Talk proposal: Expansion of the theory of metric spaces and fuzzy simplicial sets and their applications to data analysis

Lukas Silvester Barth[1] Fatemeh (Hannaneh) Fahimi[1] Parvaneh Joharinad[1]

Jürgen Jost[1] Janis Keck[1] Thomas Jan Mikhail[2]

1: Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
2: Universitat Autònoma de Barcelona, Barcelona, Spain

`Correspondence to: lukas.barth@mis.mpg.de`

Extended abstract

The contributions of our work are twofold.

(A.1) On the one hand, we show that fuzzy simplicial sets, as introduced by [Spi09], can serve as a natural combinatorial object to represent data, and expand upon the corresponding theory.

(A.2) On the other hand, we use those theoretical clarifications to put the UMAP algorithm developed by [MHM18], that heavily draws upon those category theoretical concepts, on a more solid footing, show how it is related to another prominent method called Isomap (cf. [TSL00]) and introduce a natural combination of both methods.

Since data is often given in form of a metric space, we address the first point by providing additional details about the relationship between the category **UM** of uber metric spaces and the category **sFuz** of fuzzy simplicial sets, which are defined below.

Definition 0.1. An uber-metric space (X,d) is a set X equipped with a map $d : X \times X \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ such that

1. $d(x,y) \geq 0$, and $d(x,y) = 0$ only if $x = y$;
2. $d(x,y) = d(y,x)$; and
3. $d(x,z) \leq d(x,y) + d(y,z)$.

The category of uber-metric spaces **UM** has as objects uber-metric spaces and as morphisms non-expansive maps.

A closely related concept, important in the context of [MHM18], is a so-called **extended pseudo-metric space**, the definition of which is equivalent to a uber metric space, except that the 3rd condition is amended to "3. Either $d(x,z) = \infty$ or $d(x,z) \leq d(x,y) + d(y,z)$", i.e. the triangle inequality does not have to hold for infinite distances in an extended pseudo-metric space. We denote the category of extended pseudo-metric spaces and non-expansive maps by **EPMet**.

We follow [Spi09] in defining fuzzy simplicial sets. Since the link to the publication was recently removed, we provide a rather detailed description below. Let $\mathbf{I} = [0,1]$ be a topological space in which the open sets are declared to be the intervals $[0,a)$ (for all $a \in [0,1]$) and $[0,1]$, ordered by inclusion. We

Submitted to:
ACT 2024

use the same notation for the category **I**, in which objects are open sets and morphisms are inclusions. To simplify our notation, we refer to objects $[0,a)$ simply by writing a. Consequently, we write $i_{ab} : a \to b$ for the inclusion maps $[0,a) \to [0,b)$ $(b \geq a)$.

Definition 0.2. A *fuzzy set* S is a sheaf on **I** for which all restriction maps $S(i_{ab} : a \to b) : S(b) \to S(a)$ are injections. Their category is denoted by **Fuz**.
A *fuzzy simplicial set* is a functor $\Delta^{\mathrm{op}} \to \mathbf{Fuz}$, where Δ denotes the *simplicial indexing category*. (Its objects are given by finite totally ordered sets $[n]$ with exactly $n+1$ elements (we follow the standard convention in this context to start counting at 0) and its morphisms are order preserving maps ($f : [n] \to [m]$ s.t. $f(a) \geq f(b)$ if $a \geq b$).)

One of our contributions is that we use the well-known nerve-realization procedure to view the adjunction that was provided in [Spi09] (which we define below) as one member of a family of adjunctions and investigate the consequences of different choices. To start this analysis, we prove the following theorem.

Proposition 0.1. Let $\mathbf{y} : (\Delta \times \mathbf{I})^{\mathrm{op}} \to \mathbf{sFuz}$ be the Yoneda embedding. Suppose that $\mathrm{Re}_\Delta : \mathbf{y}(\Delta \times \mathbf{I}) \to \mathbf{C}$ is any functor and that **C** is either **UM** or **EPMet**. Then the following defines a functor:

$$\begin{aligned} \mathrm{Re} : \mathbf{sFuz} &\to \mathbf{C}, \\ \mathrm{Re}(S) : &= \mathrm{colim}(D_S) \\ \text{where} \quad D_S &= \mathrm{Re}_\Delta \circ \mathbf{y} \circ P_S : \mathbf{El}(S) \to \mathbf{C}. \end{aligned} \tag{1}$$

where the category of elements $\mathbf{El}(S)$ consists of objects which are pairs (A,x) with $A \in \Delta \times \mathbf{I}$ and $x \in S(A)$ and $P_S : \mathbf{El}(S) \to \Delta \times \mathbf{I}$ is the projection functor from the category of elements to the fuzzy simplex category. Furthermore, the following functor is adjoint to Re:

$$\begin{aligned} \mathrm{Sing} : \mathbf{C} &\to \mathbf{sFuz}, \\ Y &\mapsto \mathrm{Sing}(Y) : (\Delta \times \mathbf{I})^{\mathrm{op}} \to \mathbf{Sets}, \\ &\qquad (n,a) \mapsto \mathrm{Hom}_{\mathbf{UM}}(\mathrm{Re}_\Delta(\Delta^n_{<a}), Y). \end{aligned} \tag{2}$$

To show that $\mathbf{Re}_\Delta$ is a functor only requires to define realizations and non-expansive maps on the subcategory of **sFuz** of representable functors $\Delta^n_{<a} := \mathbf{y}(n,a)$ and there are infinitely many possibilities. One choice is the realization provided in [Spi09], where Re_Δ is defined as follows:

$$\begin{aligned} \mathrm{Re}_\Delta^{\text{Spivak}} : \mathbf{y}(\Delta \times \mathbf{I}) \to \mathbf{UM}, \quad \mathrm{Re}_\Delta(\Delta^n_{<a}) &:= \left\{ x \in \mathbb{R}^{n+1} \,\middle|\, \sum_{i=1}^{n+1} x_i = -\log(a) \right\} \\ \mathrm{Re}_\Delta^{\text{Spivak}}(\mathbf{y}(\sigma, i_{ab})) : &= \frac{\log(b)}{\log(a)} \left(\sum_{i_0 \in \sigma^{-1}(0)} x_{i_1}, \cdots, \sum_{i_n \in \sigma^{-1}(n)} x_{i_n} \right). \end{aligned} \tag{3}$$

In UMAP, instead of employing **UM**, the category **EPMet** (or more precisely the subcategory of finite extended pseudo-metric spaces) is used, and they define the functor $\mathbf{Re}_\Delta$ as follows

$$\begin{aligned} \mathrm{Re}_\Delta^{\mathrm{UMAP}} : \mathbf{y}(\Delta \times \mathbf{I}) \to \mathbf{EPMet}, \quad & \Delta^n_{<a} \mapsto (\{x_0, \cdots, x_n\}, d_a), \\ \text{where} \quad & d_a(x_i, x_j) := \begin{cases} -\log(a), & \text{if } i \neq j, \\ 0, & \text{else.} \end{cases}, \\ \mathrm{Re}_\Delta^{\mathrm{UMAP}}(\sigma, i_{ab}) : \mathrm{Re}_\Delta(\Delta^n_{<a}) \to \mathrm{Re}(\Delta^m_{<b}), \quad & x_i \mapsto x_{\sigma(i)} \end{aligned} \tag{4}$$

However, this functor is also well-defined when the codomain category is **UM** instead. Furthermore, the form of the Sing-functor, defined by (2), is independent of that choice. Since only the Sing-functor is used in the UMAP method (and one could replace infinite sums that do not fulfill the triangle inequality by certain sums), it would be possible to conduct their method fully in the categories **UM** and **sFuz**. In this concrete sense, the adjunction provided in [MHM18] and the adjunction provided in [Spi09] are elements of the same family. In a similar way, one can show that the transformation provided in t-SNE (cf. [MH08]) can be expressed as a member of that family, by noting that (a slight generalization of) the inverse transformation of the student-t-distribution can replace the logarithm in (4). Nevertheless, we point out that the realization functor, as defined in (1), does depend on the choice **UM** vs **EPMet** because the colimits in those two categories are different. In fact, we proved the following theorem:

Proposition 0.2. Let $\mathbf{C}$ be either **UM** or **EPMet**. The colimit of a small diagram $D : \mathbf{I} \to \mathbf{C}$ is given by

$$\operatorname{colim}(D) = (\operatorname{colim}(FD),\ d_{\mathrm{colim}}) \tag{5}$$

where F is the forgetful functor $F : \mathbf{C} \to \mathbf{Sets}$, and $\operatorname{colim}(FD)$ is given by

$$\begin{aligned} \operatorname{colim}(FD) \simeq X/\sim, \quad & \text{where } X := \dot{\bigcup_{I \in \mathbf{I}}} FD(I) \\ \text{and } \sim \text{ is generated by } & (\ x \sim x' \quad \text{iff} \quad x' = FDu(x)\ ), \end{aligned} \tag{6}$$

(where u is a morphism in the indexing category $\mathbf{I}$ of the diagram D).
For the case $\mathbf{C} = \mathbf{UM}$, d_{colim} is defined by

$$d_\sim([x],[x']) = \inf(d_X(p_1,q_1) + \cdots + d_X(p_n,q_n)), \quad d_X(p_i,q_i) := \begin{cases} d_J(p_i,q_i), & \text{if } p_i, q_i \in FD(J) \\ \infty, & \text{else,} \end{cases} \tag{7}$$

where the infimum is taken over all pairs of sequences $(p_1, \cdots, p_n)$, $(q_1, \cdots, q_n)$ of elements of X such that $p_1 \sim x, \quad q_n \sim x'$, and $p_{i+1} \sim q_i$ for all $1 \leq i \leq n-1$, whereas, for the case $\mathbf{C} = \mathbf{EPMet}$, d_{colim} is defined by

$$d_\sim^{\mathrm{EPMet}}([x],[x']) := \begin{cases} \infty, & \text{if } \inf_{y \in [x], y' \in [x']} = \infty, \\ d_\sim \quad \text{(as in (7))}, & \text{else.} \end{cases} \tag{8}$$

The theorem about the colimits above allowed us make the description of the colimit appearing in (1) more precise, which in turn allowed us to show that, a particular member of the family, namely, $\mathrm{Re}^{\mathrm{UMAP}}$, either with codomain **UM** or **EPMet**, allows to embed the metric categories as subcategories into **sFuz**. More concretely, we proved that

$$\mathrm{Re}^{\mathrm{UMAP}} \circ \mathrm{Sing}^{\mathrm{UMAP}} \simeq \mathrm{id}, \tag{9}$$

no matter which codomain one considers for $\mathrm{Re}^{\mathrm{UMAP}}$. However, $\mathrm{Sing}^{\mathrm{UMAP}}$ is not essentially surjective and we prove another theorem that explicitly characterizes the image, providing the exact conditions under which a fuzzy simplicial set corresponds to a metric space.

We also show that the formalization of persistent homology in terms of fuzzy simplicial sets, that was introduced in [Spi09], can be carried out in a somewhat simpler way with the adjunction mentioned above, and provide an existence proof for an appropriate analogue to the related adjunction between the truncation functor and the (co-)skeleton functors known from the category of (ordinary) simplicial sets. However, we omit this somewhat lengthy theorem here.

The conclusion of those propositions is that there is in general more freedom to encode data in a fuzzy simplicial set than in an (uber or extended pseudo-) metric space, which can be useful for encoding relationships that do not satisfy all metric relations (for example in a social network graph), or to encode higher order interactions (between more than 2 points), that a metric can not account for, while providing a rich combinatorial and category theoretical structure that alternative structures, like for example hypergraphs, can not offer. One important example of such higher relationships are those that arise in the area of persistent homology (see, for example [Car09]). At the same time, our explicit descriptions of the metric realization functors allow to specify concrete algorithms that transform this combinatorial description into a metric one if desired.

The second point (A.2) is motivated by the importance of dimension reduction and data visualization methods in the area of data analysis. Prominent methods include PCA (cf. [Pea01]), LE (cf. [BN03]), MDS (cf. [Tor52], [BG05]), Isomap (cf. [TSL00]), t-SNE (cf. [MH08]) and UMAP (cf. [MHM18]).

Of those methods, UMAP is interesting from a category theoretical point of view because the authors pick up and modify the adjunction in [Spi09] for their purposes. Roughly speaking, UMAP can be described by the upper, solid pathway in the following diagram:

$$\begin{tikzcd}
\mathbf{Met} \arrow[r, "\mathrm{split}"] & \overline{\mathbf{UM}}^N \arrow[r, "\mathrm{Sing}^N"] \arrow[d, dashed, "\mathrm{merge}_{\mathbf{UM}}"'] & \overline{\mathbf{sFuz}}^N \arrow[r, "\mathrm{ctr}_1^N"] & \overline{\mathbf{c1Fuz}}^N \arrow[r, "\mathrm{merge}_{\mathbf{c1Fuz}}"] & \mathbf{c1Fuz} \arrow[d, "\mathrm{embedding}"'] \\
 & \mathbf{UM} \arrow[rrr, dashed, "\mathrm{embedding}"] & & & \mathbf{Euc}
\end{tikzcd} \tag{10}$$

where split (which is not a functor) converts the dataset X, that contains N points, into N local k-nearest neighborhood spaces with the same underlying set, $\overline{\mathbf{A}}$ (for any category $\mathbf{A}$) denote appropriately chosen subcategories (obtained via pullbacks) to which the merge functors can be applied, Sing converts the local neighbourhood spaces into fuzzy simplicial sets, ctr_1 truncates the fuzzy simplicial sets and converts them to their classical counterpart (**c1Fuz** can be thought of as directed weighted graphs), $\mathrm{merge}_{\mathbf{c1Fuz}}$ describes a procedure to merge the N different local graphs and finally a gradient-descent based scheme is used for embedding into a Euclidean space.

The merge functor is ultimately defined in terms of so-called **t-conorms** T, which are binary operations $T : [0,1] \times [0,1] \to [0,1]$, satisfying some natural conditions, that were originally developed to formalize the combination of two probabilities. Typical examples of t-conorms are the probabilistic sum, or the maximum of two probabilities. During the merge process, the t-conorm is then applied to the probabilities of each pair of simplices, that are supposed to be combined when two fuzzy simplicial sets are merged.

Note that the merge operation in the diagram above is only performed on the underlying graphs and that, even though the Sing functor is employed, the Re functor never maps the merged fuzzy simplicial set back to a metric space because the final form in **Euc** is obtained indirectly via gradient descent.[1] Therefore a natural question is whether one could make use of the Re functor to walk along the alternative path that is visualized with dashed lines in Diagram (10), where $\mathrm{merge}_{\mathbf{UM}} := \mathrm{Re} \circ \mathrm{merge}_{\mathbf{sFuz}} \circ \mathrm{Sing}^N$. In our work, we derive explicit formulas for the computations involved in this alternative path and implement an algorithm that makes use of them. This is facilitated by the explicit description of the metric realization functors that we derived and it requires to generalize the merge operation, that in [MHM18] was only defined in **c1Fuz** to a merge operation in **sFuz**. Concretely, we proved the following theorem:

Proposition 0.3. The functor

$$\mathrm{merge}_{\mathbf{UM}} := \mathrm{Re}^{\mathrm{UMAP}} \circ \mathrm{merge}_{\mathbf{sFuz}} \circ \mathrm{Sing}^N : \mathbf{UM} \times_{\mathbf{Sets}} \cdots \times_{\mathbf{Sets}} \mathbf{UM} \to \mathbf{UM}. \tag{11}$$

can be given the following explicit description:

$$\begin{aligned} &\mathrm{merge}_{\mathbf{UM}}((X,d_1),\cdots,(X,d_N)) = (X,d), \text{ where} \\ &d(x,y) := \inf_{x=x_1,\cdots,x_n=y} \sum_{i=1}^{n-1} (-\log(T_1(x_i,x_{i+1}))), \end{aligned} \tag{12}$$

where T_1 is defined recursively in terms of a t-conorm T:

$$\begin{aligned} T_{k\in\{1,\cdots,N-1\}}(x,y) &:= T(e^{-d_k(x,y)}, T_{k+1}(x,y)), \text{ and} \\ T_N(x,y) &:= e^{-d_N(x,y)}. \end{aligned} \tag{13}$$

Similarly, the functor

$$\mathrm{merge}_{\mathbf{EPMet}} := \mathrm{Re}^{\mathrm{UMAP}} \circ \mathrm{merge}_{\mathbf{sFuz}} \circ \mathrm{Sing}^N : \mathbf{EPMet} \times_{\mathbf{Sets}} \cdots \times_{\mathbf{Sets}} \mathbf{EPMet} \to \mathbf{EPMet}. \tag{14}$$

can be given the description:

$$\begin{aligned} &\mathrm{merge}_{\mathbf{EPMet}}((X,d_1),\cdots,(X,d_N)) = (X,d_{\mathrm{EPMet}}), \text{ where} \\ &d_{\mathrm{EPMet}}(x,y) := \begin{cases} \infty, & \text{if } T_1(x,y) = 0, \\ d(x,y) \quad \text{(as in (12))}, & \text{else.} \end{cases} \end{aligned} \tag{15}$$

It turns out that a special case of the general formulas obtained for the case **UM** in this way describe the process employed in the well-known method Isomap (cf. [TSL00]). This makes a previously unknown close relationship between Isomap and UMAP apparent. At the same time, our derivation shows how to naturally incorporate the local distances and t-conorms used in UMAP into the alternative pathway, which then yields a natural combination of both methods. Since our algorithm represents a combination of Isomap and UMAP, and takes place entirely in the category **UM**, we decided to call it **IsUMap**. Our preliminary simulations displayed in Figure 1 show that, even though IsUMap is less useful for clustering data in high dimensions, it is better at visualizing low-dimensional manifolds with small distortions, while

[1] Since the form of the Sing-functor is agnostic to the choice of the metric category, this is the reason why UMAP works both in **UM** and **EPMet**, as mentioned above.

retaining the capability of uniformizing the data distribution if desired. The reason why low-dimensional data visualization is more faithful, while high-dimensional data visualization delivers less intuitive results, is that UMAP during the embedding performs a series of numerical approximations, the most important of which is perhaps the so-called negative undersampling (cf. [DH21] for a detailed analysis), that heavily distorts the distances of the high-dim spaces. Since in high-dim spaces, much more points can be equally far away from each other than in low-dim spaces, this distortion seems to be necessary for obtaining the visually appealing clustering effects. However, this does not imply that the embedding is more faithful. In fact, since the negative undersampling distorts distances, the embedding is less optimal in terms of the original loss function. There is, to our knowledge, no extensive theory that establishes a rigorous criterium as to which degree of distortion is optimal. In any case, we believe that one might be able to combine our method with the ideas about functorial manifold learning and clustering described in [Shi20b] and [Shi20a], to improve the clustering capabilities of our method in future work.

References

[BG05] I. Borg & P.J. Groenen (2005): *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

[BN03] M. Belkin & P. Niyogi (2003): *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation*. *Neural Computation* 15, pp. 1373–1396.

[Car09] G. Carlsson (2009): *Topology and Data*. *Bulletin of the American Mathematical Society* 46, pp. 255–308, doi:10.1090/S0273-0979-09-01249-X.

[DH21] S. Damrich & F.A. Hamprecht (2021): *On UMAP's true loss function*. *Advances in Neural Information Processing Systems* 34, pp. 5798–5809.

[MH08] L. van der Maaten & G. Hinton (2008): *Visualizing Data using t-SNE*. *Journal of Machine Learning Research* 9(86), pp. 2579–2605. Available at `http://jmlr.org/papers/v9/vandermaaten08a.html`.

[MHM18] L. McInnes, J. Healy & J. Melville (2018): *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, doi:10.48550/ARXIV.1802.03426. Available at `https://arxiv.org/abs/1802.03426`.

[Pea01] K. Pearson (1901): *On lines and planes of closest fit to systems of points in space*. In: *Proceedings of the* 17*th ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems* (*SIGMOD*), p. 19.

[Shi20a] D. Shiebler (2020): *Functorial clustering via simplicial complexes*. In: *TDA* {\&} *Beyond*, p. N.A.

[Shi20b] D. Shiebler (2020): *Functorial manifold learning*. *arXiv preprint arXiv*:2011.07435.

[Spi09] D.I. Spivak (2009): *Metric realization of fuzzy simplicial sets*. *N.A.* https://math.mit.edu/ dspivak/-files/metric_realization.pdf.

[Tor52] W.S. Torgerson (1952): *Multidimensional scaling: I. Theory and method*. *Psychometrika* 17(4), pp. 401–419, doi:10.1007/BF02288916. Available at `https://doi.org/10.1007/BF02288916`.

[TSL00] J.B. Tenenbaum, V.d. Silva & J.C. Langford (2000): *A global geometric framework for nonlinear dimensionality reduction*. *science* 290(5500), pp. 2319–2323.

Appendix

A Figures

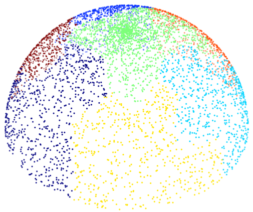
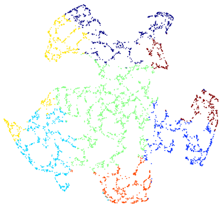
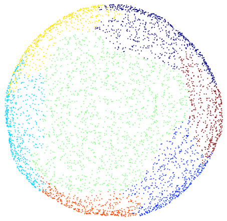
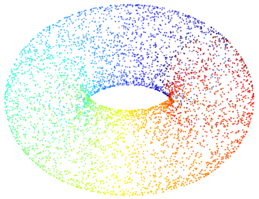
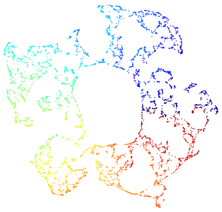
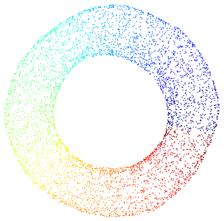
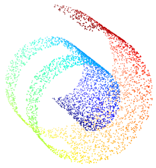
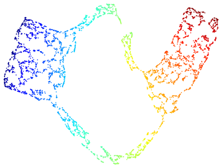
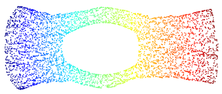
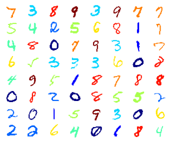
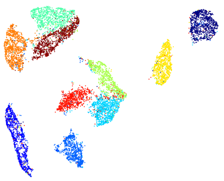
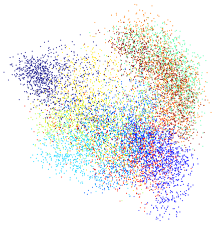
	Input data (a)	**UMAP** (b)	**IsUMap** (c)
(1) Hemisphere			
(2) Torus			
(3) Swiss role **with hole**			
(4) MNIST			

Table 1: Comparison of UMAP and IsUMap