On UMAP's true loss function

Sebastian Damrich

Fred Hamprecht

HCI/IWR at Heidelberg University

Heidelberg Collaboratory for Image Processing

Interdisciplinary Center for Scientific Computing

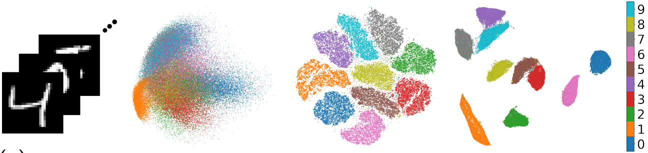
UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

30 sec summary

1. Closed form formula for UMAP's true loss function.
2. Drastically reduced repulsion strength.
3. Explains why UMAP tends to over-contract embeddings.
4. **Theoretically shows that the sophisticated UMAP weights have no benefit.**
5. This effect increases with the dataset size.

Dimension Reduction

Given $x_1, \ldots, x_n \in \mathbb{R}^D$ find layout $e_1, \ldots, e_n \in \mathbb{R}^d$ with $d \ll D$.



(a) MNIST
70000 × 786

(b) 2D PCA

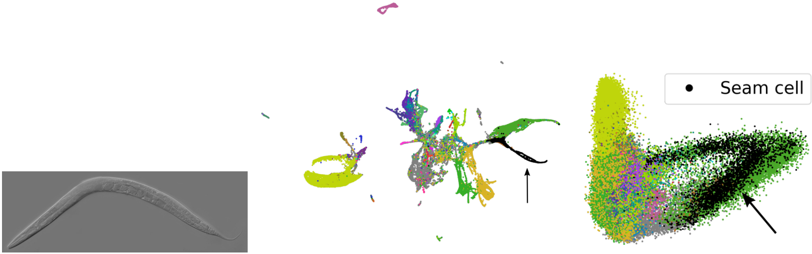
(c) 2D t-SNE

(d) 2D UMAP

Figure 1: Dimension reduction of the vectorized, unlabelled MNIST dataset

UMAP artifacts

UMAP tends to produce crisp structures even if there is variation.



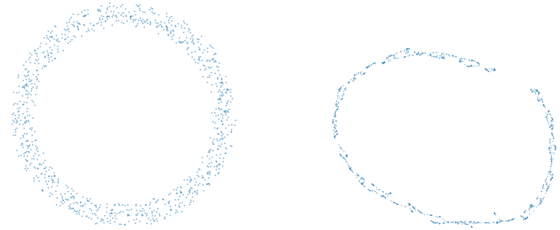
(a) C. elegans (b) UMAP (c) PCA

Figure 2: Gene expression data of 86024 cells of C.elegans [1-2].

Picture from https://en.wikipedia.org/wiki/Caenorhabditis_elegans#/media/File:Adult_Caenorhabditis_elegans.jpg

UMAP artifacts

Over-contraction even if no dimension reduction necessary.

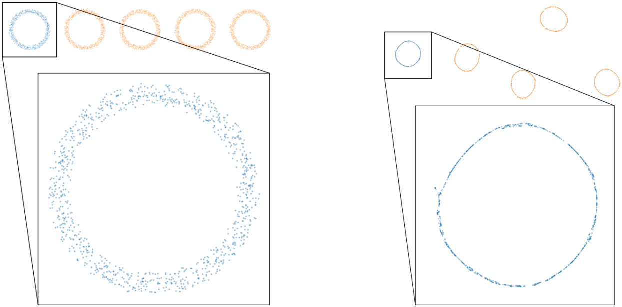


(a) Original (b) UMAP

Figure 3: 3a 1000 points from a 2D ring. 3b 2D UMAP embedding.

UMAP artifacts

The larger the dataset the stronger the over-contraction.



(a) Original (b) UMAP

Figure 4: 4a Five 2D rings with 1000 points. 4b 2D UMAP embedding.

Overview UMAP [3]

1. k nearest neighbour graph of input data.
2. Input similarities $\mu_{ij} \in [0, 1]$, non-zero only on kNN graph; embedding similarities $\nu_{ij} = \nu(||e_i - e_j||)$.
3. Loss function

$$\mathcal{L}(\{e_i\}) = -2 \sum_{1 \leq i < j \leq n} \mu_{ij} \log(\nu_{ij}) + (1 - \mu_{ij}) \log(1 - \nu_{ij}).$$

   $\Rightarrow$ Minimum at $\nu_{ij} = \mu_{ij}$.

4. Optimization via negative sampling [4].

Overview UMAP [3]

1. k nearest neighbour graph of input data.
2. **Input similarities $\boldsymbol{\mu_{ij}}$** $\in [0, 1]$, non-zero only on kNN graph; embedding similarities $\nu_{ij} = \nu(||e_i - e_j||)$.
3. Loss function

$$\mathcal{L}(\{e_i\}) = -2 \sum_{1 \leq i < j \leq n} \mu_{ij} \log(\nu_{ij}) + (1 - \mu_{ij}) \log(1 - \nu_{ij}).$$

   $\Rightarrow$ Minimum at $\nu_{ij} = \mu_{ij}$.
4. Optimization via negative sampling [4].

Overview UMAP [3]

1. k nearest neighbour graph of input data.
2. Input similarities $\mu_{ij} \in [0, 1]$, non-zero only on kNN graph; embedding similarities $\nu_{ij} = \nu(||e_i - e_j||)$.
3. Loss function

$$\mathcal{L}(\{e_i\}) = -2 \sum_{1 \leq i < j \leq n} \mu_{ij} \log(\nu_{ij}) + (1 - \mu_{ij}) \log(1 - \nu_{ij}).$$

   $\Rightarrow$ Minimum at $\nu_{ij} = \mu_{ij}$.
4. Optimization via negative sampling [4].
   $\Rightarrow$ **This changes the loss function!** [5]

Overview UMAP [3]

1. k nearest neighbour graph of input data.
2. ~~Input similarities $\mu_{ij} \in [0, 1]$, non-zero only on kNN graph; embedding similarities $\nu_{ij} = \nu(||e_i - e_j||)$.~~
3. ~~Loss function~~

$$\mathcal{L}(\{e_i\}) = -2 \sum_{1 \leq i < j \leq n} \mu_{ij} \log(\nu_{ij}) + (1 - \mu_{ij}) \log(1 - \nu_{ij}).$$

~~$\Rightarrow$ Minimum at $\nu_{ij} = \mu_{ij}$.~~

4. Optimization via negative sampling [4].

UMAP's true loss function

Theorem

The expected loss of UMAP's optimization procedure is

$$\tilde{\mathcal{L}} = -2 \sum_{1 \leq i < j \leq n} \mu_{ij} \cdot \log(\nu_{ij}) + \frac{(d_i + d_j)m}{2n} \cdot \log(1 - \nu_{ij})$$

with $d_i = \sum_{j=1}^{n} \mu_{ij}$ *and* m *the number of negative samples.*

UMAP's true loss function

Theorem

The expected loss of UMAP's optimization procedure is

$$\tilde{\mathcal{L}} = -2 \sum_{1 \leq i < j \leq n} \mu_{ij} \cdot \log(\nu_{ij}) + \boxed{\frac{(d_i + d_j)m}{2n}} \cdot \log(1 - \nu_{ij})$$

with $d_i = \sum_{j=1}^{n} \mu_{ij}$ *and* m *the number of negative samples.*

$\Rightarrow$ Dramatically reduced repulsion as $1 - \mu_{ij} = 1$ for most ij.

Difference between the loss functions

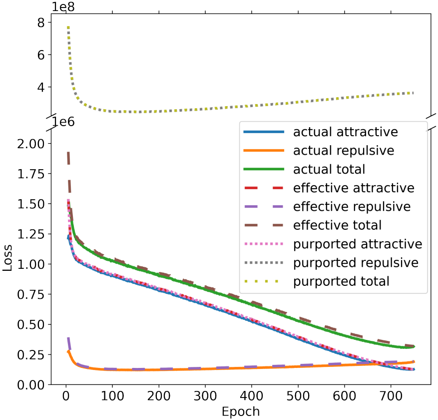


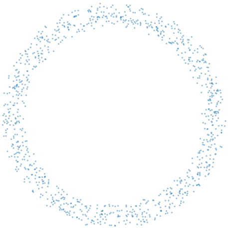
Figure 5: Loss functions for the UMAP optimization on the C.elegans dataset.

$\Rightarrow$ UMAP does not optimize its own loss function!

Target similarities

Optimal embedding similarity of true loss function is binarized μ_{ij}.

$$\nu_{ij}^* = \frac{\mu_{ij}}{\mu_{ij} + \frac{(d_i + d_j)m}{2n}} \begin{cases} = 0 \text{ if } \mu_{ij} = 0 \\ \approx 1 \text{ if } \mu_{ij} > 0. \end{cases}$$



(a) Original

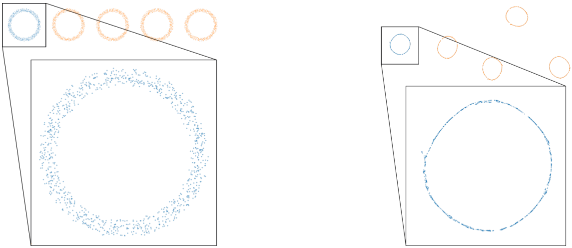
(b) UMAP

Figure 6: Since positive target similarities are close to one, UMAP embeddings tend to be over-contracted. For more details see Figure 3.

Dependence on dataset size

Binarization is stronger for larger dataset size n.

$$\nu_{ij}^* = \frac{\mu_{ij}}{\mu_{ij} + \frac{(d_i + d_j)m}{2n}} \begin{cases} = 0 \text{ if } \mu_{ij} = 0 \\ \approx 1 \text{ if } \mu_{ij} > 0. \end{cases}$$

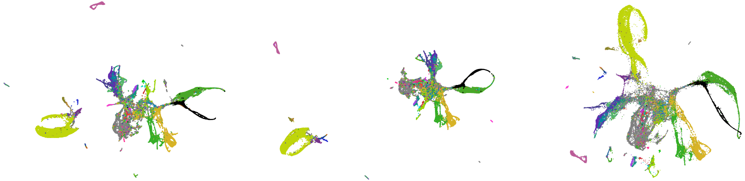


(a) Original (b) UMAP

Figure 7: The presence of additional rings decreases the repulsion further which leads to stronger over-contraction. For more details see caption of Figure 4.

Perturbed input similarities

Binarization renders exact value of input similarities unimportant.



(a) Original (b) *k*NN edge indicator (c) Inverted

Figure 8: UMAP visualizations of the C.elegans dataset are robust to severe perturbations of the input similarities.

Summary

- UMAP's sampling based optimization reduces repulsion.
- Input similarities are unimportant as they get binarized; only the kNN graph matters.
- This explains over-contraction artifacts.
- More faithful interpretation of UMAP plots in various domains.

References

[1] J. S. Packer, Q. Zhu, C. Huynh, P. Sivaramakrishnan, E. Preston, H. Dueck, D. Stefanik, K. Tan,C. Trapnell, J. Kim, et al. A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution. Science, 365(6459), 2019

[2] A. Narayan, B. Berger, and H. Cho. Assessing single-cell transcriptomic variability through density-preserving data visualization. Nature Biotechnology, pages 1–10, 2021

[3] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426., 2018

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119), 2013

[5] J. N. Böhm, P. Berens, and D. Kobak. A unifying perspective on neighbor embeddings along the attraction-repulsion spectrum. arXiv preprint arXiv:2007.08902., 2020

Thank you and
see you during the poster session!