**UMAP, theoretical background***(adapted from https://umap-learn.readthedocs.io/en/latest/how\_umap\_works.html)*

UMAP is short for *Uniform Manifold Approximation and Projection* and is a dimension reduction algorithm. The algorithm builds on three assumptions: the data under consideration is uniformly distributed on a Riemannian manifold[[1]](#footnote-1), this manifold is locally connected and its metric is locally constant. The aim is to reconstruct (or: model) this manifold and derive a lower dimensional projection which is as closely equivalent as possible.

As a first step, an open ball of fixed radius is positioned around each datapoint. The union of these balls should be an open cover of the manifold. In general, this will not be the case, since in reality, data will not look uniformly distributed. Thus, some parts will not be covered while other parts may be overly covered (by too many balls). Since we assumed the data to be uniformly distributed on the manifold, this means that the notion of distance is not constant on the manifold. Instead of fixing the radius, a unit ball will therefore be defined as stretching to the k-th nearest neighbor, where k is a hyper parameter (set by the user). The algorithm is refined by working in a fuzzy topology, meaning that being in a set is no longer binary (yes or no), but this is described as a fuzzy value between zero and one. To guarantee local connectedness (no point being completely isolated), the fuzzy confidence decays only after the first nearest neighbor.

Next, the constructed open cover of the manifold can be translated into a simplicial complex. To understand what this is, first consider a k-simplex. A k-simplex is a convex hull of k+1 independent points. For example, a 0-simplex is a point, a 1-simplex is a line and a 2-simplex is a triangle. A simplicial complex is the result of gluing k-simplices together along their faces (their corners and sides). Now, for each set in the cover, a 0-simplex is created; when the intersection between two sets is non-empty, a 1-simplex between the two is created; when the triple-intersection between three sets is non-empty, a 2-simplex between them is created; and so on. If one only allows 0- and 1-simplices, one obtains a graph with weighted edges, where each weight represents the probability of the edge existing.

The lower dimensional representation will be constructed on a Euclidian manifold, in our case a two-dimensional one. Constructing it is considered an optimalisation problem: minimize the cross entropy between the high dimensional manifold with a fuzzy topological structure () and the low dimensional Euclidian manifold (). Cross entropy is defined as follows

Where is the set of edges and is the weight of edge . The first part represents an attractive force that is large when is large and which is minimized by being as large as possible (meaning the points are close to each other); the second part represents a repellent force that is large when is small and which is minimized by being as small as possible (the points being far away from each other). Solving this problem, gives a low-dimensional embedding of the manifold.

The distance as measured on the high dimensional manifold can be any type of distance. In our analysis, we used among others Mahalanobis distance, which is a multi-dimensional generalization of measuring the amount of standard deviations between a point and a distribution mean.

1. A manifold is a topological space wherein each point has a neighborhood that is homeomorphic to an open subset of Euclidian space. A Riemannian manifold is a real, smooth manifold with a positive-definite inner product on the tangent space at each point. [↑](#footnote-ref-1)