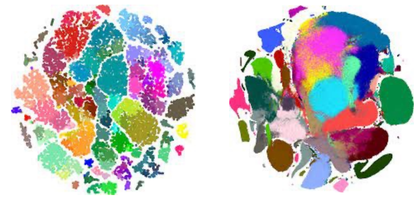
On Optimal t-distributed Stochastic Neighbor Embeddings



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 works well in practice, but it only shows us local minima of the t-SNE objective.

Goal: What can we say about *global minima* or the t-SNE objective function? Are they desirable? Can we hope to find them? Are the local minima "close enough"?

t-SNE formulation and special cases

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

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

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

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

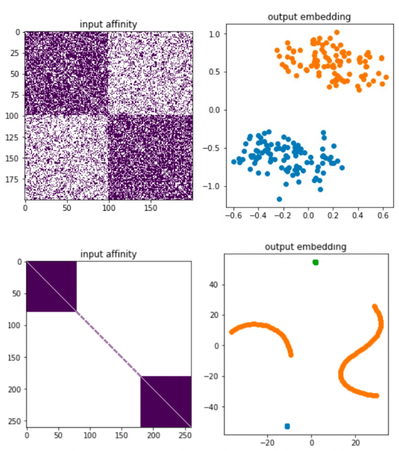
An (new) advantageous way of rewriting our loss:

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

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.

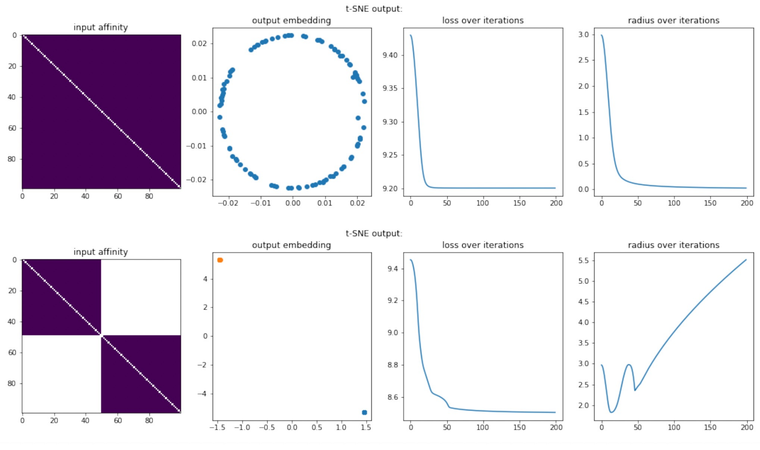


Illustration of two "pathological" cases of the t-SNE embedding, and how gradient descent optimization does not converge but rather contracts (top) and expands (bottom) indefinitely.

This motivates our first result:

Theorem 1 (Existence of minimizer). *Let P be a valid affinity matrix (symmetric, non-negative, zero diagonal, adds up to zero). Then $L_P(Y) = KL(P||Q_Y)$ attains its infimum if and only if P is irreducible (i.e. only one connected component).*

t-SNE approximates Laplacian Eigenmaps in the low-diameter regime

[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 2 (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 Taylor expansions on the loss function. The result plays out in practice because the canonical implementation of t-SNE is in a small radius of $[0.01, 0.01]^2$ The nice thing about this result is that it is independent of the optimization algorithm (whereas [CM22] only applies when we optimize t-SNE with gradient descent).

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 3 (Symmetry). *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.

Hardness and Approximation

Reducing from restricted not-all-equal 3SAT, we can show the following form of t-SNE is NP-hard.

Definition 1. *Let* RigidtSNE *be the following problem.*

- *Given an $n\times n$ affinity matrix P and $L \in \mathbb{R}$: return TRUE if there exists $Y = (y_1, \dots, y_n) \in \{0,1\}^n$ such that $L_P(Y) \leq L$. Otherwise, output FALSE.*

Theorem 4. RigidtSNE *is an* NP*-hard decision problem.*

The restriction is that we only allow points to sit at 0 or 1, rather than on, say, the entire real number line

The reduction is similar to [CD06]'s hardness proof of L1 embedding. Given some 3CNF formula on n literals, we construct an affinity matrix over 2n points (one for each literal) such that optimal t-SNE embedding of these points is a balanced clustering of positive and negative literals. With this established, it becomes clear that the cost is optimized precisely when the formula is NAE-3SAT*.

Next Step 1: How might we generalize this reduction to the usual t-SNE setup?
Proposal: Construct a gadget which forces the embedded points to sit at 0 or 1.

Next Step 2: If t-SNE optimization is NP-hard, does it admit a poly-time approximation scheme (PTAS)?

Proposal 1: Follow the lead of [DHKLU21]'s PTAS for for multi-dimensional scaling. The procedure would be as follows: first, establish a bound on the radius of optimal embeddings (one can trivially establish an $O(n^n)$ bound, but we would need a poly(n) bound for this to work). Then, discretize over a ball of this sufficiently large diameter and use some simple greedy algorithm to obtain an additive approximation of the discrete optimal. Translating this discrete optimal into the true optimal incurs further additive error, but if we can establish that this error is small enough, then the PTAS follows.

Proposal 2: Perhaps gradient descent, with enough random restarts, is indeed a PTAS for a t-SNE. Since t-SNE is very much a non-convex loss function (see Theorem 3), this would require a careful analysis of t-SNE's local minima. Are they close to the global minimum? Are they escapable (e.g. satisfying the "strict saddle property")?

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.
[CD06]: Cayton, Lawrence, and Sanjoy Dasgupta. "Robust euclidean embedding." *Proceedings of the 23rd international conference on machine learning*. 2006.

Presented at the **UMD Fall Fourier Talks 2023**. Work supported by the 2023 Pritzker Pucker Summer Funding Program.