizing the performance of t-SNE or any clustering algorithm. Also, even for simple cases where our data points lie in a low dimensional structure it is difficult to track the evolution of the t-SNE embedding as the number of steps increases.

More questions remain about the robustness of such a clustering result for real-world data that is not perfectly clustered, and the questions regarding preservation of other types of structure under t-SNE remain wide open.

References

Arora, S., Hu, W., & Kothari, P. K. (2018). An analysis of the t-sne algorithm for data visualization. *arXiv preprint arXiv:1803.01768*.

f divergence. (2019). *f-divergence — wikipedia, the free encyclopedia.* Retrieved from `https://en.wikipedia.org/wiki/F-divergence` ([Online; accessed 2-April-2019])

Hinton, G. E., & Roweis, S. T. (2003). Stochastic neighbor embedding. In *Advances in neural information processing systems* (pp. 857–864).

Im, B., Verma. (2018). Stochastic neighbor embedding under f-divergences. *arXiv preprint arXiv:1811.01247*.

Linderman, G. C., & Steinerberger, S. (2017). Clustering with t-sne, provably. *arXiv preprint arXiv:1706.02582*.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(Nov), 2579–2605.

Wattenberg, M., Viégas, F., & Johnson, I. (2016). How to use t-sne effectively. *Distill*. Retrieved from `http://distill.pub/2016/misread-tsne` doi: 10.23915/distill.00002