t-SNE and its theoretical guarantee

Ziyuan Zhong

Columbia University

July 4, 2018

Overview

Timeline:

- PCA (Karl Pearson, 1901)
- Manifold Learning(Isomap by Tenenbaum et al., LLE by Roweis and Saul,... 2000)

The following topics will be covered in today's presentation:

- SNE (Hinton and Roweis, 2002)
- t-SNE (van der Maaten and Hinton, 2008)
- Accelerate t-SNE. (van der Maaten, 2014)
- First step towards theoretical guarantee for t-SNE. (Linderman and Steinerberger, 2017)
- Accelerate t-SNE more.(Linderman et al., 2017)
- Theoretical guarantee for t-SNE. (Arora et al., 2018)
- Generalization of t-SNE to manifold (Verma et al. preprint)

Preliminaries and Notation Part1

Matrix Norm:

$$||A||_p = sup_{x \neq 0} \frac{||Ax||_p}{||x||_p}.$$

$$||A||_2 = \sqrt{\lambda_{\max}(A^* A)} = \sigma_{\max}(A)$$

$$||A||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}$$

$Diam(S)$ denotes that diameter of a bounded set $S \subset \mathbb{R}^s$,i.e., $Diam(S) := sup_{x,y \in \mathbf{s}} ||x - y||_2$.

Preliminaries and Notation Part2

In a clustering setting, $\pi : [n] \to [k]$ denotes function that maps data index $i \in [n]$ to cluster index k. $i \sim j$ means $\pi(i) = \pi(j)$

KL-divergence:

A measure of how unsimilar two distributions are.

Let p and q be two joint probability over i, j,

$KL(p||q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$

Problem Formulation(A sketch)

Given a collection of points $\mathbf{X} = \{x_1, ..., x_n\} \subset \mathbb{R}^d$, find a collection of points $\mathbf{Y} = \{y_1, ..., y_n\} \subset \mathbb{R}^{d'}$, where $d' \ll d$, such that the lower dimension embedding preserves the relationship among different points in the original space.

What is a good visualization? Illustration

Intuitive idea: Help to explore the inherent structure of the dataset(e.g. discover natural clusters, find linear relationship)

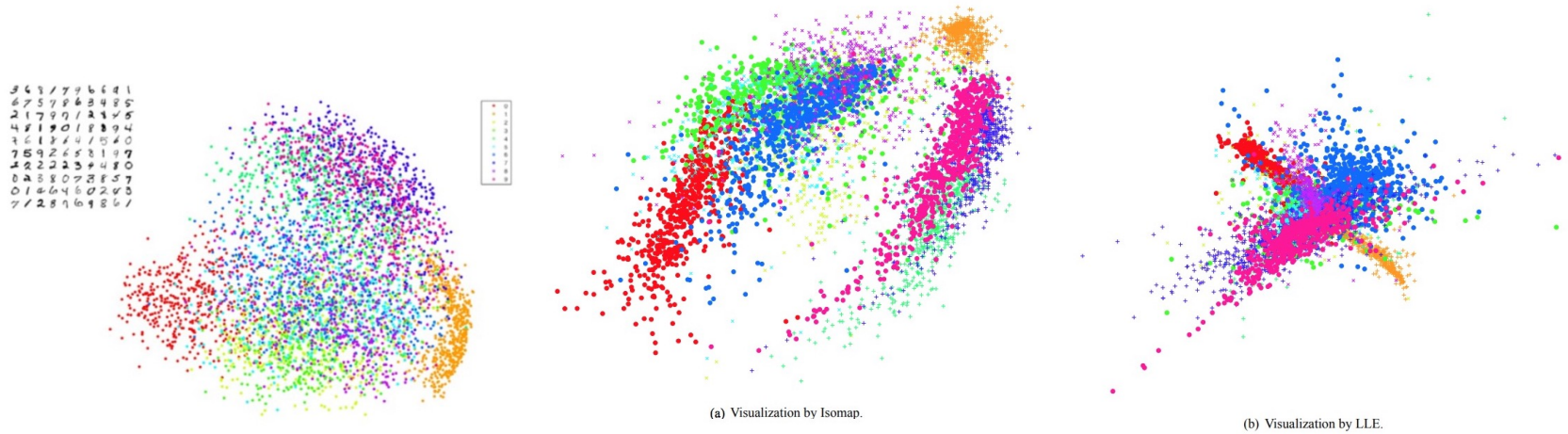
What is a good visualization for embedding? Technical Requirement

Technical requirements:

- Visualizability: (usually 2D or 3D).
- Fidelity: Relationship among points are preserved(i.e. similar points remain distinct; distinct points remain distinct).
- Scalability: Can deal with large, high-dimensional data sets.

Motivation for a better visualization algorithm

For many real-world dataset with non-linear inherent structures(e.g. MNIST), both linear methods like PCA and classical manifold learning algorithms like Isomap and LLE fail.



(a) Visualization by Isomap.

(b) Visualization by LLE.

SNE

G.E. Hinton and S.T. Roweis. Stochastic Neighbor Embedding. NIPS2002

SNE Summary

- Compute an N x N similarity matrix in the high-dimensional input space.
- Define an N x N similarity matrix in the low-dimensional embedding space.
- Define cost function - sum of KL divergence between the two probability distributions at each point.
- Iteratively learn low-dimensional embedding by minimizing the cost function using gradient descent.

SNE

Similarity matrix at high dimension:

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

where τ_i^2 is the variance for the Gaussian distribution centered around x_i.
Similarity matrix at low dimension(τ_i^2 is set to $\frac{1}{2}$ for all i):

$$q_{j|i} = \frac{exp(-||y_i - y_j||^2)}{\sum_{k \neq i} exp(-||y_i - y_k||^2)}$$

The cost function is defined as:

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

SNE focuses on local structure(Small $p_{ij} \rightarrow$ Small penalty. Large $p_{ij} \rightarrow$ Large penalty)
It has gradient: $\frac{dC}{dy_i} = 2\sum_j (y_i - y_j)(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})$

How to choose variance τ_i? Reference

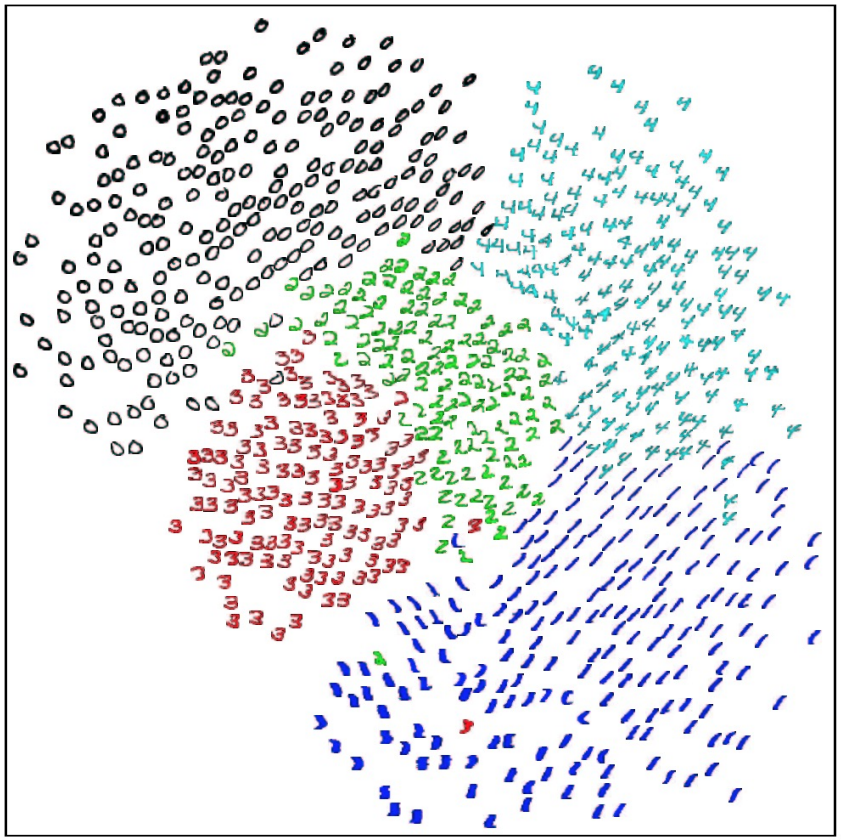
τ_i is the variance of the Gaussian that is centered around each high-dimensional datapoint x_i. In dense region, small τ_i is more appropriate.
SNE performs a binary search for the value of τ_i that produces a P_i with a fixed perplexity that is specified by the user. The perplexity can be interpreted as a smooth measure of the effective number of neighbors. The perplexity is defined as:

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

where $H(P_i)$ is the Shannon entropy of P_i measured in bits:

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

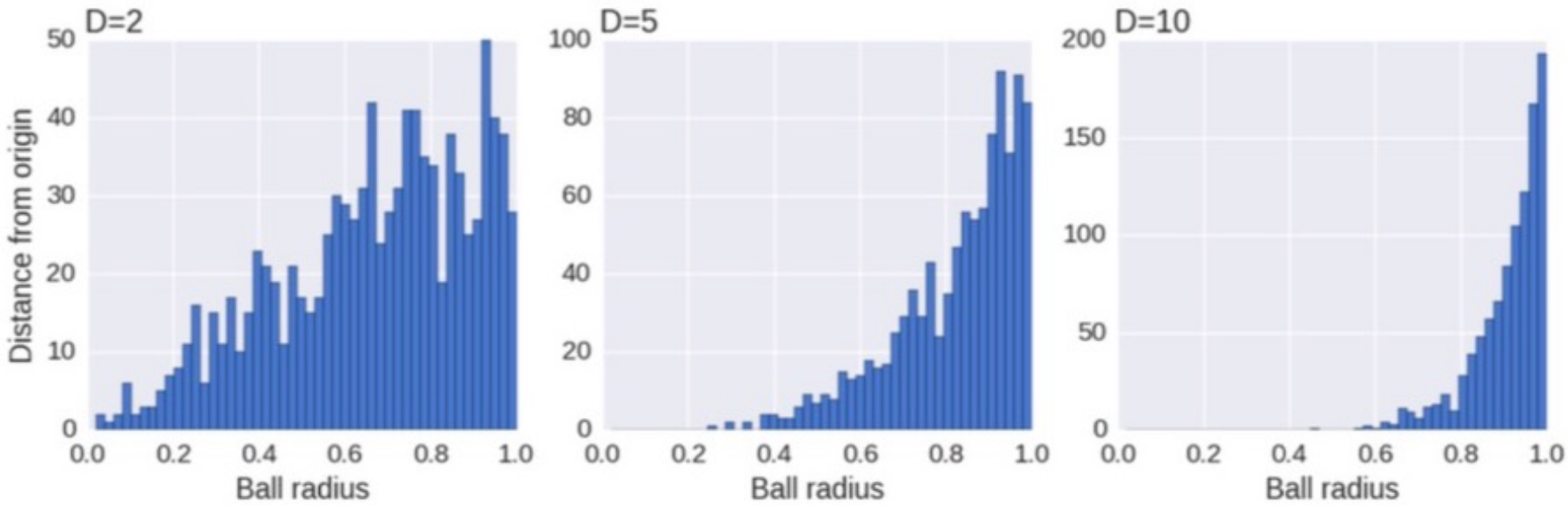
Result



The result of running the SNE algorithm on 3000 256-dimensional grayscale images of handwritten digits(Not all points are shown).

Problem with SNE: "crowding problem"

SNE suffers from the "crowding problem": The area of the 2D map that is available to accommodate moderately distant data points will not be large enough compared with the area available to accommodate nearby data points.



t-SNE

L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. JMLR2008

- A symmetrized version of the SNE cost function with simpler gradients.
- A Student-t distribution rather than a Gaussian to compute the similarity in the low-dimensional space to alleviate the crowding problem and the optimization problems of SNE.

t-SNE: Make similarity symmetric

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

We may tend to define the similarity function in the symmetric similarity matrix on high dimension as:

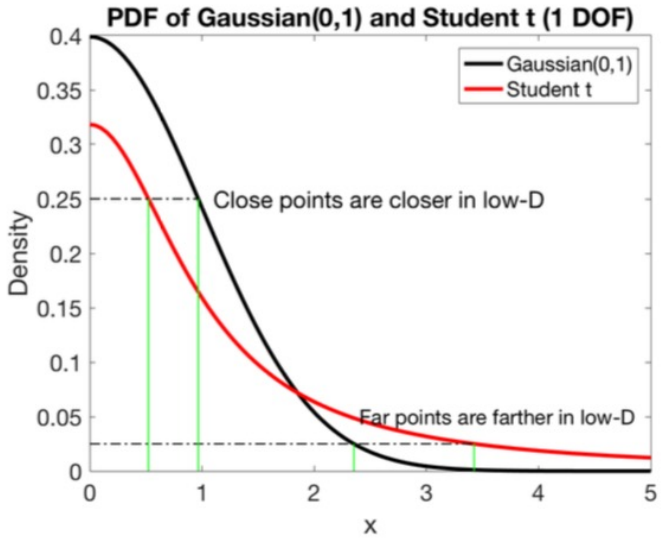
$$p_{ij} = \frac{exp(-||x_i - x_j||^2/2\tau_i^2)}{\sum_{k \neq l} exp(-||x_k - x_l||^2/2\tau^2)}$$

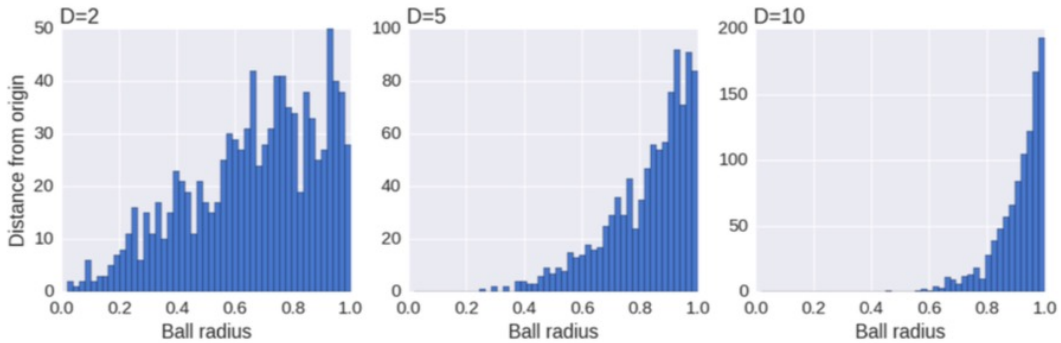
However, outlier x_i on high-dimensional space can cause problem by making p_{ij} very small for all j. Instead, define p_{ij} by symmetrizing two conditional probabilities as follows:

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

In this way, $\sum_j p_{ij} > \frac{1}{2n}$ for all data points x_i. As a result, each x_i makes a significant contribution to the cost function. The main advantage of the symmetric form is mainly simpler gradient(will be shown later).

t-SNE: Fix crowding, Illustration





t-SNE: Fix crowding

Define

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}}$$

The cost function of t-SNE is now defined as:

$$C = \sum_i KL(P_i||Q_i) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

The heavy tails of the normalized Student-t kernel allow dissimilar input objects x_i and x_j to be modeled by low-dimensional counterparts y_i and y_j that are too far apart because q_{ij} is not very small for two embedded points that are far apart.
Note: Since q is what to be learned, the outlier problem does not exist for low-dimension.

t-SNE: Gradient

The gradient of the cost function is:

$$\begin{aligned}
\frac{dC}{dy_i} &= 4 \sum_{j=1, j\neq i}^{n} (p_{ij} - q_{ij})(1 + ||y_i - y_j||^2)^{-1}(y_i - y_j) \\
&= 4 \sum_{j=1, j\neq i}^{n} (p_{ij} - q_{ij}) q_{ij} Z (y_i - y_j) \\
&= 4 \Big(\sum_{j\neq i} p_{ij} q_{ij} Z (y_i - y_j) - \sum_{j\neq i} q_{ij}^2 Z (y_i - y_j) \Big) \\
&= 4(F_{attraction} + F_{repulsion})
\end{aligned}$$

where $Z = \sum_{l,s=1, l\neq s}^{n} (1 + ||y_l - y_s||^2)^{-1}$. The derivation can be found in the appendix of the t-SNE paper.

Exercise: There are two small errors but they cancel out so the result is correct. Can you find the errors in their derivation?

After Fix

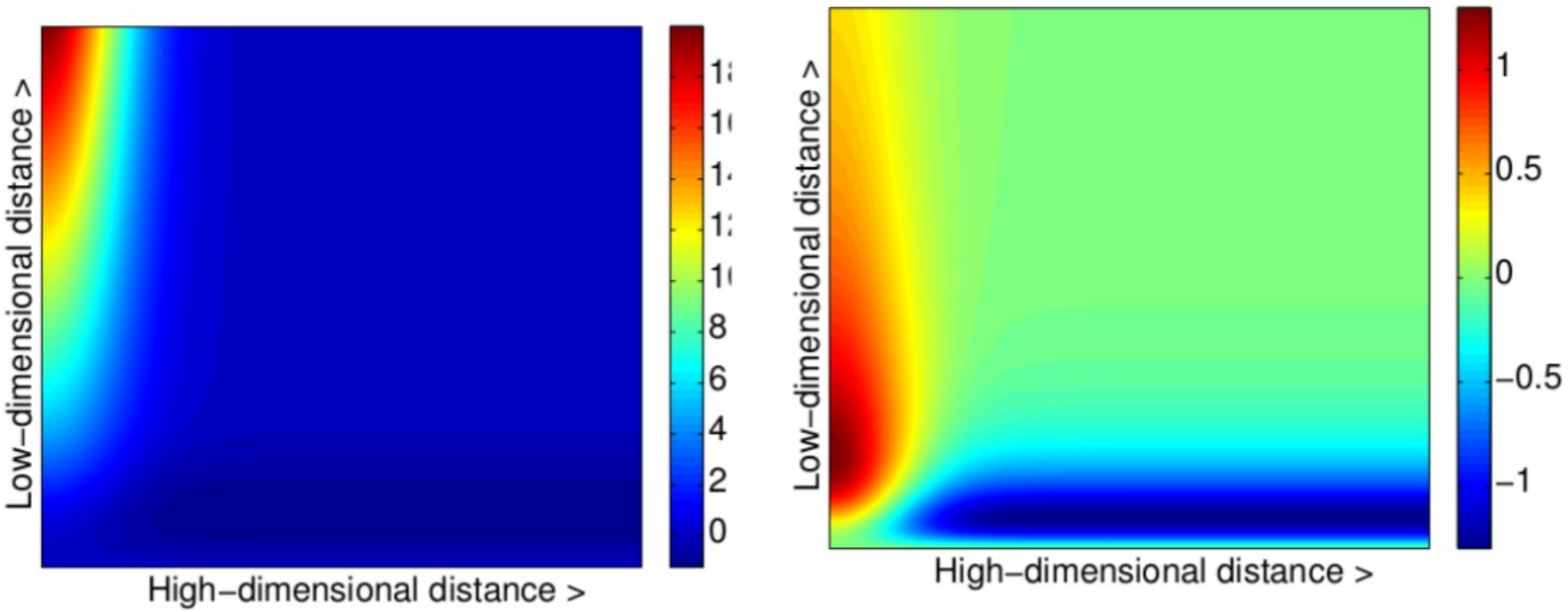
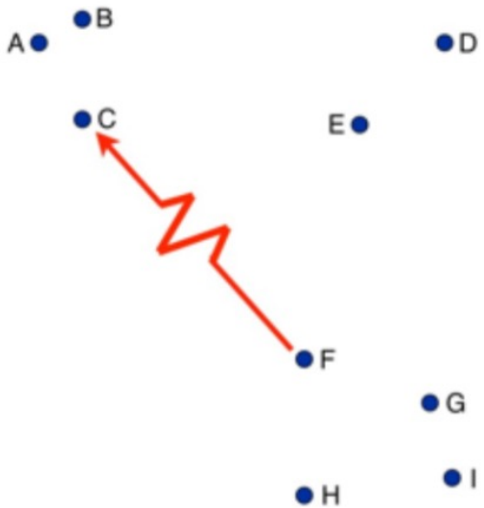
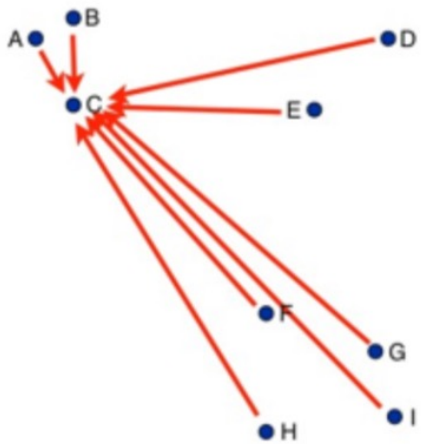


Figure : Gradient of SNE and t-SNE

t-SNE: Physics Analogy-N body system

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_{i \neq j} (p_{ij} - q_{ij}) \frac{\mathbf{y}_i - \mathbf{y}_j}{1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2}$$





Spring Analogy: $F = -k * (\mathbf{y}_i - \mathbf{y}_j)$, attraction/repulsion

Algorithm 1 t-SNE

Input: Dataset $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, Gaussian bandwidths $\tau_1, \ldots, \tau_n > 0$, exaggeration parameter $\alpha > 0$, step size $h > 0$, number of rounds $T \in \mathbb{N}$

1: Compute $\{p_{ij} : i, j \in [n], i \neq j\}$ using (1)

2: Initialize $y_1^{(0)}, y_2^{(0)}, \ldots, y_n^{(0)}$ i.i.d. from the uniform distribution on $[-0.01, 0.01]^2$

3: **for** $t = 0$ **to** $T - 1$ **do**

4: $\quad Z^{(t)} \leftarrow \sum_{i,j \in [n], i \neq j} \left(1 + \left\| y_i^{(t)} - y_j^{(t)} \right\|^2\right)^{-1}$

5: $\quad q_{ij}^{(t)} \leftarrow \dfrac{\left(1 + \left\| y_i^{(t)} - y_j^{(t)} \right\|^2\right)^{-1}}{Z^{(t)}}, \qquad \forall i, j \in [n], i \neq j$

6: $\quad y_i^{(t+1)} \leftarrow y_i^{(t)} + h \sum_{j \in [n] \setminus \{i\}} \left(\alpha p_{ij} - q_{ij}^{(t)}\right) q_{ij}^{(t)} Z^{(t)} \left(y_j^{(t)} - y_i^{(t)}\right), \qquad \forall i \in [n]$

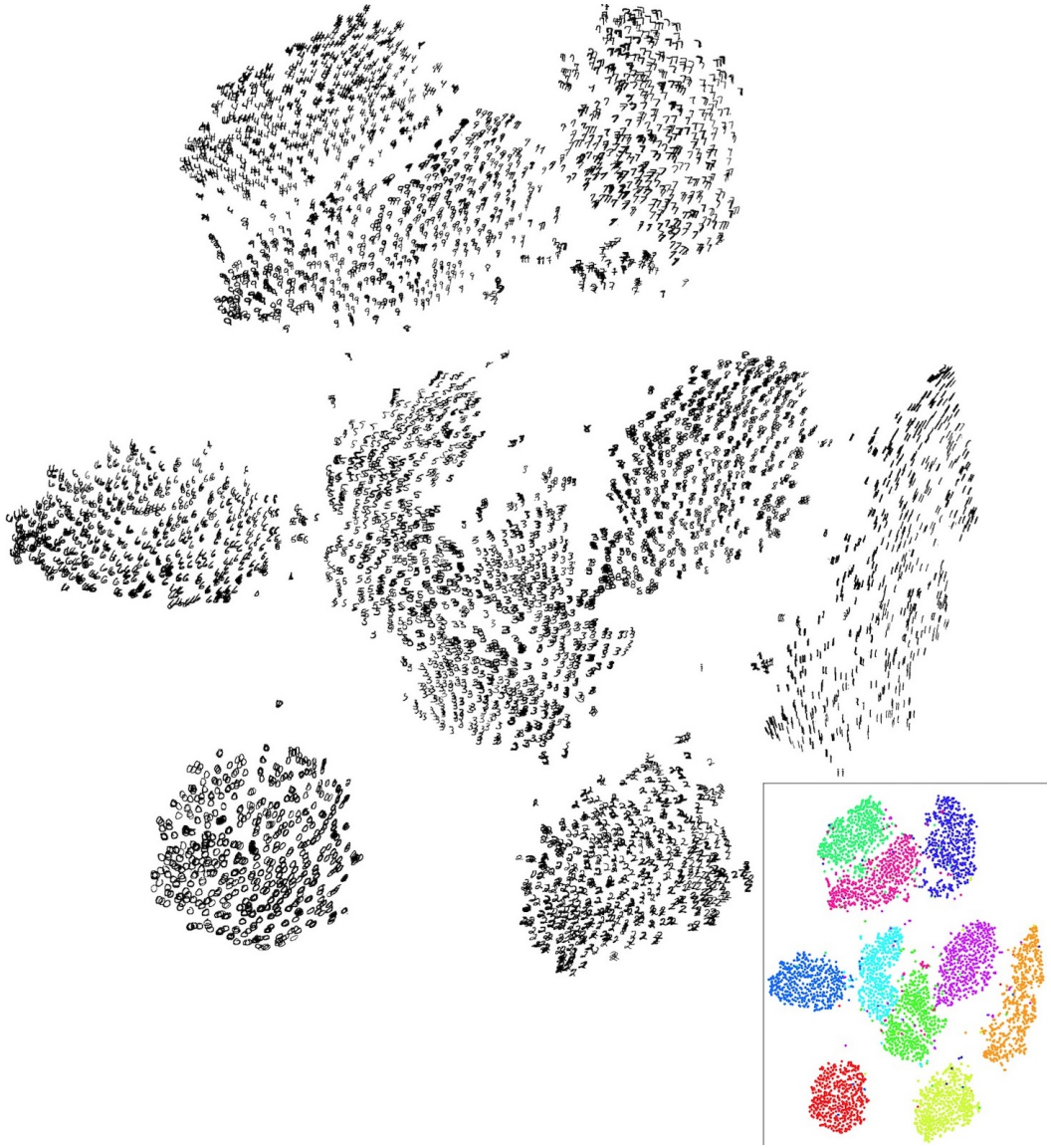
7: **end for**

Output: 2D embedding $\mathcal{Y}^{(T)} = \left\{y_1^{(T)}, y_2^{(T)}, \ldots, y_n^{(T)}\right\} \subset \mathbb{R}^2$

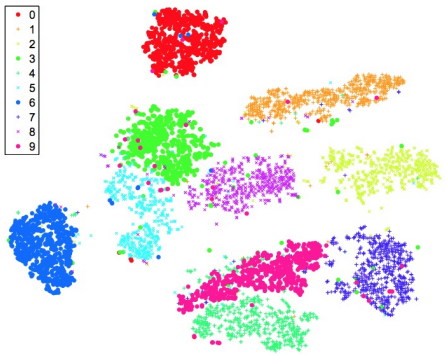
t-SNE: Early exaggeration

In the initial stage, multiply p_{ij} by a coefficient $\alpha > 1$. This encourages to focus on modeling the large p_{ij} by fairly large q_{ij}. A natural result is to form tight widely separated clusters in the map and thus makes it easier for the clusters to move around relative to each other in order to find a global organization.

Result on MNIST



Comparison on MNIST

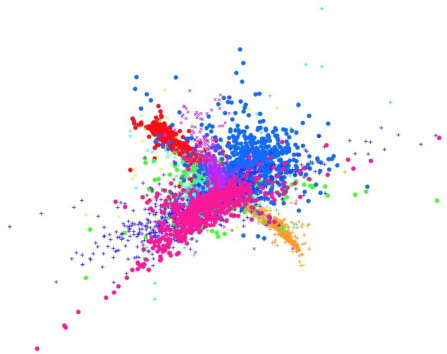


(a) Visualization by t-SNE.

(b) Visualization by Sammon mapping.

Figure 2: Visualizations of 6,000 handwritten digits from the MNIST data set.

(a) Visualization by Isomap.

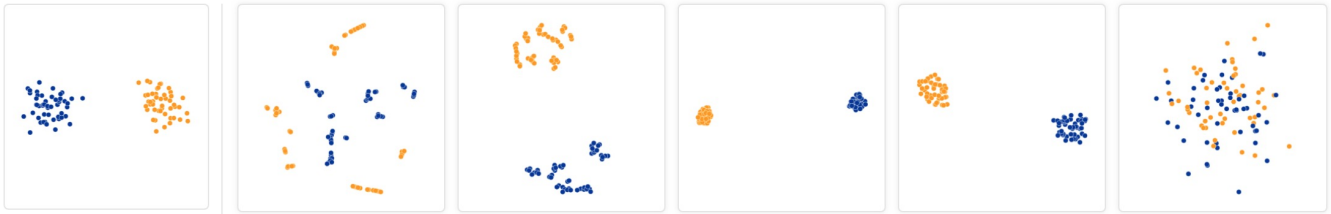


(b) Visualization by LLE.

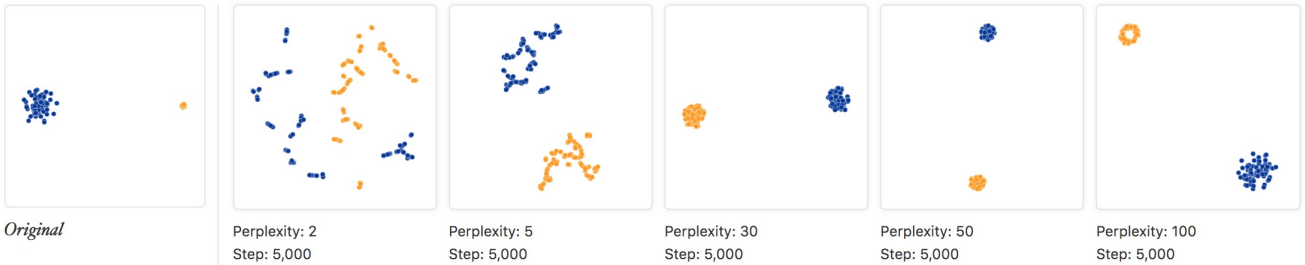
Figure 3: Visualizations of 6,000 handwritten digits from the MNIST data set.

Caveats I

Perplexity needs to be chosen carefully.

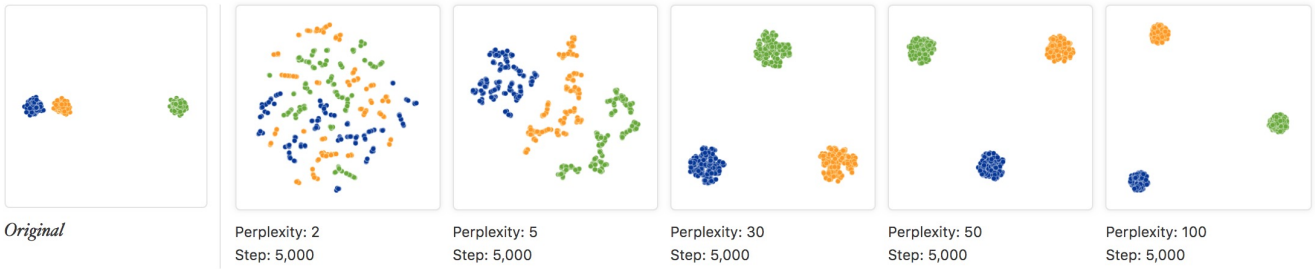


Relative size is usually not meaningful.



Caveats II

Global Structures are preserved only sometimes.



Limitations

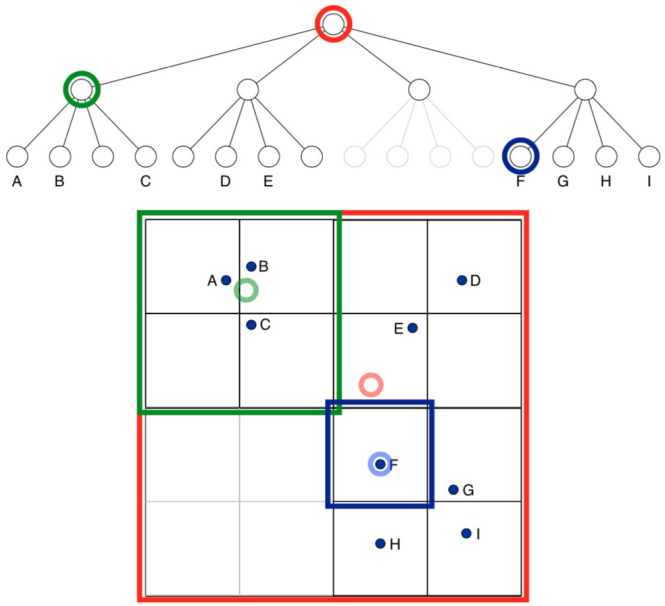
- Dimensionality reduction for other problems(due to the heavy tail of the t-distribution, it does not preserve the local structure as well if the embedded dimension is larger, say 100).
- curse of dimensionality(t-SNE employs Euclidean distances between near neighbors so it implicitly depends on the local linearity on the manifold).
- $O(N^2)$ computational complexity(the evaluation of the joint distributions involve $N(N-1)$ pairs of objects. The method is limited to 10k points).
- Perplexity number, number of iterations, the magnitude of early exaggeration parameter have to be manually chosen.

Accelerate t-SNE

L.J.P. van der Maaten. Accelerating t-SNE using Tree-Based Algorithms. JMLR2014

- Observations: Many of the pairwise interactions between points are very similar.
- Idea: Approximate similar interactions by a single interaction using a metric tree that has $O(uN)$ non-zero values.
- Result: Reduce complexity to $O(N \log N)$ via Barnes-Hut-SNE (tree-based) algorithm. The method can deal with up to tens of millions data points.

Preliminary: Quadtree



A quadtree(2d) is defined as the following: Each node represents a rectangular cell with a particular center, width, and height. It stores the the number of points inside N_{cell} and their center-of-mass y_{cell}. Non-leaf node have four children that split up the cell into four smaller cells on the current embedding. It can be constructed in $O(N \log N)$ time by inserting the points one-by-one, splitting a leaf node whenever a second point is inserted in its cell, and updating y_{cell} and N_{cell} of all visited nodes.

Approximate high-dimensional similarity

Compute the sparse approximation by finding $3u$ nearest neighbors where u is the perplexity of the conditional distribution.

$$p_{j|i} = \begin{cases} \frac{exp(-||x_i - x_j||^2/2\tau_i^2)}{\sum_{k \in N_i} exp(-||x_i - x_k||^2/2\tau_i^2)} & \text{, if } j \in N_i \\ 0 & \text{, otherwise} \end{cases}$$

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

Problem with low-dimensional repulsion gradient

Recall the t-SNE gradient as the following:

$$\frac{dC}{dy_i} = 4(F_{attraction} + F_{repulsion}) = 4\Big(\sum_{j \neq i} p_{ij} q_{ij} Z(y_i - y_j) - \sum_{j \neq i} q_{ij}^2 Z(y_i - y_j)\Big)$$

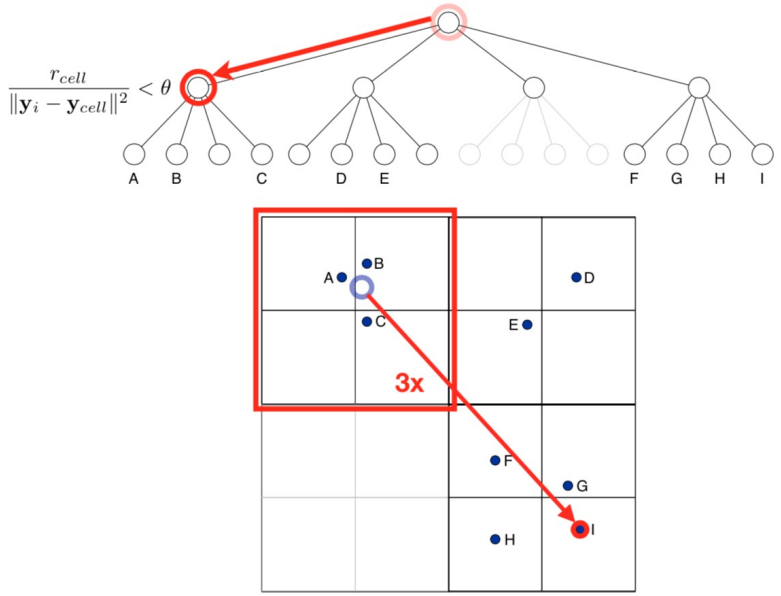
where $Z = \sum_{k \neq l} (1 + ||y_k - y_l||^2)^{-1}$ so $q_{ij} Z = (1 + ||y_i - y_j||^2)^{-1}$
Problem: $F_{attraction}$ can be done in $O(uN)$ but naive of computation of $F_{repulsion}$ is $O(N^2)$.

Barnets-Hut Approximation for low-dimensional repulsion gradient

Solution:

- construct a quadtree(2d) for the current embedding.
- Traverse the quadtree using DFS.
- At every node, decide whether the corresponding cell can be used as a "summary" for the contribution to $F_{repulsion}$ of all points in that cell.
- Replace $-q_{ij}^2 Z(y_i - y_j)$ with $-N_{cell} q_{i,cell}^2 Z(y_i - y_{cell})$.
- Z and $q_{ij}Z$ are evaluated along the way so we can compute $F_{repulsion} = \frac{q_{ij}^2 Z^2}{Z}$ in $\sim O(NlogN)$.

Illustration of BH-Approximation



The condition was proposed by Barnes and Hut(1986), where r_{cell} represents the length of the diagonal of the cell under consideration and θ is a threshold that trades off speed and accuracy (higher values of θ lead to faster but coarser approximations). Note that when $\theta = 0$, all pairwise interactions are computed(equivalent to naive t-SNE).

Experiment Part1

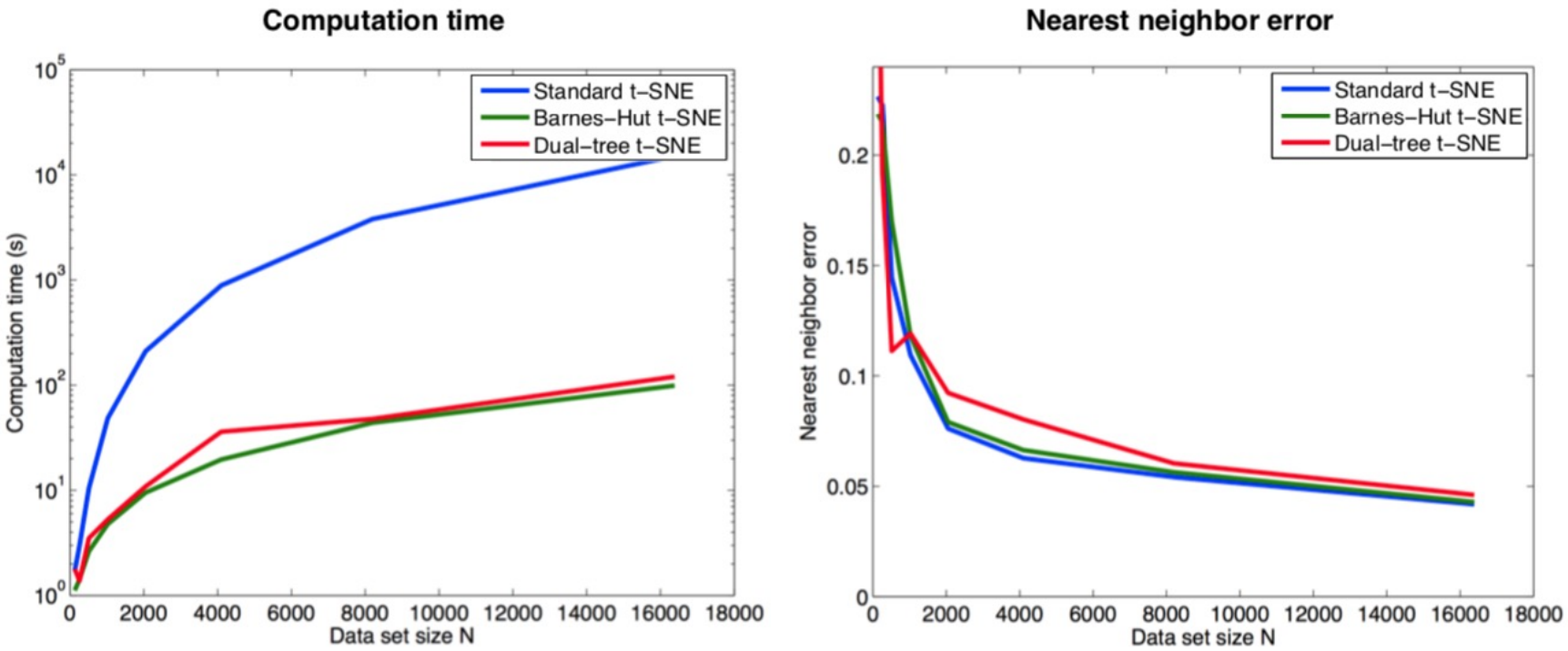
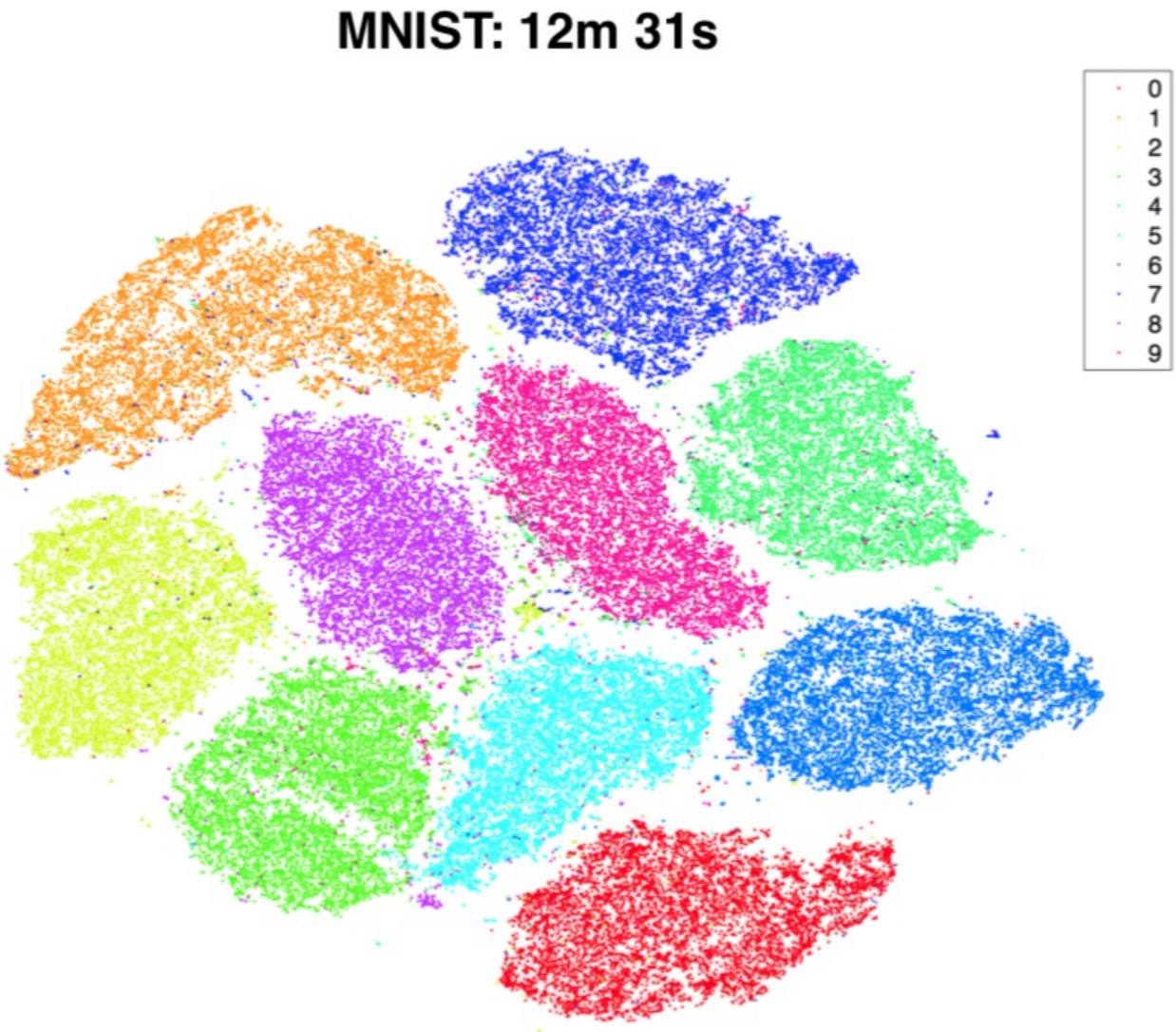


Figure 4: Compution time (in seconds) required to embed MNIST digits (left) and the 1-nearest neighbor errors of the corresponding embeddings (right) as a function of data set size N for standard t-SNE (in blue), Barnes-Hut t-SNE (in green), and dual-tree t-SNE (in red). Note that the required computation time, which is shown on the y-axis of the left figure, is plotted on a logarithmic scale.

Experiment Part2



Further speedup

Limitation: $O(N \log N)$ is still not fast enough when the dataset is very, very large (e.g. some datasets in biology have $\sim 1,000,000,000$ data points).
Efficient Algorithms for t-distributed Stochastic Neighborhood Embedding. Linderman et al. arxiv2017.
Solution: Further approximation to achieve $O(N)$.
Summary: Use *ANNOY* to search for nearest neighbors for high dimension similarity approximation and use interpolation-based methods to approximate low dimension similarity.
Results: more than 10x speedup on 1D and 2D visualization tasks.

A step towards theoretical guarantee for t-SNE

"Clustering with t-SNE, provably." George C. Linderman, Stefan Steinerberger. arxiv 2017.

Contribution:

At the high level, this paper shows that points in the same cluster move towards each other.

Limitation:

The result is insufficient to establish that t-SNE succeeds in finding a full visualization as it does not rule out multiple clusters merging into each other.

First Theoretical Guarantee for t-SNE

"An Analysis of the t-SNE Algorithm for Data Visualization" Sanjeev Arora, Wei Hu, Pravesh K. Kothari. COLT 2018

Contribution:

First provable guarantees on t-SNE for computing visualization of clusterable data. The proof is built on results from Linderman's paper.

Proof Technique:

They obtained an update rule for the centroids of the embeddings of all underlying clusters. They showed that the distance between distinct centroids remains lower-bounded whenever the data is γ-spherical and γ-well-separated. Combined with the shrinkage result for points in the same cluster, this implies that t-SNE outputs a full visualization of the data.

Preliminary

A distribution D with density function f on $\mathbb{R}^d$ is said to be log-concave if $\log(f)$ is a concave function. Example: Gaussian, Uniform on any convex set.

D is said to be isotropic if its covariance is $\mathbf{I}$.

A mixture of k log-concave distributions is described by k positive mixing weights $w_1, ..., w_k, (\sum_{l=1}^{k} w_l = 1)$ and k log-concave distribution $\mathbf{D}_1, ..., \mathbf{D}_k$ in $\mathbb{R}^d$. To sample a point from this model, we pick cluster l with probability w_l and draw x from $\mathbf{D}_l$.

Formalizing Visualization

Given a collection of points $\mathbf{X} = \{x_1, ..., x_n\} \subset \mathbb{R}^d$ and there exists a "ground-truth" clustering described by a partition $C_1, ..., C_k$ of $[n]$ into k clusters. A visualization is a 2-dimensional embedding $\mathbf{Y} = \{y_1, ..., y_n\} \subseteq \mathbb{R}^2$ of $\mathbf{X}$, where each $x_i \in \mathbf{X}$ is mapped to the corresponding $y_i \in \mathbf{Y}$.
Intuitively, a cluster C_l in the original data is visualized if the corresponding points in the 2-dimensional embedding Y are well-separated from all the rest.

Formalizing Visualization

Definition 1.1 (Visible Cluster)

Let $\mathbf{Y}$ be a 2-dimensional embedding of a dataset $\mathbf{X}$ with ground-truth clustering $C_1, ..., C_k$. Given $\epsilon \geq 0$, a cluster C_l in $\mathbf{X}$ is said to be $(1-\epsilon)$-visible in $\mathbf{Y}$ if there exist $\mathbf{P}, \mathbf{P_{err}} \subseteq [n]$ such that:
1.$|(\mathbf{P} \backslash C_l) \cup (C_l \backslash \mathbf{P})| \leq \epsilon \cdot |C_l|$ i.e. the number of False Positive points and False Negative points are relatively small compared with the size of the ground-truth cluster.
2.for every $i, i' \in \mathbf{P}$ and $j \in [n] \backslash (\mathbf{P} \cup \mathbf{P_{err}})$, $||y_i - y_{i'}|| \leq \frac{1}{2}||y_i - y_j||$ i.e. except some mistakenly embedded points, other clusters are far away from the current clusters.
In such a case, we say that $\mathbf{P}(1-\epsilon)$-visualize C_i in $\mathbf{Y}$.

Formalizing Visualization

Definition 1.2 (Visualization)

Let $\mathbf{Y}$ be a 2-dimensional embedding of a dataset $\mathbf{X}$ with ground-truth clustering $C_1, ..., C_k$. Given $\epsilon \geq 0$, we say that $\mathbf{Y}$ is a $(1-\epsilon)$-visualization of $\mathbf{X}$ if there exists a partition $\mathbf{P}_1, ..., \mathbf{P}_k, \mathbf{P}_{err}$ of $[n]$ such that:
(i) For each $i \in [k]$, $\mathbf{P}_i(1-\epsilon)$-visualizes C_i in $\mathbf{Y}$.
(ii) $|\mathbf{P_{err}}| \leq \epsilon n$ i.e. the proportion of mistakenly embedded points must be small.
When $\epsilon = 0$, we call $\mathbf{Y}$ a full visualization of $\mathbf{X}$.

Definition of Well-separated and Spherical

Definition 1.4 (Well-separated, spherical data)

Let $\mathbf{X} = \{x_1, ..., x_n\} \subset \mathbb{R}^d$ be clusterable data with $C_1, ..., C_k$ defining the individual clusters such that for each $l \in [k]$, $|C_l| \geq 0.1(n/k)$. We say that $\mathbf{X}$ is γ-spherical and γ-well-separated if for some $b_1, ..., b_k > 0$, we have:

1.γ-**Spherical**: For any $l \in [k]$ and $i, j \in C_l (i \neq j)$, we have $||x_i - x_j||^2 \geq \frac{b_l}{1+\gamma}$, and for $i \in C_l$ we have $|\{j \in C_l \backslash \{i\} : ||x_i - x_j||^2 \leq b_l\}| \geq 0.51|C_l|$i.e. for any point, points from the same cluster are not too close with it and at least half of them are not too far away.

2.γ-**Well-separated**: For any $l, l' \in [k](l \neq l'), i \in C_l$ and $k \in C_l'$, we have $||x_i - x_j||^2 \geq (1 + \gamma \log n) \max\{b_l, b_{l'}\}$i.e. for any point, points from other clusters are far away.

Main Result

Theorem 3.1

Let $\mathbf{X} = \{x_1, ..., x_n\} \subset \mathbb{R}^d$ be a γ-spherical and γ-well-separated clusterable data with $C_1, ..., C_k$ defining k individual clusters of size at least $0.1(n/k)$, where $k << n^{1/5}$. Choose $\tau_i^2 = \frac{\gamma}{4} \cdot \min_{j \in [n] \setminus \{i\}} ||x_i - x_j||^2 (\forall i \in [n])$, step size $h = 1$, and any early exaggeration coefficient α satisfying $k^2 \sqrt{n} \log n << \alpha << n$.
Let $\mathbf{Y}^{(T)}$ be the output of t-SNE after $T = \Theta(\frac{n \log n}{\alpha})$ iterations on input $\mathbf{X}$ with the above parameters. Then, with probability at least 0.99 over the choice of the initialization, $\mathbf{Y}^{(T)}$ is a full visualization of $\mathbf{X}$.

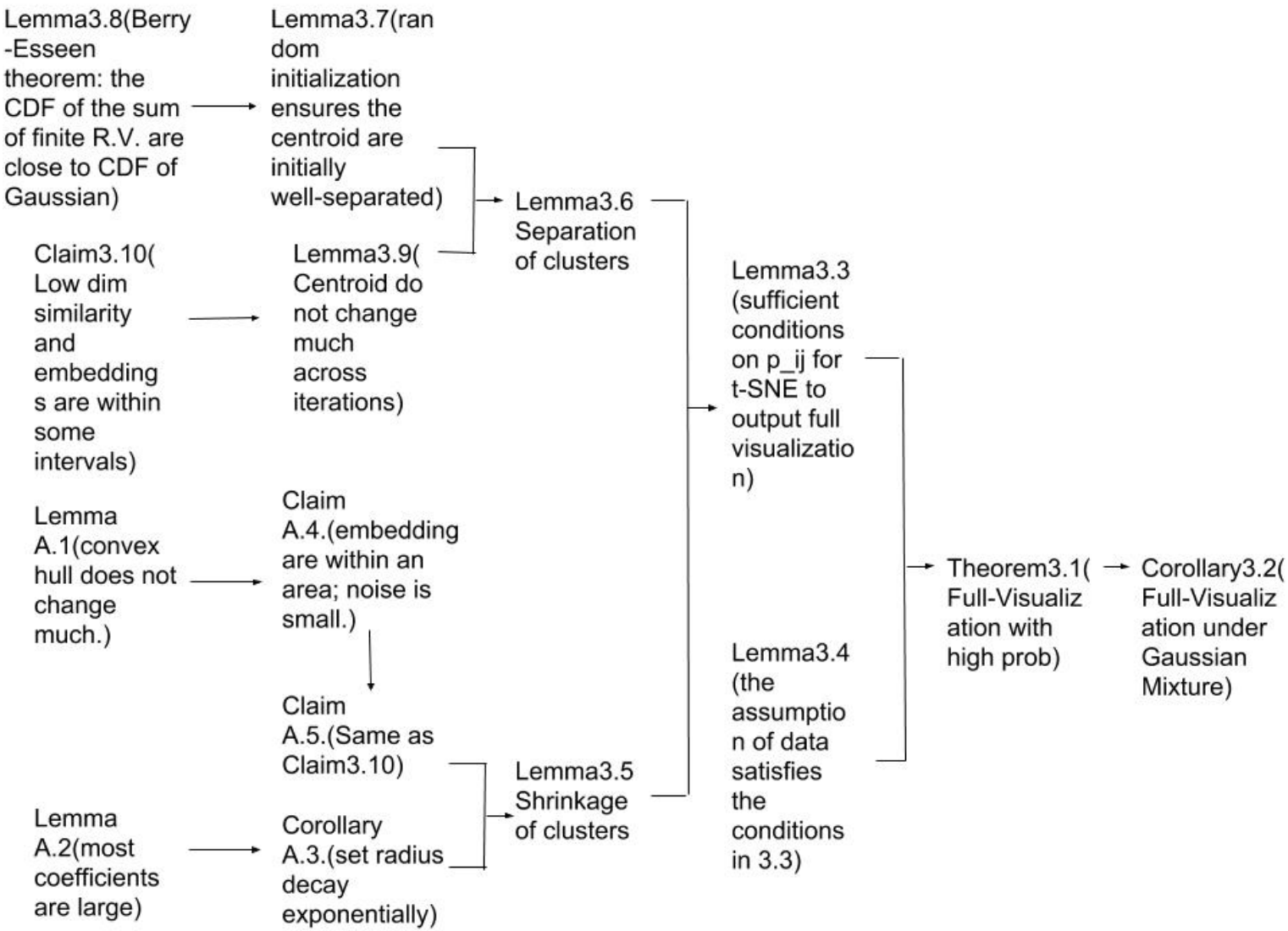
As γ becomes smaller, t-SNE requires more separations among points in the same cluster but less separation between individual clusters in $\mathbf{X}$ in order to succeed in finding a full visualization of $\mathbf{X}$.

Main Result

Corollary 3.2

Let $\mathbf{X} = \{x_1, ..., x_n\}$ be generated i.i.d. from a mixture of k Gaussians $\mathbf{N}(\mu_i, \mathbf{I})$ whose means $\mu_1, ..., \mu_k$ satisfy $||\mu_l - \mu_{l'}|| = \tilde{\Omega}(d^{1/4})$($d$ is the dimension of the embedded space) for any $l \neq l'$.
Let $\mathbf{Y}$ be the output of the t-SNE algorithm with early exaggeration when run on input $\mathbf{X}$ with parameters from Theorem 3.1. Then with high probability over the draw of $\mathbf{X}$ and the choice of the random initialization, $\mathbf{Y}$ is a full visualization of $\mathbf{X}$.

Proof Road-map



Recall: h is the step size and α is the early exaggeration coefficient.

Lemma 3.3. *Consider a dataset $\mathcal{X} = \{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{R}^d$ with ground-truth clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$ satisfying $|\mathcal{C}_\ell| \geq 0.1(n/k)$ for each $\ell \in [k]$ and $k \ll n^{1/5}$. Let $\{p_{ij} : i, j \in [n], i \neq j\}$ be the pairwise affinities in t-SNE computed by (1). Suppose that there exist $\delta, \epsilon, \eta > 0$ such that $\{p_{ij}\}$, α and h in t-SNE (Algorithm 1) satisfy:*

(i) for any cluster ℓ, and any point $i \in C_\ell$, we have $\left|\left\{j \in C_\ell : \alpha h p_{ij} \geq \frac{\delta}{|C_\ell|}\right\}\right| \geq \left(\frac{1}{2} + \eta\right)|C_\ell|$;

(ii) for any cluster ℓ, and any point $i \in C_\ell$, we have $\alpha h \sum_{j \in C_\ell \setminus \{i\}} p_{ij} \leq 1$;

(iii) for any cluster ℓ, and any point $i \in C_\ell$, we have $\alpha h \sum_{j \notin C_\ell} p_{ij} + \frac{h}{n} \leq \epsilon$;

(iv) $\frac{\epsilon \log \frac{1}{\epsilon}}{\delta \eta} \ll \frac{1}{k^2 \sqrt{n}}$.

Then, with high probability over the choice of the random initialization, the output $\mathcal{Y}^{(T)}$ of t-SNE after $T = \Theta\left(\frac{\log \frac{1}{\epsilon}}{\delta \eta}\right)$ iterations is a full visualization of $\mathcal{X}$.

Remarks: For any point, (i)most points from the same cluster are not far away.(ii)All points from the same clusters cannot be too close to it.(iii)All points from the same clusters are far away from it. (iv) for math in Lemma3.6 to work out.

Proof of Theorem 3.1

Lemma 3.4. *Let $\mathcal{X} = \{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{R}^d$ be γ-spherical and γ-well-separated clusterable data with $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ defining the individual clusters such that $|\mathcal{C}_i| \geq 0.1n/k$ for every i. Let $p_{i,j}$'s be the affinities computed by t-SNE (Algorithm 1) with parameters $\tau_i^2 = \frac{\gamma}{4} \cdot \min_{j \in [n] \setminus \{i\}} \|x_i - x_j\|^2$ $(\forall i \in [n])$, $h = 1$, and any α satisfying $k^2\sqrt{n}\log n \ll \alpha \ll n$.*

Then, p_{ij}'s satisfy (i)-(iv) in Lemma 3.3 with $\delta = \Theta(\alpha/n), \epsilon = 2/n$ and $\eta = 0.01$.

Proof.

Lemma 3.3 identifies sufficient conditions on the pairwise affinities p_{ij}'s that imply that t-SNE outputs a full visualization, and Lemma3.4 shows that $p_{i,j}$'s computed for γ-spherical, γ-well-separated data satisfy the requirements in Lemma 3.3. Therefore, combining Lemma3.3 and 3.4 gives Theorem 3.1. □

Lemmas needed for Lemma3.3

Lemma 3.5 (Shrinkage of clusters). *Under the same setting as Lemma 3.3, after running t-SNE for* $T = \Theta\left(\frac{\log \frac{1}{\epsilon}}{\delta\eta}\right)$ *rounds, we have* $\mathsf{Diam}\left(\left\{y_i^{(T)} : i \in \mathcal{C}_\ell\right\}\right) = O(\frac{\epsilon}{\delta\eta})$ *for all* $\ell \in [k]$.

In the second part, we establish that points in different clusters remain separated in the embedding if the clusters are well-separated in the input data. Concretely, let $\mu_\ell^{(t)} := \frac{1}{|\mathcal{C}_\ell|}\sum_{i\in\mathcal{C}_\ell} y_i^{(t)}$, which is the *centroid* of $\left\{y_i^{(t)} : i \in \mathcal{C}_\ell\right\}$. The following lemma says that the centroids of all clusters will remain separated in the first $O\left(\frac{\log \frac{1}{\epsilon}}{\delta\eta}\right)$ rounds.

Lemma 3.6 (Separation of clusters). *Under the same setting as Lemma 3.3, if t-SNE is run for* $T = O\left(\frac{\log \frac{1}{\epsilon}}{\delta\eta}\right)$ *iterations, with high probability we have* $\left\|\mu_\ell^{(T)} - \mu_{\ell'}^{(T)}\right\| = \Omega\left(\frac{1}{k^2\sqrt{n}}\right)$ *for all* $\ell \neq \ell'$.

We can finish the proof of Lemma 3.3 using the above two lemmas.

Proof of Lemma3.3

- Using Lemmas3.5 and 3.6, we know that after $T = \Theta(\frac{\log \frac{1}{\epsilon}}{\delta\eta})$ iterations, for any $i, j \in [n]$ we have:
- if $i \sim j$, then $||y_i^{(T)} - y_j^{(T)}|| \leq Diam(\{y_l^{(t)} : l \in C_{\pi(i)}\}) = O(\frac{\epsilon}{\delta\eta})$
- if $i \not\sim j$, then

$$\begin{aligned}
||y_i^{(T)} - y_j^{(T)}|| &\geq ||\mu_{\pi(i)}^{(T)} - \mu_{\pi(j)}^{(T)}|| - ||y_i^{(T)} - \mu_{\pi(i)}^{(T)}|| - ||y_j^{(T)} - \mu_{\pi(j)}^{(T)}|| \\
&\geq ||\mu_{\pi(i)}^{(T)} - \mu_{\pi(j)}^{(T)}|| - Diam(\{y_l^{(t)} : l \in C_{\pi(i)}\}) \\
&\quad - Diam(\{y_l^{(t)} : l \in C_{\pi(j)}\}) \\
&\geq \Omega(\frac{1}{k^2\sqrt{n}}) - O(\frac{\epsilon}{\delta\eta}) - O(\frac{\epsilon}{\delta\eta}) \\
&= \Omega(\frac{1}{k^2\sqrt{n}}) \\
&>> O(\frac{\epsilon}{\delta\eta})
\end{aligned}$$

Recall Lemma3.6

Lemma 3.6 (Separation of clusters). *Under the same setting as Lemma 3.3, if t-SNE is run for* $T = O\left(\frac{\log \frac{1}{\epsilon}}{\delta\eta}\right)$ *iterations, with high probability we have* $\left\|\mu_\ell^{(T)} - \mu_{\ell'}^{(T)}\right\| = \Omega\left(\frac{1}{k^2\sqrt{n}}\right)$ *for all* $\ell \neq \ell'$.

We can finish the proof of Lemma 3.3 using the above two lemmas.

Lemmas needed for Lemma3.6

Random initialization ensures that the cluster centroids are initially well-separated with high probability.

Lemma 3.7. *Suppose* $|\mathcal{C}_\ell| \geq 0.1(n/k)$ *for all* $\ell \in [k]$. *If* $y_i^{(0)}$*'s are generated i.i.d. from the uniform distribution over* $[-0.01, 0.01]^2$, *then with probability at least* 0.99 *we have* $\left\|\mu_\ell^{(0)} - \mu_{\ell'}^{(0)}\right\| = \Omega\left(\frac{1}{k^2\sqrt{n}}\right)$ *for all* $\ell \neq \ell'$.

The centroid of each cluster will move no more than ϵ in each of the first $\frac{0.01}{\epsilon}$ iterations.

Lemma 3.9. *Under the same setting as Lemma 3.3, for all* $t \leq \frac{0.01}{\epsilon}$ *and all* $\ell \in [k]$ *we have* $\left\|\mu_\ell^{(t+1)} - \mu_\ell^{(t)}\right\| \leq \epsilon$.

Proof of Lemma3.6

By condition (iv) in Lemma3.3 ($\frac{\epsilon \log \frac{1}{\epsilon}}{\delta\eta} << \frac{1}{k^2\sqrt{n}}$), We have $O(\frac{\log \frac{1}{\epsilon}}{\delta\eta}) << \frac{1}{k^2\sqrt{n}\epsilon} < \frac{0.01}{\epsilon}$. Hence we can apply Lemma3.9 for all $t \leq T$.
Lemma3.7 says that the initial distance $\mu_l^{(0)}$ and $\mu_{l'}^{(0)}$ is at least $\Omega(\frac{1}{k^2\sqrt{n}})$ with high probability.
Lemma3.9 says that after each iteration every centroid moves by at most ϵ so the distance between any two centroids changes by at most 2ϵ.
We know that after T rounds, with high probability, we have

$$\begin{aligned} ||\mu_l^{(T)} - \mu_{l'}^{(T)}|| &\geq \Omega(\frac{1}{k^2\sqrt{n}}) - T \cdot 2\epsilon \\ &= \Omega(\frac{1}{k^2\sqrt{n}}) - O(\frac{\epsilon \log \frac{1}{\epsilon}}{\delta\eta}) \\ &= \Omega(\frac{1}{k^2\sqrt{n}}) \end{aligned}$$

, where the last step is due to condition (iv) in Lemma3.6.

Recall Lemma3.7

Lemma 3.7. *Suppose* $|\mathcal{C}_\ell| \geq 0.1(n/k)$ *for all* $\ell \in [k]$. *If* $y_i^{(0)}$*'s are generated i.i.d. from the uniform distribution over* $[-0.01, 0.01]^2$, *then with probability at least* 0.99 *we have* $\left\|\mu_\ell^{(0)} - \mu_{\ell'}^{(0)}\right\| = \Omega\left(\frac{1}{k^2\sqrt{n}}\right)$ *for all* $\ell \neq \ell'$.

It is suffices to prove that $|(\mu_l)_1 - (\mu_{l'})_1| = \Omega(\frac{1}{k^2\sqrt{n}})$ for all $l \neq l'$.

Lemma needed for Lemma3.7

Lemma 3.8 (Berry-Esseen theorem (Berry, 1941; Esseen, 1942)). *Suppose that $X_1, X_2, \ldots, X_m$ are i.i.d. random variables with $\mathbb{E}[X_1] = 0$, $\mathbb{E}[X_1^2] = \sigma^2 < \infty$ ($\sigma > 0$) and $\mathbb{E}[|X_1|^3] = \zeta < \infty$. Let $Y_m := \frac{1}{m}\sum_{i=1}^{m} X_i$, F_m be the cumulative distribution function (CDF) of $\frac{Y_m\sqrt{m}}{\sigma}$, and Φ be the CDF of the standard normal distribution $\mathcal{N}(0,1)$. Then there exists a universal constant C such that for all $x \in \mathbb{R}$ and all $m \in \mathbb{N}$ we have*

$$|F_m(x) - \Phi(x)| \leq \frac{C\zeta}{\sigma^3\sqrt{m}}.$$

Remark: This is similar to Central Limit Theorem. It says that the CDF of the average of i.i.d random variables is close to the standard normal's CDF.

Proof of Lemma3.7 Part1

Consider a fixed $I \in [k]$. Note that $(y_i)_1$'s i.i.d. with uniform distribution over $[-0.01, 0.01]$, which clearly has zero mean and finite second and third absolute moments.
Since $(\mu_I)_1 = \frac{1}{|C_I|}\sum_{i \in C_I}(y_i)_1$, using the Berry-Esseen theorem we know that

$$|F(x) - \phi(x)| \leq O(1/\sqrt{|C_I|})$$

where F is the CDF of $\frac{(\mu_I)_1\sqrt{|C_I|}}{\sigma}$ (σ is the standard deviation of the uniform distribution over $[-0.01, 0.01]$), and ϕ is the CDF of $N(0, 1)$.

Proof of Lemma3.7 Part2

It follows that for any fixed $a \in \mathbb{R}$ and $b > 0$, we have:

$$
\begin{aligned}
Pr\Big[|(\mu)_1 - a| \leq \frac{b}{k^2\sqrt{|C_l|}}\Big] &= Pr\Big[\Big|\frac{(\mu)_1\sqrt{|C_l|}}{\sigma} - \frac{a\sqrt{|C_l|}}{\sigma}\Big| \leq \frac{b}{k^2\sigma}\Big] \\
&= F\Big(\frac{\alpha\sqrt{|C_l|} + b/k^2}{\sigma}\Big) - F\Big(\frac{\alpha\sqrt{|C_l|} - b/k^2}{\sigma}\Big) \\
&\leq \Phi\Big(\frac{\alpha\sqrt{|C_l|} + b/k^2}{\sigma}\Big) - \Phi\Big(\frac{\alpha\sqrt{|C_l|} - b/k^2}{\sigma}\Big) \\
&+ O\Big(\frac{1}{\sqrt{|C_l|}}\Big) \\
&= \int_{\frac{\alpha\sqrt{|C_l|}-b/k^2}{\sigma}}^{\frac{\alpha\sqrt{|C_l|}+b/k^2}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx + O(\sqrt{k/n})
\end{aligned}
$$

Remark: The last equality is by the definition of CDF for standard normal distribution and the assumption that $|C_l| \geq 0.1\frac{n}{k}$.

Proof of Lemma3.7 Part3

$$\leq \int_{\frac{\alpha\sqrt{|C_I|}-b/k^2}{\sigma}}^{\frac{\alpha\sqrt{|C_I|}+b/k^2}{\sigma}} \frac{1}{\sqrt{2\pi}} dx + O(\sqrt{k/n})$$

$$= \frac{1}{\sqrt{2\pi}} \frac{2b}{k^2\sigma} + O(\sqrt{k/n})$$

From $k << n^{1/5}$ we have $\sqrt{k/n} << 1/k^2$. Therefore, letting b be a sufficiently small constant, for any $a \in \mathbb{R}$, we can ensure

$$Pr\left[|(\mu_I)_1 - a| \leq \frac{b}{k^2\sqrt{|C_I|}}\right] \leq \frac{0.01}{k^2}$$

Proof of Lemma3.7 Part4

For any $l' \neq l$, because $(\mu_l)_1$ and $(\mu_{l'})_1$ are independent, we can let $a = (\mu_{l'})_1$,

$$Pr\left[|(u_l)_1 - |(u_{l'})_1| \leq \frac{b}{k^2\sqrt{|C_l|}}\right] \leq \frac{0.01}{k^2}$$

The above inequality holds for any $l, l' \in [k](l \neq l')$. Taking a union bound over all l and l', we know that with probability at least 0.99 we have $|(\mu_l)_1 - (\mu_{l'})_1 \geq \frac{b}{k^2\sqrt{|C_l|}} = \Omega(\frac{1}{k^2\sqrt{n}})$ for all $l, l' \in [k](l \neq l')$ simultaneously(Recall $|C_l| \geq 0.1(n/k)$).

Claim needed for Lemma3.9

Claim 3.10 (Same as Claim A.5). *Under the same setting as Lemma 3.3, for all* $t \le \frac{0.01}{\epsilon}$, *we have* $y_i^{(t)} \in [-0.02, 0.02]^2$ *and* $\left\|\epsilon_i^{(t)}\right\| \le \epsilon$ *for all* $i \in [n]$, *as well as* $\left\|y_i^{(t)} - y_j^{(t)}\right\| \le 0.06$, $0.9 \le q_{ij}^{(t)} Z^{(t)} \le 1$ *and* $\frac{0.9}{n(n-1)} \le q_{ij}^{(t)} \le \frac{1}{0.9n(n-1)}$ *for all* $i, j \in [n]$ $(i \ne j)$.

where

$$\epsilon_i^{(t)} := \alpha h \sum j \not\sim i p_{ij} q_{ij}^{(t)} Z^{(t)} (y_j^{(t)} - y_i^{(t)}) - h \sum_{j \ne i} (q_{ij}^{(t)})^2 Z^{(t)} (y_j^{(t)} - y_i^{(t)})$$

Proof of Lemma3.9 Part1

Taking the average of $y_i^{(t+1)} \forall\ i \in C_l$, we obtain:

$$\begin{aligned}
\frac{1}{|C_l|}\sum_{i\in C_l} y_i^{(t+1)} &= \frac{1}{|C_l|}\sum_{i\in C_l} y_i^{(t)} + \frac{h}{C_l}\sum_{i\in C_l}\sum_{j\neq i}(\alpha p_{ij} - q_{ij}^{(t)})q_{ij}^{(t)} Z^{(t)}(y_j^{(t)} - y_i^{(t)}) \\
&= \frac{1}{|C_l|}\sum_{i\in C_l} y_i^{(t)} \\
&+ \frac{h}{C_l}\sum_{i\in C_l}\sum_{j\in C_l, j\neq i}(\alpha p_{ij} - q_{ij}^{(t)})q_{ij}^{(t)} Z^{(t)}(y_j^{(t)} - y_i^{(t)}) \\
&+ \frac{h}{C_l}\sum_{i\in C_l}\sum_{j\notin C_l}(\alpha p_{ij} - q_{ij}^{(t)})q_{ij}^{(t)} Z^{(t)}(y_j^{(t)} - y_i^{(t)}) \\
&= \frac{1}{|C_l|}\sum_{i\in C_l} y_i^{(t)} + \frac{h}{C_l}\sum_{i\in C_l}\sum_{j\notin C_l}(\alpha p_{ij} - q_{ij}^{(t)})q_{ij}^{(t)} Z^{(t)}(y_j^{(t)} - y_i^{(t)})
\end{aligned}$$

Proof of Lemma3.9 Part2

Thus we have:

$$||\mu_l^{t+1} - \mu_l^t|| = ||\frac{h}{C_l}\sum_{i\in C_l}\sum_{j\notin C_l}(\alpha p_{ij} - q_{ij}^t)q_{ij}^{(t)}Z^{(t)}(y_j^{(t)} - y_i^{(t)})||$$

$$\leq \frac{h}{|C_l|}\sum_{i\in C_l}\sum_{j\notin C_l}(\alpha p_{ij} + q_{ij}^{(t)}q_{ij}^{(t)}Z^{(t)}||y_j^{(t)} - y_i^{(t)}||)$$

Since $t \leq \frac{0.01}{\epsilon}$ and $\alpha h\sum_{j\notin C_l} p_{ij} + \frac{h}{n} \leq \epsilon$ for all $i \in C_l$ (condition(iii) in Lemma 3.3), we can apply Claim 3.10 and get:

$$||\mu_l^{t+1} - \mu_l^t|| \leq \frac{h}{C_l}\sum_{i\in C_l}\sum_{j\notin C_l}(\alpha p_{ij} + \frac{1}{0.9n(n-1)})\cdot 1\cdot 0.06$$

$$\leq \frac{h}{C_l}\sum_{i\in C_l}(\alpha h\sum_{j\notin C_l} p_{ij} + \frac{h}{0.9n})\cdot 0.06$$

$$\leq \frac{0.06}{|C_l|}\sum_{i\in C_l}\frac{\epsilon}{0.9} \leq \epsilon$$

Results on mixture of Log-Concave Distributions

The results on mixture of isotropic Gaussian or isotropic log-concave distributions can be generalized to mixture of (non-isotropic) log-concave distributions.

Corollary 3.11. *Let $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ be i.i.d. samples from a mixture of k log-concave distributions with means $\mu_1, \mu_2, \ldots \mu_k$, covariances $\Sigma_1, \Sigma_2, \ldots, \Sigma_k$ and all mixing weights at least $\frac{0.2}{k}$. Let $\sigma_\ell := \sqrt{\|\Sigma_\ell\|_2}$ for every $\ell \in [k]$. Suppose that the radius of each of the k components is equal to $R = \sqrt{\mathrm{tr}(\Sigma_\ell)}$ and suppose $R/\sigma_\ell > d^\eta$ for some constant $\eta > 0$ for every $\ell \in [k]$. Finally, suppose that the means have separation $\|\mu_\ell - \mu_{\ell'}\| \gg Rd^{-\eta/6} \log^{2/3} n$ for all $\ell \neq \ell'$. Let $\mathcal{Y}$ be the output of the t-SNE (Algorithm 1) with the same parameter choices as in Theorem 3.1 when run on $\mathcal{X}$. Then, with probability at least 0.99 over the choice of the random initialization and the draw of $\mathcal{X}$, $\mathcal{Y}$ is a full visualization of $\mathcal{X}$.*

Results on partial visualization via t-SNE

Theorem 4.1. *Suppose $n > 4d$. Let $\mathcal{C}_1, \mathcal{C}_2$ be a partition of $[n]$ such that $|\mathcal{C}_1| = |\mathcal{C}_2| = n/2$. Let $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ be generated from the mixture of two Gaussians $\mathcal{N}(0, \sigma_1^2)$ and $\mathcal{N}(0, \sigma_2^2)$ such that $1.5 \leq \frac{\sigma_2}{\sigma_1} \leq 10$,[2] where x_i is generated from $\mathcal{N}(0, \sigma_\ell^2)$ if $i \in \mathcal{C}_\ell$ $(\ell = 1, 2)$. Choose $\tau_i^2 = \frac{1}{\sqrt{2}} \cdot \min_{j \in [n] \setminus \{i\}} \|x_i - x_j\|^2$ $(\forall i \in [n])$, $h = 1$, and $\alpha = \rho n$ for a sufficiently small constant $\rho > 0$.*

Let $\mathcal{Y}^{(T)}$ be the output of t-SNE (Algorithm 1) after $T = \Theta(\log d)$ iterations on input $\mathcal{X}$ with the above parameters. Then, with high probability over the choice of the initialization, $\mathcal{Y}^{(T)}$ is a $(1 - d^{-\Omega(1)})$-partial visualization of $\mathcal{X}$ where $\mathcal{C}_1$ is $(1 - d^{-\Omega(1)})$-visible.

Limitations

- $k << n^{1/5}$ can be potentially improved.
- It has implicit assumption that $k^2\sqrt{n} > 100$

Generalize t-SNE

Stochastic Neighbor Embedding under f-divergences. Verma et al. Preprint.

Table 1: ft-SNE

$D_f(P\|Q)$	$f(t)$	ft-SNE objective	Emphasis
Kullback-Leibler (KL)	$t \log t$	$\sum p_{ij} \left(\log \frac{p_{ij}}{q_{ij}} \right)$	Local
Chi-square ($\mathcal{X}^2$ or CS)	$(t-1)^2$	$\sum \frac{(p_{ij}-q_{ij})^2}{q_{ij}}$	Local
Reverse-KL (RKL)	$-\log t$	$\sum q_{ij} \left(\log \frac{q_{ij}}{p_{ij}} \right)$	Global
Jensen-Shannon (JS)	$(t+1) \log \frac{2}{(t+1)} + t \log t$	$\frac{1}{2}(J_{\text{KL}} + J_{\text{RKL}})$	Both
Hellinger distance (HL)	$(\sqrt{t}-1)^2$	$\sum (\sqrt{p_{ij}} - \sqrt{q_{ij}})^2$	Both

Results

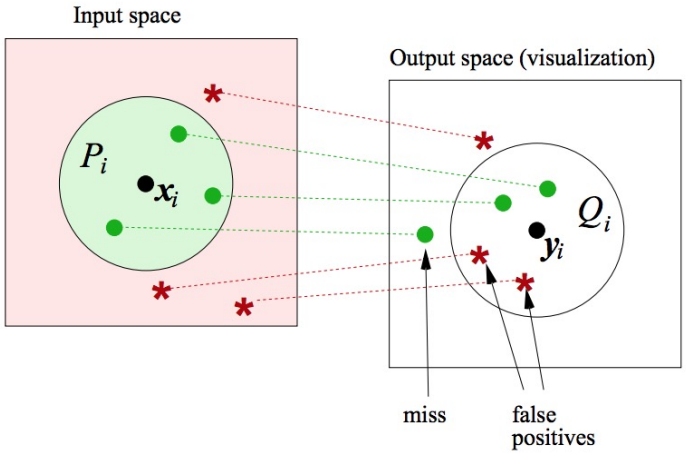


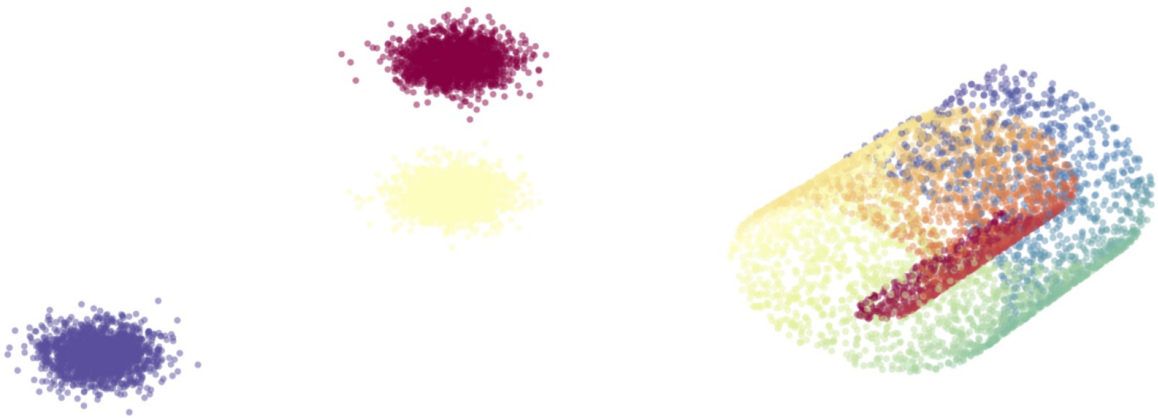
Figure 1: Diagram of the types of errors in visualization.

(i) $J_{KL} \propto \left(\sum_i n^i_{FN}/r_i \right)$ *and hence KL-SNE maximizes for recall.*

(ii) $J_{RKL} \propto \left(\sum_i n^i_{FP}/k_i \right)$ *and hence RKL-SNE maximizes for precision.*

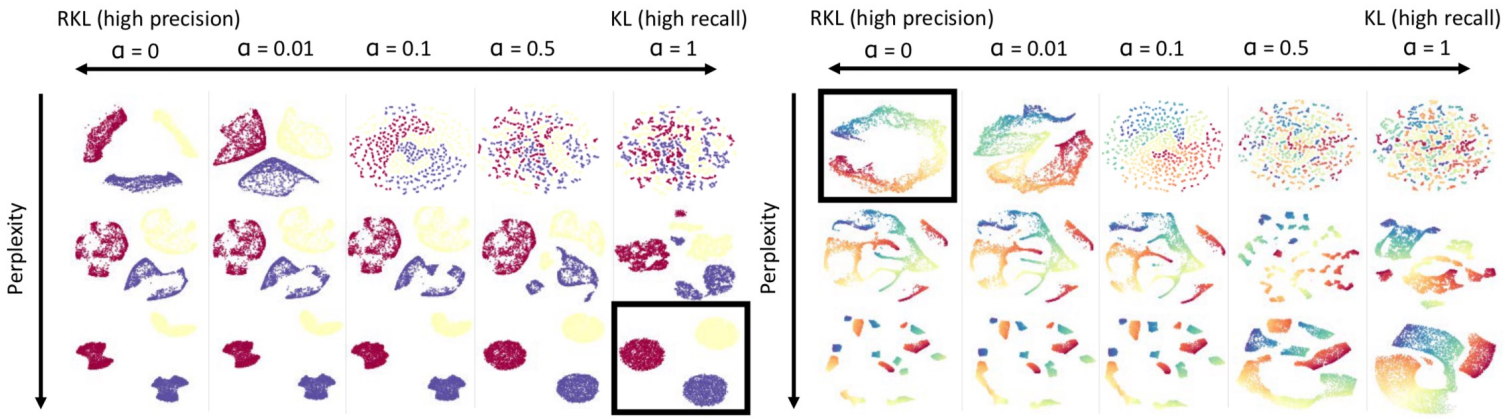
(iii) $J_{JS} = \frac{1}{2}(J_{KL} + J_{RKL})$ *and hence JS-SNE balances the precision and recall rates.*

Results



(a) Three Gaussian clusters

(b) Swiss Roll



(a) 3 well-separated Gaussian clusters

(b) Swiss roll manifold

Figure 4: ft-SNE embeddings obtained with interpolated divergences between KL and RKL.

Results

Table 2: Best ft-SNE method for each dataset and criterion, according to maximum F-score in Figure 8.

		Data-Embeddings			Class-Embedings
Data	Type	K-Nearest	K-Farthest	F-Score on X-Y	F-Score on Z-Y
MNIST (Digit 1)	Manifold	RKL	RKL	RKL	-
Face	Manifold	HL,RKL	RKL	RKL	JS
MNIST	Clustering	KL	KL	CS	KL
GENE	Clustering	KL	KL	KL	KL
20 News Groups	Sparse & Hierachical	CS	CS	CS	HL

Summary

- SNE $\rightarrow$ tSNE(crowding problem)
- t-SNE $\rightarrow$ BH-tSNE($O(N^2) \rightarrow O(N \log N)$)
- BH-SNE $\rightarrow$ FIt-tSNE($O(N \log N) \rightarrow O(N)$)
- t-SNE guarantee(clusters shrink)
- More t-SNE guarantee(centroids keep well-separated $\Rightarrow$ for mixtures of well-separated log-concave distributions $\rightarrow$ full-visualization and for general settings $\rightarrow$ partial-visualization).
- t-SNE with KL is for dataset with cluster-like structure. RKL is for dataset with manifold-like structure.

Future Work

- $k << n^{1/5}$ can potentially be improved.
- improve the visualization of t-SNE using the insight gained from its theoretical guarantee.
- automatically adjust the early/late exaggeration parameter α and choose the appropriate perplexity number.
- modify t-SNE for general dimensionality reduction problems(not just for visualization).
- Theoretical guarantee for t-SNE with RKL on manifold data.