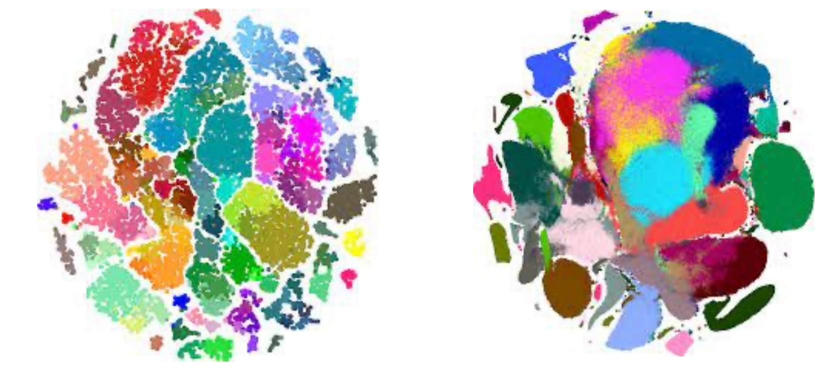
On Optimal t-distributed Stochastic Neighbor Embeddings



Noah Bergam (noah.bergam@columbia.edu), Nakul Verma (verma@cs.columbia.edu)

Background: t-distributed stochastic neighbor embedding (t-SNE), created by [MH08, HS02], is a nonlinear dimensionality reduction algorithm, provably good at visualizing cluster structure in high-dimensional data [AHK21, LS19].

Problem: Gradient-based optimization (practical implementation) only shows us local minima of the t-SNE objective.

Goal: What can we say about *global minima* or the t-SNE objective function?

t-SNE formulation and pathological cases...

We think of t-SNE as a graph embedding problem (more general than its original formulation as a metric embedding).

- Given: an N×N "affinity" matrix (P_{ij}) with zero diagonal, symmetric, non-negative, and all entries sum to 1.
- Construct low-dimensional points (y_i) and a corresponding affinity matrix (Q_{ij}), computed as follows

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

- Find (y_i) which minimizes the Kullback-Leibler divergence (relative entropy) of (P_{ij}) with respect to (Q_{ij}).

$$KL(P||Q) = \sum_{i=1}^{N}\sum_{j=1}^{N} P_{ij}\ln(P_{ij}/Q_{ij})$$

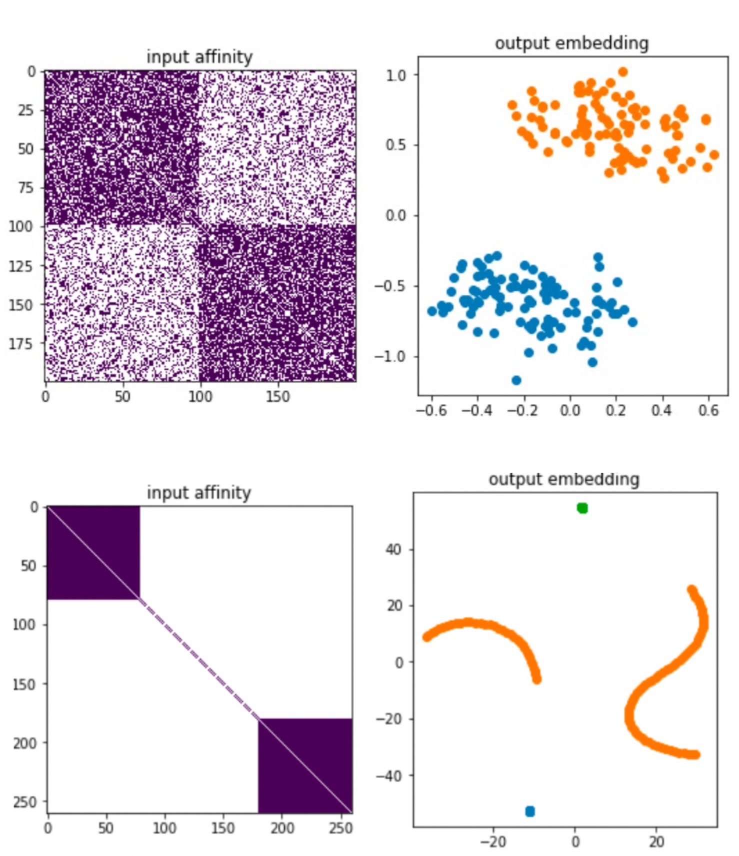
An (new) advantageous way of rewriting our loss:

$$\underset{y_1,\dots,y_n\in\mathbb{R}^d}{\arg\min}\left[\underbrace{\sum_{i\neq j} p_{ij}\ln(1+||y_i-y_j||^2)}_{\text{contraction}} + \underbrace{\ln\left(\sum_{i\neq j}\frac{1}{1+||y_i-y_j||^2}\right)}_{\text{repulsion}}\right]$$

Observation 1: Non-metric embeddable graphs such as stochastic block models and "clique-path" graphs are still well-clustered by t-SNE.

Observation 2: However, this generalization admits simple examples cases where:

- The optimal embedding is **trivial**
- No optimal embedding exists (i.e. the infimum of the objective isn't attained)



t-SNE on non-metric graphs. Capable of clustering planted partition graphs with p as low as 0.55

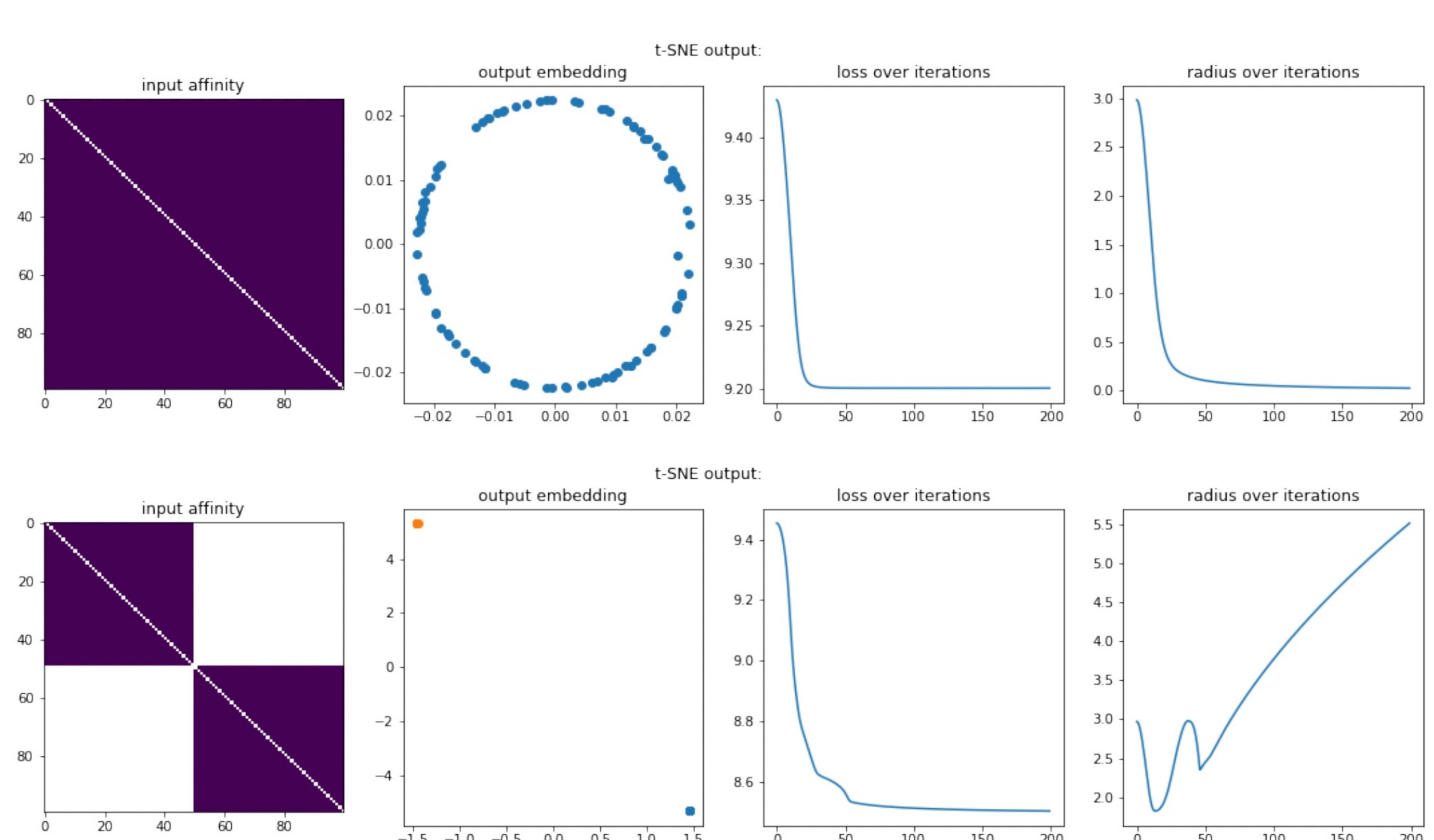


Illustration of two "pathological" cases of the t-SNE embedding, and how gradient descent optimization does not converge but rather contracts (top) and expands (bottom) indefinitely.

Supported by the Pritzker-Pucker Summer 2023 Funding Program.

Low-diameter t-SNE approximates Laplacian Eigenmaps: A new demonstration

[CM22] established a rigorous connection between gradient-optimized t-SNE and spectral clustering. We build upon this connection, showing that the t-SNE objective, in low-diameter regimes, is approximately equal to the objective of Laplacian eigenmaps, a spectral method.

Theorem 10 (Approximate Spectral Clustering). *Let $\boldsymbol{y} = (y_1, \dots y_n) \in \mathbb{R}^{n\times 1}$ and a modified (but still equivalent for optimization purposes) t-SNE objective function $L_P(\boldsymbol{y}) = KL(P||Q(\boldsymbol{y})) - H(P) - \ln(n^2-n)$. If $diam(\boldsymbol{y}) := d_{\boldsymbol{y}} < 1$, then:*

$$\left|\boldsymbol{y}^T L(P - H_n)\boldsymbol{y} - L_P(\boldsymbol{y})\right| = O(n^2 d_{\boldsymbol{y}}^4)$$

where $L(\cdot)$ is the graph Laplacian of an $n\times n$ matrix and $H_n = \frac{1}{n^2-n}(\boldsymbol{1}\boldsymbol{1}^T - I_n)$.

The proof is simple and involves mostly Taylor expansions on the loss function. It is relevant because the canonical implementation of t-SNE is in a small radius of $[0.01, 0.01]^2$,

A slight symmetry, as an artifact of normalization

We say the loss function has a *symmetry* if a non-identity transformation of embedding leaves the loss value the same. We found an infinite family of symmetries.

Theorem 1. *For almost every $(y_1, \dots, y_n) = Y \in \mathbb{R}^{dn}$, there exists an $\epsilon > 0$ and an infinite family of embeddings $\{Y_\alpha\}_{\alpha\in\mathcal{A}} \subset \mathbb{R}^{dn}$ such that:*

$$||D(Y) - D(Y_\alpha)||_\infty \in (0, \epsilon) \quad \text{and} \quad L_P(Y) = L_P(Y_\alpha) \quad \forall\alpha\in\mathcal{A}$$

where $D(Y)$ is the matrix of squared distances, $[D(Y)]_{ij} = ||y_i - y_j||^2$.

It is easy to find a non-identity transformation of the distance matrix which preserves the objective. The hard part of the proof is showing that this transformation of the distances is still Euclidean-embeddable. This involves the use of Gram matrices and some topological reasoning about the positive-semidefinite cone.

Diameter bound and open questions

We prove that a P matrix with non-zero off-diagonal entries will always yield an optimal t-SNE embedding. Furthermore, we find that this optimal embedding will occur within a finite radius dependent on the number of points *n* and the smallest off-diagonal entry in P.

Proposition 7 (Diameter Bound). *Given a P matrix with $\min_{i\neq j} P_{ij} \geq \frac{1}{Cn^2}$ for $C \in \mathbb{R}_{>0}$, there exists an optimal embedding Y^* with diameter $O(n^{n/c})$. Specifically:*

$$\exists Y^* \in \mathbb{R}^{dn} \ s.t. \quad \max_{y,y'\in Y^*} ||y - y'|| \leq 2n^{nC}$$

where $L_P(Y^*) \leq L_P(Y)$ for all $Y \in \mathbb{R}^{dn}$.

This bound is likely not tight. In well-clustered settings, can show a much tighter n^2 dependence.

Remaining questions are centered on the **hardness and approximability of t-SNE.**

- Is t-SNE optimization NP-hard? Perhaps reduce from NAE-3SAT* or multi-terminal cuts?
- Can we develop a poly-time approximation scheme (PTAS) for t-SNE, in a similar vein as [DHKLU21]'s PTAS for multi-dimensional scaling? The crucial first step towards this is <u>a tighter diameter bound</u>––this would allow us to efficiently discretize the input space and turn this continuous problem into a discrete one, upon which we can apply greedy methods.

Works Cited:

[HS02]: Geoffrey Hinton and Sam Roweis. Stochastic neighbor embedding. Advances in Neural Information Processing Systems (2002).

[AHK21]: Sanjeev Arora, Wei Hu, and Pravesh K Kothari. An analysis of the t-SNE algorithm for data visualization. In Conference on learning theory, pages 1455–1462. PMLR, 2018.

[LR19]: George C Linderman and Stefan Steinerberger. Clustering with t-SNE, provably. SIAM Journal on Mathematics of Data Science, 1(2):313–332, 2019.

[CM22]: T Tony Cai and Rong Ma. Theoretical foundations of t-SNE for visualizing high-dimensional clustered data. The Journal of Machine Learning Research, 23(1):13581–13634, 2022.

[MH08]: Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. The Journal of Machine Learning Research, 9(11), 2008.

[DHKLU21]: Demaine, Erik, et al. "Multidimensional scaling: Approximation and complexity." *International Conference on Machine Learning*. PMLR, 2021.