UNIVERSITY OF AMSTERDAM

MSc Artificial Intelligence
Master Thesis

Simplicial AutoEncoders

A connection between Algebraic Topology and Probabilistic Modelling

---

by

Jose Daniel Gallego Posada

11390689

August, 2018

36EC
February - August, 2018

Supervisor:
Dr Patrick Forré

Assessor:
Dr Max Welling

Informatics Institute

Abstract

Within representation learning and dimensionality reduction, there are two main theoretical frameworks: probability and geometry. Unfortunately, there is a lack of a formal definition of a statistical model in most geometry-based dimension reduction works, which perpetuates the division.

We introduce a statistical model parameterized by geometric simplicial complexes, which allows us to interpret the construction of an embedding proposed by UMAP as an approximate maximum a posteriori estimator. This is a step towards a theory of unsupervised learning which unifies geometric and probabilistic methods.

Finally, based on the the notion of structure preservation between simplicial complexes we define Simplicial AutoEncoders. Along with the construction of a probabilistic model for the codes in the latent space, Simplicial AutoEncoders provide a parametric extension of UMAP to a generative model.

Acknowledgement

I owe a debt of gratitude to the many people that helped through comments and discussions during the development of this project.

First, I would like to thank Marco Federici, Dana Kianfar for the frequent and challenging discussions which made the last two years so much more enjoyable; and Taco Cohen, for steering the course of my research into what turned out to be an incredibly enriching experience.

I would also like to thank Max Welling for agreeing to assess my work. Special thanks to Patrick Forré for his generous supervision during the last semester, and specially for the abundant advice and motivation during our long meetings.

Finalmente, a mi familia, por su apoyo incondicional.

Contents

1 Introduction

> *God created the integers, all the rest is the work of man*".
>
> — **Leopold Kronecker**

The performance of machine learning algorithms is strongly influenced by the representation of the data on which they are applied. A simple yet revealing example of this is shown in Figure 1.1. The change of coordinate dramatically affects the linear separability of this dataset. One could argue that the dataset *is* linearly separable, just not in the original *representation* as two concentric circles.

(a) Dataset formed by two circles. **(b)** Same dataset in polar coordinates.

Fig. 1.1: A simple change of representation can drastically affect the performance of a machine learning algorithm.

The immediate question is then, given a dataset, what is a *good* representation for it? As it is clear from the example, the goodness of a representation is directly linked to the type of algorithm we are applying on it: if our model was a radial basis function, the circular representation could be more useful. Also, depending on the learning task at hand, the preference of one representation over another might change.

In the last decade, *feature engineering* and the associated data preprocessing pipelines, accounted for a large part of the effort in the deployment of machine learning algorithms. During the early 2010s, a paradigm shift occurred. The focus went from manufacturing features, to *learning* them. In the words of Bengio et al. (2013), we want to learn "representations of the data that make it easier to extract useful information when building classifiers or other predictors".

Real world datasets arise from interactions between many sources. The interaction between these components create *entanglements*, which in turn account for the complexity present in datasets like audio, images or text. From a causality point of view, a good representation would be one which successfully disentangles such factors of variation.

At the same time, we would like our representation to distinguish and be equivariant with respect to relevant features, as well as to remain unchanged under transformations in less fundamental aspects. From the conjunction of disentanglement and invariance *dimension reduction* appears: "disentangle as many factors as possible, discarding as little information about the data as is practical" (Bengio et al., 2013).

The idea of dimension reduction also has some physical justification. Real datasets arise as the measurement (be it pictorial, sonorous, numeric, etc.) of variables in a physical system. Lin et al. (2017) argue that the locality, symmetry and low-order Hamiltonians characteristics of physical systems imply that the degrees of freedom of such systems are usually fairly low.

The machine learning embodiment of this idea is the so-called *manifold hypothesis*, according to which "real world data presented in high-dimensional spaces is likely to concentrate in the vicinity of non-linear sub-manifolds of much lower dimensionality" (Rifai et al., 2011b). This means that we can gain insights about the probability distribution in the high-dimensional space by studying the properties of those sub-manifolds around which it concentrates.

As we will see later, simplicial complexes are geometric constructions which can be represented combinatorially and which can be used to obtain reliable approximations of smooth manifolds. The combinatorial structure of simplicial complexes makes them much more amenable for computations than general smooth manifolds. For these reason they will be our central object of study.

Within representation learning and dimensionality reduction, there are two main theoretical frameworks: probability and geometry. Unfortunately, there is a lack of a formal definition of a statistical model in most geometry-based dimension reduction techniques, which perpetuates the division. Our construction of a statistical model parameterized by simplicial complexes is an attempt to close this gap.

Among the most notable examples of probabilistic approaches to representation learning we can count Probabilistic PCA (Tipping and Bishop, 1999), Variational Autoencoders (VAEs) (Kingma and Welling, 2013) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). An advantage of probabilistic approaches

over probabilistic ones is that they naturally induce a generative model, from which new data can be sampled.

Geometric methods can roughly be categorized as parametric or non-parametric. Non-parametric methods usually involve the construction of a (nearest-neighbor) graph and a random walk in the graph governed by some Markov chain. The most prominent example is the state-of-the-art, tSNE (Maaten and Hinton, 2008).

On the other hand, methods such as Contractive (Rifai et al., 2011a) or Denoising autoencoders (Vincent et al., 2010) try to learn a representation by imposing conditions such as robustness or smoothness on the a parametric embedding.

More recently, UMAP (McInnes and Healy, 2018) proposes a theoretical framework for manifold learning based in Riemannian geometry and algebraic topology, which is competitive with t-SNE. In short, it builds a non-parametric embedding of a dataset by minimizing the difference between the fuzzy topological representation of the data and the embedding. This work is the cornerstone in our theoretical and practical developments.

The main contributions of this thesis are:

- An equivalence theorem between fuzzy sets and a class of non-increasing set-valued random variables.
- A statistical model parameterized by geometric simplicial complexes.
- An interpretation of the UMAP algorithm as an approximate a posteriori estimator over random simplicial complexes.
- The introduction of Simplicial AutoEncoders as a parametric extension of UMAP and a principled generalization of *mixup* (Zhang et al., 2017).
- The extension of the representation induced by UMAP to a generative model.

This rest of this thesis is structured as follows. In Section 2 we provide a brief overview of the mathematical theories involved in work: category theory, (algebraic) topology, measure theory and fuzzy sets. In Section 3 we describe the theoretical foundations and inner workings of UMAP; prove an equivalence theorem between fuzzy sets and a special types of random variables; and use this result to interpret UMAP as the solution of a maximum a posteriori problem. In Section 4 we introduce the notions of simplicial regularization and simplicial autoencoders. Finally, Sections 5 and 6 contain our results and conclusions.

2 Mathematical Preliminaries

> *There is no royal road to geometry*".
>
> **— Euclid**
> (when asked if there was a shorter road to
> learning geometry than through the *Elements*)

In this chapter we provide a brief introduction to the several branches of mathematics on which this thesis is based. The starting point is a brief overview of category theory. We then define topological spaces, list some of their properties and illustrate how manifolds and simplicial complexes arise as particular examples. Additionally, we present (persistent) homology as a group-valued invariant on a topological space. Using category-theoretic tools, we extend the notion of simplicial complexes and introduce simplicial sets as their straightforward generalization. Finally, we describe random variables and fuzzy sets, as well as clarify the notation and terminology regarding deep generative networks.

2.1 Category Theory

Perhaps the most common theme in mathematics is that of studying classes of objects by considering transformations which "preserve structure" between said objects. Famous examples of this idea include sets and functions; vector spaces and linear transformations; posets and order-preserving maps, groups and homomorphisms; metric spaces and non-expansive maps; and topological spaces and continuous transformations.

Given the broad range of topics which can be formalized under the language of category theory, we provide a short introduction to its main concepts: categories, functors and natural transformations. This initial effort will be compensated with a general framework in which we can extend simplicial complexes to simplicial sets, define fuzzy sets and describe the theory underlying UMAP.

Definition 1: Category

A category $\mathbf{C}$ consists of:

- a *class*[1] of *objects* $\mathrm{Ob}(\mathbf{C})$,
- for every pair of objects c, d a set of *morphisms* $\mathrm{Hom}_{\mathbf{C}}(c, d)$,
- a binary operation $\circ$, called *composition of morphisms*, such that for every $f \in \mathrm{Hom}_{\mathbf{C}}(c, d)$ and $g \in \mathrm{Hom}_{\mathbf{C}}(d, e)$, there is an element $f \circ g \in \mathrm{Hom}_{\mathbf{C}}(c, e)$.

satisfying the following axioms:

- for all $f \in \mathrm{Hom}_{\mathbf{C}}(a, b)$, $g \in \mathrm{Hom}_{\mathbf{C}}(b, c)$ and $h \in \mathrm{Hom}_{\mathbf{C}}(c, d)$, we have that $h \circ (g \circ f) = (h \circ g) \circ f$, and
- for every object c, there exists a morphism $\mathrm{id}_c \in \mathrm{Hom}_{\mathbf{C}}(c, c)$ such that for every morphism $f \in \mathrm{Hom}_{\mathbf{C}}(c, d)$ and every morphism $g \in \mathrm{Hom}_{\mathbf{C}}(e, c)$ we have $f \circ \mathrm{id}_c = f$ and $\mathrm{id}_c \circ g = g$.

Let us look at some concrete examples categories (see Figure 2.1):

- Any set be regarded as a category whose only morphisms are the identity morphisms. Note that the conditions on composition are vacuously true. Such categories are called *discrete*.

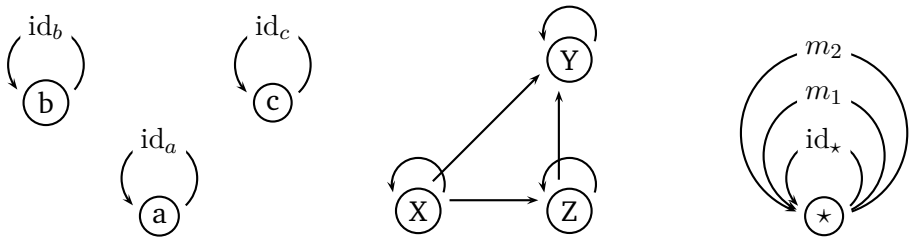
- For every directed graph we can construct a category, called the *free* category generated by the graph. The objects are the vertices of the graph, and the morphisms are the paths in the graph and where composition of morphisms is concatenation of paths.

- A *monoid* is an algebraic structure with a single associative binary operation and an identity element, e.g. $(\mathbb{N}, +, 0)$. We can view any monoid as a category with a single object $\star$. Every element m in the monoid corresponds to a morphism $m : \star \to \star$, the identity morphism $\mathrm{id}_\star$ comes from the identity of the monoid, and the composition of morphisms is given by the monoid operation.

[1] A class is an expression of the type $\{x \mid \phi(x)\}$, where φ is a formula with the free variable x. Informally, a *proper* class is a collection of objects which is too large to be a set under a given axiomatic set theory system, while a class that is a set is called a *small* class.

Note 1

In a slight abuse of notation we often declare an object as an element $c \in \mathbf{C}$ rather than $c \in \mathrm{Ob}(\mathbf{C})$. Whenever the category can be inferred from the context, we denote the morphisms from c to d by $\mathrm{Hom}(c, d)$ and a generic morphism by $f : c \to d$.



(a) A discrete category **(b)** A directed graph **(c)** A monoid

Fig. 2.1: Set, graph and monoid viewed as categories.

The importance of the previous Definition 1 lies on its ability to accommodate plenty of major mathematical constructions:

- The category $\mathbf{Set}$ has the class of all sets as objects together with all functions between them as morphisms and usual function composition as the composition of morphisms,
- $\mathbf{Top}$ is the category whose objects are topological spaces and whose morphisms are continuous maps,
- The category $\mathbf{Vect}_{\mathbb{F}}$ has all vector spaces over a fixed field $\mathbb{F}$ as objects and $\mathbb{F}$-linear transformations as morphisms
- $\mathbf{Man}^{\infty}$ is the category which has all smooth manifolds as objects and smooth maps between them as morphisms.
- Given a category $\mathbf{C}$, we can define the *opposite* or *dual* category $\mathbf{C}^{\mathbf{op}}$ by keeping the same set of objects and reversing the morphisms.

It is evident that a category constitutes a mathematical structure by itself. We can start building a new level of abstraction by studying which kind of maps are those which appropriately preserve structure between categories. Note how such a construction would allows us to consider the category $\mathbf{Cat}$ which consists of all small categories, and the structure-preserving maps between them as morphisms. We have arrived at the notion of a functor.

Definition 2: Functor

Let $\mathbf{C}$ and $\mathbf{D}$ be categories. A functor F from $\mathbf{C}$ to $\mathbf{D}$ consists of:

- a mapping $F : \mathrm{Ob}(\mathbf{C}) \to \mathrm{Ob}(\mathbf{D})$, and
- for all $c, c' \in \mathbf{C}$, a mapping between $\mathrm{Hom}_{\mathbf{C}}(c, c')$ and $\mathrm{Hom}_{\mathbf{D}}(F(c), F(c'))$

such that the following conditions hold:

- for every object $c \in \mathbf{C}$, $F(\mathrm{id}_c) = id_{F(c)}$,
- for every $f : c \to c'$, $g : c' \to \tilde{c}$ in $\mathbf{C}$, $F(g \circ_{\mathbf{C}} f) = F(g) \circ_{\mathbf{D}} F(f)$.

With respect to a reference category $\mathbf{C}$, a functor $F : \mathbf{C} \to \mathbf{D}$ is called *covariant*, while $F : \mathbf{C}^{\mathbf{op}} \to \mathbf{D}$ is called *contravariant*.

In other words, functors are the transformations between categories which preserve identities and composition, see Figure 2.2.

$$\begin{array}{ccc} \mathbf{C} \ni c & \longmapsto & F(c) \in \mathbf{D} \\ \mathrm{Hom}_{\mathbf{C}}(c, c') \ni f \downarrow & & \downarrow F(f) \in \mathrm{Hom}_{\mathbf{D}}(F(c), F(c')) \\ \mathbf{C} \ni c' & \longmapsto & F(c') \in \mathbf{D} \end{array}$$

Fig. 2.2: Graphical representation of a functor.

Some examples of functors are in order:

- $\Delta_d : \mathbf{C} \to \mathbf{D}$ is the *constant* functor which maps every object of $\mathbf{C}$ to a fixed object $d \in \mathbf{D}$ and every morphism in $\mathbf{C}$ to the identity morphism on d.
- The functor $P : \mathbf{Set} \to \mathbf{Set}$ maps a set to its power set and each function $f : X \to Y$ to the map $U \mapsto f(U)$ for each $U \subseteq X$.
- The functor $\pi_1 : \mathbf{Top}. \to \mathbf{Grp}$ maps a topological space with basepoint to its fundamental group based at the given basepoint.
- Different categories are be capable of encoding different structural refinements, and thus the application of a functor might cause information to be lost. For example, the functor $U : \mathbf{Grp} \to \mathbf{Set}$ which maps a group to its underlying set and a homomorphism to its underlying function of sets is a *forgetful* functor.
- The *free* functor $F : \mathbf{Set} \to \mathbf{Grp}$ sends every set to the free group generated by it and functions are associated to group homomorphisms between free groups.

Let us recall what the development has been so far. We started by considering a collection of mathematical objects (say the elements of a group) and translated the information contained in their relations (existence of an identity, existence of inverses, associativity, etc.) to construct a category. Then we realized how by gathering together all similar collections of objects and considering the relations between them (in terms of structure preservation) we could construct a new category (in the running example, $\mathbf{Grp}$). Taking this process one step further, the concept of a natural transformation is revealed.

Definition 3: Natural Transformation

Let F and G be functors between the categories $\mathbf{C}$ and $\mathbf{D}$. A natural transformation $\alpha : F \to G$ is a family of morphisms $\{\alpha_c\}_{c \in \mathbf{C}}$ such that:

- for every object $c \in \mathbf{C}$, $\alpha_c : F(c) \to G(c)$ is a choice of a morphism between objects in $\mathbf{D}$.
- for every morphism $f : c \to c'$ in $\mathbf{C}$, we have that $\alpha_{c'} \circ F(f) = G(f) \circ \alpha_c$.

The last predicate in the previous definition is called the *naturality condition*. This can be conveniently expressed by means of the commutative diagram in Figure 2.3. Note that given functors as in the definition, a natural transformation might not exist. This might occur if, for example, $\mathrm{Hom}_{\mathbf{D}}(F(c), G(c))$ is empty for some $c \in \mathbf{C}$.

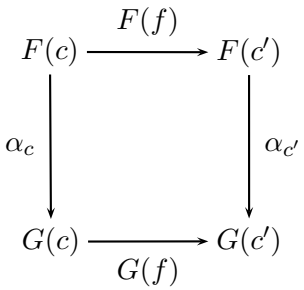


Fig. 2.3: Commutative diagram expressing the naturality of a transformation.

Let us define the category $[\mathbf{C}, \mathbf{D}]$, whose objects are functors from $\mathbf{C}$ to $\mathbf{D}$ and whose morphisms are natural transformations between said functors. As we will see later, a simplicial complex is a functor from the simplicial category $\hat{\mathbf{\Delta}}$ to $\mathbf{Set}$, and the structure preserving transformations, called simplicial mappings, correspond to natural transformations between said functors. We have built a language in which we can make sense of the statement: "the category of simplicial complexes $\mathbf{SCx}$ is the category of functors$[\hat{\mathbf{\Delta}}, \mathbf{Set}]$".

The central theoretical contribution of UMAP is the construction of two adjoint functors which allow to "translate" back and forth between the categories of metric spaces and fuzzy simplicial sets.

Definition 4: Adjunction

An adjunction between teo categories $\mathbf{C}$ and $\mathbf{D}$ consists of two functors $F : \mathbf{D} \to \mathbf{C}$ and a natural isomorphism

$$\mathbf{\Phi} : \mathrm{Hom}_{\mathbf{C}}(F-, -) \to \mathrm{Hom}_{\mathbf{D}}(-, G-).$$

This specifies a family of bijections

$$\mathbf{\Phi}_{cd} : \mathrm{Hom}_{\mathbf{C}}(Fd, c) \to \mathrm{Hom}_{\mathbf{D}}(d, Gc),$$

for all objects $c \in \mathbf{C}$ and $d \in \mathbf{D}$.

We say F is left adjoint to G (resp., G is right adjoint to F) and write $F \dashv G$.

A common view of adjoint functors is related to the construction of "optimal solutions" to certain problems. Let us illustrate this by means of an example. Consider the following procedure for turning a set into S a group:

- Let $G = \varnothing$.
- For every element $s \in S$, add an element s^{-1}. Now, we have $G = \{ a, a^{-1}, b, b^{-1}, \ldots \}$.
- Adjoin an special element λ, called the empty word, which will act as the group identity. At this stage, $G = \{\lambda, a, a^{-1}, b, b^{-1}, \ldots\}$.
- Define a pre-word to be any finite sequence of elements in G, i.e., the group operation is concatenation of strings. A typical pre-word is $abaa^{-1}bbabcc^{-1}$.
- Extend the elements of G to be all *reduced* pre-words, by removing expressions of the form aa^{-1} in every pre-word, and add the corresponding inverse word.

Note that we do not impose any relations between the elements of G which are not forced by the axioms of a group. Intuitively, this the "most efficient" construction of a group out of S. This is, of course, nothing but the free group generated by S we introduced earlier. Similarly, the "most efficient" way to turn a group into a set is by forgetting the group structure and returning the underlying set. Adjoint functors are in a sense, "conceptual inverses" between categories. Pairs of free and forgetful constructions are common examples of adjunctions.

2.2 Topology

We have first mentioned the concept of a manifold as the standard term used to describe regions of (usually) low dimensionality in the data space in which the probability density is highly concentrated. In this section we display manifolds and simplicial complexes as special types of topological spaces. We argue why simplicial complexes are an adequate tool to approximate manifolds, and how a geometric model based on simplicial complexes allows for greater generality.

Topological Spaces

Topology can be considered a qualitative study of shape. It is the analysis of those properties of spaces which are preserved under "continuous" deformations, such as stretching and bending, but not tearing or gluing. For instance, the surfaces of a disk and square share many properties: they are both "two-dimensional" objects with no "holes" and only "one piece", see Figure 2.4. It is the language of topology that will let us drop the quotes.

Fig. 2.4: The surfaces of a disk and a square can be continuously deformed into each other.

Definition 5: Topological Space

A topological space is a tuple $(X, \mathcal{T})$, where X is a set and $\mathcal{T} \subset 2^X$, satisfying the following conditions:

- X and $\varnothing$ belong to $\mathcal{T}$,
- $\mathcal{T}$ is closed under finite intersections,
- $\mathcal{T}$ is closed under arbitrary unions.

The elements of $\mathcal{T}$ are called *open sets*.

There are plenty of widely used examples of topological spaces:

- Any metric space (X, d) can be endowed with a topology which is generated by the *open balls* $B_r(x) = \{y \in X \mid d(x, y) < r\}$.

- The topology generated by the open balls on $\mathbb{R}^n$ with the Euclidean distance is called the standard topology on $\mathbb{R}^n$.

- Given a set S the collections $\{\varnothing, S\}$ and 2^X are topologies on S, called the trivial and discrete topologies.

- Given a topological space $(X, \mathcal{T})$ and a subset U of X, the collection given by $\{U \cap O \mid O \in \mathcal{T}\}$ is a topology on U.

Note 2

Whenever the topology is clear from the context, we refer to X alone as a topological space. Unless stated otherwise we consider every Euclidean space, or any subset thereof, as endowed with the standard topology.

The vague notion of "stretching and bending, but not tearing or gluing" mentioned earlier is formalized in the context of topological spaced by the central notion of continuous transformations. We emphasize the fact that the continuity of a mapping is directly related to the topologies chosen on both the domain and codomain. This means that a morphism of sets $f : X \to Y$ can be continuous with respect to $(X, \mathcal{T}_1)$ and $(Y, \mathcal{O}_1)$, but fail to be continuous with respect to a choice $(X, \mathcal{T}_2)$ and $(Y, \mathcal{O}_2)$.

Definition 6: Continuous Mapping

A mapping $f : (X, \mathcal{T}) \to (Y, \mathcal{O})$ is called continuous if for every $O \in \mathcal{O}$, the preimage $f^{-1}(O) \in \mathcal{T}$.

It is easy to see that for functions on $\mathbb{R}^n$, this definition of continuity is equivalent to the standard "epsilon-delta" definition. However, Definition 6 allows us to see clearly the way in which a continuous function induces a mapping between the collections of open sets by sending $O \in \mathcal{O}$ to $f^{-1}(O) \in \mathcal{T}$. When this mapping between the open sets is a bijection, we obtain an equivalence of topological spaces.

(a) A donut. **(b)** Still a donut.

Fig. 2.5: The surfaces of a donut and a mug are topologically equivalent.

Definition 7: Homeomorphism

A mapping $f : (X, \mathcal{T}) \to (Y, \mathcal{O})$ is called a homeomorphism if:

- f is continuous,
- f is bijective,
- the inverse function f^{-1} is continuous.

We say two topological spaces $(X, \mathcal{T})$ and $(Y, \mathcal{O})$ are homeomorphic, i.e., topologically equivalent, if there exists a homeomorphism between them. We denote this by $(X, \mathcal{T}) =_{\mathbf{Top}} (Y, \mathcal{O})$.

Homeomorphisms induce an equivalence class in the category of topological spaces, and thus, as far as the topology is concerned, we might regard the surface of a donut and that of a coffee mug to be identical, see Figure 2.5. Alternative ways to categorize topological spaces are related to equivalence classes or groups of loops. We proceed to define the first homotopy group, and postpone the definition of homology groups to its simplicial version in the next section.

Simplicial Complexes

Definition 8: Geometric Simplex

A geometric k-simplex in $\mathbb{R}^n$ is the convex set spanned by $k+1$ geometrically independent points $\{x_0, \ldots, x_k\}$. The points x_i are called *vertices*, and the convex set spanned by any non-empty subset of these vertices is called a *face* of the k-simplex. The *standard* geometric k-simplex, denoted Δ^k, is the convex hull of the canonical basis of $\mathbb{R}^n$.

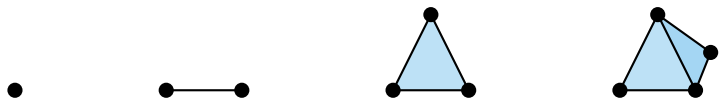


Fig. 2.6: Examples of simplices for dimensions zero to three.

As a subset of $\mathbb{R}^n$ we can endow a simplex with a topology induced from the ambient space. In particular, it is easy to see that a k-simplex is homeomorphic to a k-ball, i.e., a filled k-sphere. This property is crucial for the results regarding the approximation of surfaces using simplicial complexes.

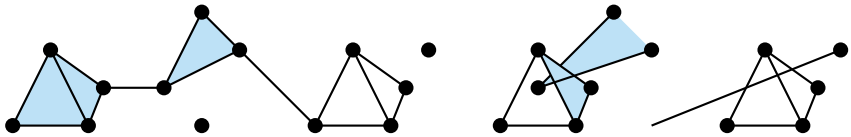
Note 3

Throughout this text we consider all simplices to be *ordered*, i.e, we assume that every set of vertices carries a total order. This implies that the symbol $[x_{i_0}, \ldots, x_{i_k}]$ may stand for a simplex if and only if $x_{i_j} < x_{i_l}$ whenever $j < l$. Note how a face of an ordered simplex corresponds to a totally ordered subset of vertices.

Definition 9: Geometric Simplicial Complex

A geometric simplicial complex K in $\mathbb{R}^n$ is a collection of simplices, of possibly various dimensions, in $\mathbb{R}^n$ such that:

- every face of a simplex of K is in K, and
- the intersection of any two simplices of K is a face of each of them.



(a) A simplicial complex. **(b)** **Not** a simplicial complex.

Fig. 2.7: Example and non-example of a simplicial complex.

Intuitively, we can think of a simplicial complex K as made up of copies of standard simplices of several dimensions, glued together among some common faces. We can organize the relevant information about a simplicial complex into the skeleta K^k, for $k = 0, 1, \ldots$, so that K^k is the set of all k-simplices of K. This purely combinatorial view on a simplicial complex yields the notion of an *abstract* simplicial complex.

Definition 10: Abstract Simplicial Complex

An abstract simplicial complex K consists of a set of vertices K^0 and for each positive integer k, a set K^k consisting of subsets of K^0 of cardinality $k + 1$, with the condition that every $j + 1$-element subset of K^k is an element of K^j. The elements of K^k are called the k-simplices of K.

As we will see later, a manifold is the generalization of the concept of surface to higher dimensions. The standard definition of a manifold implies a *global* choice of dimension for the surface, which in the case of practical datasets might not be appropriate. The intrinsic hierarchical structure between simplices of different dimensions allows simplicial complexes to represent such a diversity in a straightforward way.

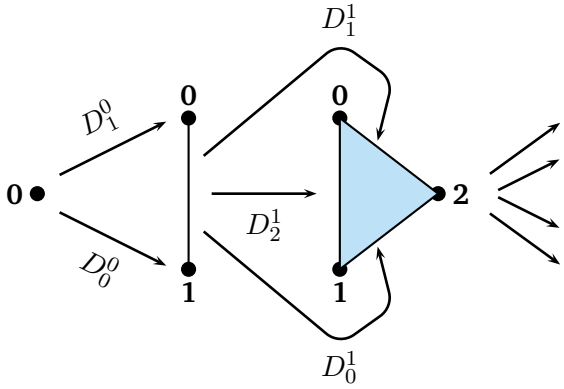


Fig. 2.8: Partial illustration of the category $\hat{\mathbf{\Delta}}$.

Consider the three natural inclusions of a 1-simplex into the 2-simplex. Note how each of these corresponds to an order preserving map $[1] \to [2]$. For instance, the inclusion of the 1-simplex as the face opposite to the vertex 1, symbolized in Figure 2.8 by the arrow D_1^1 which sends $0 \mapsto 0$ and $1 \mapsto 2$. This pattern of objects and arrows between them is the archetypal situation for category theory.

Definition 11: Simplicial Category

The category $\hat{\mathbf{\Delta}}$ has as objects the finite ordered sets $[n] = [0, 1, \ldots, n]$ and as morphisms the strictly order-preserving functions $[m] \to [n]$.

Recall that given a category, its opposite category is formed by the same collection of objects, but with the morphisms reversed. By considering the reversed version of D_1^1, $d_1^1 : [2] \to [1]$, we are effectively obtaining an association between the 2-simplex and its 1-face missing the vertex 1. Now, since the previous definition of a simplicial complex could be interpreted as a collection of sets which are consistent with the face operation, we can recast our definition in terms of a functor.

Definition 12: Simplicial Complex (Categorical Definition)

A simplicial complex is a contravariant functor $K : \hat{\mathbf{\Delta}} \to \mathbf{Set}$, i.e., a functor $\hat{\mathbf{\Delta}}^{\mathbf{op}} \to \mathbf{Set}$.

Once again, we invoke our structure-preserving motto. The adequate notion of a morphism between two simplicial complexes is a simplicial mapping. Such maps will play an important role in our attempt to regularize a parametric autoencoding architecture.

Definition 13: Simplicial Map

Let K and L be geometric simplicial complexes. A simplicial map $f : K \to L$ is given by a function $f : K^0 \to L^0$ and its extension by convex interpolation on each simplex in K.

Algebraically, if a point $x \in K$ can be represented using barycentric coordinates $\{t_j\}$ inside the simplex spanned by $\{x_{i_j}\}$, we have that $f(x) = f\left(\sum_{j=1}^{m} t_j x_{i_j}\right) = \sum_{j=1}^{m} t_j f(x_{i_j})$.

Equivalently, a simplicial map is a natural transformation between the simplicial complexes regarded as functors.

The maps D_i^k represented an inclusion of the standard k-simplex as the i-th face of the $k+1$-simplex. However, consider the simplicial map $\pi : [2] \to [1]$ defined on the vertices as $\pi(0) = 0$ and $\pi(1) = \pi(2) = 1$. This represents a *collapse* of the 2-simplex into the 1-simplex, and thus the image of $[2]$ under π is an example of a *degenerate* simplex, i.e., a simplex that does not have the "correct" number of dimensions. We would like to be able to detect the "hidden" 2-simplex living inside $\pi([2])$. For this, we need to extend our notions of simplex and simplicial category.

Definition 14: Degenerate Simplex

A degenerate k-simplex is a collection $[x_{i_0}, \ldots, x_{i_k}]$ in which $x_{i_j} \leq x_{i_l}$ whenever $j < l$ such that the x_{i_j} are *not* all distinct.

The addition of degeneracy to our view of simplices translates in a straightforward manner to the simplicial category, by allowing maps $[m] \to [n]$ to be non-necessarily strict order-preserving.

Definition 15: Extended Simplicial Category

The category $\mathbf{\Delta}$ has as objects the finite ordered sets $[n] = [0, 1, \ldots, n]$ and as morphisms the order-preserving functions $[m] \to [n]$.

The category-theoretic language developed before allows us to present the generalization of simplicial complexes to simplicial sets elegantly.

Definition 16: Simplicial Set

A simplicial set is a contravariant functor $K : \mathbf{\Delta} \to \mathbf{Set}$.

Note 4

Every ordered simplicial complex K can be "completed" into a simplicial set $\bar{K}$ by adjoining all possible degenerate simplices: for every simplex $[x_{i_0}, \ldots, x_{i_k}] \in K$, we have in $\bar{K}$ all simplices of the form $[x_{i_0}, \ldots, x_{i_0}, x_{i_1}, \ldots, x_{i_1}, \ldots, x_{i_k}]$ for any number of repetitions of each vertex.

We have developed a full theory of simplicial complexes and are now ready to present yet another way to classify topological spaces. Simplicial homology formalizes the idea of the number of holes of a given dimension in a simplicial complex, and can be algorithmically and efficiently computed.

Definition 17: Homology Group

The group C_k of *k-chains* on a simplicial complex K is the free abelian group of finite formal sums with integer coefficients $\left\{\sum_{i=1}^{M} c_i \sigma_i\right\}$, generated by the k-simplices in K.

The *boundary operator* $\partial_k : C_k \to C_{k-1}$ is the homomorphism, defined by the action on the basis of C_k:

$$\partial_k(\sigma) = \sum_{i=0}^{k} (-1)^i [x_0, \ldots, \hat{x}_i, \ldots, x_k].$$

The k-th homology group of a simplicial complex K is the quotient abelian group $H_k(K) = Z_k / B_k = \ker \partial_k / \operatorname{im} \partial_{k+1}$.

The *k-th Betti number* of K is defined as the rank of $H_k(k)$.

The theory of singular homology is defined for all topological spaces and is much more common among the broader mathematical community. Fortunately, singular and simplicial homology agree (Hatcher (2001), Theorem 2.27) for spaces which can be *triangulated*, i.e., spaces homeomorphic to a simplicial complex.

The homology groups for familiar spaces are listed next:

- Let G be a connected graph with spanning tree T and let m be the number of edges of G not in T. The first homology group of G is $\mathbb{Z}^m$. Since the graph is connected, the zeroth homology group has rank 1; and since G is a 1-dimensional simplicial complex, the higher homology groups are trivial.

- In particular, note that a "shallow" 2-simplex has a spanning tree with one edge left out, and thus its first homology group is $\mathbb{Z}$, corresponding to the 1-dimensional hole enclosed by it.

- The kth homology group of the n-sphere is trivial if $k \neq n$ or $\mathbb{Z}$ if $k = n$. This is consistent with the intuition that the sphere encloses an n-dimensional hole, and has no holes of any other dimension.

- The k-th homology group of the n-torus is the free abelian group $\mathbb{Z}^{\binom{n}{k}}$. In particular, for the 2-torus, the ranks of H_0, H_1 and H_2 are 1 (connected component), 2 ("vertical" and "horizontal" cycles), and 1 (2-dimensional hole enclosed by the surface of the torus), respectively.

In practical settings, we often are only provided with a sample $\{x_i\}_{i=1}^N$ of points embedded in a metric space (X, d) and not a simplicial complex. The idea of persistent homology is to build a filtration (growing sequence) of simplicial complexes indexed by some scale parameter, and study the topological properties of the underlying space by computing the homology for all values of the scale parameter. One important example of a filtration is that induced by a sequence of Čech complexes.

Definition 18: Nerve

Consider a collection of open sets $U = \{U_i\}_{i \in I}$, where I is an index set. The nerve of U is the abstract simplicial complex whose k-simplices correspond to all subsets of cardinality k of I, such that the intersection of the corresponding open sets is non-empty.

Definition 19: Čech Complex

Given a set of points $\mathcal{D} = \{x_i\}_{i=1}^N$ in a metric space X and $\epsilon > 0$ we define the Čech complex $\check{\mathrm{C}}_\epsilon(\mathcal{D})$ as the nerve of the collection of open balls $\{B_\epsilon(x_i)\}_{i=1}^N$.

Clearly, given a sample $\mathcal{D}$, the inclusions $\check{\mathrm{C}}_\epsilon(\mathcal{D}) \subset \check{\mathrm{C}}_{\epsilon'}(\mathcal{D})$ holds for $\epsilon \leq \epsilon'$. Note how the scale parameter ϵ acts as global filter on the features we can detect on the topological space. For instance, for a very small ϵ, we get a space with a discrete topology since each point is disconnected from the rest, while for large values of ϵ, we get a fully connected simplicial complex, with trivial topology.

Definition 20: Good Cover

Let X be a topological space. A good cover U is a cllection of open sets $U = \{U_i\}_{i \in I}$, where I is an index set, such that $\bigcup_{i \in I} U_i = X$, and for every finite subset σ of I, the intersection $\bigcap_{i \in \sigma} U_i$ is contractible, i.e., is homotopy-equivalent to a point.

The goal of a construction like the Čech complex is to capture topological information about the underlying space or distribution from which the point cloud is drawn from. The following theorem guarantees that the topological properties we observe in the complex are consistent with those of the underlying space.

Theorem 1: Nerve Lemma (Ghrist, 2014)

A topological space X is homotopy-equivalent to every finite good cover.

Smooth Manifolds

In the introduction we mentioned a special kind of topological spaces which locally resemble an Euclidean space as part of the fundamental hypothesis of dimension reduction, which states that even though the samples we obtain might be embedded in a high-dimensional space, for real data, most of the probability density concentrates around low dimensional regions.

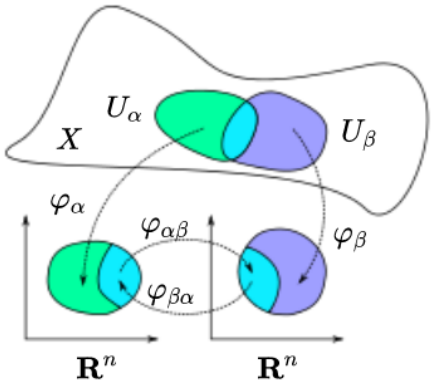


Fig. 2.9: Charts on a manifold.[2]

Definition 21: Manifold

Let $(\mathcal{M}, \mathcal{T})$ be a topological space. A tuple (U, φ), where $U \in \mathcal{T}$, and $\varphi : U \to V$ is a homeomorphism to an open set V in $\mathbb{R}^d$, is called a *chart* on $\mathcal{M}$ and the mapping φ is called a *coordinate system* on U.

An *atlas* on $\mathcal{M}$ is a collection $\mathcal{A} = \{U_\alpha, \varphi_\alpha\}$ such that $\{U_\alpha\}$ is an open cover of $\mathcal{M}$. The homeomorphisms $\varphi_{\alpha\beta} \coloneqq \varphi_\beta \circ \varphi_\alpha^{-1} : \varphi_\alpha(U_\alpha \cap U_\beta) \to \varphi_\beta(U_\alpha \cap U_\beta)$ are called *transition maps*.

A smooth d-dimensional manifold is a topological space $(\mathcal{M}, \mathcal{T})$ enriched with an atlas $\mathcal{A}$, such that all coordinate systems are homeomorphisms with images in $\mathbb{R}^d$ and all transition maps are smooth homeomorphisms, i.e., all partial derivatives exist and are continuous.

[2] Taken from wikipedia.org/wiki/Differentiable_manifold.

This definition of a smooth manifold is described as intrinsic as it makes no reference to an ambient space in which the manifold might be embedded and highlights the importance of charts as the additional structure compared to topological spaces. Figure 2.9 illustrates the consistency condition imposed on transition functions.

Intuitively, if two charts are covering the same region of a manifold ($U_\alpha \cap U_\beta \neq \varnothing$), we would like to translate smoothly between the coordinates of points in the intersection given by the systems φ_α and φ_β. One can consider a point in the manifold to be an equivalence class of points which are mapped to each other by transition maps. The following theorem guarantees that the intrinsic chart-based construction, or the view of a manifold as a surface in a Euclidean space are equivalent.

Theorem 2: Whitney's Embedding Theorem (Whitney, 1944)

Any smooth real d-dimensional manifold can be smoothly embedded in $\mathbb{R}^{2d}$.

Riemannian manifolds are bridges between topological and metric spaces. They are enriched with a metric, which makes it possible to define various geometric notions, such as angles, lengths of curves, volumes, and curvature. In some sense, a topology determines the *shape* of a space, while a Riemannian metric specifies its *geometry*.

Definition 22: Riemannian Manifold

A smooth Riemannian manifold $(\mathcal{M}, g)$ is a smooth manifold $\mathcal{M}$ equipped with an inner product g_p on the tangent space $T_p\mathcal{M}$ at each point p that varies smoothly from point to point. The family g_p of inner products is called a Riemannian *metric tensor*.

Definition 23: Geodesic

Let $(\mathcal{M}, g)$ be a Riemannian Manifold. Let $\gamma(t) : [0,1] \to \mathcal{M}$ be a smooth curve on $\mathcal{M}$. For every $t \in (0,1)$, the inner product $g_{\gamma(t)}$ induces a norm $||\cdot||_{\gamma(t)}$ on the tangent space $T_{\gamma(t)}\mathcal{M}$, and thus on the tangent vector $\gamma'(t) \in T_{\gamma(t)}\mathcal{M}$.

The *length* of the curve α is defined as the integral

$$L(\gamma) = \int_0^1 ||\gamma'(t)||_{\gamma(t)} \mathrm{d}t = \int_0^1 \sqrt{g_{\gamma(t)}(\gamma'(t), \gamma'(t))} \mathrm{d}t.$$

A geodesic between two points $p, q \in \mathcal{M}$ is smooth curve between them which minimizes the energy functional $E(\gamma) = \frac{1}{2}\int_0^1 g_{\gamma(t)}(\gamma'(t), \gamma'(t))\mathrm{d}t$.

As we have argued previously, certain dimension considerations regarding simplicial complexes make them more desirable as an inductive bias for practical applications. However, in order to ensure that we can restrict our attention to simplicial complexes, we should be able to represent any manifold by a simplicial complex. That is precisely the content of the following theorem, illustrated in Figure 2.10.

Theorem 3: Triangulation of a Manifold (Cairns, 1961)

Every smooth manifold $\mathcal{M}$ admits a triangulation (K, h) consisting of a simplicial complex K and a homeomorphism $h : K \to \mathcal{M}$.

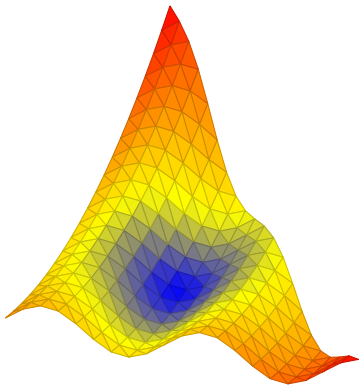


Fig. 2.10: Approximation of a smooth manifold in $\mathbb{R}^3$ with a simplicial complex.

The following result by Niyogi et al. (2008) provides conditions under which a Čech complex constructed from a randomly sampled point cloud is homotopy equivalent to the underlying manifold. The *injectivity radius* τ of a Riemannian manifold $\mathcal{M}$ is the largest number for which all rays orthogonal to $\mathcal{M}$ of length τ are mutually non-intersecting. Intuitively, the notion of *deformation retraction* formalizes the idea of continuously shrinking a space into a subspace.

Theorem 4: Manifold Approximation by a Random Sample

Let $\mathcal{M}$ be a smooth compact submanifold of $\mathbb{R}^n$ with injectivity radius τ. Let $\mathcal{D}$ be a collection of points on $\mathcal{M}$ such that the minimal distance from any point of $\mathcal{M}$ to $\mathcal{D}$ is less than $\epsilon/2$ for $\epsilon < \tau\sqrt{3/5}$, then the Čech complex $\check{\mathrm{C}}_{2\epsilon}(\mathcal{D})$ deformation retracts to $\mathcal{M}$.

2.3 Measure Theory

Measure theory is the study of measures, i.e., systematic ways to assign a "size" to each suitable subset of a set in a way that generalizes the concepts of length, area, and volume. Measure theory is also a framework which allows to unify the common

notions of continuous and discrete random variables as examples of variables which admit densities (in the sense of a Radon-Nikodym derivative) with respect to the Lebesgue or counting measure, respectively.

It would be desirable to assign a size to every subset of a space Ω, but it is in general not possible to do so. For example, the construction of the Vitali sets via the axiom of choice shows that the power set of Ω is "too large" to assign a size to each of its elements in a consistent and non-trivial manner when Ω is uncountable. For this reason, one considers instead a smaller collection of privileged subsets of Ω, a σ-algebra, which is closed under the operations of taking complements and countable unions, and whose elements are called *measurable sets*.

Definition 24: Measure Space

Let Ω be a set. A *σ-algebra* on Ω is a collection $\mathcal{F} \subseteq 2^{\Omega}$ which satisfies:

- $\varnothing \in \mathcal{F}$,
- for all $A \in \mathcal{F}$, $A^C \in \mathcal{F}$,
- for every sequence $(A_n)_{n\in\mathbb{N}} \subseteq \mathcal{F}$, $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$.

A *measure* on a σ-algebra $\mathcal{F}$ is a function $\mu : \mathcal{F} \to [0, \infty]$ such that:

- $\mu(\varnothing) = 0$, and
- $\mu\left(\bigsqcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mu(A_n)$ for every sequence of disjoint sets $(A_n)_{n\in\mathbb{N}} \subseteq \mathcal{F}$.

A tuple $(\Omega, \mathcal{F})$ is called a *measurable space*, while a *measure space* is a tuple $(\Omega, \mathcal{F}, \mu)$. If $\mu(\Omega) = 1$, μ is called a *probability measure* and is usually denoted by $\mathbb{P}$. In that case, $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space*.

Let us examine several examples of the measurable and measure spaces:

- For any countable set S, it is customary to take as σ-algebra the power set of S and as measure, the counting measure $\tau(B) = |B|$, corresponding to the cardinality of the subset B.

- It is easy to verify that the arbitrary intersection of σ-algebras is still a σ-algebra. Given a collection $\mathcal{S}$ of subsets of Ω, we define the σ-algebra generated by $\mathcal{S}$ as the intersection of all σ-algebras which contain $\mathcal{S}$

- In a similar fashion as before, the collection of open balls of a topological space S generates a σ-algebra, called *Borel σ-algebra* on S, denoted $\mathscr{B}(S)$.

- Consider the interval $[0,1]$ endowed with the subset topology from $\mathbb{R}$. Let λ be the 1-dimensional Lebesgue measure defined on the intervals $\lambda([a,b]) = b - a$ for $a \leq b$ (and Carathéodory-extended to $\mathscr{B}([0,1])$). The tuple $([0,1], \mathscr{B}([0,1]), \lambda)$ forms a probability space.

- The space $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \mathbb{P}^{\mathcal{N}})$, where $\mathbb{P}^{\mathcal{N}}(B) = \int_B \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}\lambda(x)$ is a probability space. The measure $\mathbb{P}^{\mathcal{N}}$ is called the standard Gaussian distribution, and the integrand is called the density of this distribution with respect to the Lebesgue measure on $\mathbb{R}$.

It should not be a surprise that we consider structure-preserving maps between measurable spaces, called measurable mappings. In the context of probability theory, those maps are called random variables. Note the similarity between the following defintion and that of continuous mappings.

Definition 25: Measure Mapping

A mapping between measurable spaces $f : (\Omega, \mathscr{F}) \to (\Psi, \mathscr{G})$ is called measurable if for every $B \in \mathscr{G}$, the preimage $f^{-1}(B) \in \mathscr{F}$.

A Ψ-valued *random variable* is a measurable mapping $X : (\Omega, \mathscr{F}, \mathbb{P}) \to (\Psi, \mathscr{G})$.

The remarkable aspect of random variables is that their structure-preserving property allows us to "transport" or "push-forward" the measure on the domain probability space to the target measurable space.

Definition 26: Distribution of a Random Variable

A Ψ-valued random variable $X : (\Omega, \mathscr{F}, \mathbb{P}) \to (\Psi, \mathscr{G})$ induces a measure on $\mathscr{G}$ by the *pushforward* of $\mathbb{P}$ under X, defined for $B \in \mathscr{G}$ by:

$$\mathbb{P}^X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in B\}).$$

The measure $\mathbb{P}^X$ is called the distribution or *law* of the random variable X.

So far, our description of random variables requires our ability to construct a domain probability space. The following theorem guarantees that for every sufficiently well-behaved probability measure on a metric space, there exists a random variable whose law matches our prescribed measure.

Theorem 5: Skorokhod's representation theorem (Dudley, 1968)

Let $\mathbb{Q}$ be a probability measure on a metric space Ψ with separable support. Then there exist a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a Ψ-valued random variable X defined on it such that $\mathbb{P}^X = \mathbb{Q}$.

We close this section with a family of measures which will be very important in our treatment of probabilistic models based on random simplicial complexes. A characteristic of the n-dimensional Lebesgue measure is that any m-manifold in $\mathbb{R}^n$, with $m < n$, has measure zero.

However, we would like to describe that within a simplicial complex in $\mathbb{R}^n$, the 1-simplices have length, the 2-simplices have area, etc. The d-dimensional Hausdorff measure provides such a generalization in a way that coincides exactly with the Lebesgue measure for Euclidean spaces.

Definition 27: Haussdorff Measure

Let (S, ρ) be a metric space. The *diameter* of a subset A of S is defined by $\operatorname{diam} A = \sup\{\rho(x, y) \mid x, y \in A\}$. The d-dimensional Haussdorff (outer) measure of a subset U of S is defined by

$$\mathscr{H}^d(U) = \lim_{\delta \to 0} H^d_\delta(U) \coloneqq \lim_{\delta \to 0} \inf\left\{ \sum_{i=1}^{\infty} (\operatorname{diam}\ A_i)^d : \bigcup_{i=1}^{\infty} A_i \supseteq U, \operatorname{diam}\ A_i < \delta \right\}.$$

2.4 Fuzzy Sets

Under the Zermelo-Fraenkel axioms for set theory, the membership of an element in a set is assessed in a binary fashion: the element belongs to the set or not. In fuzzy set theory, this condition is relaxed, allowing for a gradual assessment of the membership in terms of a real number in the interval $[0, 1]$. Note that a real valued membership is a natural way to encode *uncertainty* in the structure of a set.

Let $\mathbf{I}$ be the unit interval $(0, 1]$ with open sets the intervals $(0, a)$ for $a \in (0, 1]$. We consider $\mathbf{I}$ as a category of open sets, with morphisms given by inclusion.

Definition 28: Fuzzy Set

A fuzzy set is set S enriched with a function $\mu : S \to [0,1]$. Given fuzzy sets (S,μ) and (T,ν) a morphism of between them is a function $f : S \to T$ such that for all $s \in S$, $\mu(s) \leq \nu(f(s))$.

Equivalently, a fuzzy set can be defined as a contravariant functor $\mathscr{P} : \mathbf{I}^{\mathbf{op}} \to \mathbf{Set}$ such that all morphisms $\mathscr{P}(a \leq b)$ are injections. We denote the category of fuzzy sets and morphisms betwen them by **Fuz**.

Intuitively, one can think about the action of the functor $\mathscr{P}$ on the element $(0,a)$ as selecting the set of elements whose membership is at least a, i.e., in terms of the membership function, the super-level set $\{\mu \geq a\}$. For every $a, b \in (0,1]$ with $a \leq b$, one gets an inclusion between the super-level sets $\{\mu \geq a\} \supseteq \{\mu \geq b\}$ which justifies the requirement for injectivity in the definition. Note how the inversion $\leq \mapsto \supseteq$ relates to the definition of a fuzzy set as a contravariant functor.

It should be clear that fuzzy sets represent a generalization of classical (crisp) sets, for which the membership is an indicator function. Similarly, there are ways to define operations between fuzzy sets which resemble their crisp counterparts.

Definition 29: De Morgan triplet

A *strong negator* $\neg$ is a monotonous decreasing involutive function with $\neg 0 = 1$ and $\neg 1 = 0$.

A *t-norm* is a symmetric function $\top : [0,1]^2 \to [0,1]$ satisfying:

- $\top(a,b) \leq \top(c,d)$ whenever $a \leq c$ and $b \leq d$,
- $\top(a, \top(b,c)) = \top(\top(a,b), c)$, and
- $\top(1,a) = a$.

Given a t-norm $\top$, its complementary *conorm* under the negator $\neg$ is defined by $\bot(a,b) = \neg\top(\neg a, \neg b)$.

A De Morgan triplet is a triple $(\top, \bot, \neg)$ where $\top$ is a t-norm, $\bot$ is the associated t-conorm, $\neg$ is a strong negator and for all $a, b \in [0,1]$ one has that $\neg\bot(a,b) = \top(\neg a, \neg b)$.

The most common example of a De Morgan triplet is the one formed by $\top_{\text{prod}}(a,b) = ab$, $\bot_{\text{sum}} = a + b - ab$ and $\neg(a) = 1 - a$. Note how the t-norm and t-conorm express the probability of intersection and union of independent events. Another important example arises by taking $\top_{\text{min}}(a,b) = \min(a,b)$, $\bot_{\text{max}} = \max(a,b)$.

Definition 30: Operations on Fuzzy Sets

Let $(\top, \bot, \neg)$ be a De Morgan triplet. Let U be a set and let μ and ν be membership functions on U.

The *complement* of μ given by the function $\neg \circ \mu$.

We define the *intersection* of μ and ν as the function $\tau_{\mu \cap \nu}(\cdot) = \top(\mu(\cdot), \nu(\cdot))$.

Th *union* of μ and ν is the membership function $\tau_{\mu \cup \nu}(\cdot) = \bot(\mu(\cdot), \nu(\cdot))$.

Given two fuzzy membership functions on a common set U, we can define a notion of dissimilarity between them by means of the fuzzy set cross entropy .

Definition 31: Fuzzy Set Cross Entropy

The cross entropy between two fuzzy sets μ and ν on a common carrier set U is defined as

$$C_U(\mu, \nu) = \sum_{u \in U} \mu(u) \log \left(\frac{\mu(u)}{\nu(u)} \right) + (1 - \mu(u)) \log \left(\frac{1 - \mu(u)}{1 - \nu(u)} \right).$$

For every fuzzy set, one can construct a family of distributions $\{\text{Ber}(\mu(u)) \mid u \in U\}$. Note that the fuzzy cross entropy can be rewritten in terms of a sum of pointwise Kullback-Leibler divergences

$$C_U(\mu, \nu) = \sum_{u \in U} \mathbb{KL}(\text{Ber}(\mu(u)) \,||\, \text{Ber}(\nu(u))).$$

Not very surprisingly, just as we could construct the category **Set** of sets and function between them, there is a category **Fuz** of fuzzy sets and fuzzy set morphisms between them. With this category in mind, we can state a final generalization of simplicial complexes and sets.

Definition 32: Fuzzy Simplicial Complexes and Sets

A fuzzy simplicial complex is a functor $K : \hat{\mathbf{\Delta}}^{\mathbf{op}} \to \mathbf{Fuz}$. A fuzzy simplicial set is a functor $K : \mathbf{\Delta}^{\mathbf{op}} \to \mathbf{Fuz}$.

We denote the category of fuzzy simplicial sets and natural transformations between them by $\mathbf{sFuz}$.

2.5 Generative Models

We assume familiarity of the reader with concepts related to Deep Learning. For completeness we define graphical models, neural networks, and autoencoders in this section. Goodfellow et al. (2016) provide a good overview of the field. In particular, we refer the interested reader to chapters 6, 14 and 20.

Suppose we are given a dataset of points coming from a probability distribution $\mathbb{P}$ on $\mathbb{R}^n$. If $\mathbb{P}$ is, for instance, the distribution of pictures of cars, a concise description of it is, for all practical matters, non-existent. Thus, we need to resort to alternative ways to gain insights about $\mathbb{P}$.

According to Bishop (2006), "producing synthetic observations form a generative model can prove informative in understanding the form of the probability distribution represented by that model". Additionally, being able to sample new points from a distribution which *resembles* $\mathbb{P}$ would allows us, among other things, to estimate intractable sums, speed up training, or provide the raw material on which to train a model to solve a particular task.

Graphical Model

Definition 33: Graphical Model

A graphical model is a probabilistic model which expresses conditional (in)dependence relations between random variables by means of a graph. In the case of a directed acyclic graph, the model represents the factorization of the joint distribution of all random variables given by:

$$\mathbb{P}(X_1, \dots, X_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \mid \text{pa}_i),$$

where pa_i is the set of parents of node X_i.

Let us illustrate the concept of a graphical model by means of an example. Suppose that we are given a dataset $\{x_i\}_{i=1}^N$. Assume that these observations are independent and follow a Gaussian distribution with variance 1 but with an unknown mean μ. We can, in turn, reflect our uncertainty about μ by selecting a prior $\mathcal{N}(\mu \mid \mu_0, \sigma_0^2)$, for some μ_0, σ_0. All the previous dependence (between and observation x and μ) and independence (between two different observations) relations can be concisely represented by the graphical model in Figure 2.11.

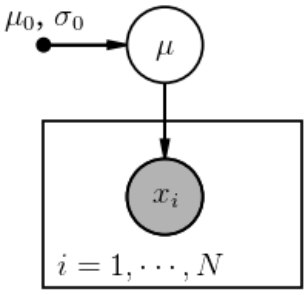


Fig. 2.11: Graphical model for an iid sequence of Gaussian random variables.

According to the definition, and our selection of Gaussian distributions for the prior and observation model, we can factorize the joint distribution of the observations $\{x_i\}$ and unknown parameter μ by:

$$p(\{x_i\}_{i=1}^N, \mu) = \mathcal{N}(\mu \mid \mu_0, \sigma_0^2) \prod_{i=1}^n \mathcal{N}(x_i \mid \mu, 1)$$

Given this representation, we can readily postulate maximum likelihood or maximum a posteriori estimates for the parameter μ. Additionally, we can *generate* new datapoints $\{\hat{x}_i\}$ by *ancestral sampling*: sample $\hat{\mu} \sim \mathcal{N}(\mu_0, \sigma_0^2)$ and then sample $\{\hat{x}_i\}$ iid according to $\mathcal{N}(\hat{\mu}, 1)$. If the model were a perfect representation of the data, then the probability distribution of $\{\hat{x}_i\}$ would coincide with the real distribution.

Neural Networks

Consider the function $f : [0, \ldots, 255]^{28\times 28} \to \{\text{cat}, \text{dog}, \text{none}\}$, which receives as input a 28-by-28 pixels grayscale image and determines whether it is a cat or a dog, or neither. There are $3^{256\cdot 28^2} \approx 10^{95760}$ possible such functions. In comparison, the estimated number of atoms in the universe is around 10^{80}. It seems like a very hopeless situation to find one such function. And it is indeed!

Definition 34: Neural Network

A neural network from $\mathbb{R}^{n_{\text{in}}}$ to $\mathbb{R}^{n_{\text{out}}}$ is a function of the form

$$f = \sigma_L \circ \mathbf{A}_L \circ \ldots \sigma_2 \circ \mathbf{A}_2 \ldots \sigma_1 \circ \mathbf{A}_1,$$

where $\{\mathbf{A}_i\}$ are affine transformations between Euclidean spaces of consistent dimensions and σ_i is a non-linear, non-polynomial function, applied element-wise. In some cases the final *activation function* σ_L is taken to be an identity map.

Neural networks are special types of functions which try to *approximate* a desired behavior by successively composing affine and non-linear transformations. This means that we give up on the goal to find *the* perfect classification function f, but rather focus on a particular family of functions, which are hopefully broad enough to approximate f adequately.

Note that neural networks are inherently hierarchical. Every *layer* (a pair $\sigma_i \circ \mathbf{A}_i$) builds on top of the representation provided by the previous layer. The following theorem ensure that this hierarchical structure makes the class of neural networks rich enough for us to approximate any desired continuous behavior arbitrarily well.

Theorem 6: Universal Approximation (Hornik, 1991)

Let X be a compact subset of $\mathbb{R}^k$ and let $\mathscr{C}(X)$ be the class of continuous functions on X. Let $\mathcal{N}_k^m$ be the class of functions from $\mathbb{R}^k$ to $\mathbb{R}$ which can be implemented by neural networks with one m-dimensional hidden layer and activation function σ.

If σ is continuous, bounded and non-constant, then the class $\bigcup_{m\in\mathbb{N}} \mathcal{N}_k^m$ is dense in $\mathscr{C}(X)$.

AutoEncoders

An autoencoder is a pair of neural networks which are trained to be mutual inverses. We hope that by training the joint system, we can learn useful properties of the input data. For this reason an autoencoder which acts as an identity function over the whole input space is not particularly useful. Instead, autoencoders are constrained in ways which do not allow for perfect invertibility of the individual components.

Definition 35: AutoEncoder

Let X and Z be Euclidean spaces and $\mathcal{M}$ a submanifold embedded in X . An autoencoder for $\mathcal{M}$ is a pair of functions $e : X \to Z$ and $d : Z \to X$ such that $d \circ e|_{\mathcal{M}} \approx \mathrm{id}_{\mathcal{M}}$. The images of e are usually called *codes*, and the images of d, *reconstructions*.

For instance, *undercomplete* autoencoders involve a dimensionality bottleneck, which destroys the invertibility and forces the autoencoder to capture the most relevant characteristics of the data in the available dimensions. Alternatively, *overcomplete* autoencoders, rather than constraining the architecture, impose conditions on the learned codes by, for example, encouraging sparsity or smoothness.

3 UMAP as Approximate MAP

> *Mathematics in general is fundamentally the science of self-evident things*".
>
> — **Felix Klein**

In this section we provide an interpretation of UMAP as an approximate maximum a posteriori estimator on a statistical model parameterized by simplicial complexes. First, we provide a brief account of the theoretical foundations of UMAP. We then prove an equivalence result between fuzzy sets and a class of random variables. We introduce the notion of a K-parameterized statistical model and introduce the Haussdorff distribution on a simplicial complex. We show that in the limit of a large dataset, under certain conditions, the true underlying topological space can be recovered by maximum likelihood. Finally, we cast UMAP as an approximate maximum a posteriori estimator via a Lagrangian relaxation of a constrained maximum likelihood problem.

3.1 UMAP

Consider a dataset $\mathscr{D} = \{x_i\}_{i=1}^N$ of samples in $\mathbb{R}^n$. Recall the construction of the Čech complex $\check{\mathrm{C}}(\mathscr{D})$ as the nerve of the collection of open balls of radius ϵ centered at the points in $\mathscr{D}$. The guarantee provided by the Nerve Lemma (Theorem 1) is conditioned on the collection $\{B_\epsilon(x_i)\}_{i=1}^N$ to be a *good cover* of the underlying topological space, i.e., all intersection of such balls has to be contractible. One way to ensure this, is to endow the underlying manifold with a Riemannian metric such that our sample is approximately uniformly distributed with respect to that metric, which is in general different from that inherited from the ambient space.

The main idea of UMAP is to construct a custom metric for each x_i so as to ensure the validity of the uniformity assumption. Then translate each of these metric spaces into fuzzy simplicial sets, in such a way that the topological information is filtered but preserving information about the metric structure. Finally, merge these individual fuzzy sets by taking a fuzzy union between them. This provides a fuzzy topological representation of $\mathscr{D}$, denoted by $K_{\mathscr{D}}$.

If we are interested in finding a low-dimensional embedding for $\mathscr{D}$, we can start with a (randomly) initialized embedding $\mathscr{Z}$, compute its fuzzy topological representation $K_{\mathscr{Z}}$ and iteratively optimize $\mathscr{Z}$ so as to minimize the fuzzy set cross-entropy between the fuzzy topological representations, $C(K_{\mathscr{D}}, K_{\mathscr{Z}})$. In practice, this is done by computing the fuzzy set cross-entropy betwen the 1-skeletons of each of the simplicial sets considered as fuzzy sets of edges.

Theorem 7: UMAP Adjunction

Let **FinEPMet** be the category of finite extended pseudometric spaces with non-expansive maps as morphisms. For $a \in (0, 1]$ define the metric d_a by $d_a(x, x) = 0$ and $d_a(x, y) = \log\left(\frac{1}{a}\right)$ for $x \neq y$. Given a simplicial set K, let $K^k_{<a}$ be the set $K([k], (0, a))$, in other words, the set of k-simplices with membership at least a.

Define the functor $\mathrm{FinReal} : \mathbf{sFuz} \to \mathbf{FinEPMet}$ by:

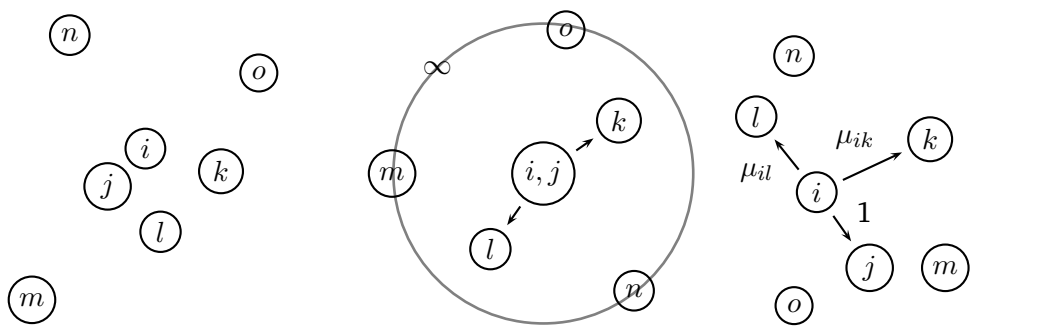
$$\mathrm{FinReal}(\Delta^k_{<a}) = (\{\star_1, \ldots, \star_k\}, d_a) \qquad \mathrm{FinReal}(K) = \operatorname*{colim}_{\Delta^k_{<a} \to K} \mathrm{FinReal}(\Delta^k_{<a}).$$

Define the functor $\mathrm{FinSing} : \mathbf{FinEPMet} \to \mathbf{sFuz}$ by:

$$\mathrm{FinSing}(Y)^k_{<a} = \mathrm{Hom}_{\mathbf{FinEPMet}}(\mathrm{FinReal}(\Delta^n_{<a}), Y)$$

The functors $\mathrm{FinReal}$ and $\mathrm{FinSing}$ form an adjunction $\mathrm{FinReal} \dashv \mathrm{FinSing}$.

Around every datapoint x_i UMAP constructs a finite extended pseudo-metric space. This construction is based on an approximation to the geodesic distance from x_i to its neighbors by normalizing the distances with respect to the distance to the $\tilde{n}$-th nearest neighbor of x_i (McInnes and Healy (2018), Lemma 1). In practice, this is performed by constructing a fuzzy set of edges from x_i to its $\tilde{n}$ nearest neighbors such that the cardinality of the set is equal to $\tilde{n}$. This is related to the choice of a target entropy for the conditional distribution around a point in t-SNE.



(a) $(X, d) \in \mathbf{FinMet}$ **(b)** $(X, d_i) \in \mathbf{FinEPMet}$ **(c)** $\mathrm{FinSing}(X, d_i) \in \mathbf{sFuz}$

Fig. 3.1: Image of the singular functor in the context of UMAP.

Let us study the action of the singular functor on the example metric space shown in Figure 3.1. For every datapoint i we construct a finite extended pseudometric space (X, d_i) around it by considering the distances from i to its $\tilde{n} = 3$ nearest neighbors *beyond* the distance to the nearest neighbor j. For this reason, $d_i(i, j) = 0$ even though $i \neq j$, and those points which are not within the $\tilde{n}$ nearest neighbors of i are considered to be infinitely far way. This has an important consequence in terms of the size of the fuzzy set of edges, since it restricts the complexity from $O(N^2)$ to $O(N\tilde{n})$. Besides, we take $d_i(p, q) = \infty$ if neither p or q are equal to i.

Upon applying the singular functor, we end up with a fuzzy simplicial set (which in this case corresponds to a fuzzy set of edges centered at i). Note how there is a full-strength connection between i and its nearest neighbor. We highlight the fact that the adjunction is indeed a *weak* form of equivalence between the categories **FinEPMet** and **sFuz**. For instance, trying to reconstruct the metric space from the fuzzy simplicial set, we only know that i and j should be nearest neighbors, but we have lost information regarding the exact distance between them.

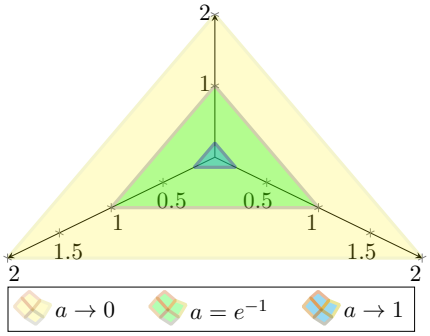


Fig. 3.2: Image of the metric realization functor on objects of the type $([\mathbf{2}], [0, a))$.

Figure 3.2 illustrates the action of the metric realization functor on the representable functors of objects of the type $([\mathbf{2}], (0, a))$. Note that this corresponds to the finite extended pseudometric space formed by the corners of the standard geometric 2-simplex, scaled according to the membership strength. Thus, 2-simplices with lower membership (a close to zero) result in corners being placed far from each other, while 2-simplices with strong membership (a close to one) induce a mapping to a shrunk geometric simplex.

For completeness, we present Algorithms 1 and 2, which summarize the computational pipeline for UMAP. In practice the fuzzy topological representation $K_{\mathscr{X}}$ is not

fully computed, but rather the objective is optimized via negative sampling on the edges of $K_{\mathscr{D}}$ and then the corresponding memberships in $\mathscr{Z}$ are calculated.

Algorithm 1: `FuzzyTop` - Fuzzy topological representation of a dataset.

Data: Dataset $\mathscr{D} = \{x_i\}_{i=1}^N \subset \mathbb{R}^n$, number of neighbors $\tilde{n}$.

Result: Fuzzy topological representation of $\mathscr{D}$ given by $K_{\mathscr{D}}$.

for $i = 1, \ldots, N$ **do**
  Compute $(\mathscr{D}, d_i) \in$ **FinEPMet**
  $K_i = \mathrm{FinSing}(\mathscr{D}, d_i)) \in$ **sFuz**
end
$K_{\mathscr{D}} = \perp_{i=1}^N K_i$

Algorithm 2: `UMAP` - Uniform Manifold Approximation and Projection.

Data: Dataset $\mathscr{D} = \{x_i\}_{i=1}^N \subset \mathbb{R}^n$, number of neighbors $\tilde{n}$, embedding dimension d, maximum iterations $i_{\max}$, learning rate α.

Result: Low dimensional embedding of $\mathscr{D}$ given by $\mathscr{Z} = \{z_i\}_{i=1}^N \subset \mathbb{R}^d$.

$K_{\mathscr{D}} = \texttt{FuzzyTop}(\mathscr{D}, \tilde{n})$
Initialize $\mathscr{Z} \subset \mathbb{R}^d$
$K_{\mathscr{Z}_0} = \texttt{FuzzyTop}(\mathscr{Z}_0)$
for $\tau = 1, \ldots, i_{\max}$ **do**
  $l = C(K_{\mathscr{D}}, K_{\mathscr{Z}_{\tau-1}})$
  $\mathscr{Z}_\tau = \mathscr{Z}_{\tau-1} - \alpha \nabla_{\mathscr{Z}_{\tau-1}} l$
  $K_{\mathscr{Z}_\tau} = \texttt{FuzzyTop}(\mathscr{Z}_\tau)$
end

3.2 A correspondence between random variables and fuzzy sets

We now present an equivalence between the fuzzy sets defined on a set S and a special kind of set-valued random variables. This result will be crucial in our probabilistic interpretation of UMAP. Let us begin with a definition.

Definition 36: Set-valued Random Variable

Let S be a set, $(\Omega, \mathscr{F}, \mathbb{P})$ a probability space and $(2^S, \Sigma)$ a measurable space. A mapping $X : (\Omega, \mathscr{F}, \mathbb{P}) \to (2^S, \Sigma)$ is called an S-set-valued random variable.

In the case Ω is a totally ordered set, we say X is non-increasing if for all $\omega, \omega' \in \Omega$ with $\omega \leq \omega'$, $X(\omega) \supseteq X(\omega')$.

Theorem 8

Let $\mathscr{F}(S)$ be the set of fuzzy sets with common carrier set S. There exists a bijection between $\mathscr{F}(S)$ and the class non-increasing S-set-valued random variables.

Proof. Let $(\Omega, \Sigma, \mathbb{P})$ be the probability space $([0,1], \mathscr{B}, \lambda)$. Recall that $\mathscr{F}(S)$ is, by definition, the set of functions $\mathrm{Hom}_{\mathbf{Set}}(S, [0,1])$. For each μ in $\mathscr{F}(S)$ define the mapping $M_\mu : \Omega \to 2^S$ by:

$$M_\mu(\omega) = \{s \in S \,|\, \mu(s) \geq \omega\}.$$

Let Σ be the largest σ-algebra on 2^S for which the all mappings in $\{M_\mu \,|\, \mu \in \mathscr{F}(S)\}$ are measurable. It is clear that each M_μ is, by construction, non-increasing, and thus, by definition of Σ, an S-set-valued random variable. Let $\downarrow(S)$ be the class of non-increasing S-set-valued random variables.

Consider the map $\mathscr{N} : S \to 2^{2^X}$ which sends $s \in S$ to the set $\{W \in 2^S \mid s \in W\}$. Given $M \in \downarrow(S)$, we would like to define a membership function on S by:

$$\mu_M(s) = \mathbb{P}(M \ni s) = \mathbb{P}(M \in \mathscr{N}(s)) = \mathbb{P}^M(\mathscr{N}(s)).$$

However, for this construction to make any sense at all, we need to ensure that the pre-images of the sets $\mathscr{N}(s)$ are measurable sets in $\mathscr{B}$. For this purpose, let $\overline{M}_s = \sup\{\omega \in \Omega : s \in M(\omega)\}$. Since M is non-increasing, for every $\omega \in [0, \overline{M}_s]$, $s \in M(\omega)$, which implies that $M(\omega) \in \mathscr{N}(s)$. On the other hand, for all $\omega \in (\overline{M}_s, 1]$, $s \notin M(\omega)$, and thus there is no $N \in \mathscr{N}(s)$ such that $M(\omega) = N$. In conclusion, for every $s \in S$, $M^{-1}(\mathscr{N}(s)) = [0, \overline{M}_s($ (where the right end-point might be open or closed), and therefore a Borel-measurable set.

We are now certain that the definition of μ_M is sound, and we even have a simple expression for it given by $\mu_M(s) = \mathbb{P}\left(M^{-1}\left(\mathscr{N}(s)\right)\right) = \lambda\left([0, \overline{M}_s(\right) = \overline{M}_s$.

We now show that the proposed mappings establish a bijective correspondence between $\mathscr{F}(S)$ and $\downarrow(S)$. Let $\mu \in \mathscr{F}(S)$. Recall that the condition $s \in M_\mu(\omega)$ is equivalent to $\mu(x) \geq \omega$. Therefore we have pointwise equality between μ and μ_{M_μ}:

$$\mu_{M_\mu}(s) = \overline{M}_{\mu_s} = \sup\{\omega \in \Omega : s \in M_\mu(\omega)\} = \sup\{\omega \in \Omega : \mu(s) \geq \omega\} = \mu(s).$$

Finally, let $M \in \downarrow(S)$. Since $s \in M(\omega)$ if and only if $\overline{M}_s \geq \omega$, we have pointwise equality between M and M_{μ_M}:

$$M_{\mu_M}(\omega) = \{s \in S : \mu_M(s) \geq \omega\} = \{s \in S : \overline{M}_s \geq \omega\} = M(\omega).$$

□

3.3 UMAP as Approximate MAP

K-parameterized statistical models

Definition 37: Statistical Model

A statistical model is a pair $(S, \mathbb{M})$, where S is the set of possible observations and $\mathbb{M}$ is a set of probability measures on S.

We say $\mathbb{M}$ is Θ-*parameterized* if every measure in $\mathbb{M}$ can be parameterized by some $\theta \in \Theta$, i.e., $\mathbb{M} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$. The parametrization is called *identifiable* or *injective* if $\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2}$ implies $\theta_1 = \theta_2$.

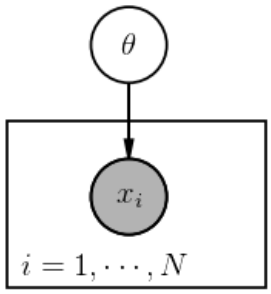


Fig. 3.3: Graphical model representing the generative process of observations x under a statistical model parameterized by θ.

In the usual setting, we interpret the generative process of the data as being determined by a distribution given by a "true" value of the parameter θ_{true}. This corresponds to the graphical model presented in Figure 3.3. The key step in our extension is to consider distributions which are parameterized by simplicial complexes. For this, we introduce the Hausdorff distribution induced by a simplicial complex.

Definition 38: Hausdorff Distribution on a Simplicial Complex

Let K be a simplicial complex in $\mathbb{R}^n$. For every simplex $\sigma \in K$, define the probability measure δ_σ on $\mathscr{B}(\mathbb{R}^n)$ by:

$$\delta_\sigma(\cdot) = \frac{\mathscr{H}^{\dim(\sigma)}(\cdot \cap \sigma)}{\mathscr{H}^{\dim(\sigma)}(\sigma)}$$

We define the Hausdorff probability distribution on $\mathscr{B}(\mathbb{R}^n)$ induced by K as

$$\mathbb{H}(\cdot|K) = \frac{1}{\#K} \sum_{\sigma \in K} \delta_\sigma(\cdot),$$

where $\#K$ is the number of simplices in K

Note 5

Throughout this work, and in practice while training models, one only considers a finite dataset, and thus a complex with a *finite* number of simplices, for which the induced Hausdorff measure is well-defined.

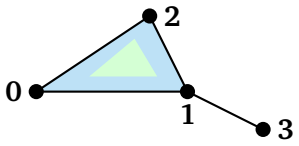


Fig. 3.4: An example of a simplicial complex.

Consider the simplicial complex shown in Figure 3.4. Let us examine the value of the Hausdorff distribution induced by K for several Borel-sets in $\mathbb{R}^2$. Note that K contains 9 simplices: 4 0-simplices, 4 1-simplices and 1 2-simplex.

- $B = [\mathbf{0}]$. We only need to consider those simplices with which B has a non-empty intersection, i.e., $[\mathbf{0}]$, $[\mathbf{0}, \mathbf{1}]$, $[\mathbf{0}, \mathbf{2}]$ and $[\mathbf{0}, \mathbf{1}, \mathbf{2}]$.

$$\mathbb{H}(B|K) = \frac{1}{9}\left(\delta_{[\mathbf{0}]}([\mathbf{0}]) + \delta_{[\mathbf{0},\mathbf{1}]}([\mathbf{0}]) + \delta_{[\mathbf{0},\mathbf{2}]}([\mathbf{0}]) + \delta_{[\mathbf{0},\mathbf{1},\mathbf{2}]}([\mathbf{0}])\right) = \frac{1}{9},$$

  since any Hausdorff measure of dimension greater than zero vanishes at $[\mathbf{0}]$.

- Similarly, for $B = \{\mathbf{0}, \mathbf{1}, \mathbf{2}, \mathbf{3}\}$, $\mathbb{H}(B|K) = \frac{4}{9}$.

- Let $B = [\mathbf{0}, \mathbf{1}, \mathbf{2}]$. Note that if $\sigma \subseteq B$, then $\delta_\sigma(B) = 1$.

$$\begin{aligned}\mathbb{H}(B|K) = \frac{1}{9}(&\delta_{[\mathbf{0}]}(B) + \delta_{[\mathbf{1}]}(B) + \delta_{[\mathbf{2}]}(B) + \cancel{\delta_{[\mathbf{3}]}(B)} + \delta_{[\mathbf{0},\mathbf{1}]}(B) \\ &+ \delta_{[\mathbf{0},\mathbf{2}]}(B) + \delta_{[\mathbf{1},\mathbf{2}]}(B) + \cancel{\delta_{[\mathbf{1},\mathbf{3}]}(B)} + \delta_{[\mathbf{0},\mathbf{1},\mathbf{2}]}(B)) = \frac{7}{9}\end{aligned}$$

  Note that $\delta_{[\mathbf{3}]}(B)$ vanishes since the intersection is empty. $\delta_{[\mathbf{1},\mathbf{3}]}(B)$ is also null since, even though the intersection $[\mathbf{1}, \mathbf{3}] \cap B = [\mathbf{1}]$, $[\mathbf{1}, \mathbf{3}]$ is a 1-simplex and therefore, the 1-dimensional Hausdorff measure of $[\mathbf{1}]$ is zero.

- If B is the green triangle, $\mathbb{H}(B|K)$ corresponds to $\frac{1}{9}$ times the proportion of the area of $[\mathbf{0}, \mathbf{1}, \mathbf{2}]$ covered by B.

Let $\mathfrak{H} = \{\mathbb{H}(\cdot|K) \mid K \text{ is a simplicial complex in } \mathbb{R}^n\}$ and consider the statistical model $(\mathbb{R}^n, \mathfrak{H})$. Note that $(\mathbb{R}^n, \mathfrak{H})$ is injectively parameterized by the set of simplicial complexes in $\mathbb{R}^n$. The corresponding graphical model is presented in Figure 3.5.

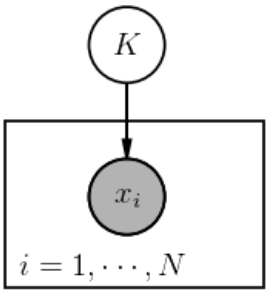


Fig. 3.5: Graphical model for a K-parameterized statistical model on $\mathbb{R}^n$.

Theorem 9

Let $\mathscr{D} = \{x_i\}_{i=1}^N$ be a sample of points in $\mathbb{R}^n$. The distribution induced by the maximum likelihood estimator corresponds to the empirical distribution of the data.

Geometric Proof. Let K be an arbitrary simplicial complex in $\mathbb{R}^n$ and let $\mathbb{H}(\cdot \mid K)$ be its induced Hausdorff distribution. Recall that the value of measures $\delta_\sigma(\{x\})$ vanishes for all σ of dimension greater than one.

$$\mathbb{H}(\mathscr{D} \mid K) = \sum_{i=1}^{N} \mathbb{H}(\{x_i\} \mid K) = \sum_{i=1}^{N} \frac{1}{\#K} \mathbb{1}_{\{x_i \in K^0\}} = \frac{|K^0 \cap \mathscr{D}|}{\#K}.$$

Considering $\mathbb{H}(\mathscr{D} \mid K)$ as a function of K, we see that we can maximize the numerator by taking $K^0 = \mathscr{D}$ and minimize the denominator by not adding any high order simplices, i.e, $K^{k \geq 1} = \varnothing$. For such K, $\mathbb{H}(\cdot \mid K)$ is clearly the empirical distribution of the data. □

Probabilistic Proof. Recall that maximum likelihood estimation can be casted as a minimization of the Kullback-Leibler divergence between the empirical distribution of the data and the model distribution. Since we can represent the empirical data distribution with $\mathbb{H}(\cdot \mid K)$ by choosing K to be $K^0 = \mathscr{D}$ and $K^{k \geq 1} = \varnothing$, the properties of the Kullback-Leibler divergence and the injectivity of the parametrization of $\mathfrak{H}$ imply the desired result. □

Definition 39: Hausdorff Distance

Let (S, d) be a metric space and A, B be subsets of S. The Hausdorff distance is defined as

$$d_{\mathrm{H}}(A, B) = \max \left\{ \sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b) \right\}.$$

Intuitively, the Hausdorff distance is the largest of all the distances from a point in one set to the closest point in the other set.

Theorem 10

Let $\mathscr{K}(\mathbb{R}^n)$ of $2^{\mathbb{R}^n}$ be the set non-empty compact subsets of $\mathbb{R}^n$. $\mathscr{K}(\mathbb{R}^n)$ can be endowed with the structure of a separable metric space.

Proof. For simplicity, we show the proof for the case $n = 1$. Take the Hausdorff metric on $\mathscr{K}(\mathbb{R})$. For a proof that d_{H} is a metric on $\mathscr{K}(\mathbb{R})$, see (Burago et al., 2001).

Recall that $\mathbb{Q}$ is dense in $\mathbb{R}$. Consider the set $\mathscr{F}(\mathbb{Q}) = \{A \subset \mathbb{Q} \mid 0 < |A| < \infty\}$. Note that every such A is finite and thus compact, i.e., $A \in \mathscr{K}(\mathbb{R})$. It is easy to see that $\mathscr{F}(\mathbb{Q})$ is countable. We show that $\mathscr{F}(\mathbb{Q})$ is dense in $\mathscr{K}(\mathbb{R})$.

Take $E \in \mathscr{K}(\mathbb{R})$ and $\epsilon > 0$, arbitrary. The open ball of radius ϵ around E in $\mathscr{K}(\mathbb{R})$ is given by $B_\epsilon(E) = \{W \in \mathscr{K}(\mathbb{R}) \mid d_{\mathrm{H}}(W, E) < \epsilon\}$.

Consider the open cover $E = \bigcup_{e \in E} B_{\frac{\epsilon}{2}}(e)$. Since E is compact, there exists a finite sub-cover $E = \bigcup_{i=1}^{N} B_{\frac{\epsilon}{2}}(e_i)$.

Therefore, for every $e \in E$, there is an e_i such that $d(e, e_i) < \frac{\epsilon}{2}$. Now, for every i choose a $q_i(e_i) \in B_{\frac{\epsilon}{4}}(e_i)$. By the triangle inequality, for all $e \in E$, there is a $q_i(e_i)$ such that $d(e, q_i(e_i)) \leq d(e, e_i) + d(e_i, q_i) = \frac{\epsilon}{2} + \frac{\epsilon}{4} = \frac{3}{4}\epsilon < \epsilon$. Let $A = \{q_i(e_i)\}_{i=1}^{N}$.

Since the largest of distance from a point in E to A is bounded by $\frac{3}{4}\epsilon$ and the largest distance from a point in A to the closest point in E is bounded by $\frac{\epsilon}{2}$, then $d_{\mathrm{H}}(A, E) < \epsilon$. And thus, $\mathscr{F}(\mathbb{Q})$ is a countable and dense set in $\mathscr{K}(\mathbb{R})$.

Therefore $\mathscr{K}(\mathbb{R})$ can be made into a separable metric space. □

Theorem 11

Let $\mathcal{D} = \{x_i\}_{i=1}^N$ be an iid sample of points which follows a distribution $\mathbb{H}(\cdot \mid K_{\text{true}})$, where K_{true} is a simplicial complex in $\mathbb{R}^n$. In the limit as the number of samples N goes to infinity, the maximum likelihood estimator K_{ML} converges (in probability) to K_{true}.

Proof. Since every simplex in $\mathbb{R}^n$ is closed and bounded, by the Heine–Borel theorem, it is compact. Every finite simplicial complex in $\mathbb{R}^n$ is also compact since it is a finite union of simplices. And thus each finite simplicial complex lives in the (by the previous theorem) separable metric space $\mathscr{K}(\mathbb{R}^n)$.

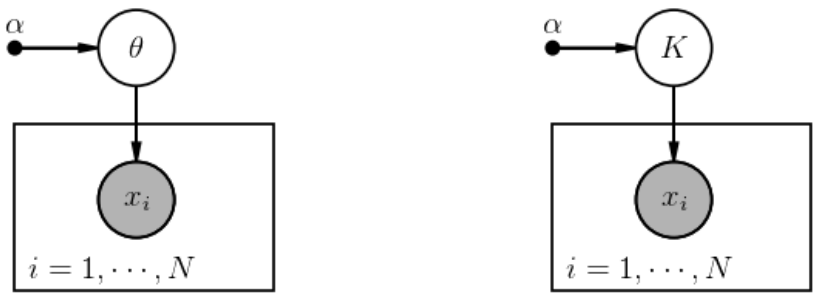
The result follows directly from the consistency of maximum likelihood estimators, the injectivity of the parametrization $\mathfrak{H}$ and the fact that the true distribution can be represented by an element in $\mathfrak{H}$. □

Random Simplicial Complexes

The next step in our characterization of the UMAP method as a posterior optimization problem is to adopt a Bayesian approach and treat the parameters of the distributions in our statistical model, i.e., the simplicial complexes, as random variables. This leads us to the notion of a random simplicial complex.

Definition 40: Random Simplicial Complex

A random simplicial complex on $\mathbb{R}^n$ is an $\mathbb{R}^n$-set-valued non-increasing random variable K such that every realization of K is a simplicial complex in $\mathbb{R}^n$.



(a) Specify a prior on θ parameterized by α. **(b)** Inductive bias as a prior on K.

Fig. 3.6: Graphical models for Bayesian generative models.

The graphical models presented in Figure 3.6 prescribe a factorization on the joint distribution of the variables $\{x_i\}$ and θ (or K). The generating process of $x \mid \theta$ is

given by a Θ-parameterized statistical model. The prior (parameterized by α) on θ has the effect of placing higher probability at regions in Θ which are preferred a priori, for example, because they induce simpler distributions (Goodfellow et al., 2016). We refer to this preference for certain types of solutions as an inductive bias.

The UMAP principles induce a series of constraints on the types of admissible random simplicial complexes K. In other words, our prior on K is an inductive bias which arises from the UMAP construction. Note how the non-increasing nature of K allows us to express most these constraints as *boundary conditions*.

(i) Each 0-simplex should belong to K with membership 1: $K(0)^0 = K(1)^0$, or quantitatively, $d_{\mathrm{H}}(K(0)^0, K(1)^0) = 0$.

(ii) The manifold is locally connected in the sense that every 0-simplex is connected to its nearest neighbor via a 1-simplex with membership 1. Let $K(\omega)^{\mathrm{nn}}$ be the 1-nearest neighbor graph induced by $K(\omega)^0$, considered as a simplicial complex in $\mathbb{R}^n$. We have the constraint $K(0)^{\mathrm{nn}} = K(1)^{\mathrm{nn}}$.

(iii) The size of the neighborhood around every point (in terms of the cardinality of the fuzzy set of edges starting at said point) should be $\tilde{n}$. Recall that $\bar{K}(\psi)$ is equivalent to the membership of ψ in K, considered as a fuzzy set. Thus, we can state the constraint as:

$$\sum_{\sigma' \in K(0)^0} \bar{K}([\sigma, \sigma']) = \tilde{n} \quad \forall \sigma \in K(0)^0.$$

(iv) Finally, we require that if the 0-simplex σ' is further than $\tilde{\sigma}$ to σ, the corresponding membership should be weaker. Therefore, for all $\sigma, \sigma', \tilde{\sigma} \in K(0)^0$, if $d(\sigma, \sigma') \geq d(\sigma, \tilde{\sigma})$, then include the constraint $\bar{K}([\sigma, \sigma']) \leq \bar{K}([\sigma, \tilde{\sigma}])$.

Theorem 12

Let $\mathcal{D} = \{x_i\}_{i=1}^N$ be an iid sample of points which follows a distribution $\mathbb{H}(\cdot \mid K_{\text{true}})$, where K_{true} is a random simplicial complex in $\mathbb{R}^n$.

The random variable associated to the simplicial complex constructed by UMAP can be interpreted as an approximate maximum a posteriori estimate.

Proof. Consider the constrained maximum log-likelihood optimization problem:

$$\begin{aligned} \underset{K}{\text{maximize}} \quad & \log \mathbb{H}(\mathscr{D} \mid K) \\ \text{subject to} \quad & (\mathrm{i}) - (\mathrm{iv}) \end{aligned}$$

Then, its Lagrangian relaxation is given by:

$$K_{\text{UMAP}} = \arg\max_{K} \log \mathbb{H}(\mathscr{D} \mid K) + \mathscr{L}(K),$$

where $\mathscr{L}(K)$ be the Lagrangian associated to the finite set of constraints $(\mathrm{i}) - (\mathrm{iv})$ and can be interpreted as a log-prior on K. □

Quantifying Uncertainty about K_{true}

We have constructed a theory based on simplicial complexes and their random extensions. A central question now is: given a data sample, and an (ML or MAP) estimate $\hat{K}$ for the underlying simplicial complex, how can we quantify our uncertainty about the true location of the simplicial complex? In other words, given a point $x^* \in \mathbb{R}^n$, we would like to measure our confidence on whether x^* belongs or not to K_{true}. The answer to this question arises naturally from our equivalence between random variables and fuzzy sets.

Consider the membership function on $\mathbb{R}^n$ given by $\mu_{\hat{K}}(x) = \mathbb{P}(\hat{K} \ni x)$. Note that this is equivalent to the largest of the memberships of the simplicial complexes in $\hat{K}$ (considered as a fuzzy set) which contain x.

A disadvantage of this definition is that the complexity for finding (if it exists) the simplex of maximum strength which contains a given test point grows in a combinatorial fashion with the size of the dataset. One would desire to have constant time evaluation of such membership. We propose an *adversarially learned membership* function.

Consider a distribution $\mathbb{P}$ of points in $\mathbb{R}^n$ with compact support Ξ. One can learn a *discriminator* $\mu_{\text{Adv}} : \mathbb{R}^n \to [0, 1]$ by training a neural network under a cross-entropy loss in which samples from $\mathbb{P}$ are considered true, and thus should attain a value of μ_{Adv} close to 1, and inputs sampled uniformly on Ξ are considered fake.

This naive approach suffers from the curse of dimensionality as n gets larger. Suppose now we had access to a trained autoencoder (e, d) on Ξ, such that the dimension m of the latent space is much smaller than n. We can alleviate the problem by learning

the discriminator on the distribution induced on the latent space by $\mathbb{P}$ and e. Then we can define $\mu^n_{\text{Adv}}(x) = \mu^m_{\text{Adv}}(e(x))$ for every $x \in \mathbb{R}^n$.

Let us now consider whether this construction would let us recover the underlying simplicial complex. In other words, in the case when K_{true} is almost surely constant, does $\mu_{\hat{K}}(x)$ converge to the indicator function on K_{true}, as the number of datapoints goes to infinity?

In practice, we perform computations with the 1-skeleton of the fuzzy simplicial set. By definition (Spivak, 2009), this can be used to construct the fuzzy simplicial set by recursively defining the membership of the non-degenerate simplices as the minimum of the memberships of its faces. Since the minimum t-norm is an upper bound for every other t-norm, this is an "optimistic" view on the membership of higher-order simplices.

Recall that the construction of the nerve of a cover included a finite subset if and only if the intersection of all the balls centered at the points corresponding to the subset was non empty. In our setting: a simplex has positive membership if and only if the minimum membership of its faces is positive. This is, of course, equivalent to the condition that all the vertices of the simplex are connected to each other with positive strength.

This, would yield a fuzzy construction analog to that of a Vietoris-Rips complex rather than a Čech complex. Fortunately, Vietoris-Rips complexes also possess convenient theoretical properties regarding homotopy-equivalence under certain conditions (Latschev, 2001). Future research can be devoted to formalize this connection and thus provide a theoretical proof about the recovery of the true manifold/simplicial complex based on the fuzzy topological representation created by UMAP. In the experiments section we verify this claim empirically for several manifolds in 2 and 3 dimensions.

4 Simplicial AutoEncoders

> *Experience [...] tells us not which is the truest geometry, but which is the most convenient".*
>
> **— Henri Poincaré**

In this section we propose two extensions of UMAP: first, we introduce simplicial autoencoders as simplicial maps between the high-dimensional data and the low-dimensional embedding created by UMAP; additionally, we extend our parametric autoencoder with a model for the distribution of codes in the latent space, based on a mixture of uniform distributions over ellipsoids. Joining these two ideas we obtain a parametric generative model.

4.1 Simplicial regularization and autoencoders

The richness of the class of neural networks as function approximators given by Theorem 6 comes at the expense of the need to regularize such kind of functions. Recent works by Zhang et al. (2017) and Verma et al. (2018) have considered the regularization induced by convex combinations of latent representations of data samples. In spite of the successful results, the use of the *mixup* regularization scheme is justified heuristically. We now introduce *simplicial regularization*, a scheme that arises naturally from our consideration of simplicial complexes and which directly generalizes *mixup* regularization.

Recall our definition of a simplicial map. We start with two (geometric) simplicial complexes $K \subset \mathbb{R}^n$ and $L \subset \mathbb{R}^m$ and a function $f : K \to L$. We say f is a simplicial map if it is "linear under convex combinations on the simplices". This means we do *not* require f to be a (globally) linear transformation, but only to commute with convex combinations of vertices which form a simplex. Algebraically, if σ is a k-simplex in K, we require that for every convex coefficient vector, λ:

$$f\left(\sum_{j=1}^{k} \lambda_j \sigma_j^0\right) = \sum_{j=1}^{k} \lambda_j f(\sigma_j^0)$$

In practice, enforcing the equality is not trivial. Thus, we can measure "how far a given function between the complexes is from being a simplicial map":

$$\mathscr{L}(f,K,\alpha)=\sum_{\sigma\in K}\mathbb{E}_{\lambda_j\sim\mathrm{Dir}(\dim(\sigma),\alpha)}\; l\left(f\left(\sum_{j=1}^{\dim(\sigma)}\lambda_j\sigma_j^0\right),\sum_{j=1}^{\dim(\sigma)}\lambda_j f(\sigma_j^0)\right),$$

In the previous equation, $\mathrm{Dir}(k,\alpha)$ denotes the symmetric Dirichlet distribution of order k with density $\frac{\Gamma(\alpha k)}{\Gamma(\alpha)^k}\prod_{i=1}^k x_i^{\alpha-1}$. In the case $\alpha=1$, this corresponds to a uniform distribution on the standard $(k-1)$-simplex. When $\alpha\to 0$, the distribution is very peaked at the corners of the simplex, while for $\alpha>1$, the distribution becomes more dense towards the center of the simplex. In other words, α inversely controls the level of interpolation which takes place.

However, recall that simplicial maps are the structure-preserving morphisms in the category of simplicial complexes. Thus, choosing $\mathscr{L}$ as a loss function is equivalent encouraging the map f to preserve the simplicial structure from K to L.

Algorithm 3: Training of a Simplicial AutoEncoder.

Data: Dataset $\mathscr{D}=\{x_i\}_{i=1}^N\subset\mathbb{R}^n$, UMAP parameters θ, maximum iterations $i_{\max}$, learning rate η, Dirichlet interpolation coefficients α_τ.

Result: Parametric autoencoder trained with a simplicial regularization.

$K_{\mathscr{D}},K_{\mathscr{Z}},\mathscr{Z}=\texttt{UMAP}(\mathscr{D},\theta)$
Initialize encoder e and decoder d
for $\tau=1,\dots,i_{\max}$ **do**
    $l_e=\mathscr{L}(e,K_{\mathscr{D}},\alpha_\tau)$
    $l_d=\mathscr{L}(d,K_{\mathscr{Z}},\alpha_\tau)$
    Gradient step on l_e and l_d with learning rate η
end

We highlight the following aspects from Algorithm 3:

- Since K contains its vertices as 0-simplices, the setting $\alpha\to 0$ corresponds to the the standard loss on the training set. Also, the framework applies to high dimensional simplices without any modification.

- It is possible to set up a schedule $\{\alpha_\tau\}$for the Dirichlet interpolation coefficient such that α_τ starts at 0 and increases to 1 with τ. This schedule ensures that the mapping between the vertices of the simplicial complexes is properly trained at an early stage and then regularized during future iterations.

- The training of the encoder and decoder become disentangled and thus, in principle, possible to execute in parallel.

- Contrary to previous works, we do not mix between random points, but rather start with the carefully constructed fuzzy topological representation generated by UMAP and regularize both the encoder and decoder to be approximately simplicial maps on their corresponding domain simplicial complexes.

- Note that the definition of $\mathscr{L}$ only depends on the domain simplicial complex and not on the target. The definition of manifold mixup for (semi-)supervised training by Verma et al. (2018) is clearly particular case of $\mathscr{L}$.

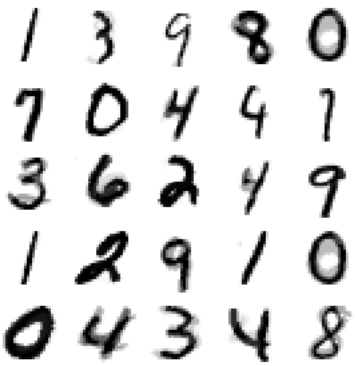


Fig. 4.1: Training inputs sampled from an interpolating Dirichlet distribution with $\alpha = 1$.

Figure 4.1 shows a subset of a training batch for $\alpha = 1$. Since the interpolation is done on simplices belonging to the fuzzy topological representation of the data, the resulting inputs are approximately possible variations in the tangent space of the underlying data manifold at a given point. Intuitively, this regularization can also be interpreted as a graph-induced data augmentation process.

4.2 Mixture of ellipsoids

One of main differences between simplicial autoencoders and other types of generative autoencoders, like VAEs, is the lack of probabilistic model for the codes in the latent space. For this reason, in order to complete our description of a generative model, we need to define a suitable distribution from which we can draw codes.

Since we assume we already have constructed a representation $\{z_i\}_{i=1}^N$, our task is equivalent to constructing a generative model for *low-dimensional data*. However, $\{z_i\}_{i=1}^N$ was optimized in such a way that its fuzzy simplicial set was close (in the set of fuzzy set cross entropy) to the fuzzy topological representation of the data.

Recall that we defined a custom metric around every point to make sure our uniformity assumption about the distribution of the data on the manifold was valid. Thus, it should be no surprise that the distribution of $\{z_i\}_{i=1}^N$ is in practice approximately uniform (at least at a connected component-level).

Therefore, we decide to fit a multivariate mixture of Gaussians to the dataset $\{z_i\}_{i=1}^N$ and then construct a *confidence ellipsoid* based on each component of the mixture. The resulting model is a mixture of uniform distributions on each ellipsoid, with coefficients given by the weights fitted by the mixture of Gaussians.

Definition 41: Confidence Ellipsoid

Consider a p-dimensional Gaussian distribution centered at the origin with covariance matrix $\mathbf{\Sigma}$. The $(1-\alpha)\cdot 100\%$ confidence ellipsoid is the ellipsoid centered at μ and whose j-th half-axis corresponds to the vector $\sqrt{\lambda_j \chi^2_{p,\alpha}}\mathbf{e}_j$, where $(\lambda_j, \mathbf{e}_j)$ is the j-th eigenpair of $\mathbf{\Sigma}$, and $\chi^2_{p,\alpha}$ is the α statistic for a chi-squared distribution with p degrees of freedom.

In principle one could use any type of generative model, like a GAN or a VAE to generate codes. However, the mixture of ellipsoids directly exploits the low-dimensionality and local uniformity on the latent space, and avoids the training and mode-collapse challenges of GANs.

5 Experiments

> *The above proposition is occasionally useful".*
>
> — **Bertrand Russell**
> (comment after the proof that 1+1=2,
> completed in *Principia Mathematica*)

In this section we display our experimental results. We start with an empirical verification of the ability of the fuzzy topological representation induced by the singular functor from UMAP to recover the topological space underlying a random sample. Next, we illustrate the effect of the simplicial regularization introduced in the previous section. We then compare the results of UMAP and our parametric approximation for the MNIST (LeCun et al., 1998), Fashion MNIST (Xiao et al., 2017) and Frey Faces (*Frey Faces Dataset*) datasets.

5.1 Inferring topological spaces from samples

As mentioned earlier, even though we do not provide theoretical guarantees for the fuzzy to converge to the true manifold, we test this conjecture empirically. For this, we have constructed a fuzzy topological representation for several 1 and 2-dimensional manifolds embedded in $\mathbb{R}^2$ and $\mathbb{R}^3$, for several numbers of samples. Figures 5.1 - 5.2 show that in the limit as the number of samples goes to infinity, the true structure of the underlying manifold is recovered in all of the experiments.

5.2 Simplicial regularization on a synthetic task

We would like to assess the effect of the simplicial regularization on the generalization error achieved by models trained with and without it. We start by sampling a dataset 100 points $\{x_i\}_{i=1}^{100}$ from a 20-dimensional isotropic Gaussian distribution. Given a radius $r = 1.4$, we build the Vietoris-Rips complex associated to the dataset. The Vietoris-Rips complex is similar to the Čech complex mentioned earlier, but with a slighltly weaker non-emptiness condition. Then we sample a random trans-

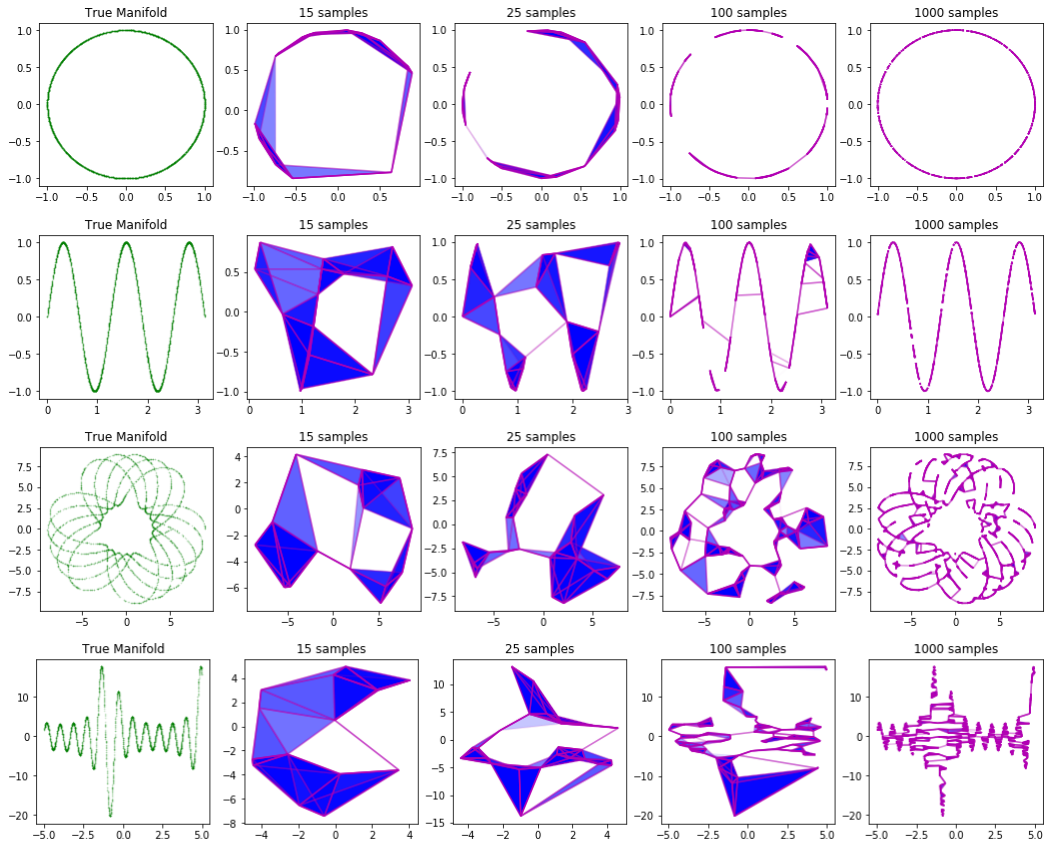


Fig. 5.1: Fuzzy topological representation of 1-dimensional manifolds embedded in $\mathbb{R}^2$.

formation $T : \mathbb{R}^{20} \to \mathbb{R}^2$, and define $z_i = T(x_i)$. We now have a "labelled" dataset $\{(x_i, z_i)\}_{i=1}^{100}$.

Note that, in general, $\{z_i\}$ does not form a *geometric* simplicial complex in $\mathbb{R}^2$, see Figure 5.4. Nevertheless, we would like to find an encoder function which preserves as much of the simplicial structure of the original complex as possible.

Figure 5.5 shows the performance of the regularized and unregularized encoders. The use of the simplicial regularization does not affect the performance on the training set significantly.

However, the behavior on unseen datapoints coming from interpolations inside the simplicies of the 20-dimensional simplicial complex is drastically different. Figure 5.6 shows the distribution of the generalization error for several tasks. In the legend, Δ represents simplicial regularization and the number after the hyphen is the parameter α of the symmetric Dirichlet distribution used to sample the *test* set. Note that difficulty of the generalization task increases with α. We can see that the regularized models have a better performance, which is statistically significant for all configurations, according to the p-values in Table 5.1.

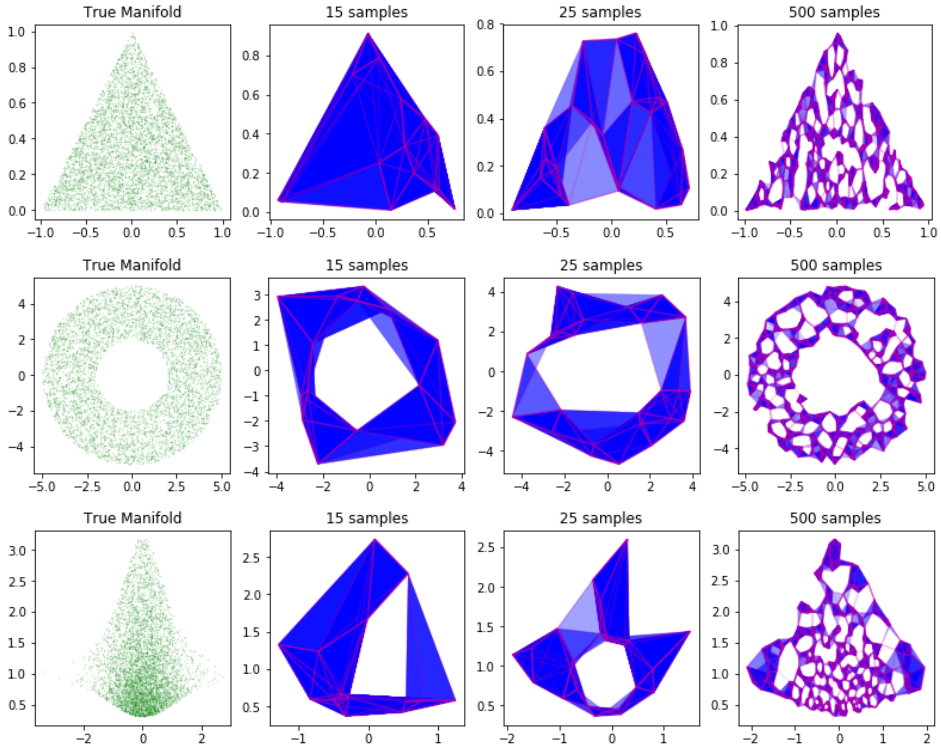


Fig. 5.2: Fuzzy topological representation of 2-dimensional manifolds embedded in $\mathbb{R}^2$.

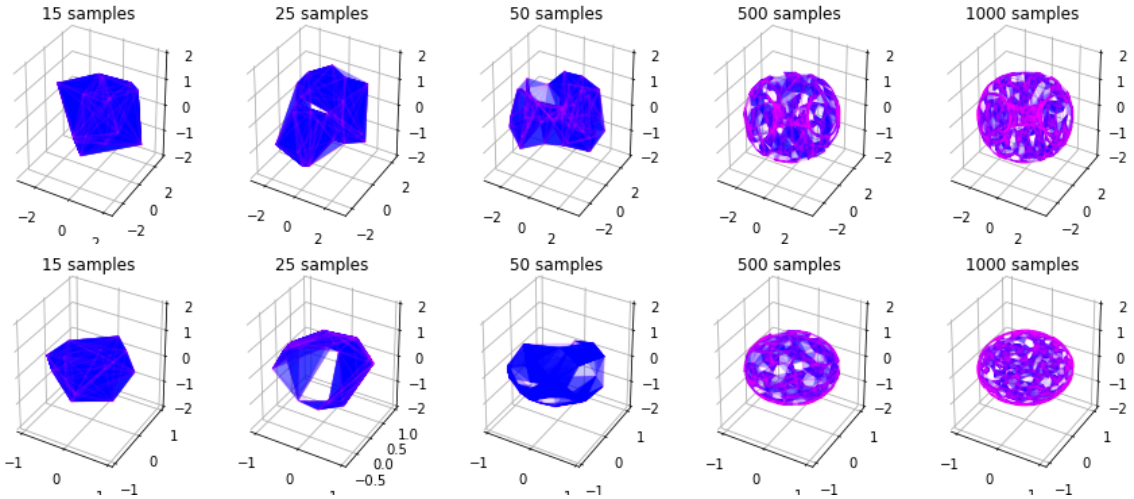


Fig. 5.3: Fuzzy topological representation of 2-dimensional manifolds embedded in $\mathbb{R}^3$. Top row: torus, bottom row: sphere.

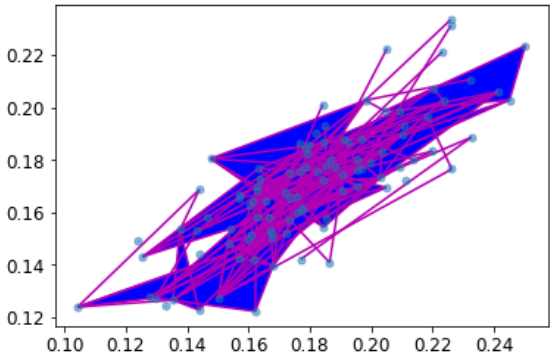
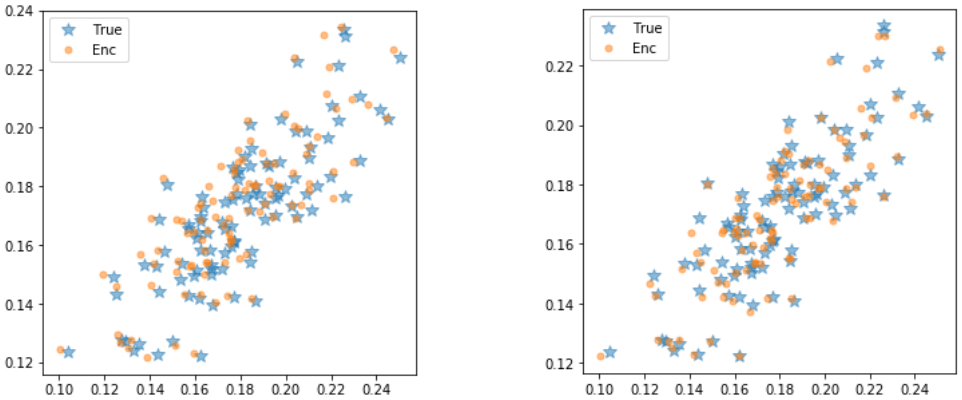


Fig. 5.4: Random 2-dimensional projection of complex originally embedded in $\mathbb{R}^{20}$.



(a) Without simplicial regularization. **(b)** With simplicial regularization.

Fig. 5.5: Encoding performance on the set of vertices for the encoders with and without simplicial regularization.

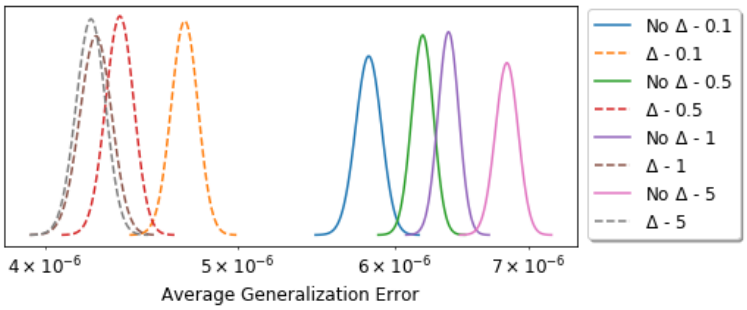


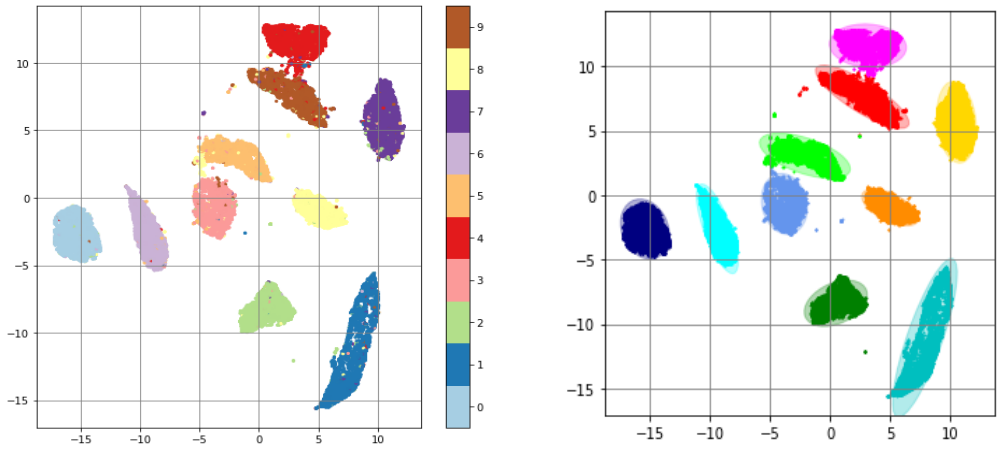
Fig. 5.6: Effect of simplicial regularization on synthetic random simplicial complex.

Dirichlet	Weight Decay	Weight Decay + Simplicial	p-value
0.1	$(5.820 \pm 0.087) \cdot 10^{-6}$	$(4.702 \pm 0.072) \cdot 10^{-6}$	$9.76 \cdot 10^{-51}$
0.5	$(6.196 \pm 0.077) \cdot 10^{-6}$	$(4.362 \pm 0.071) \cdot 10^{-6}$	$1.25 \cdot 10^{-65}$
1	$(6.384 \pm 0.076) \cdot 10^{-6}$	$(4.242 \pm 0.078) \cdot 10^{-6}$	$5.76 \cdot 10^{-69}$
5	$(6.831 \pm 0.090) \cdot 10^{-6}$	$(4.218 \pm 0.072) \cdot 10^{-6}$	$7.90 \cdot 10^{-70}$

Tab. 5.1: Comparison of the generalization error on the synthetic simplicial complex encoding task between a standard encoder and an encoder with a simplicial regularization. 95% confidence intervals on 30 trials. The p-value corresponds to a t-test regarding equality of means between both systems.

5.3 Real datasets

We now conduct an in-depth exploration of the performance of the proposed methods for the MNIST dataset. The corresponding results for the Fashion MNIST and Frey faces datasets can be found in Section 5.5.



(a) Embedding generated by UMAP. Colors correspond to the true class labels.

(b) Mixture of ellipsoids on the latent space. Colors correspond to the mixture component with largest probability.

Fig. 5.7: UMAP embedding and mixture model on MNIST.

Figure 5.7 shows the UMAP-generated embedding and the trained mixture of ellipsoids. The embedding in Figure 5.7a captures the local structure of each class as a connected component, as well as the global structure of the manifold by creating clearly separated clusters. As mentioned before, note how within each connected component, the distribution of the datapoints in the embedding is approximately uniform. The fitted mixture of ellipsoids is displayed in Figure 5.7b.

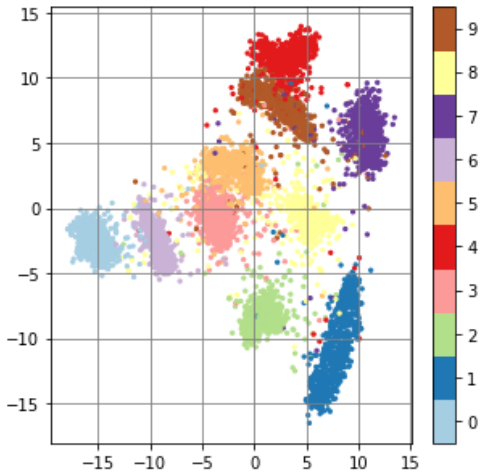


Fig. 5.8: Parametric approximation to the UMAP embedding.

Figure 5.8 shows the embedding results of a parametric embedding constituted by a Simplicial AutoEncoder. The structure of the networks correspond to a neural network with 1 hidden layer containing 100 units. Note how, overall, the separation properties of the representation generated by UMAP are preserved. The fact that the resulting embedding is more noisy than the one in Figure 5.7a is a consequence of the random initialization, continuity and bounded capacity of the encoder.

Figure 5.9 shows the manifold learned by our generative model. In this pictures the strength of each grid point is determined by the density/membership according to the corresponding model. In particular, the adversarial density model was trained to predict the membership of every point in the fuzzy simplicial complex constructed from the embedding.

(a) The strength of each image is proportional to its density under the mixture model. Compare colors with Figure 5.7b.

(b) The strength of each image is proportional to the adversarially learned membership.

Fig. 5.9: Learned MNIST manifold.

In Figure 5.10 we show generated samples for each of the datasets and in Figure 5.11 we show the effect of increasing the dimension of the latent space in the quality of the generated samples.

Figure 5.12 shows linear interpolations between MNIST digits. For this, we create a sequence of evenly spaced points between the codes corresponding to the endpoints, and then project them to the data space using the decoder.

(a) MNIST (b) Fashion MNIST (c) Frey faces

Fig. 5.10: Generated samples for several datasets using a mixture of ellipsoids model on a 2-dimensional latent space.

(a) 2-dim latent space (b) 5-dim latent space (c) 10-dim latent space

Fig. 5.11: Generated MNIST samples for several dimensionalities of the latent space.

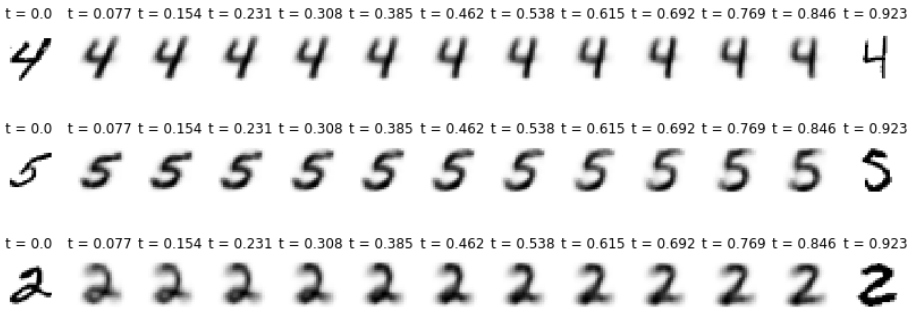


Fig. 5.12: Linear interpolation between samples on MNIST.

Given that we have a representation of the topological space as a fuzzy graph, we can examine an approximation to the geodesic distance between two points by considering shortest paths on the graph between said points. This has two advantages with respect to the linear interpolation: every point in our interpolating path is an actual sample; and whenever points are in different connected components, no interpolating path exists, i.e., interpolation between classes which are not linked is forbidden. For instance, interpolating between a zero and a one is impossible, while there might exist a path between a four and a nine, according to the embedding in Figure 5.7.

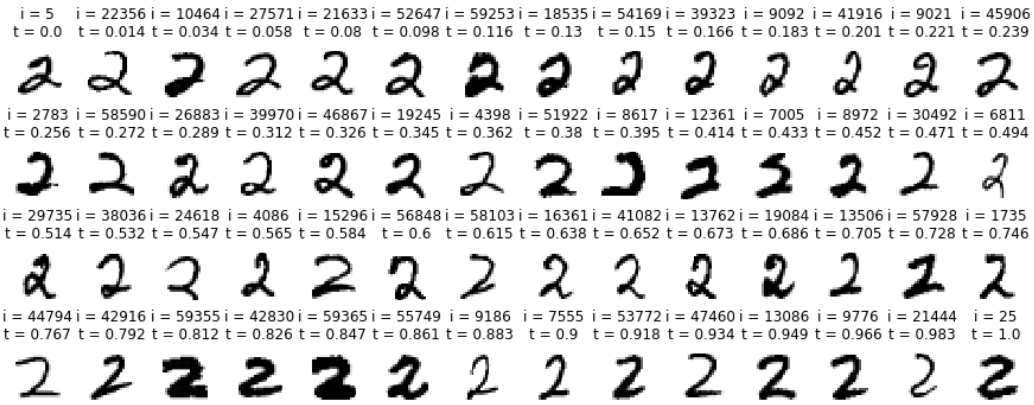
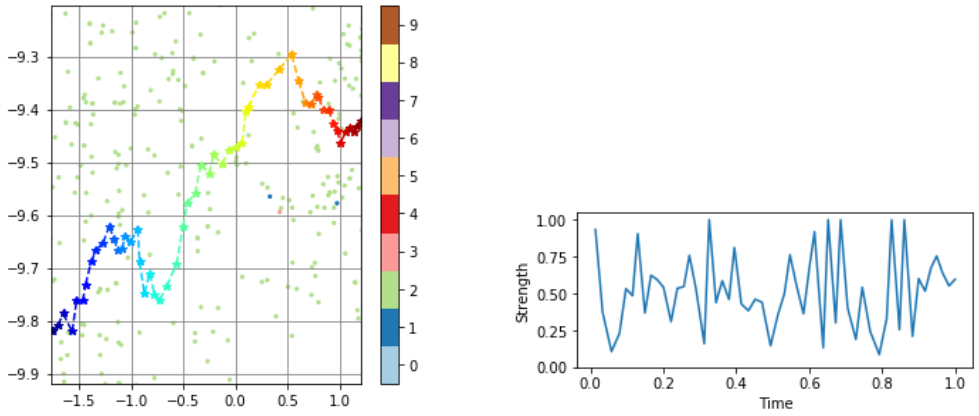


Fig. 5.13: Approximate geodesic interpolation on MNIST.

For the particular example above, we can observe the geodesic path on the graph in Figure 5.14. Starting point ($t = 0$) is blue and endpoint ($t = 1$) is red. The figure on the right shows the membership of every edge corresponding to a transition between samples at every time step.



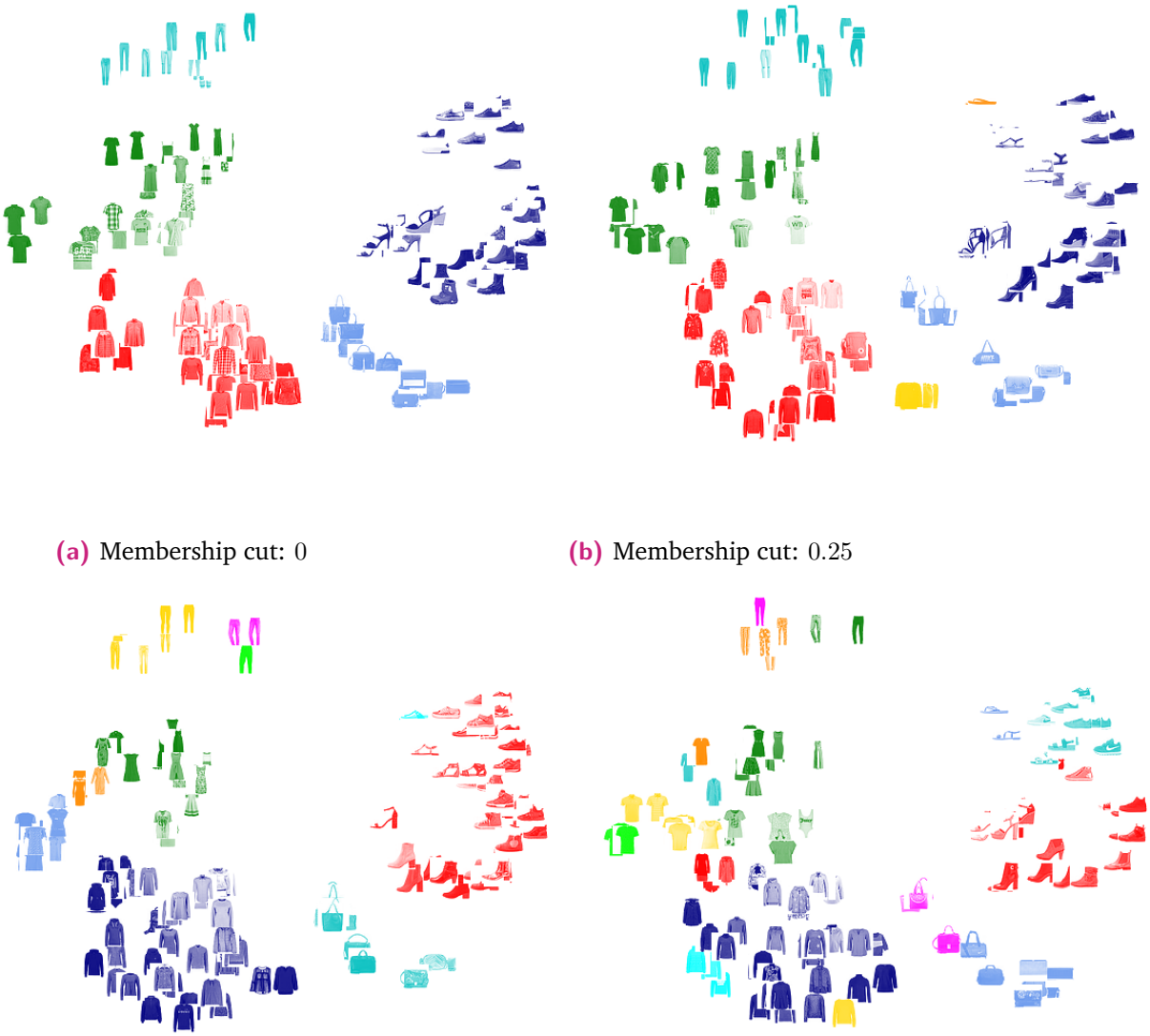
(a) Interpolation path on the latent space. **(b)** Strength of the edge for each transition.

Fig. 5.14: Path of an approximate geodesic interpolation on MNIST.

5.4 Cut-induced compositional representation

Recall how cuts of the membership function played a major role in the construction of our random variables in the theoretical section before. If we consider a sequence of cuts to the 1-skeleton of a fuzzy simplicial complex, performed by deleting edges with membership less than the value of the threshold, we get a hierarchical fragmentation of the original connected components.

The interesting aspect of this sequence of cuts is that, in practice, they induce a sense of *compositionality* in the learned representation. Let us see this in idea in practice for the case of the Fashion MNIST dataset.



(a) Membership cut: 0

(b) Membership cut: 0.25

(c) Membership cut: 0.47

(d) Membership cut: 0.5

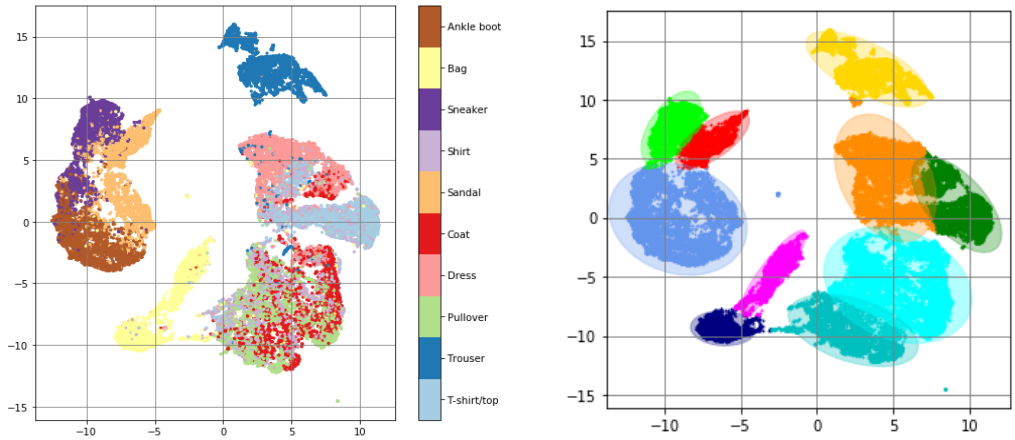
Fig. 5.15: Compositionality arises naturally by cut-induced connected components.

In Figure 5.15a we have the initial fuzzy simplicial complex. We observe 4 major connected components: trousers, footwear, bags, shirts/pullovers and t-shirts/dresses.

When we make the first cut at 0.25, some new small components appear: orange sandals and yellow bags. At 0.47, the big connected component of shirts/dresses breaks into three components: t-shirts, long sleeve dresses, and sleeve-less dresses. Similarly, at 0.5, we get a separation of footwear in 3 connected components: blue sandals, bluegreen sports shoes and red formal footwear. The bags also get separated between those with and without handle.

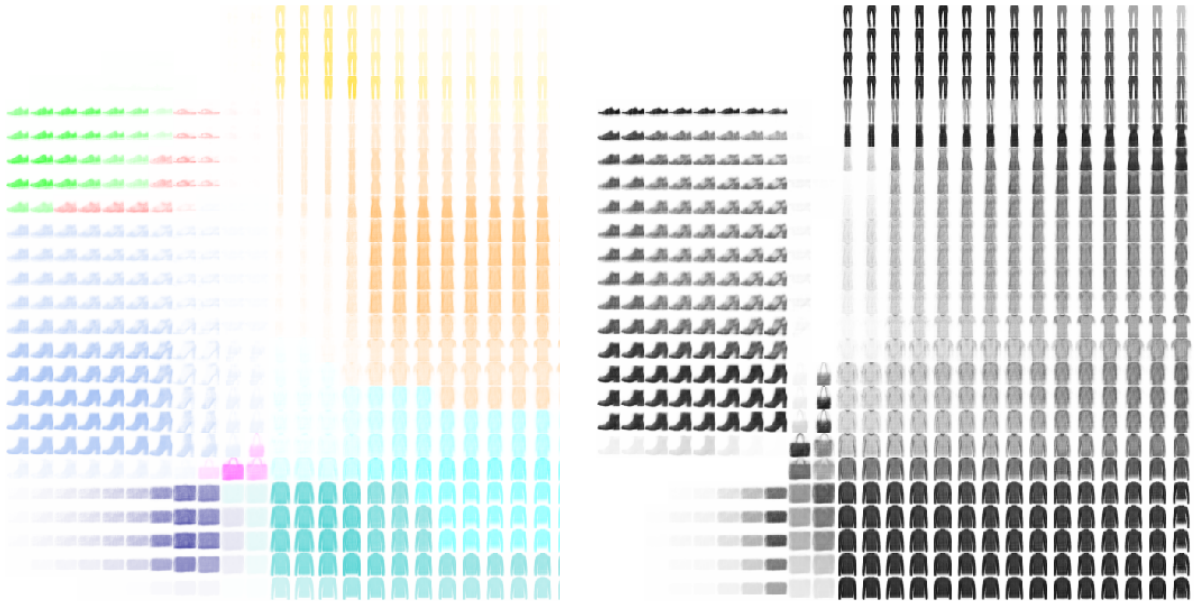
5.5 Complementary results

Fashion MNIST



(a) Embedding generated by UMAP. **(b)** Mixture of ellipsoids on the latent space.

Fig. 5.16: UMAP embedding and mixture model on MNIST.

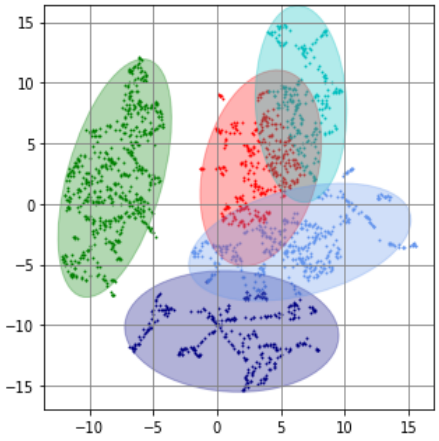


(a) Mixture of ellipsoids, c.f. Figure 5.16b. **(b)** Adversarially learned density.

Fig. 5.17: Learned Fashion MNIST manifold.

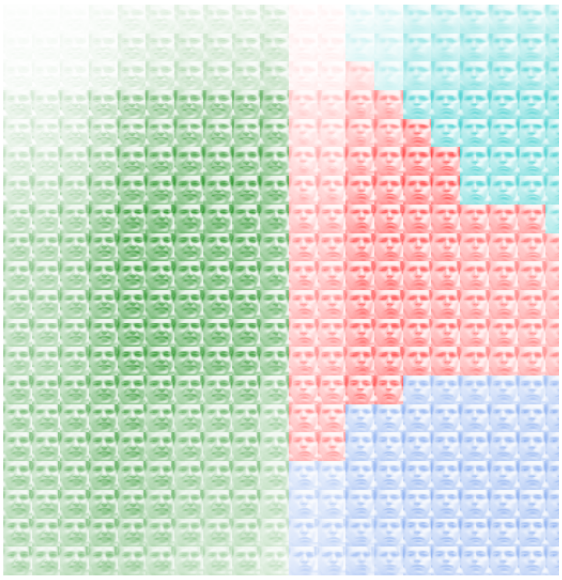
Frey Faces

(a) Embedding generated by UMAP.

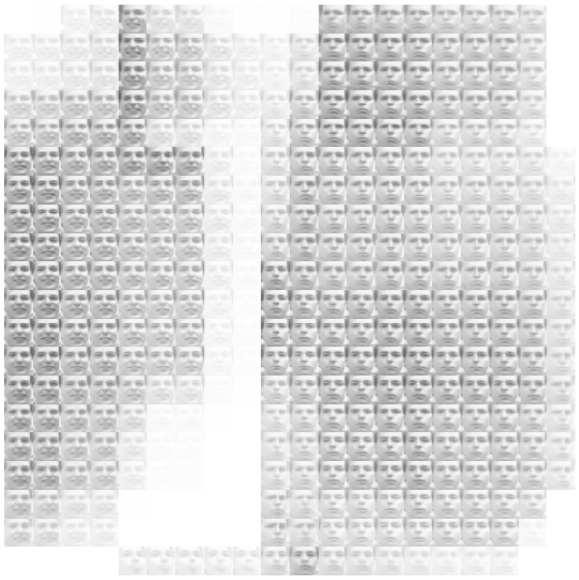


(b) Mixture of ellipsoids on the latent space.

Fig. 5.18: UMAP embedding and mixture model on MNIST.



(a) Mixture of ellipsoids, c.f. Figure 5.18b.



(b) Adversarially learned density.

Fig. 5.19: Learned Frey faces manifold.

Linear Interpolations

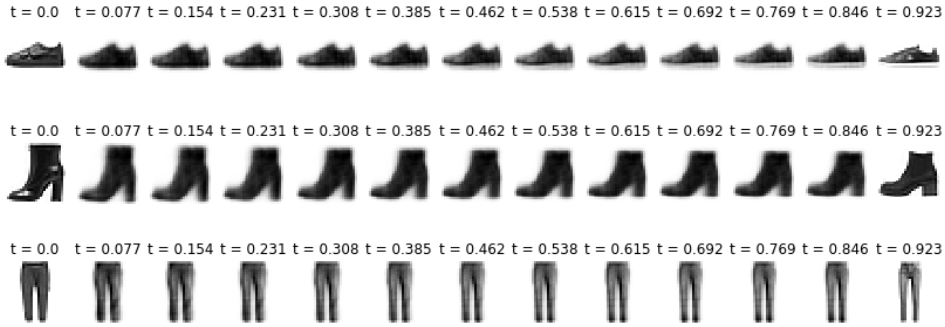


Fig. 5.20: Linear interpolation between samples on Fashion MNIST.

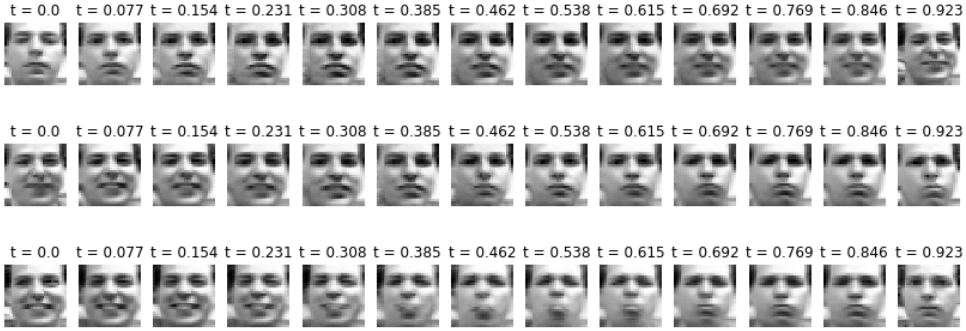


Fig. 5.21: Linear interpolation between samples on the Frey faces dataset.

Conclusions and Future Work

6

> *'The surprising thing about this paper is that a man who could write it–would' ".*
>
> **— John Littlewood**
> (repeating a joke without attribution)

We have presented the construction of an embedding proposed by UMAP as an approximate maximum a posteriori estimator. This is a step in the direction of a theory of unsupervised learning which unifies geometric and probabilistic methods.

We showed how the notion of structure preservation between simplicial complexes gives rise to Simplicial AutoEncoders. By adjoining the construction of a mixture of ellipsoids, we turned the non-parametric embedding of UMAP into a generative model and tested it successfully against several datasets.

A fully Bayesian treatment would require our ability to asses uncertainty by "integrating out" the parameter, i.e., the random simplicial complex. A refinement of our Lagrangian or a more tangible construction of a prior would be a useful step towards a complete Bayesian analysis of UMAP.

The membership of high-dimensional simplices in a fuzzy simplicial set depends on a particular choice of a t-norm. An adequate choice and a formalization of the notion of fuzzy Vietoris-Rips complex could lead to a result which guarantees homotopic equivalence between the fuzzy topological representation constructed by UMAP and the topological space underlying the sample points.

Bibliography

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013). "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8, pp. 1798–1828 (cit. on pp. 1, 2).

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag (cit. on p. 27).

Burago, Dmitri, Yuri Burago, and Sergei Ivanov (2001). *A course in metric geometry*. Vol. 33. American Mathematical Soc. (cit. on p. 39).

Cairns, Stewart S. (1961). "A simple triangulation method for smooth manifolds". In: *Bull. Amer. Math. Soc.* 67.4, pp. 389–390 (cit. on p. 21).

Dudley, Richard M. (1968). "Distances of Probability Measures and Random Variables". In: *Ann. Math. Statist.* 39.5, pp. 1563–1572 (cit. on p. 24).

Ghrist, Robert W. (2014). *Elementary applied topology*. 1st ed. CreateSpace (cit. on p. 19).

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, et al. (2014). "Generative Adversarial Networks". In: *ArXiv e-prints*. arXiv: 1406.2661 [stat.ML] (cit. on p. 2).

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press (cit. on pp. 27, 41).

Hatcher, Allen (2001). *Algebraic topology*. Cambridge University Press (cit. on p. 17).

Hornik, Kurt (1991). "Approximation capabilities of multilayer feedforward networks". In: *Neural networks* 4.2, pp. 251–257 (cit. on p. 29).

Kingma, D. P and M. Welling (2013). "Auto-Encoding Variational Bayes". In: *ArXiv e-prints*. arXiv: 1312.6114 [stat.ML] (cit. on p. 2).

Latschev, J. (2001). "Vietoris-Rips complexes of metric spaces near a closed Riemannian manifold". In: *Archiv der Mathematik* 77.6, pp. 522–528 (cit. on p. 43).

LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324 (cit. on p. 49).

Lin, H. W., M. Tegmark, and D. Rolnick (2017). "Why Does Deep and Cheap Learning Work So Well?" In: *Journal of Statistical Physics* 168, pp. 1223–1247. arXiv: 1608.08225 [cond-mat.dis-nn] (cit. on p. 2).

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing data using t-SNE". In: *Journal of machine learning research* 9.Nov, pp. 2579–2605 (cit. on p. 3).

McInnes, Leland and John Healy (2018). “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”. In: *ArXiv e-prints*. arXiv: 1802.03426 [stat.ML] (cit. on pp. 3, 32).

Niyogi, Partha, Stephen Smale, and Shmuel Weinberger (2008). “Finding the homology of submanifolds with high confidence from random samples”. In: *Discrete & Computational Geometry* 39.1-3, pp. 419–441 (cit. on p. 21).

Rifai, Salah, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio (2011a). “Contractive auto-encoders: Explicit invariance during feature extraction”. In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. Omnipress, pp. 833–840 (cit. on p. 3).

Rifai, Salah, Yann N Dauphin, Pascal Vincent, Yoshua Bengio, and Xavier Muller (2011b). “The Manifold Tangent Classifier”. In: *Advances in Neural Information Processing Systems 24*. Ed. by J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger. Curran Associates, Inc., pp. 2294–2302 (cit. on p. 2).

Spivak, David I (2009). “Metric realization of fuzzy simplicial sets”. In: (cit. on p. 43).

Tipping, Michael E and Christopher M Bishop (1999). “Probabilistic principal component analysis”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3, pp. 611–622 (cit. on p. 2).

Verma, V., A. Lamb, C. Beckham, et al. (2018). “Manifold Mixup: Encouraging Meaningful On-Manifold Interpolation as a Regularizer”. In: *ArXiv e-prints*. arXiv: 1806.05236 [stat.ML] (cit. on pp. 45, 47).

Vincent, Pascal, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol (2010). “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion”. In: *Journal of machine learning research* 11.Dec, pp. 3371–3408 (cit. on p. 3).

Whitney, Hassler (1944). “The Self-Intersections of a Smooth n-Manifold in 2n-Space”. In: *The Annals of Mathematics* 45.2, p. 220 (cit. on p. 20).

Zhang, H., M. Cisse, Y. N. Dauphin, and D. Lopez-Paz (2017). “mixup: Beyond Empirical Risk Minimization”. In: *ArXiv e-prints*. arXiv: 1710.09412 [cs.LG] (cit. on pp. 3, 45).

List of Figures