The mathematics of UMAP

Adele Jackson

May 10, 2019

1 Introduction

UMAP (Uniform Manifold Approximation and Projection) is a new dimension reduction technique [2], currently implemented [5, 7] for both labelled and unlabelled data. Figure 1 shows a comparison of UMAP embeddings with some other standard dimension reduction algorithms. It gives similarly good outputs for visualisation as t-SNE, with a substantially better runtime, and may capture more of the global structure of the data. The current implementation also allows for embedding of new data into an existing model, so can be used as a preprocessing step in data analysis. Some example images, compared to those from t-SNE, are available at [6], and runtime benchmarking is at [4]. Excitingly, it is based on strong and general mathematical theory. This has allowed for its implementation for both unsupervised and semi-supervised dimension reduction and for datasets with custom metrics (such as categorical data).

The mathematics behind UMAP is quite abstract, and most presentations of it assume a substantial background in abstract topology and category theory. We will give a brief exposition of the theory without (hopefully) assuming too much prior knowledge. However, we would highly recommend reading further in category theory and introductory algebraic topology if you are interested in these types of techniques (see Further Reading). Before reading this document, we would recommend watching [3] for a more intuitive motivation of the algorithm. For a direct explanation of the algorithm that does not explain the mathematics behind it see [2, Section 3].

2 Approximating the underlying manifold

UMAP is an algorithm to find a representation of a given dataset D in $\mathbb{R}^N$ in a lower-dimensional space $\mathbb{R}^m$. We think of the datapoints as being drawn from some Riemannian manifold[1] M, then mapped into $\mathbb{R}^N$ by some embedding $\phi : M \hookrightarrow \mathbb{R}^N$.[2] See Figure 2 for an illustration of the setup.

One thought we might have is to reconstruct M, then find a good map from M into $\mathbb{R}^m$. To do this, we assume that D is uniformly drawn from M, as in the example in Figure 2. (Note that parts of M might be stretched out or compressed under the embedding into $\mathbb{R}^N$, so this does **not** imply that the data is uniformly distributed in $\mathbb{R}^N$.) This assumption means that D approximates M well. (This is also where the 'U' in UMAP comes from.) We will also assume that M is locally connected and that there are enough points in D that no point in D is isolated in its own connected component. These connectivity assumptions imply that every point in D is connected in M to its nearest neighbour in $D \hookrightarrow \mathbb{R}^N$. This assumption will be used in Section 5.

Finally, we will assume that the metric[3] on M is locally constant, as this property gives us

[1] A Riemannian manifold, for our purposes, is a space that locally looks like Euclidean space, in which we have well-defined notions of distances, angles, and volumes. For example, the surface of a unit sphere is a two-dimensional Riemannian manifold.

[2] The hooked arrow $\hookrightarrow$ indicates that the map ϕ is injective.

[3] A *metric* on a space X is a function $d : X \times X \to \mathbb{R}$ such that $d(x, y) \geq 0$, $d(x, y) = d(y, x)$, $d(x, y) = 0$ if and only if $x = y$, and $d(x, z) \leq d(x, y) + d(y, z)$. We view $d(x, y)$ as the distance from x to y.

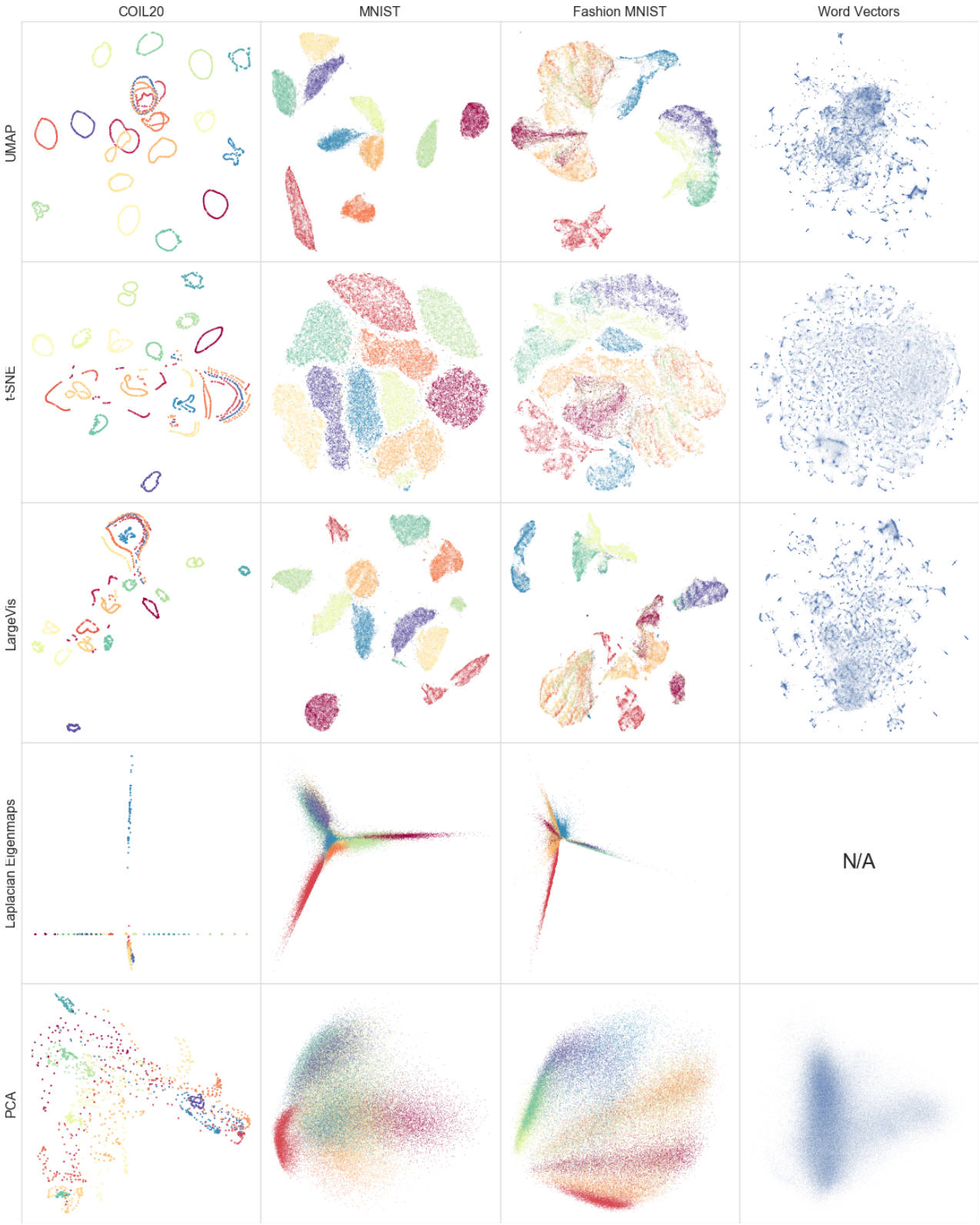
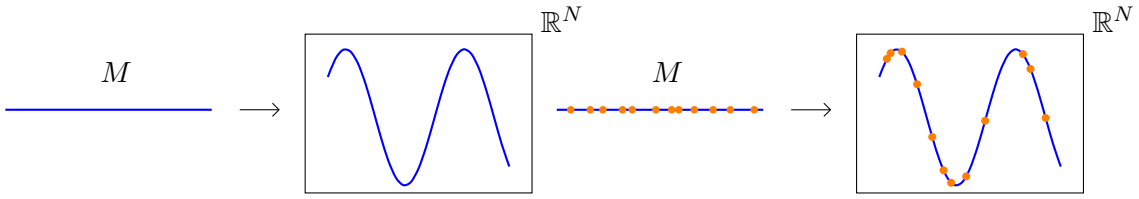
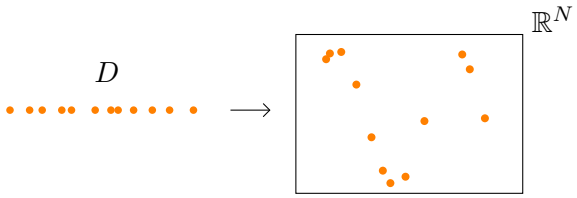


Figure 1: Embeddings from UMAP, t-SNE, LargeVis, Laplacian Eigenmaps and PCA on some standard datasets. Image taken from [2].



(a) The manifold M embedded in $\mathbb{R}^N$. (b) The dataset D is uniformly drawn from M.



(c) We see only D.

Figure 2: The dataset is drawn from a manifold embedded in $\mathbb{R}^N$.

the following lemma, which will let us approximate distances in M between points in D that are close enough in $\mathbb{R}^N$. Note that, as M is embedded in $\mathbb{R}^N$, we can measure distance and volume within M in two ways: that of the intrinsic metric on M, and that of the metric induced by $\mathbb{R}^N$.

Lemma ([2]). *Let (M, g) be a Riemannian manifold*[4] *embedded in $\mathbb{R}^n$. Let $p \in M$ be a point. Suppose that g is locally constant.*[5] *Let B be a ball in M, containing p, whose volume is $\frac{\pi^{n/2}}{\Gamma(n/2+1)}$ with respect to the metric on M. Then the distance of the shortest path in M from p to a point $q \in B$ is $\frac{1}{r} d_{\mathbb{R}^n}(p, q)$, where r is the radius of B in $\mathbb{R}^n$ and $d_{\mathbb{R}^n}(p, q)$ is the distance from p to q in $\mathbb{R}^n$.*

Its consequence for us is that we can approximate distances around M close to our datapoints by scaling the distance in $\mathbb{R}^N$. We do this as follows. As we have assumed our data is uniformly distributed on M, any ball of a fixed volume R on M should contain the same number of datapoints. Working backwards, let $N_k(x)$ be the ball in $\mathbb{R}^N$ around a datapoint x that contains its k nearest neighbours in $\mathbb{R}^N$ (with respect to the distance in $\mathbb{R}^N$). Then for any datapoint x_i, consider the neighbourhood of M that is sent to $N_k(x_i)$ in $\mathbb{R}^N$ (that is, consider $\phi^{-1}(N_k(x_i))$). This ball should have the same volume as if we followed the same procedure for any other datapoint x_j. By the lemma, for k small enough, we can approximate distances in M from x_i to one of its k nearest neighbours x_j as follows. Fix k as a hyperparameter, and write $\{x_{i_1}, \ldots, x_{i_k}\}$ for the k nearest neighbours of x_i. Then, from the lemma, we can derive that the distance in M from x_i to x_j is approximately $\frac{1}{r_i} d_{\mathbb{R}^N}(x_i, x_j)$, where r_i is the distance to the k^{th} nearest neighbour of x_i. To smooth this value, and reduce the impact of happening to have the k^{th} nearest neighbour be very far away while the $(k-1)^{th}$ nearest neighbours are clustered close to x_i, we take r_i to be the value such that

$$\sum_{j=1}^{k} \exp\left(\frac{-|x_i - x_{i_j}|}{r_i}\right) = \log_2(k).$$

[4] In this, M is the manifold and g is a two-form that gives us measures of distance, volumes and angles.

[5] To be precise, we assume that there is an open neighbourhood U with $p \in U$ on which g is locally constant, such that g is a constant diagonal matrix in the ambient coordinates.

This setup presents us with a problem. The distance we get from x_i to x_j using this method will in general be different to that from x_j to x_i, as $r_j \neq r_i$. We can interpret this as the fact that, while the x_i are uniformly drawn from the manifold, they are not all the same distance apart. We need a technique for combining a family of locally-defined finite metric spaces[6] to get a global structure, where we have some idea of uncertainty on the metric spaces. For this, we will use fuzzy simplicial sets.

Note that, although we have used the $\mathbb{R}^N$ metric to develop this theory, it still holds with any other metric. UMAP can be used with custom distance measures and so can handle categorical data and other measures of distance between datapoints.

At this point, we wish to set up an elegant correspondence between fuzzy simplicial sets (combinatorial presentations of a topological space, with probabilities on it) and finite extended-pseudo-metric spaces (metric spaces where we allow infinite distances). To do this, we need to define the classes of objects involved, and introduce a little category theory.

3 Categories, functors and adjunctions

Category theory is a branch of mathematics that unifies common concepts across different parts of mathematics. For example, one can take the product of two vector spaces, or the product of two sets, or two groups. Using ideas from category theory, we can give one unified definition of a "product" and prove results about it that are valid in all these contexts.

For the purposes of UMAP, category theory gives us an adjunction between fuzzy simplicial sets and finite extended-pseudo-metric spaces. An adjunction is a translation between different domains of discourse – for example, there is an adjunction between sets and vector spaces that we will discuss later in the section. We will now define some fundamental category theory concepts.[7]

A *category* $\mathscr{C}$ is a collection of *objects*, $\mathrm{Obj}(\mathscr{C})$, and between each $X, Y \in \mathrm{Obj}(\mathscr{C})$, a collection of *morphisms* $\mathrm{Hom}_{\mathscr{C}}(X, Y)$, satisfying the following conditions. We write $f : X \to Y$ to mean $f \in \mathrm{Hom}_{\mathscr{C}}(X, Y)$. First, we can *compose* morphisms: if $f : X \to Y$ and $g : Y \to Z$, there is a specified composite morphism $gf : X \to Z$. Second, we have *identity morphisms*: for each object X, there is a specified identity morphism $1_X \in \mathrm{Hom}_{\mathscr{C}}(X, X)$ (that is, $1_X : X \to X$). Third, the identity morphisms act as the identity under composition: for any $f : X \to Y$, $1_Y f = f 1_X = f$. Finally, composition is *associative*: let $f : X \to Y$, $g : Y \to Z$ and $h : Z \to W$ be morphisms. Then $h(gf) = (hg)f$.

Two familiar examples of categories are **Vect**, whose objects are real vector spaces and morphisms are linear maps, and **Set**, whose objects are sets and morphisms are functions between sets.

Given a category $\mathscr{C}$, we can define the opposite category $\mathscr{C}^{op}$, which we will use in the next section. This is the category whose objects are $\mathrm{Obj}(\mathscr{C})$, and whose morphisms are defined by $\mathrm{Hom}_{\mathscr{C}^{op}}(X, Y) = \mathrm{Hom}_{\mathscr{C}}(Y, X)$, with composition given by $g^{op} f^{op} = (fg)^{op}$. That is, for each morphisms $f : X \to Y$ in $\mathscr{C}$, there is an opposite morphism $f^{op} : Y \to X$ in $\mathscr{C}^{op}$. (Note that these opposite morphisms are formal maps, and there is in general not an answer to the question of where f^{op} sends some $x \in X$. For our purposes, we will usually be taking the opposite category where the morphisms are inclusion maps. In this case, we interpret the opposite morphisms as restrictions.) If we imagine each morphism $f : X \to Y$ in $\mathscr{C}$ as an arrow between its domain X and codomain Y, $\mathscr{C}^{op}$ is the category where we "turn all the arrows around".

We can now define maps between categories. A *functor* between two categories $F : \mathscr{C} \to \mathscr{D}$ is a map from the objects and morphisms of $\mathscr{C}$ to those of $\mathscr{D}$ that sends the identity 1_X to $1_{F(X)}$, and preserves composition (that is, $F(gf) = F(g)F(f)$). For example, we can define a

[6] A metric space is a set of points with an associated metric.

[7] See [9] for a more detailed exposition of this material aimed at scientists. We follow [8] for these definitions.

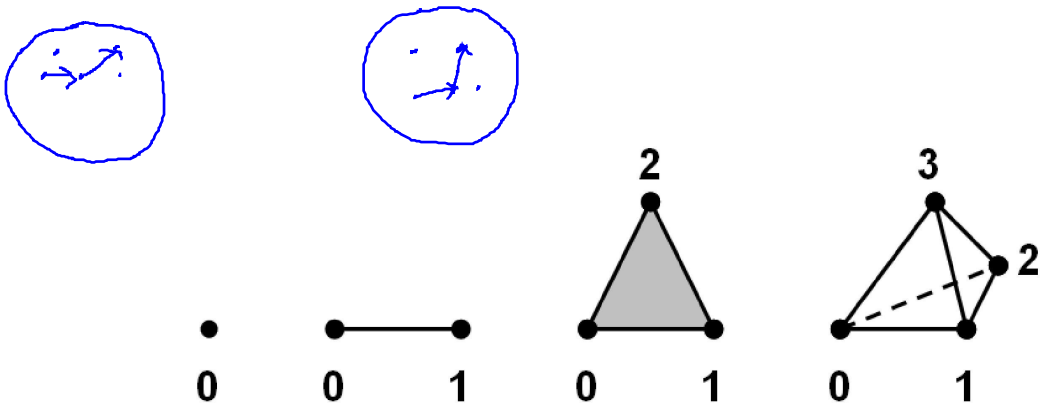


Figure 3: Examples of n-simplices for $0 \leq n \leq 3$. Image taken from [1].

functor **Vect** $\to$ **Set** taking a vector space to its underlying set of points, and taking a linear map between vector spaces to the induced map on the set of points. Note that, for any category $\mathscr{C}$, we have an identity functor $1_{\mathscr{C}}$ that is the identity on all objects and morphisms.

We can also define maps between functors themselves. Let $\mathscr{C}$ and $\mathscr{D}$ be categories, with functors $F, G : \mathscr{C} \to \mathscr{D}$. A *natural transformation* $\alpha : F \to G$ is a morphism $\alpha_X : FX \to GX$ for each $X \in \mathrm{Obj}(\mathscr{C})$ such that for any $f : X \to Y$ in $\mathscr{C}$, $Gf \circ \alpha_X = \alpha_Y \circ Ff : FX \to GY$.

We can compose natural transformations: if $F, G, H : \mathscr{C} \to \mathscr{D}$, with α a natural transformation from F to G and β one from G to H, then the map $\beta\alpha$ defined by $(\beta\alpha)_X = \beta_X \circ \alpha_X$ is a natural transformation from G to H. Thus, one can check that the set of all functors $\mathscr{C} \to \mathscr{D}$ forms a category whose morphisms are the natural transformations.

Finally, we define an adjunction. Suppose we wish to move between two categories $\mathscr{C}$ and $\mathscr{D}$, using two functors $F : \mathscr{C} \to \mathscr{D}$ and $G : \mathscr{D} \to \mathscr{C}$. We could ask for an equivalence of categories, which would be if $GF = 1_{\mathscr{C}}$ and $FG = 1_{\mathscr{D}}$. This is generally too strong a requirement. An adjunction is a weakening of the equivalence idea, and gives a translation between two categories. One example of an adjunction goes between **Vect** and **Set** (though there is no equivalence between them). We have a forgetful functor $F :$ **Vect** $\to$ **Set** that sends a vector space to its underlying set of points. We also have a functor $G :$ **Set** $\to$ **Vect** that sends a set S to the free vector space on S. This is clearly not an equivalence: $FG(S)$ is the set of all finite formal sums with real coefficients of elements of S. However, F and G form an adjunction.

An *adjunction* is a pair of functors $F : \mathscr{C} \to \mathscr{D}$ and $G : \mathscr{D} \to \mathscr{C}$ such that there are natural transformations $\eta : 1_{\mathscr{C}} \to GF$ and $\epsilon : FG \to 1_{\mathscr{D}}$ satisfying some naturality conditions.[8] We write $F \dashv G$ to represent this situation, and say that F and G form an adjunction with F the *left adjoint* and G the *right adjoint*. Note that the definition is not symmetric: $F \dashv G$ does not imply that $G \dashv F$.

The principal mathematical result behind UMAP is that there is an adjunction between finite fuzzy simplicial sets and finite extended-pseudo-metric spaces.

4 Fuzzy simplicial sets

A simplicial complex describes a topological space[9] in a combinatorial way.[10]

A (geometric) *n-simplex* is the convex hull spanned by a set of $n+1$ linearly independent vertices $\{x_0, \ldots, x_n\}$ in Euclidean space. That is, a geometric n-simplex is $\{\sum_{i=0}^{n} t_i x_i \,|\, t_i \geq 0, \sum_{i=0}^{n} t_i = 1\}$. A 0-simplex is a single point; a 1-simplex an interval; a 2-simplex a triangle. See Figure 3 for a depiction. Note that the convex hull spanned by a n-vertex subset of the $\{x_i\}$ is itself an $(n-1)$-dimensional simplex. We call this a *face* of the n-simplex.

[8] The naturality conditions are that for all $X \in \mathscr{C}$, $1_{FX} = \epsilon_{FX} F\eta_X$ and for all $Y \in \mathscr{D}$, $1_{GY} = G\epsilon_Y \eta_{GY}$.

[9] A topological space is a space with associated data consisting of all open sets in the space. Manifolds and metric spaces are topological spaces with some more structure on them.

[10] This section follows the exposition in [1], which is an excellent geometrically-motivated explanation of simplicial sets and simplicial homotopy theory. See that document for more detail, motivation and examples of these definitions.

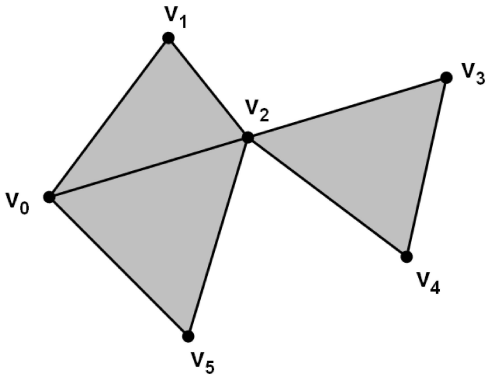


Figure 4: A simplicial complex X. Image taken from [1].

In topology, we often consider spaces up to *homeomorphism*. Let X and Y be topological spaces. A map $f : X \to Y$ is a *homeomorphism* if it is a continuous bijection with a continuous inverse. For example, there is a homeomorphism between a circle and an ellipse, but not between a circle and an interval. Note that homeomorphisms consider only the open sets of the space, and are not affected by any metric or other structure on the space. Geometric n-simplices have the property that for any two simplices T and U, T and U are homeomorphic.

We can describe manifolds as simplicial complexes by decomposing them into simplices. A *geometric simplicial complex* X is a collection of simplices in $\mathbb{R}^N$ such that (a) for any simplex in X, all of its faces are also in X, and (b) for any two simplices in X, their intersection is either empty or is a face of both of them. Note that, up to homeomorphism, we can describe X by listing the vertices of each simplex. Common vertices then allow us to recover which simplices share faces.

Generalising this idea, as simplices are defined by their vertices, an (abstract) *simplicial complex* X is a series of sets X^i $(i \geq 0)$ such that the elements of X^n (the n-simplices) are $(n+1)$-element sets that satisfy the following condition: for any $\{x_i\}_{i=0}^n \in X^n$, any n-element subset of this set is in X^{n-1}. Note that different simplices can have common elements, which indicates they have vertices, edges, or other sub-faces in common. For example, we can describe a square S, formed by gluing two triangles together, by

$$S^0 = \{\{a\},\{b\},\{c\},\{d\}\} \qquad S^1 = \{\{a,b\},\{a,c\},\{a,d\},\{b,c\},\{c,d\}\}$$
$$S^2 = \{\{a,b,c\},\{a,c,d\}\}.$$

We can describe the simplicial complex X depicted in Figure 4 by

$$X^0 = \{\{v_0\},\{v_1\},\{v_2\},\{v_3\},\{v_4\},\{v_5\}\}$$
$$X^1 = \{\{v_0,v_1\},\{v_0,v_2\},\{v_0,v_5\},\{v_1,v_2\},\{v_2,v_3\},\{v_2,v_4\},\{v_2,v_5\},\{v_3,v_4\}\}$$
$$X^2 = \{\{v_0,v_1,v_2\},\{v_0,v_2,v_5\},\{v_2,v_3,v_4\}\}.$$

To abstract this idea further, let all abstract simplices come with an ordering on its vertices. Write an (ordered) n-simplex as $[x_0,\ldots,x_n]$. Note that we can characterise an n-simplex in a simplicial complex by its $n+1$ *face maps*, where the i^{th} face map sends the simplex to the face $[x_0,\ldots,x_{i-1},x_{i+1},\ldots,x_n]$; that is, it sends the simplex to the face formed by removing the i^{th} vertex from the original simplex. Let d_i be the i^{th} face map.[11] One can check that we have the relation that, for $i \leq j$, $d_i d_j = d_{j-1} d_i$. Now we can give the following definition: a *Delta complex* X is a collection of sets X^i with, for each $n \geq 0$ and $0 \leq i \leq n$, a map $d_i : X^n \to X^{n-1}$ satisfying the relation $d_i d_j = d_{j-1} d_i$ for all $i \leq j$.

[11] Note that d_i is not a map of topological spaces. It is a formal map that to a simplex assigns its i^{th} face. Properly, we should write d_i^n for the i^{th} face map on n-simplices. We omit the n as it is almost always clear from the context.

We interpret the elements of X^n as the n-simplices of the Delta complex. The difference between Delta complexes and simplicial complexes is that in simplicial complexes, simplices are identified by their vertex set, while in Delta complexes, two distinct simplices may share the same vertex set. For example, the cone depicted in Figure **??** is a Delta complex but not a simplicial complex.

We can define Delta complexes in category theoretic language. Let $\hat{\Delta}$ be the category whose elements are the finite ordered sets $[n] = [0, 1, \ldots, n]$ and morphisms are strictly order-preserving maps $[m] \to [n]$. Then a Delta complex is a functor $X : \hat{\Delta}^{op} \to \mathbf{Set}$.

We can translate between our two definitions of Delta complex as follows. The set $X([n])$ is the n-simplices of the Delta complex. A map in $\hat{\Delta}^{op}$ from $[n]$ to $[n-1]$, which is the opposite of an order-preserving injection, is then a face map. We can write maps from $[m] \to [n]$ in $\hat{\Delta}$ as compositions of maps $[k] \to [k+1]$, so in general the images of the morphisms under X are compositions of face maps.

Now we will define a simplicial set. Let Δ be the category whose elements are the finite ordered sets $[n]$, and whose morphisms are order-preserving maps $[m] \to [n]$. (Compare to $\hat{\Delta}$ where the morphisms are strictly order-preserving maps.) A *simplicial set* is a functor $X : \Delta^{op} \to \mathbf{Set}$.

The difference between a Delta complexes and a simplicial set is that in a simplicial set, we allow degenerate simplices. For example, in Δ, there is a map $[0, 1, 2] \to [0, 1]$ that takes $0 \mapsto 0$, $1 \mapsto 1$ and $2 \mapsto 1$. The image of this map under a simplicial set X is a map $s : X^1 \to X^2$ that takes 1-simplex (an edge) $e \in X^1$ to some "2-simplex" in X^2. We can interpret this 2-simplex as a *degenerate* 2-simplex (a triangle) that has been collapsed into an edge. A simplicial set carries the information of all its degenerate simplices.

A *fuzzy set* is a generalisation of a set where, rather than elements being either in the set or not, there is a continuous membership function which one can think of as a probability. It is a set of objects A and a function $\mu : A \to [0, 1]$, where if $\mu(a) = 1$, a is definitely in the fuzzy set. The category **Fuzz** of fuzzy sets has fuzzy sets as objects, and maps of sets $f : A \to B$ such that $f \circ \mu(a) \geq \mu(a)$ as morphisms.[12] (That is, the maps are functions that take elements of A to elements of B of the same or higher membership strength.) Then a *fuzzy simplicial set* is a functor $X : \Delta^{op} \to \mathbf{Fuzz}$.[13]

5 Converting between metric spaces and fuzzy simplicial sets

Let x_i be a fixed datapoint in the dataset D. From Section 2, using the manifold metric based at x_i, we can approximate the distance from x_i to any other point x_j in D by $d_{x_i}(x_i, x_j) = \frac{1}{r_i} d_{\mathbb{R}^N}(x_i, x_j)$. This gives us a partly defined metric space, where we know the distance between x_i and x_j for all j, but not between x_j and x_k for $j, k \neq i$. Let ρ_i be the distance from x_i to its nearest neighbour in D in $\mathbb{R}^N$. As we have assumed M is locally connected, to force x_i to be connected to its nearest neighbour, for $j \neq i$, set

$$d_{x_i}(x_i, x_j) = \frac{1}{r_i}(d_{\mathbb{R}^N}(x_i, x_j) - \rho_i).$$

Now, this definition means that x_i and its nearest neighbour are distance 0 apart. As, for $j, k \neq i$, we do not know the distances between x_j and x_k relative to x_i, we set this to infinity.

[12] We can also define fuzzy sets categorically. Let I be the interval $[0, 1]$ with the topology whose open sets are generated by intervals $[0, a)$. Let $\mathscr{I}$ be the category of open subsets of I where the morphisms are inclusion maps. Then a fuzzy set S is a functor $S : \mathscr{I}^{op} \to \mathbf{Set}$ satisfying the following conditions. We interpret $S([0, a))$ to be the set of elements of S of membership strength at least a. Let $\rho_{b,a}$ be the inclusion map $[0, a) \to [0, b)$ for $b \geq a$. Then S is a fuzzy set if (a) it is a sheaf and (b) the restriction maps $S(\rho_{b,a})$ are injections. See [11] for the definition of a sheaf in the category theoretic language we want here, where you replace the category of R-modules with with **Set**.

[13] One nice property of these fuzzy simplicial sets is that the membership strength of the face of a simplex is at least the membership strength of the simplex.

In summary,

$$d_{x_i}(x_j, x_k) = \begin{cases} \frac{1}{r_i}(d_{\mathbb{R}^N}(x_j, x_k) - \rho_i) & \text{if } j = i \text{ or } k = i \\ \infty & \text{otherwise.} \end{cases}$$

We now no longer have a metric space: instead we have the following generalisation.

An *extended-pseudo-metric space* is a set X and a function $d : X \times X \to \mathbb{R} \cup \{\infty\}$ such that (a) $d(x, y) \geq 0$, (b) $d(x, x) = 0$, (c) $d(x, y) = d(y, x)$ and (d) either $d(x, z) = \infty$ or $d(x, z) \leq d(x, y) + d(y, z)$. Note that this allows for infinite distances, and also for $d(x, y) = 0$ when $x \neq y$.

Let **EPMet** be the category of extended-pseudo-metric spaces where the morphisms are non-expansive maps. Let **FinEPMet** be the sub-category of **EPMet** whose objects are finite extended-pseudo-metric spaces. Note that each of our approximations of distances within D is an element of **FinEPMet**.

Let **sFuzz** be the category of fuzzy simplicial sets (which are functors $\Delta^{op} \to$ **Fuzz**) whose morphisms are natural transformations between the functors. Let **Fin-sFuzz** be the sub-category of **sFuzz** consisting of the fuzzy simplicial sets with a finite number of non-degenerate simplices, defined as follows. Let X be a fuzzy simplicial set. An element of $X([n])$ is *degenerate* if its geometric realisation is an $(n-1)$-simplex. The number of non-degenerate n-simplices of X is the number of non-degenerate elements of $X([n])$ with positive membership strength. Thus X is in **Fin-sFuzz** if it has a finite number of non-degenerate n-simplices.

The main theorem of [2] gives a translation between these two categories.

Theorem. *There is an adjunction between* ***FinEPMet*** *and* ***Fin-sFuzz*** *given by* $FinSing$: ***FinEPMet*** $\to$ ***Fin-sFuzz*** *and* $FinReal$: ***Fin-sFuzz*** $\to$ ***FinEPMet****, with* $FinReal \dashv FinSing$.

The functors in this theorem are the following.[14] The functor $FinReal$: **Fin-sFuzz** $\to$ **FinEPMet** takes a finite simplicial set X to the finite metric space T whose elements are the vertices of X. The metric on T is defined as follows. Let μ be the membership strength function for the simplices of X.[15] For $x, y \in T$, let A be the set of all subsets of T containing both x and y. Then $d_T(x, y) = \min_{U \in A} - \log(\mu(U))$.

The functor $FinSing$: **FinEPMet** $\to$ **Fin-sFuzz** takes an finite extended-pseudo-metric space Y to a finite fuzzy simplicial set, which is a functor from $\Delta^{op} \to$ **Fuzz**. It acts as follows: $FinSing(Y)([n])$ is the fuzzy set of $(n+1)$-vertex subsets of the datapoints $\{x_{k_0}, \ldots, x_{k_n}\}$ where the subset has membership strength

$$\mu(\{x_{k_0}, \ldots, x_{k_n}\} = \min_{i,j} e^{-d(x_{k_i}, x_{k_j})}.$$

Given this translation, we can convert a set of datapoints D to a family of elements of **FinEPMet**, and from that to a family of finite fuzzy simplicial sets $FinSing(D, d_{x_i})$ for $x_i \in D$. Now, set the *fuzzy topological representation* of D to be

$$\bigcup_{i=1}^{n} FinSing(D, d_{x_i})$$

where the union is some choice of a union of fuzzy sets. We know that each of the $FinSing(D, d_{x_i})$ has the same set of objects, which are all simplices whose vertices are in D. Now, if (A, μ) and (A, ν) are two fuzzy sets with the same underlying set of objects, one reasonable definition of the union $(A, \mu) \cup (A, \nu)$ is $(A, (\mu \cup \nu))$ where $(\mu \cup \nu)(a) = \mu(a) \perp \nu(a)$ for $\perp$ some t-conorm. The current implementation of UMAP uses $x \perp y = x + y - xy$, which is the obvious t-conorm to use if you interpret $\mu(a)$ and $\nu(a)$ as probabilities of the simplex a existing, assume these are independent between the different local metric spaces, and do not care about higher-dimensional simplices (as, for reasons of computational complexity, we will not).

[14] See [2] for a categorical definition of these functors that makes it easier to prove they are adjoint.

[15] We do not define this precisely. See the remark after Definition 1.1 of [10] for a rigorous definition; Spivak's characteristic form is our membership strength function.

6 Finding a good low-dimensional representation

We now have a method for constructing a fuzzy simplicial set from a given set of points in $\mathbb{R}^n$. Let the dataset D be in $\mathbb{R}^N$. Let E be a low-dimensional representation of our dataset D in $\mathbb{R}^m$, for $m < N$. To evaluate how good E is as a representation of D, we compare the fuzzy simplicial set X constructed from D to one constructed from E. In constructing a fuzzy simplicial set Y from E, note that we already know the metric of the underlying manifold as it is $\mathbb{R}^m$ itself. Thus, $Y = FinSing((E, d))$ where d is the Euclidean metric on $\mathbb{R}^m$.

Consider the sets of edges in X and Y as fuzzy sets. Note that they have the same underlying set of elements, which is all edges whose vertices are labelled by elements of D, and differ only in the membership strength of the simplices. We define the cross-entropy C of two fuzzy sets with the same underlying elements set, (A, μ) and (A, ν), as follows [2, Definition 10]:

$$C((A,\mu),(A,\nu)) = \sum_{a\in A}\left(\mu(a)\log\left(\frac{\mu(a)}{\nu(a)}\right) + (1-\mu(a))\log\left(\frac{1-\mu(a)}{1-\nu(a)}\right)\right).$$

Note that, in our case, μ is fixed. We can view this formula as follows: $\mu(a)\log\left(\frac{\mu(a)}{\nu(a)}\right)$ provides the attractive force, as it is minimised if short edges in D correspond to short edges in E, since the length of the edge is small if $\nu(a)$ is large. Then $(1-\mu(a))\log\left(\frac{1-\mu(a)}{1-\nu(a)}\right)$ provides the repulsive force, as it is minimised if long edges in D correspond to long edges in E. We can then optimise the embedding using stochastic gradient descent. Note that X and Y contain many simplices of high dimension. For reasons of computational cost, the current implementation of UMAP only looks at the cross-entropy of the one-dimensional simplices in X and Y.

7 Further reading

The paper presenting UMAP, [2], gives a good explanation of the algorithm and implementation separate from its mathematical foundations.

For a geometrically-motivated definition of simplicial sets, and the realization and singular functors in the classical (non-fuzzy) context, see [1].

For an introduction to category theory aimed at non-mathematicians, see [9]. To better understand the material discussed here, you need familiarity with adjunctions (Definition 3.70).

Spivak's proof of the adjunction between the realization and singular functors in the fuzzy set context in [10] is much more explicit than that in the UMAP paper. Both assume familiarity with adjunctions.

References

[1] G. Friedman (2008). *An elementary illustrated introduction to simplicial sets*, preprint, arXiv: 0809.4221.

[2] L. McInnes, J. Healy and J. Melville (2018). *UMAP: Uniform manifold approximation and projection for dimension reduction*, preprint, arXiv:1802.03426.

[3] L. McInnes (2018). *Topological Approaches for Unsupervised Learning*, talk at Machine Learning Prague, accessed at `https://slideslive.com/38913519/topological-approaches-for-unsupervised-learning`.

[4] L. McInnes (2018). *Performance Comparison of Dimension Reduction Implementations*, accessed at `https://umap-learn.readthedocs.io/en/latest/benchmarking.html`.

[5] L. McInnes (2019). *UMAP: Uniform manifold approximation and projection*, accessed at `https://github.com/lmcinnes/umap`.

[6] J. Melville (2018). *UMAP Examples*, accessed at `https://jlmelville.github.io/uwot/umap-examples.html`.

[7] J. Melville (2019). *UWOT: An R package implementing the UMAP dimensionality reduction method*, accessed at `https://github.com/jlmelville/uwot`.

[8] E. Riehl (2014). *Category Theory in Context*, accessed at `http://www.math.jhu.edu/~eriehl/context.pdf`.

[9] B. Fong and D. Spivak (2018). *Seven Sketches in Compositionality: An Invitation to Applied Category Theory*, accessed at `http://math.mit.edu/~dspivak/teaching/sp18/7Sketches.pdf`.

[10] D. Spivak. *Metric realization of fuzzy simplicial sets*, accessed at `http://math.mit.edu/~dspivak/files/metric_realization.pdf`.

[11] D. Weng. *A categorical introduction to sheaves*, accessed at `http://www.math.uchicago.edu/~may/VIGRE/VIGRE2011/REUPapers/WengD.pdf`.