THE TREE-SNE TREE EXISTS

JACK KENDRICK

Abstract. The clustering and visualisation of high-dimensional data is a ubiquitous task in modern data science. Popular techniques include nonlinear dimensionality reduction methods like t-SNE or UMAP. These methods face the 'scale-problem' of clustering: when dealing with the MNIST dataset, do we want to distinguish different digits or do we want to distinguish different ways of writing the digits? The answer is task dependent and depends on scale. We revisit an idea of Robinson & Pierce-Hoffman that exploits an underlying scaling symmetry in t-SNE to replace 2-dimensional with (2+1)-dimensional embeddings where the additional parameter accounts for scale. This gives rise to the t-SNE tree (short: tree-SNE). We prove that the optimal embedding depends continuously on the scaling parameter for all initial conditions outside a set of measure 0: the tree-SNE tree exists. This idea conceivably extends to other attraction-repulsion methods and is illustrated on several examples.

1. Introduction and Results

1.1. **The problem of clustering.** From fields such as computer vision to computational linguistics, large sets of high dimensional data are increasingly commonplace. Thus, it is necessary to develop tools that can be used for their analysis and visualisation. While traditional linear techniques are often remarkably powerful, the visualisation problem poses a new type of challenge. Even if data in 1000 dimensions is intrinsically close to a 5-dimensional submanifold (the manifold hypothesis), embedding five dimensions into two or even three dimensions poses a considerable challenge. One popular dimensionality reduction technique is t-distributed stochastic neighbour embedding (t-SNE), first introduced by van der Maarten and Hinton in [MH08]. As a completely non-linear method, t-SNE does not try to produce an isometric or low-distortion embedding. Instead, it tries to group the data into clusters. This works remarkably well in practice and has been very widely used. Indeed, it has become apparent [BBK22] that a broad range of methods (including Laplacian eigenmaps [BN03], ForceAtlas2 [Jac+14], UMAP [MHM18; KL19] and t-SNE [MH08]) can be seen as special cases of a general attraction-repulsion method: we start with points in $\mathbb{R}^2$ that represent the high-dimensional data and then induce each point to move towards other points in $\mathbb{R}^2$ that represent 'similar' high-dimensional data. Simultaneously, points are repelled by all of the other data points. If the underlying data is highly clustered, then the same cluster structure in $\mathbb{R}^2$ is a steady state of that dynamical system, see [LS17; AHK18; CM22].

1.2. **The problem of scale.** Since t-SNE embeddings exhibit cluster structures that are representative of those seen in the high-dimensional data, a typical use case for t-SNE is producing a 2D visualisation that aids with identification of clusters within a dataset. However, the question of scale is often overlooked in this clustering process. Although often treated as a static problem, clustering may be viewed as a

dynamic problem with a hierarchical structure. As a motivating example, consider the MNIST dataset which contains 70,000 images of handwritten numerical digits. Typically, these images are clustered based on what digits they represent. However, this clustering is not helpful if we are, for example, trying to distinguish different handwriting styles within the dataset. In this case, a more granular clustering is needed. Since many digits can be written in multiple ways, we expect there to be subclusters corresponding to different handwriting styles within the clusters corresponding to each digit. Figure 1 gives a schematic of example clusters and subclusters within the MNIST dataset.

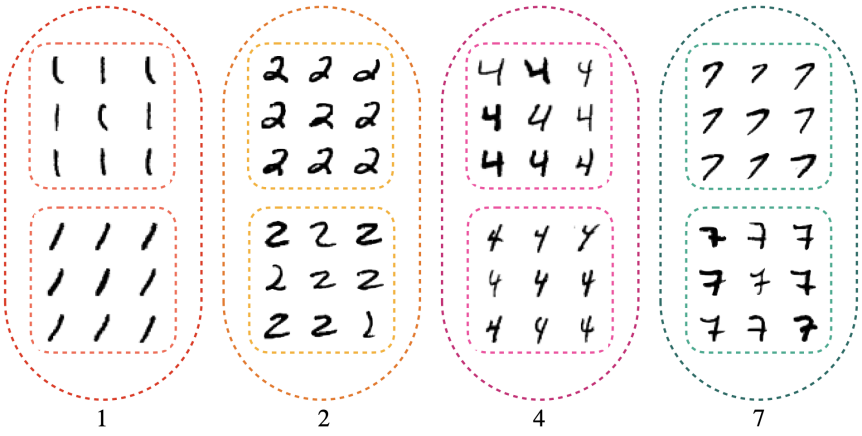

FIGURE 1. A schematic of subclusters in the MNIST dataset.

One solution to the problem of scale are *hierarchical clustering* methods, see [MC12; Nie16] for an overview. These methods are often computationally expensive and so an alternative method is desirable. It has been observed that variants of t-SNE using kernels with heavier tails tend to produce clusters with finer granularity, see for example [Kob+20]. However, there is currently no good way to choose a kernel that corresponds to a particular granularity outside of trial and error. Thus, the problem of scale persists within t-SNE embeddings and clusterings. Note that this is not exclusive to t-SNE and is a problem with *all* clustering methods yet is often ignored in practice.

1.3. **Tree-SNE.** To address the problem of scale, we explore a hierarchical variant of t-SNE known as tree-SNE that was introduced in [RPH20]. The idea is to exploit a one-parameter family [Kob+20] of kernels that produce t-SNE-type embeddings with finer and finer granularity. We begin with generating an ordinary t-SNE embedding of the data. Then, additional layers are generated by changing the parameter in the kernel and optimizing a certain loss function. Each embedding forms a layer of the tree-SNE embedding and the points in each layer can be interpolated to plot smooth trajectories of each of the data points. Figure 2 contains slices of a tree-SNE embedding of the MNIST dataset, along with the tree that comes from interpolating each layer.

We see in Figure 2 that each of the main branches of the tree splits into smaller branches as the layers increase. These correspond to different ways of writing

the same digit. In Figure 3, we show example branches and subbranches that correspond to the clusters and subclusters in Figure 1. Each branch is labelled with the average image of elements belonging to the cluster.

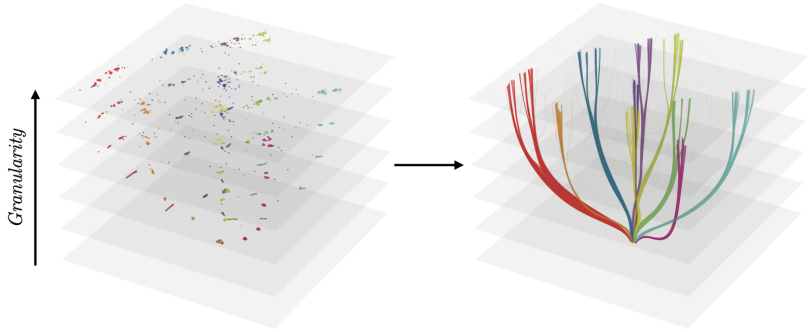

FIGURE 2. Slices of the MNIST tree-SNE embedding are interpolated to create a tree. Each color corresponds to a digit.

In particular, we see that as the layers of the embedding increase there are clusters that reveal structure with finer granularity and there is a continuous structure to the tree: the points follow a continuous trajectory and do not jump around from layer to layer. As has been previously observed in [Kob+20], t-SNE embeddings generated with heavier tailed kernels tend to produce finer clusters and so it is expected that clusters at the top of the tree-SNE embedding are more granular than those at the bottom. However, the continuous structure of the tree is non-trivial and will be the main focus of this paper.

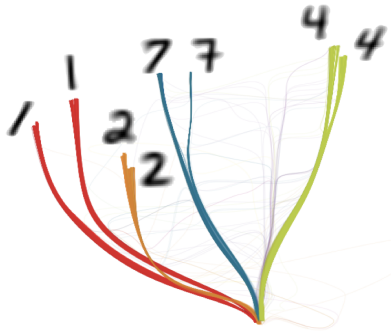

FIGURE 3. Subbranches of the MNIST tree correspond to different handwriting styles. Images are the average of elements in the corresponding cluster.

1.4. **Main Result.** Our main result concerns the continuous structure of tree-SNE embeddings, as observed in Figure 2. We prove that this structure is not coincidental and is a consequence of the tree-SNE algorithm's design. Recall that each layer of a tree-SNE embedding is generated using a kernel from a one parameter family. We show that as the parameter varies continuously, so does the embedding.

Theorem (The Tree-SNE Tree exists)**.** *Suppose* $(y_1, \dots, y_n) \in \mathbb{R}^{nd}$ *is a* t-SNE *embedding in* $\mathbb{R}^d$ *with parameter* α*, Under mild conditions, for each* α' *in some neighbourhood of* α *there is a* t-SNE *embedding* $(y'_1, \dots, y'_n) \in \mathbb{R}^{nd}$ *in a neighbourhood of* $(y_1, \dots, y_n)$*.*

To be more specific, we view the point $(\alpha, y_1, \dots, y_n) \in \mathbb{R}^{1+nd}$ as belonging to the zero set of the function F. The function F is the gradient of the loss function minimized to find t-SNE embeddings when the parameter α is not treated as a constant. The 'mild conditions' referred to in our theorem ensure that F is smooth and generically has a constant rank. This will allow us to conclude that the zero set of F is generically a smooth manifold of a dimension $1 + d(d+1)/2$ around $(\alpha, y_1, \dots, y_n)$, with one of these dimensions accounting for changes in the parameter α. These 'mild conditions' are necessary but are not restrictive. Indeed, they are satisfied by generic datasets and initializations. We provide a detailed discussion of these conditions in Section 3.

Organisation. In Section 2, we recount the necessary background information on t-SNE and describe the tree-SNE algorithm. In Section 3, we prove our main theorem to show that the tree-SNE algorithm indeed produces tree-like embeddings for generic datasets and generic initialisations. Finally, in Section 4 we use tree-SNE to visualise two different datasets from the HathiTrust library and examine the structures revealed in the embeddings.

2. Preliminaries

In this section, we give a brief overview of both t-SNE and tree-SNE. Throughout this section, we fix a dataset $\{x_1, \dots, x_n\} \subset \mathbb{R}^D$, where on may think of the dimension D as large. We begin by describing dimensionality reduction and data visualisation with t-SNE.

2.1. T-SNE.

Given a dataset $\{x_1, \dots, x_n\} \subset \mathbb{R}^D$, the central goal of t-SNE is to find a low-dimensional set of points $\{y_1, \dots, y_n\} \subset \mathbb{R}^d$ that preserves the overall structure of the data. In particular, if x_i and x_j are "close" in $\mathbb{R}^D$, then y_i and y_j should be "close" in $\mathbb{R}^d$ with high probability. Typically, the dimension d is chosen such that $d = 2$ or $d = 3$ and so the t-SNE embedding $\{y_1, \dots, y_n\}$ can be used to visualise the high-dimensonal dataset. The t-SNE method consists of two stages. First, *affinities* p_{ij} are determined between each pair of points x_i and x_j. The affinity p_{ij} between the points x_i and x_j is a probability that represents the similarity of x_i and x_j within the dataset: a high affinity is associated to pairs that are close together, whereas low affinities are assigned to pairs of points that are far apart. The affinity p_{ij} is determined using *directional affinities* $p_{j|i}$ and $p_{i|j}$. For $j \neq i$, the directional affinity $p_{j|i}$ is defined as

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

where the variance σ_i^2 is a user determined parameter. The directional affinity between x_i and itself is set to $p_{i|i} = 0$ for all i. Note that for each value of i, the directional affinities determine a probability distribution on $\{x_1, \dots, x_n\}$. The

variance σ_i^2 is chosen such that the *perplexity* of this distribution, defined

$$\mathcal{P} = \exp\left(-\log 2 \cdot \sum_{j \neq i} p_{j|i} \log_2 p_{j|i}\right),$$

has some pre-specified value. Typical values of the perplexity $\mathcal{P}$ are between 5 and 50. For each pair i, j the affinity p_{ij} is the normalised average of the directional affinities $p_{j|i}$ and $p_{i|j}$,

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

Once affinities between all pairs of points x_i and x_j have been determined, a low dimensional embedding of the data $\{y_1, \ldots, y_n\} \subset \mathbb{R}^d$ is found. For each pair of points $y_i, y_j \in \mathbb{R}^d$ the affinity q_{ij} is defined

$$q_{ij} = \frac{K(\|y_i - y_j\|)}{\sum_{k \neq \ell} K(\|y_k - y_\ell\|)}$$

where K is a kernel. Traditionally, the kernel K corresponds to the t-distribution with one degree of freedom, $K(d) = (1+d^2)^{-1}$. However, more general t-distributions can be considered. As in [Kob+20], we consider kernels given by

$$K_\alpha(d) = \frac{1}{(1 + d^{2/\alpha})^\alpha}$$

where $\alpha > 0$ is a parameter. This corresponds to a scaled t-distribution with $2\alpha - 1$ degrees of freedom. Note that choosing $\alpha = 1$ yields the standard t-distribution kernel and the taking the limit as $\alpha \to \infty$ gives the Gaussian kernel $K_\infty(d) = \exp(-d^2)$. This corresponds to the SNE method of Hinton and Roweis [HR02]. Moreover, $\alpha < 0.5$ corresponds to a t-distribution with negative degrees of freedom and so represents a distribution with heavier tails than any (non-scaled) t-distribution. Note that when $\alpha < 0.5$ the resulting distribution is no longer a probability distribution. A common misconception about the t-SNE method is that the kernel must correspond to a probability distribution, however this is not correct. Since we only consider a finite number of points, distributions may always be renormalized and it is not, for instance, necessary for any particular integral to be finite. As in the high-dimensional space, the affinity between y_i and itself is set as $q_{ii} = 0$. The embedding $\{y_1, \ldots, y_n\}$ is chosen so that the Kullbeck-Leibler (KL) divergence

$$\mathcal{L}(y_1, \ldots, y_n) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

is minimized. The gradient of $\mathcal{L}$ is given by

$$\frac{\partial \mathcal{L}}{\partial y_i} = \sum_{j=1}^{n} (p_{ij} - q_{ij}) K_\alpha(\|y_i - y_j\|)^{1/\alpha} (y_i - y_j)$$

and so an embedding that minimizes the loss function may be found using a random initialisation and applying first order methods such as gradient descent.

2.2. **Tree-SNE.** We now describe how tree-SNE builds on t-SNE to generate collections of embeddings that reveal finer substructures within data. The central idea of tree-SNE is that iteratively decreasing the degrees of freedom in the t-distribution kernel yields embeddings that capture the structure of data in increasing granularity. This sequence of embeddings can then be stacked atop one another and interpolated to reveal a tree-like structure with different branches corresponding to different subclusters in the data. Tree-SNE takes in a sequence of parameters $\{a^{(i)}\}_{i=1}^m$ and produces a sequence of embeddings $\{(y_1^{(i)}, \dots, y_n^{(i)})\}_{i=1}^m$. As with t-SNE, each embedding is produced in two stages. First, affinities between each pair x_i, x_j of data points in $\mathbb{R}^D$ are calculated so that the directional affinities at each point have some perplexity $\mathcal{P}^{(i)}$ that is dependent on $\alpha^{(i)}$. Then, the embedding $(y_1^{(i)}, \dots, y_n^{(i)})$ is a minimizer of the loss function $\mathcal{L}$. Each embedding $(y_1^{(i)}, \dots, y_n^{(n)})$ forms a layer of the tree-SNE embedding. The parameters $\{\alpha^{(i)}\}_{i=1}^m$ are chosen such that $\alpha^{(1)} = 1$ and $\alpha^{(i+1)} < \alpha^{(i)}$ for all i. Similarly, the perplexities $\mathcal{P}^{(i)}$ decrease as the layer of the embedding increases. Thus, as the layers of the tree-SNE embedding increase, the t-distribution has heavier and heavier tails and this leads to finer structures of the data being revealed.

Typically, t-SNE embeddings are found using a random initialisation. However, since the loss function $\mathcal{L}$ is highly non-convex, different random initialisations lead to different embeddings. To ensure that the tree-SNE embedding has a continuous structure from one layer to the next, we use the embedding $\{y_1^{(i)}, \dots, y_n^{(i)}\}$ as the initialisation for layer $i+1$. Assuming that the difference between $\alpha^{(i)}$ and $\alpha^{(i+1)}$ is small, we expect that if $\{y_1^{(i)}, \dots, y_n^{(i)}\}$ minimizes $\mathcal{L}$ with parameter $\alpha^{(i)}$ then there should be a minimizer for $\mathcal{L}$ with parameter $\alpha^{(i+1)}$ in some small neighbourhood of $\{y_1^{(i)}, \dots, y_n^{(i)}\}$. We prove that this is the case in Section 3. Note that the random initialisation of the t-SNE embedding in the first layer may serve as an intrinsic sanity check: since different initialisations produce different trees, any structures that persist across multiple initialisations is likely to be trustworthy. This idea is used in the work [Gre+20].

3. Proof of Continuous Structure

In this section, we prove that, under mild assumptions, the tree-SNE method described in Section 2.2 produces a sequence of embeddings with a continuous structure. Although tree-SNE is a discrete process with a finite set of parameters producing a finite set of embeddings, we consider a continuous version of the problem and treat the embedding $y_1(\alpha), \dots, y_n(\alpha)$ as a function of the parameter α which may take any value in $(0, 1]$. Similarly, the perplexity $\mathcal{P}(\alpha)$ is also a function of the parameter α. Our goal is to prove that each function $y_i(\alpha)$ is continuous.

We consider the map $F : \mathbb{R}^{1+nd} \to \mathbb{R}^{nd}$ defined by

$$F(\alpha, y_1, \dots, y_n) = \left(\frac{\partial \mathcal{L}}{\partial y_1}, \dots, \frac{\partial \mathcal{L}}{\partial y_n} \right).$$

Since the embedding $y_1(\alpha), \dots, y_n(\alpha)$ is a minimizer of $\mathcal{L}$, in particular the condition $\nabla \mathcal{L} = 0$ is satisfied and so every embedding lies on the zero-set of F. As the function $\mathcal{L}$ is not convex, it is not true that every point in the zero-set of F corresponds to a minimizer of $\mathcal{L}$. However, given an appropriate choice of parameters such as step size, first order methods like gradient descent return the point in the

zero-set of F closest to their initialisation. We restate a refined version of our main result in the following theorem.

Theorem (The Tree-SNE tree exists - refined)**.** *Suppose $(\alpha, y_1, \dots, y_n)$ is a generic point in the zero set of F. Then, the zero set of F is a smooth, properly embedded, $1 + d(d+1)/2$ dimensional submanifold of $\mathbb{R}^{1+nd}$ $(\alpha, y_1, \dots, y_n)$. Moreover, in this neighbourhood the parameter α is not constant.*

We argue that this is sufficient to prove the contuous structure of a tree-SNE embedding. Indeed, fix a parameter α and suppose that $(y_1, \dots, y_n)$ is the corresponding layer of the tree-SNE embedding. The result of Section 3 tells us that for each α' in some neighbourhood of α there exists a set of points $(y'_1, \dots, y'_n)$ in a neighbourhood of $(y_1, \dots, y_n)$ so that $(\alpha', y'_1, \dots, y'_n)$ is in the zero set of F. So, by initializing the α' layer of the tree-SNE embedding at $(y_1, \dots, y_n)$, gradient descent will return the embedding $(y'_1, \dots, y'_n)$, or some other set of points in the neighbourhood of $(y_1, \dots, y_n)$. It follows that, when the parameters $\{a^{(i)}\}_{i=1}^m$ are chosen appropriately and each layer is initialized using the preceding layer, the tree-SNE embedding will have a continuous structure. The main tool that we use is the *constant-rank level-set theorem*. This is a standard result in the theory of manifolds. For details, see for example [Lee13].

Theorem 3.1 (Constant-Rank Level Set)**.** *Let M and N be smooth manifolds, and $\varphi : M \to N$ a smooth map with constant rank r in some neighbourhood of p. Each level set of φ is locally a properly embedded submanifold of codimension r in M.*

To prove Section 3, we will show that F is smooth and generically has rank $nd - d(d+1)/2$. Then, using the constant rank level-set theorem, we show that at a generic point, the zero set of F is locally a properly embedded submanifold of $\mathbb{R}^{1+nd}$ of dimension $1 + d(d+1)/2$ where α is not constant. First, we fix the following set of assumptions.

Assumption 3.2. We assume the following conditions are satisfied.

(A) The dataset $\{x_1, \dots, x_n\} \subset \mathbb{R}^D$ is fixed.
(B) The perplexity $\mathcal{P}(\alpha)$ is a smooth function of α
(C) Given the affinities p_{ij} calculated using perplexity $\mathcal{P}(\alpha)$, there is at least one set of points $y_1, \dots, y_n \in \mathbb{R}^d$ such that $\nabla^2 \mathcal{L}$ has rank $nd - d(d+1)/2$.

Assumptions (A) and (B) will ensure that the map F is smooth, whereas Assumption (C) will ensure that F generically has rank $nd - d(d+1)/2$. This is the key to ensuring that the constant-rank level set theorem applies in our setting.

3.1. **A Discussion of Assumption (C).** Note that since the loss function $\mathcal{L}$ is defined using pairwise distances, it is invariant under all rigid transformations of $\mathbb{R}^d$. It follows that the rank of $\nabla \mathcal{L}^2$ is at most $nd - d(d+1)/2$ since the space of all rigid transformations is $d(d+1)/2$ dimensional. Since this is the only invariance exhibited by the function, we expect the Hessian $\nabla \mathcal{L}^2$ to achieve this rank at any generic embedding $y_1, \dots, y_n$. While Assumption (C) is generically satisfied, it is indeed possible for the Hessian $\nabla^2 \mathcal{L}$ to have rank lower than $nd - d(d+1)/2$, as shown in the following proposition.

Proposition (Rank Deficient Hessian)**.** *Let $n = 3$ and $d = 2$. Suppose that the set $\{y_1, y_2, y_3\} \subset \mathbb{R}^2$ is the vertex set of an equilateral triangle with side length r. Then, the rank of $\nabla^2 \mathcal{L}$ is at most $nd - d(d+1)/2 - 1$.*

Proof. Recall that the rank of $\nabla^2\mathcal{L}$ is at most $nd - d(d+1)/2$ since $\mathcal{L}$ is invariant under all rigid transformations of $\mathbb{R}^d$. Note that any additional invariance of $\mathcal{L}$ causes the rank of the Hessian $\nabla^2\mathcal{L}$ to drop. Since the set $\{y_1, y_2, y_3\} \subset \mathbb{R}^2$ is the vertex set of an equilateral triangle, in particular we have that for each choice of $i \neq j$, the distance between y_i and y_j is $\|y_i - j_y\| = r$. Note that each of the probabilities q_{ij} is given by

$$q_{ij} = \frac{K(r)}{\sum_{k\neq\ell} K(r)} = \frac{1}{6}$$

and so is independent of the side length r. It follows that $\mathcal{L}(y_1, y_2, y_3)$ is invariant under scalings of the side length r. This is not a rigid transformation of $\mathbb{R}^2$ and so the rank of $\nabla^2\mathcal{L}$ is at most $nd - d(d+1)/2 - 1$. □

Although Assumption (C) only requires that the desired rank be achieved for one set of points, if this condition is satisfied, the Hessian at any generic embedding $y_1, \ldots, y_n$ will have rank $nd - d(d+1)/2$ due to the lower semicontinuity of rank. This the content of the following lemma.

Lemma 3.3. *Suppose Assumption (C) is satisfied. Then, for any generic set of points $y_1, \ldots, y_n$ the Hessian $\nabla^2\mathcal{L}$ has rank $nd - d(d+1)/2$.*

Proof. For any set of points $y_1, \ldots, y_n$, the rank of $\nabla^2\mathcal{L}$ is at most $r = nd - d(d+1)/2$ since $\mathcal{L}$ is invariant under all rigid transformations of $\mathbb{R}^d$. Suppose the rank is equal to r at $y_1, \ldots, y_n$. Then, there exists an $r \times r$ minor $m(y_1, \ldots, y_n)$ of $\nabla^2\mathcal{L}$ that is non-zero at $y_1, \ldots, y_n$. Note that $\mathcal{L}$ is analytic and so this minor is also analytic. Thus, the zero set of $m(y_1, \ldots, y_n)$ is measure zero. Moreover, the zero set of all $r \times r$ minors of $\nabla^2\mathcal{L}$ is measure zero and so a generic embedding does not land in this set. Since the rank function is lower semicontinuous, $\operatorname{rank}\nabla^2\mathcal{L}$ may only increase in a neighbourhood of $y_1, \ldots, y_n$. If the rank of the Hessian $\nabla^2\mathcal{L}$ is equal to $nd - d(d+1)/2$ at $y_1, \ldots, y_n$, it follows that $\operatorname{rank}\nabla^2\mathcal{L}$ is constant in a neighbourhood of $y_1, \ldots, y_n$. It follows that at a generic set of points, the rank of $\nabla^2\mathcal{L}$ is exactly $nd - d(d+1)/2$. □

Note that if Assumption (C) is not satisfied, then it will be satisfied for a random small perturbation of the affinities. This corresponds to a small perturbation of the perplexity of each distribution. Since t-SNE is known to be robust to changes in perplexity [MH08], this does not have a significant effect on the embedding.

Lemma 3.4. *Suppose for a set of affinities p_{ij} that Assumption (C) is not satisfied. Then, replacing p_{ij} with $p_{ij} + \varepsilon_{ij}$ with ε_{ij} sampled i.i.d from $N(0, \sigma^2)$, satisfies Assumption (C) with probability 1.*

Proof. Since p_{ij} is not a function of the embedding $y_1, \ldots, y_n$, each entry of $\nabla^2\mathcal{L}$ is linear in p_{ij}. Note that $\nabla^2\mathcal{L}$ has rank strictly less than $nd - d(d+1)/2$ if and only if all of its $nd - d(d+1)/2$ minors are equal to 0. The set of all p_{ij} satisfying these polynomial constraints is measure zero. Thus, any random perturbation of the affinities will not land in this set with probability 1. □

3.2. **Proof of Theorem.** We now show that, in the setting of Assumption 3.2, the map $F : \mathbb{R}^{1+nd} \to \mathbb{R}^{nd}$ is smooth and has constant rank $nd - d(d+1)/2$ in some neighbourhood of a generic point $(\alpha, y_1, \ldots, y_n) \in \mathbb{R}^{1+nd}$.

Lemma 3.5. *The map $F : \mathbb{R}^{1+nd} \to \mathbb{R}^{nd}$ is smooth and has constant rank $nd - d(d+1)/2$ around a generic point $(\alpha, y_1, \ldots, y_n)$.*

Proof. Recall that the partial derivative of $\mathcal{L}$ with respect to y_i is given by

$$\frac{\partial \mathcal{L}}{\partial y_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij}) K_\alpha(\|y_i - y_j\|)^{1/\alpha} (y_i - y_j).$$

Assuming that $y_i \neq y_j$ for all $i \neq j$, note that q_{ij} and $K_\alpha(\|y_i - y_j\|)$ are smooth functions of $y_1, \ldots, y_n$. It follows that $\frac{\partial \mathcal{L}}{\partial y_i}$ is a smooth function of $y_1, \ldots, y_n$. We now justify that $\frac{\partial \mathcal{L}}{\partial y_i}$ is a smooth function of α. Assuming that $\alpha \neq 0$, it is clear that q_{ij} and $K_\alpha(\|y_i - y_j\|)^{1/\alpha}$ are smooth functions of α. Thus, we need only justify that p_{ij} is a smooth function of α. Recall that the affinities p_{ij} are smooth functions of the bandwidths σ_i chosen to assure that the distribution around each data point has the perplexity $\mathcal{P}(\alpha)$. Note that it is sufficient to prove that the bandwidths σ_i are smooth functions of α. By assumption, the perplexity $\mathcal{P}(\alpha)$ is a smooth function of α. Consider the function $G : \mathbb{R}^{1+n} \to \mathbb{R}^n$ defined by

$$G(\alpha, \sigma_1, \ldots, \sigma_n) = \left[\mathcal{P}(\alpha) - \exp(-\log 2 \textstyle\sum_{j \neq i} p_{j|i} \log_2 p_{j|i})\right]_{i=1}^n$$

where we view each directional affinity $p_{j|i}$ as a smooth function of the bandwidth σ_i. The set of parameters and bandwidths used to find the tree-SNE embeddings are in the zero set of G. It is clear the G is a smooth function. Suppose that for a given α^*, $\sigma_1^*, \ldots, \sigma_n^*$, the equality $G(\alpha^*, \sigma_1^*, \ldots, \sigma_n^*) = 0$ holds. Then, by the implicit function theorem, $\sigma_1, \ldots, \sigma_n$ are a continuous function of α in a neighbourhood of $(\alpha^*, \sigma_1^*, \ldots, \sigma_n^*)$ if the matrix

$$J = \left[\tfrac{\partial}{\partial \sigma_k}(\mathcal{P}(\alpha) - \exp(-\log 2 \textstyle\sum_{j \neq i} p_{j|i} \log_2 p_{j|i}))\right]_{i,k=1}^n$$

is invertible. Note that J is diagonal and so is not invertible whenever the equality

$$\frac{\partial}{\partial \sigma_i}(\mathcal{P}(\alpha) - \exp(-\log 2 \sum_{j \neq i} p_{j|i} \log_2 p_{j|i})) = 0$$

for some value of i. Generically, the equality does not hold and so the implicit function theorem applies. It follows that F is a smooth function as desired. We now show that F has constant rank $nd - d(d+1)/2$ in some neighbourhood of a generic point $(\alpha, y_1, \ldots, y_n)$. Recall that the rank of F is the rank of its Jacobian ∇F. The Jacobian ∇F is given by

$$\nabla F = \begin{bmatrix} \frac{\partial^2 \mathcal{L}}{\partial y_1 \partial \alpha} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial y_n \partial \alpha} \\ & \nabla^2 \mathcal{L} & \end{bmatrix}$$

and so the rank of F is at least the rank of the Hessian $\nabla \mathcal{L}^2$. Note that ∇F is an $(nd+1) \times nd$ matrix and so its rank cannot exceed nd. However, like $\mathcal{L}$, the function F is invariant under all rigid transformations of $\mathbb{R}^d$. The space of all rigid transformations of $\mathbb{R}^d$ is $d(d+1)/2-$dimensional and so it follows that $\operatorname{rank} \nabla F \leq nd - d(d+1)/2$. By assumption, $\nabla^2 \mathcal{L}$ generically has rank $nd - d(d+1)/2$ and so we must have that $\operatorname{rank} \nabla F = nd - d(d+1)/2$. By the lower semicontinuity of rank, F has constant rank of $nd - d(d+1)/2$ in a neighbourhood of a generic point, as desired. □

We now apply the constant-rank level-set theorem to prove our main theorem.

Proof. Choose a generic point $(\alpha, y_1, \ldots, y_n)$ in the zero set of F. Then, as shown in Lemma 3.5, the hypothesis of the constant-rank level-set theorem is satisfied. In particular, we have that the zero-set of F around $(\alpha, y_1, \ldots, y_n)$ is a properly embedded, $1+d(d+1)/2$ dimensional submanifold of $\mathbb{R}^{1+nd}$. It remains to argue that the parameter α is not constant in this neighbourhood. Indeed, since F is invariant under all rigid transformations of $\mathbb{R}^d$, any level set of F is at least $d(d+1)/2$ dimensional. Suppose that α is constant in this neighbourhood. Then, there exists some nontrivial vector $(v_1, \ldots, v_n)^\top \in \mathbb{R}^{nd}$ such that $(\alpha, y_1 + v_1, \ldots, y_n + v_n) = 0$. However, we would then have that $(v_1, \ldots, v_n)^\top$ is in the nullspace of $\nabla^2 \mathcal{L}$. Since $\nabla^2 \mathcal{L}$ has rank $nd - d(d+1)/2$ by assumption, this is a contradiction. It follows that α is not constant, as desired. □

4. Examples

W use tree-SNE to visualize two high-dimensional datasets constrcuted from the HathiTrust Library. The full HathiTrust dataset consists of 13.6 million works described by millions of features that correspond to word counts on each page of the work. We use the preprocessed dataset from [Sch18] that reduces each work to a 100-dimensional vector. For each example, we generate a tree-SNE embedding using the default parameters described in [RPH20]. Each layer of the embeddings was found using FIt-SNE [Lin+19] with default parameters. Code to reproduce these examples, as well as the MNIST example seen in Section 1, can be found at `https://github.com/j4ck-k/tree-sne`.

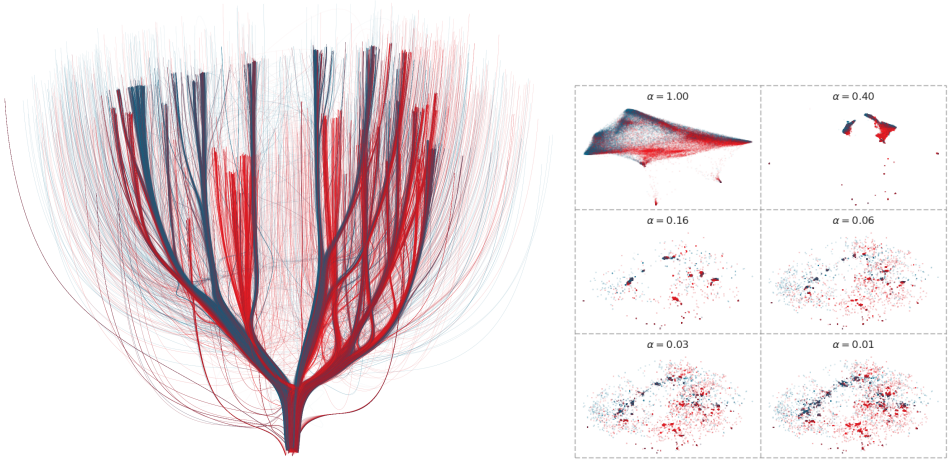

Figure 4. Tree-SNE embedding of 100,000 works of American and English literature. The red points correspond to English literature and blue points correspond to American literature.

4.1. English and American Literature.

For our first example, we use tree-SNE to visualize a set of 100,000 works of English and American literature from the HathiTrust library dataset. This dataset was constructed from the full HathiTrust dataset by filetering works classified as either PR (the Library of Congress code for English literature) and PS (the Library of Congress code for American literature)

and taking the first 100,000 works for computational simplicity and efficiency. Our dataset is approximately 55% English and 45% American literature.

We generates a 50 layer embedding with degrees of freedom between 1 and 0.01. Figure 4 shows the tree-SNE embedding of the dataset along with slices of several layers. Points are color coded depending on Library of Congress code. In each plot, red points are works classified as PR and blue points are works classified as PS. When $\alpha = 1$ we see that, although the two classifications are being pushed to opposite sides of the embedding, there is no separation between works of English and American literature. Clusters begin to appear in subsequent layers of the plot.

To illustrate how the tree-SNE embedding captures the evolution of clusters within the data, we track the paths of 109 works that form a single cluster roughly halfway into the tree-SNE embedding. This cluster was chosen because it contains a mix of American (80 works) and English (29 works) literature and its relatively small size allows us to easily examine the works that appear as well as subclusters that form within the data. The trajectories of these works, as well as slices showing the evolution of the cluster throughout the embedding, are shown in Figure 5.

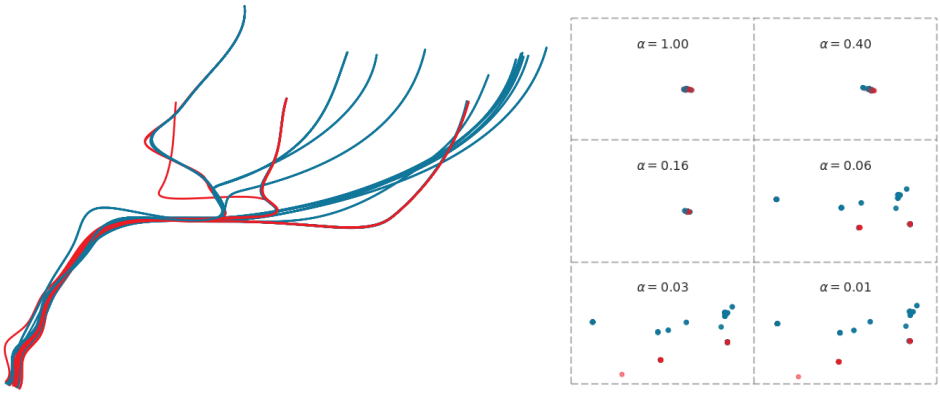

FIGURE 5. The trajectories of 109 works in the tree-SNE embedding in Figure 4. These works form a cluster in the tree approximately halfway through the embedding and subsequently split into more informative subclusters towards the top of the tree.

The 109 works in this cluster are mostly pieces of travel writing and journals by writers from the 19th century. Given that these works center a common theme, a macro-level cluster is expected. We see that as the cluster evolves within the tree, additional branches begin to appear. Notably, in the final layer of the embedding, all except one subcluster is formed entirely of either English or American literature. The one exception is a cluster containing 13 works of English literature and 18 works of American literature. However, the American literature in this subcluster consists entirely of travel writings from the novelist Nathaniel Hawthorne during his time in the United Kingdom and continental Europe.

4.2. **The Anti-Stratfordian Theory.** Although it is widely accepted that the plays and poems attributed to Shakespeare were indeed written by William Shakespeare of Stratford-upon-Avon, there is a fringe theory amongst literary historians that these works were produced either by another author or a group of writers.

This is known as the Anti-Stratfordian Theory. Motivated by the cluster formed by Shakespeare and his contemporaries in Figure 4, we apply tree-SNE to explore the Shakespearean authorship question.

Two prominent alternative candidates for Shakespearean authorship are the philospher Sir Francis Bacon and the playwright Christopher Marlowe. We constructed a dataset of 2419 English language works by Shakespeare (68 works), Bacon (87 works), Marlowe (114 works), and additional contemporaries from the HathiTrust database. Many of these contemporaries have also been proposed as candidate authors for Shakespeare's work yet are not as popular as Marlowe and Bacon. Since this dataset is relatively small, we generate a tree-SNE embedding with only 10 layers corresponding to degrees of freedom between 1 and 0.25. The embedding and slices of three layers are in Figure 6.

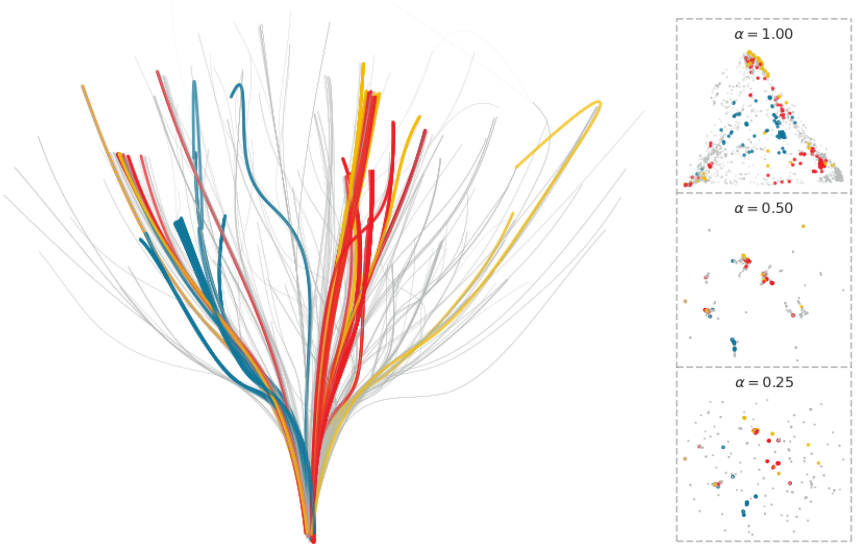

FIGURE 6. Tree-SNE embedding of Shakespeare and contemporaries. Gold points are works attributed to Shakespeare, red points are works by Marlowe, and blue points are works by Bacon.

We applied DBSCAN to various layers in the embedding to determine clusters within the data. In the initial t-SNE embedding, there is little separation between any of the included authors, although we see that works by Bacon are more central and works by Marlowe and Shakespeare are pushed towards the corners of the embedding. In particular, the works of Shakespeare are pushed to the north whereas works of Marlowe are scattered throughout. As the layers of the embedding increase, we see that clusters consisting of works attributed to a single author start to appear. In the final layer of the embedding, we see that there are two larger clusters containing works of Shakespeare. More than half of the works attributed to Shakespeare appear in one large cluster of 304 works. This cluster mostly consists of plays and many authors including Marlowe, but not Bacon, are represented. Other than Shakespeare, ten authors (including Marlowe) appear at least ten times in this cluster. Bacon and Shakespeare only appear together in one cluster that contains 6 works of Bacon, 7 works of Shakespeare, and 256 works total. Thus, the

vast majority of the works attributed to these authors do not appear in the same clusters within the data. Within the confines of this small experiment, it appears to be less likely that Sir Francis Bacon is the true author of the works attributed to Shakespeare than Christopher Marlowe although there is little to no evidence that Marlowe was the true author.

5. Conclusion

We have successfully shown that the tree-SNE procedure introduced in [RPH20] generates continuous structures that produce finer and finer clusterings within high-dimensional data. Moreover, by creating tree-SNE embeddings of three high-dimensional datasets, we observed that these clusterings are indeed meaningful and elucidate both macro- and micro-level structures within data. We conclude by describing two interesting directions for further research:

(1) The tree-SNE procedure can conceivably be adapted to other attraction-repulsion methods to produce trees of other embeddings. We leave the structure of these trees as an interesting direction to explore.
(2) While we have shown that tree-SNE embeddings have continuous strucutre, it is yet unclear why and when different branches form within the tree. To fully understand this, a greater understanding of the inner workings of t-SNE is needed. An interesting question is to uncover these threshold degrees of freedom for highly structured data, such as points sampled from a mixture of mixtures of Gaussians.

Acknowledgements. The author is grateful to Stefan Steinerberger for introducing him to this problem and for providing valuable guidance throughout the creation of this manuscript.

References

[AHK18] Sanjeev Arora, Wei Hu, and Pravesh K Kothari. "An analysis of the t-sne algorithm for data visualization". In: *Conference on learning theory*. PMLR. 2018, pp. 1455–1462.

[BBK22] Jan Niklas Böhm, Philipp Berens, and Dmitry Kobak. "Attraction-repulsion spectrum in neighbor embeddings". In: *Journal of Machine Learning Research* 23.95 (2022), pp. 1–32.

[BN03] Mikhail Belkin and Partha Niyogi. "Laplacian eigenmaps for dimensionality reduction and data representation". In: *Neural computation* 15.6 (2003), pp. 1373–1396.

[CM22] T Tony Cai and Rong Ma. "Theoretical foundations of t-sne for visualizing high-dimensional clustered data". In: *Journal of Machine Learning Research* 23.301 (2022), pp. 1–54.

[Gre+20] P. Greengard et al. *Factor Clustering with T-SNE*. SSRN, 2020. URL: https://books.google.com/books?id=Jxb2zgEACAAJ.

[HR02] Geoffrey E Hinton and Sam Roweis. "Stochastic Neighbor Embedding". In: *Advances in Neural Information Processing Systems*. Ed. by S. Becker, S. Thrun, and K. Obermayer. Vol. 15. MIT Press, 2002. URL: https://proceedings.neurips.cc/paper_files/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf.

[Jac+14] Mathieu Jacomy et al. "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software". In: *PloS one* 9.6 (2014), e98679.

[KL19] Dmitry Kobak and George C Linderman. "UMAP does not preserve global structure any better than t-SNE when using the same initialization". In: *BioRxiv* (2019), pp. 2019–12.

[Kob+20] Dmitry Kobak et al. "Heavy-Tailed Kernels Reveal a Finer Cluster Structure in t-SNE Visualisations". In: *Machine Learning and Knowledge Discovery in Databases*. Springer International Publishing, 2020, 124–139. ISBN: 9783030461508. DOI: 10.1007/978-3-030-46150-8_8. URL: http://dx.doi.org/10.1007/978-3-030-46150-8_8.

[Lee13] John M. Lee. *Introduction to smooth manifolds*. Second. Vol. 218. Graduate Texts in Mathematics. Springer, New York, 2013, pp. xvi+708. ISBN: 978-1-4419-9981-8.

[Lin+19] George C. Linderman et al. "Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data". In: *Nature Methods* 16.3 (Feb. 2019), 243–245. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0308-4. URL: http://dx.doi.org/10.1038/s41592-018-0308-4.

[LS17] George C. Linderman and Stefan Steinerberger. *Clustering with t-SNE, provably*. 2017. arXiv: 1706.02582 [cs.LG]. URL: https://arxiv.org/abs/1706.02582.

[MC12] Fionn Murtagh and Pedro Contreras. "Algorithms for hierarchical clustering: an overview". In: *WIREs Data Mining and Knowledge Discovery* 2.1 (2012), pp. 86–97. DOI: https://doi.org/10.1002/widm.53. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.53. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.53.

[MH08] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[MHM18] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[Nie16] Frank Nielsen. "Hierarchical Clustering". In: Feb. 2016, pp. 195–211. ISBN: 978-3-319-21902-8. DOI: 10.1007/978-3-319-21903-5_8.

[RPH20] Isaac Robinson and Emma Pierce-Hoffman. *Tree-SNE: Hierarchical Clustering and Visualization Using t-SNE*. 2020. arXiv: 2002.05687 [cs.LG]. URL: https://arxiv.org/abs/2002.05687.

[Sch18] Benjamin Schmidt. "Stable Random Projection: Lightweight, General-Purpose Dimensionality Reduction for Digitized Libraries". In: *Journal of Cultural Analytics* 3.1 (2018). DOI: 10.22148/16.025.