AUTORIZACIÓN TUTOR Y DIRECTORES

Ezequiel López Rubio, Catedrático de Universidad de Ciencia de la Computación e Inteligencia Artificial, perteneciente al Departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga, informa: Que ha tutorizado la Tesis Doctoral titulada "Self-organized maps", realizada por Antonio Díaz Ramos.

Ezequiel López Rubio, Catedrático de Universidad de Ciencia de la Computación e Inteligencia Artificial, perteneciente al Departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga, y Esteban José Palomo Ferrer, Profesor Titular de Universidad de Lenguajes y Sistemas Informáticos, perteneciente al Departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga, informan: Que han dirigido la Tesis Doctoral titulada "Self-organized maps", realizada por Antonio Díaz Ramos.

Finalizada la investigación que ha llevado a la conclusión de la citada Tesis Doctoral, autorizan su presentación por considerar que reúne los requisitos formales y científicos legalmente establecidos para la obtención del título de Doctor en Tecnologías Informáticas. En particular, las publicaciones que avalan dicha tesis no han sido utilizadas en tesis anteriores.

Y para que así conste y surta los efectos oportunos expiden y firman el presente informe en Málaga a 29 de enero de 2020.

Fdo.:Ezequiel López Rubio

Fdo.: Esteban José Palomo Ferrer

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña ANTONIO DÍAZ RAMOS
Estudiante del programa de doctorado TECNOLOGÍAS INFORMÁTICAS de la Universidad de Málaga, autor/a de la tesis, presentada para la obtención del título de doctor por la Universidad de Málaga, titulada: SELF-ORGANIZED MAPS

Realizada bajo la tutorización de EZEQUIEL LÓPEZ RUBIO y dirección de EZEQUIEL LÓPEZ RUBIO Y ESTEBAN JOSÉ PALOMO FERRER

DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a 27 de ENERO de 2020

Fdo.: ANTONIO DÍAZ RAMOS

Self-Organized Maps

Antonio Díaz Ramos

Departamento de Álgebra, Geometría y Topología, Facultad de Ciencias, Universidad de Málaga, Campus de Teatinos, s/n, 29071 Málaga, Spain.

Email address: adiazramos@uma.es

Contents

CHAPTER 1

Resumen de la Tesis Doctoral en español.

Los mapas auto-organizados o redes de Kohonen (*SOM* por sus siglas en inglés, *self-organizing map*) fueron introducidos por el profesor finlandés Teuvo Kalevi Kohonen en los artículos [**33, 34**]. Un mapa auto-organizado es una herramienta que analiza datos en muchas dimensiones con relaciones complejas entre ellos y los reduce o representa en, usualmente, una, dos o tres dimensiones. La propiedad más importante de una SOM es que preserva las propiedades topológicas de los datos, es decir, que datos próximos aparecen próximos en la reducción o representación.

La literatura relacionada con los mapas auto-organizados y sus aplicaciones es muy diversa y numerosa [**30, 46, 52**]. Las neuronas en un mapa auto-organizado clásico están distribuidas en una topología (o malla) bidimensional cuadrada o hexagonal y las distancias entre ellas son distancias euclídeas. Una de las disciplinas de investigación en SOM consiste en la modificación y generalización del algoritmo SOM. Esta Tesis Doctoral se centra en esta línea de investigación.

En concreto, los objetivos desarrollados han sido el estudio de topologías bidimensionales alternativas, el estudio comparativo de topologías de una, dos y tres dimensiones y el estudio de variaciones para la distancia y movimientos euclídeos. Estos objetivos se han abordado mediante el método científico a través de las siguientes fases: aprehensión de resultados conocidos, planteamiento de hipótesis, propuesta de métodos alternativos, confrontación de métodos mediante experimentación, aceptación y rechazo de las diversas hipótesis mediante métodos estadísticos.

Se han obtenido los siguientes resultados:

(1) Estudio de topologías bidimensionales alternativas. El trabajo [**38**] demuestra la importancia de topología alternativas basadas en áreas ajenas como las teselaciones.
(2) Estudio comparativo de topologías en una, dos y tres dimensiones. El trabajo [**18**] revela la influencia de la dimensión en el funcionamiento de una SOM a escala local y global.
(3) Estudio de alternativas al movimiento euclídeo. En el trabajo [**19**] se propone y presenta la alternativa FRSOM al algoritmo SOM clásico. En FRSOM, las neuronas esquivan barreras predefinidas en su movimiento.

Las conclusiones más relevantes que emanan de esta Tesis Doctoral son las siguientes:

(1) La calidad del clustering y de la preservación topológica de una SOM puede ser mejorada mediante el uso de topologías alternativas y también evitando regiones prohibidas que no contribuyen significativamente al Error Cuadrático Medio (ECM).

(2) Como era de esperar, la dimensión de la SOM que obtiene mejores resultados es la propia dimensión intrínseca de los datos. Además, en general, valores más bajos para la dimensión de la SOM producen mejores resultados en términos del ECM, y valores altos ocasionan mejor aprendizaje de la estructura de los datos.

CHAPTER 2

Publicaciones que avalan esta Tesis Doctoral.

Esta Tesis Doctoral se presenta por compendio de publicaciones y está respaldada por los trabajos enumerados a continuación. Copias de estos trabajos pueden leerse al final de este documento.

(1) [**38**]: López-Rubio E., Díaz Ramos A., *Grid topologies for the self-organizing map*, Neural Networks, Volume 56, August 2014, Pages 35–48.

Indicios de calidad: JCR (2014), categoría: Computer Science: Artificial Intelligence, posición **18/123**, indice de impacto 2.708, primer cuartil (**Q1**).

Copia del artículo aquí 1.

(2) [**18**]: Antonio Díaz Ramos, Esteban J. Palomo, Ezequiel López-Rubio, *The role of the lattice dimensionality in the self-organizing map*, 2018, Neural Network World 28(1), pp. 57–85.

Indicios de calidad: JCR (2018), categoría: Computer Science: Artificial Intelligence, posición **110/134**, indice de impacto 0.957, cuarto cuartil (**Q4**).

Copia del artículo aquí 2.

(3) [**19**]: Antonio Díaz Ramos, Ezequiel López Rubio, Esteban J. Palomo, *The Forbidden Region Self-Organizing Map neural network*, IEEE Transactions on Neural Networks and Learning Systems 31(1), Jan. 2020, pp. 201– 211.

Indicios de calidad: JCR (2018), categoría: Computer Science: Artificial Intelligence, posición **2/134**, indice de impacto 11.683, primer cuartil (**Q1**).

Copia del artículo aquí 3.

CHAPTER 3

Introduction

Artificial Intelligence (AI) might perhaps be characterized as the only field to attempt to build machines that will function autonomously in complex, changing environments [**55**, p.18]. Nevertheless, AI may be defined in different ways, depending on the point of view. For instance, in [**55**, 1.1], eight different definitions are gathered from different sources and classified into four categories: Thinking Humanly, Acting Humanly, Thinking Rationally and Acting Rationally. In that book, as well as in this work, we focus on the last perspective, i.e., Acting Rationally.

This viewpoint concentrates on studying and building a rational agent, which is an agent that acts to achieve the best expected outcome to a given problem or situation. The agent's behaviour is described by the agent function or algorithm, which maps the percepts from the environment to an action. Hence, we may consider this algorithm as a black box that is fed an input or stimulus and returns an output in response.

$$\text{INPUT}\Longrightarrow\boxed{\text{ALGORITHM}}\Longrightarrow\text{OUTPUT} \tag{1}$$

The Acting Rationally approach to AI looks for an algorithm that optimizes the output with respect to some performance measure. Although AI is still far away from emulating humans in the broad sense, in specific tasks some algorithms have already outperformed humans. This is so in the recent famous cases of board games Go [**66**] and chess [**67**], but also in more specialized fields as traffic sign visual pattern recognition and neuronal membranes image segmentation [**57**].

Machine Learning (ML) is an approach to AI in which part of the algorithm (the black box) is adjusted by a learning process or training phase that involves several learning steps or iterations. The algorithm contains explicit rules for how to learn on each step, but not for the final state of the machine after it has learnt. Once the learning phase of the Machine Learning algorithm has finished, the device works in recall phase: new inputs are presented and the algorithm produces outputs without further internal adjustment. For most practical applications, the original input variables are typically preprocessed to transform them into some new space of variables where, it is hoped, the problem will be easier to solve [**8**, p.2].

This approach to AI contrasts, for instance, with Expert Systems, a strategy in which the rules of the algorithm are fixed. Expert Systems were initially developed in the 1970s [**55**, 1.3.5], had their first commercial success in the early 1980s [**55**, 1.3.6] and are still employed [**55**, 16.7].

Within Machine Learning, there exist several methods like the following ones.

- Regression Analysis [**55**, 18.6], [**8**, 3]: curve fitting for different families of basis functions by finding a minimum of the loss function by gradient-descent methods.
- Bayesian networks [**55**, 13,14,15,16], [**8**, 8.1]: dependencies among random variables are represented in a graph, over which inference and learning processes are implemented.
- Decision Trees [**55**, 18.3], [**8**, 14.4]: a tree is created that models the value/answer to a set of fixed variables/questions, and then this tree is employed to predict the value of new variables.
- Genetic Algorithms [**55**, 4.1.4]: function optimization via a population of candidates-to-optimum values that change via crossover and mutations.
- Support Vector Machines [**55**, 18.9], [**8**, 7]: finding a separating hyperplane/hypersurface by mapping data to higher dimensional space (Kernel trick).
- Artificial Neuronal Networks (ANNs) [**55**, 18.7], [**8**, 5]: ANNs emulate biological neural networks and consist of a set of artificial neurons that are connected via links with different "weights". During the iterative learning process, these "weights" are modified as to best fit the network to the input data.

Artificial Neuronal Networks sprang back in the mid-1980s with the reinvention by several research groups of the back-propagation algorithm [**55**, 1.3.7], nowadays one of the most widespread learning algorithms for ANNs. From the point of view of the learning process, there is a dichotomy among ANNs:

(1) Supervised learning: a *target function* is provided a priori, and the quality of the output of the algorithm is measured by evaluating this function on the output.

(2) Unsupervised learning: no explicit mechanism is provided to assess the quality of the output. In this case, the learning process works by minimizing an ad hoc *cost function* that replaces the missing target function.

A self-organized map (SOM) is a kind of artificial neural network with unsupervised learning. They were introduced by Finnish Professor Teuvo Kalevi Kohonen in [**33**] and [**34**], see also [**35**]. There are thousands of published papers reporting applications of SOMs to many fields as satellite images, medical imaging, speech analysis, word recognition, robot navigation, full-text analysis, and traveling-salesman problem, to name just a few. A compendium of these works may be found in [**30**] for the period 1981-1997, [**46**] for the period 1998-2001 and [**52**] for the period 2002-2005.

A particular discipline within SOM research is modifications and generalizations of the SOM algorithm. Again, there exist hundreds of developed versions of alternative SOMs, see [**30**], [**46**] and [**52**] for a digest and [**35**, Chapter 5] for detailed descriptions of some of these variants. The present work is devoted to the study of some of the aspects of SOMs. We focus on variations or alternatives to some of the mathematical notions involved in SOMs. See Section 3.1 below for a precise description of the problem we tackle in this Ph.D. Thesis.

Although ANNs are originally inspired by biological concepts, the intrinsic study of the algorithms without further reference to the biological background, as done in this work, is a current field of research. For instance, in the reference [**55**, 1.3.7, p.25], we find,

> Modern neural network research has bifurcated into two fields, one concerned with creating effective network architectures and algorithms and understanding their mathematical properties, the other concerned with careful modeling of the empirical properties of actual neurons and ensembles of neurons.

and, in [**8**, p.226],

> From the perspective of practical applications of pattern recognition, however, biological realism would impose entirely unnecessary constraints.

Recall that, so far, we have mentioned the following concepts, which we represent below in decreasing order of scope, from top to bottom.

Artificial Intelligence (IA)
↓
Rational Agent
↓
Machine Learning
↓
Artificial Neuronal Network (ANN)
↓
Self-Organized Map (SOM).

In order to introduce SOMs, consider the example application of Figure 1 below. It consists of the outcome of a SOM applied to a point cloud of 126 points in dimension 39. This data corresponds to 39 quality of life indicators of 126 different countries from World Bank statistics from 1992. Adjacent countries in the SOM have similar values for their indicators. Colours have been added to stress the spatial adjacency. The example is taken from Helsinki University of Technology.

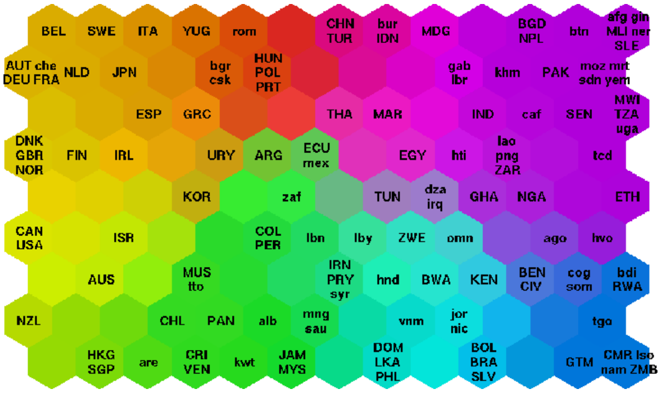


FIGURE 1. Quality of life values of different countries organized by a SOM. Source: HUT

Already in this example, we can discern a few elements involved in a SOM:

(a) Input space: this is the space where the data cloud lies. In this case, the 126 data points belong to the Euclidean space $\mathbb{R}^{39}$.
(b) Internal structure: the neuron or prototypes of the SOM are arranged according to some internal structure. In this case, there are 13×9 neurons located at the vertices of a hexagonal grid.
(c) Topological preservation: the main feature of SOMs is that they map data points that are close in the input space (neighbours) to neurons that are close in the SOM grid (neighbours).

After we have introduced some basic facts about SOMs, in the next section we explain in detail the concrete problem we face in this Ph.D. Thesis.

3.1. Ph.D. Thesis problem.

In this section we introduce the problem studied in this work, its state of the art and the proposed solutions. The description of the problem revolves around the items a and b listed above. In fact, many of the aforementioned variations of SOMs modify these items. However, this myriad of alternatives has paid little attention to some concrete aspects which, we think, are relevant, specially from the point of view of the mathematical insight. Hence, the problem of this Ph.D. Thesis is the following:

> Mathematically analyze the classical and alternative topologies for SOMs as well as pose new possibilities.

Here, by topology we mean both the topology of the input space a and the topology or internal structure of the SOM b. Although topological preservation c is not intrinsic part of the problem, we use it to asses SOMs, see Subsection 5.1.

The concrete aspects of SOMs we think deserve further consideration are the following:

(1) Internal topology of the SOM b: The standard choice for the internal topology of a SOM is the square topology, with few exceptions which correspond to the hexagonal topology [**3**], [**60**]. This observation raises two important questions:
- whether square and hexagonal topologies are special;
- whether there are other suitable topologies.

There are many self-organizing models that consider a dynamic topology instead of a fixed one, for instance the growing self-organizing map (GSOM) [**2**] and others, see [**42**] and [**41**]. We further discuss these variants in Subsection 5.2.2. Nevertheless, we do not consider them when answering the two questions above because of its dynamic nature.

Our reply to these questions is contained in the work [**38**], and there we find that alternative topologies arising from tessellations can improve the efficiency of SOMs. A copy of this work may be found at the end of this document 1. A summary of this paper may be found in Section 6.1. Further notions on tessellations may be found in Section 5.2.1 and in Appendix C.

(2) Dimension of the internal topology of the SOM b: In the example above (see Figure 1) the internal topology is two-dimensional. The topology of SOM in applications is usually one, two or three-dimensional, but the two-dimensional version is by far the most used, probably because of its

easy representation and visualization. Three-dimensional topologies are employed in applications for 3D data visualization [**47**], 3D data processing [**31**] or other tasks without 3D data involved [**4**]. The use of one-dimensional topologies is frequently linked to data that is known to lie in a curve [**29**], although in some cases this constraint does not hold [**39**]. Although formal results about one-dimensional SOMs are relatively abundant [**21, 20, 12**], there are no works about which lattice dimensionality should be chosen for a particular application. In the work [**18**], we answer this question, investigating the interplay between the data inner dimension a and the SOM internal dimension b. We do this from both the theoretical and empirical approaches. A copy of this work may be found at the end of this document 2. A summary of this paper may be found in Section 6.2.

(3) Topology of input space a: For some data sets and applications, it is known beforehand that some regions of the input space cannot contain any samples. Moreover, there are applications of SOM to this sort of data, for instance, to geographic data or geospatial analysis: In [**37**], a SOM-based method for invariant shape identification [**40**] was successfully applied to hydrographic zones partitioning of Shandong province in China. In [**50**], a SOM was used for patterning and predicting aquatic insect species richness in running waters. A classification of land usage in mountain grassland bovine areas in the Massif Central, France, was proposed in [**49**]. In [**59**], a SOM was applied for visualizing demographic trajectories based on census observations. Despite this wide variety of SOM-based methods applied to spatial data analysis, the possibility of including forbidden regions in the SOM has not been further explored. To fill this gap, in the work [**19**], we present FRSOM, a variant of SOM whose prototype vectors avoid predefined forbidden regions in the input space. A copy of this work may be found at the end of this document 3. A summary of this paper may be found in Section 6.3.

The solutions proposed in the three points above are partial answers to what we think are deficiencies in the current development of the SOM algorithm and its variants. Our answers, on the one hand, improve the overall efficiency of SOMs and, on the other hand, enlarge the field of its applications, either by providing advise on which lattice dimension to use or by giving a new tool for data with forbidden regions. This way, we contribute to the vast world of SOM ANNs in the direction of better efficiency and applicability.

3.2. Outline.

In this section we delineate the contents of this document: It is divided into 7 chapters, 3 appendices, and copies of the papers that endorse this Ph.D. Thesis. Chapters 1 and 2 are written in Spanish and contain, respectively, a summary of the work and a list of the publications that support this Ph.D. Thesis by compendium of publications.

In Chapter 4 we start introducing ANNs. Then we describe SOMs as a special type of ANN and we discuss the SOM algorithm as well as the usual cost function (MSE). We give a mathematical definition of SOM in Definition 4.3.1. There, the input space a and the internal arrangement b discussed above will take a precise form.

In Chapter 5, we discuss the biological foundations of the notion of topological preservation stated in point c above. We then comment on how topological preservation is employed as a measure of the quality of the output of SOM. We also compare this concept with the mathematical notion of continuity. In addition, we describe some known variations of the SOM algorithm as motivation and preamble for the next chapter.

Chapter 6 contains a summary of the three works developed by the Ph.D. student during the last 5 years; see [**38**], [**18**] and [**19**]. At the beginning of this chapter, there is a thorough description of the elements of the SOM involved in each work. For each of the three papers, we have also included subsections for experiments, results, and discussion. The conclusions for each of the three works, as well as the overall conclusions of this report, are all gathered in Chapter 7.

These three works have been developed under the scientific method, which seems to be the current approach in AI as stated in [**55**, 1.3.8, p.25],

> In terms of methodology, AI has finally come firmly under the scientific method. To be accepted, hypotheses must be subjected to rigorous empirical experiments, and the results must be analyzed statistically for their importance. It is now possible to replicate experiments by using shared repositories of test data and code.

Finally, the foundations of Artificial Intelligence involving mathematics are logic, computation, and probability [**55**, 1.2.2]. For the description of the probability and statistical methods supporting the implementation of the scientific method, we have included an Appendix on Statistical Methods B. Besides these three fields, namely: logic, computation, and probability, we have employed mathematical notions arising from the area of Geometry and Topology. For that reason, we have also incorporated two appendices with basic notions on Topology A and Tessellations C.

CHAPTER 4

Self-organized maps

In the next two sections, we give a general background on artificial neural networks. Then, in Sections 4.3, 4.4, 4.5 and 4.6, we describe in detail self-organized maps (SOMs) as follows:

(1) In Section 4.3, we describe the neurons' layout of a SOM and the connections between these neurons. Moreover, we provide a formal definition of SOM (Definition 4.3.1) and the usual way of constructing a SOM from a grid (Example 4.3.3).
(2) In Section 4.4, we interpret SOMs in the general setup of ANNs and derive the basic rules that govern SOMs.
(3) In Section 4.5, we give a full description of the SOM algorithm for the learning phase.
(4) Finally, in Section 4.6, we give a detailed account of SOM in recall phase, introducing also the Mean Square Error (MSE, Equation 15) as cost function for unsupervised learning in SOMs.

4.1. Artificial Neuronal Network (ANN)

An Artificial Neuronal Network (ANN) is a machine learning computer system that is inspired in biology. More precisely, neurons, their connections and their learning method are the main source of inspiration for ANNs.

In a biological neuron, the synapse is the structure that allows the incoming signal from the axon of another neuron to pass to the dendrite of the neuron. The signals coming from different neurons are combined and, if this aggregate exceeds a threshold, the neuron generates a signal which is sent through its axon, see Figure 2.

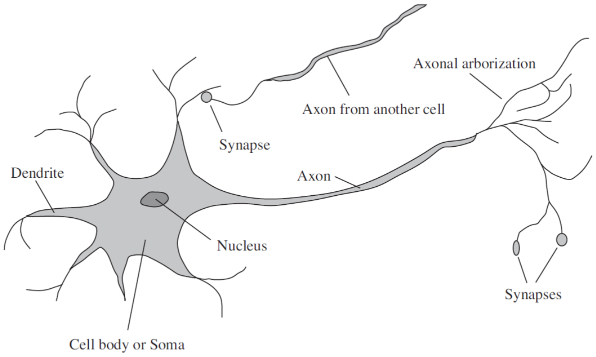


Figure 2. A biological neuron. Source: [**55**, Figure 1.2].

This is usually simplified as follows to give a model of an artificial neuron: The inputs from incoming connections of other neurons correspond to real numbers $x_1, x_2, \ldots, x_n$, and each connection has an associated weight $w_i \in \mathbb{R}$. This data is mixed in a linear combination and compared to the threshold $\theta \in \mathbb{R}$ to produce an output. Hence, the output of the neuron may be written as $\sigma(\sum_{i=1}^{n} w_i x_i, \theta)$, where σ is the so-called activation function, see Figure 3.

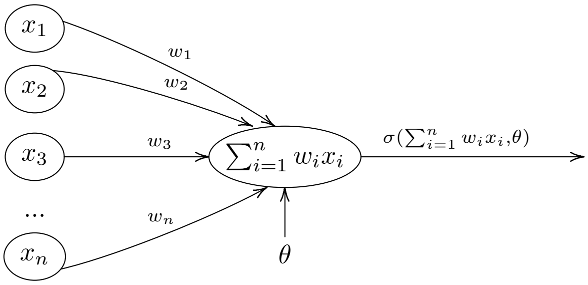


Figure 3. An artificial neuron.

The activation function can be, for instance, the sign function, or a derivable substitute as the sigmoid function, the hyperbolic tangent or the inverse of the tangent function, see Figure 4.

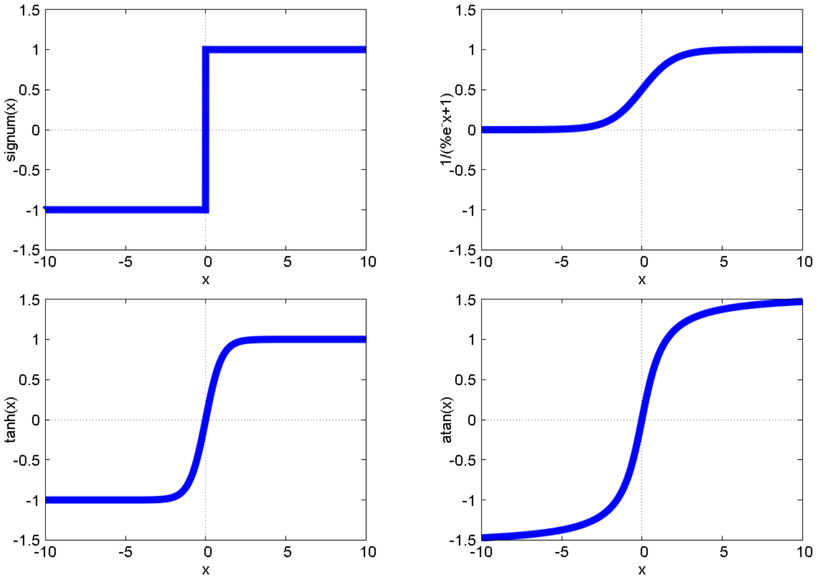


Figure 4. Sign function, sigmoid function, hyperbolic tangent and inverse of tangent function.

The brain contains billions of neurons and each one of them is connected up to $10,000$ other neurons. In the artificial setup, the artificial neurons are laid into layers, with neurons from one layer connected to neurons in the next later. The first layer (leftmost) is the input layer, and the last layer (rightmost) is the output layer. The rest of the layers are termed hidden layers. Each layer may have a different number of neurons. See Figure 5.

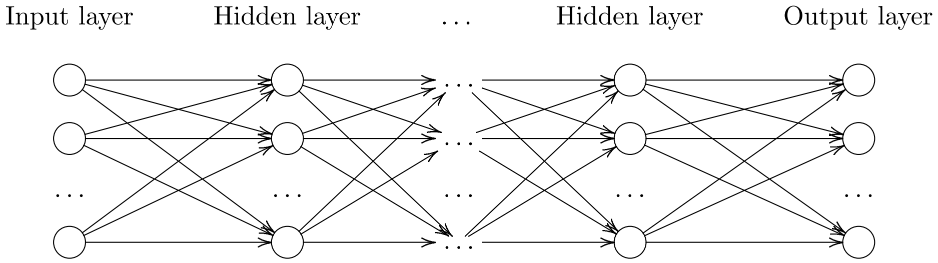


Figure 5. An artificial neural network.

Here we have described feedforward neural networks. In the more general setting of recurrent neural networks, cycles are admitted in the directed graph of connections among neurons. Learning in biological neurons is based on Hebb's principle [**26**, p.62]:

A NEUROPHYSIOLOGICAL POSTULATE:

> Let us assume then that the persistence or repetition of a reverberatory activity (or "trace") tends to induce lasting cellular changes that add to its stability. The assumption can be precisely stated as follows: When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.

This principle translates in turn into concrete algorithms for both supervised and unsupervised learning. In the former case, the most spread learning algorithm is backpropagation: the weights w_{ij} are adjusted in the direction of the gradient of the target function in an error-correction fashion. It requires the activation functions to be derivable, and it proceeds from the last layer backwards, until the weights in the first layer are adjusted. The multilayer perceptron is an example of an artificial neural network using this learning technique.

In supervised learning, Hebb's principle is often implemented as competitive learning or winner-take-all strategy: for each input, neurons compete to be the one generating the output. Self-organizing maps are an example of artificial neural network employing this learning method, and they will be described in Sections 4.3, 4.4, 4.5 and 4.6.

4.2. ANN as a function

After the details provided in the preceding section 4.1, the black-box algorithm scheme (1) for ANNs can be refined as follows.

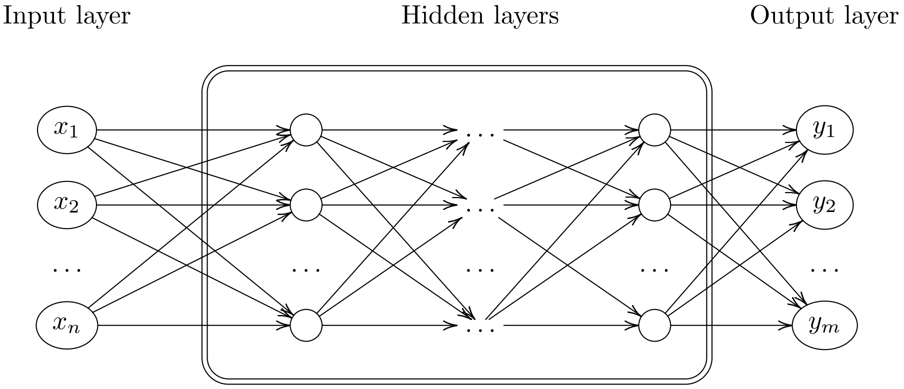


Figure 6. An artificial neural network.

Here, we have written n for the number of input variables x's and m for the number of output variables y's. Thus, the ANN can be seen as a function

$$\mathbf{y} = f(\mathbf{x}, \mathbf{w}),$$

where $\mathbf{y} = (y_1, \dots, y_m)$ and $\mathbf{x} = (x_1, \dots, x_n)$. Moreover, $\mathbf{w} = \{w_{l,i,j}\}$ is the set of weights between neurons (i's and j's) of consecutive layers (l's) and f is a composition of functions as the one described in Figure 3. Slightly unfolding this definition, we have

$$\begin{aligned} y_1 &= f_1(x_1, \dots, x_n, \mathbf{w}), \\ y_2 &= f_2(x_1, \dots, x_n, \mathbf{w}), \\ \dots \\ y_m &= f_m(x_1, \dots, x_n, \mathbf{w}), \end{aligned}$$

where again each function f_i is a composition of function as the one described in Figure 3.

The variables x's and y's will be, in most cases, of real, integral or binary type. We denote the *input space* accordingly as $\mathbb{R}^n$, $\mathbb{Z}^n$ or $\{0,1\}^n$ and write $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \in \mathbb{Z}^n$ or $\mathbf{x} \in \{0,1\}^n$. Similarly, we use term *output space* and write $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{Z}^m$ or $\mathbf{y} \in \{0,1\}^m$. This is from the theoretical point of view of course. Once the ANN has a hardware or software implementation, the real, integral and binary types are interpreted as floating point, integer or binary types respectively. Unless stated otherwise, we assume throughout this work that the input space is the Euclidean space $\mathbb{R}^n$ and that the output space is the Euclidean space $\mathbb{R}^m$.

In the learning phase, the weights $\mathbf{w}$ will be modified following some procedure. As commented in the introduction, there are two main learning types: supervised and unsupervised.

4.2.1. Supervised learning in ANNs. For supervised learning and as commented above, a common method is back-propagation: Assume $\{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1,\dots,e}$ is a set consisting of e training pairs. This means that, for each $1 \leq i \leq e$, if the algorithm is presented the input $\mathbf{x}^i = (x_1^i, \dots, x_n^i)$, we expect the output of the algorithm to be as close as possible to $\mathbf{y}^i = (y_1^i, \dots, y_m^i)$. Mathematically, this may be formulated as searching for $\mathbf{w}$ such that the sum of the squared errors over the training samples is as small as possible:

$$(2) \qquad E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{e} ||\mathbf{y}^i - f(\mathbf{x}^i, \mathbf{w})||^2 = \frac{1}{2}\sum_{i=1}^{e}\sum_{j=1,\ldots,m} |y_j^i - f_j(x_1^i, \ldots, x_n^i, \mathbf{w})|^2.$$

If this expression is differentiable in the variables $\mathbf{w}$, then a gradient-descent method may be used to look for a (local) minimum of the error $E(\mathbf{w})$. This is known as back-propagation learning phase for the artificial neural network (ANN). See [**53**, Section 3.9] or [**10**, Section 8.4] for more details.

4.2.2. Unsupervised learning in ANNs. Besides supervised learning (Subsection 4.2.1), the other learning paradigm for ANNs is unsupervised learning. This term is more or less synonymous with vector quantization and clustering [**27**, 1.4], and below we describe some clustering algorithms.

So let $\mathbf{x}^1, \ldots, \mathbf{x}^M$ with $\mathbf{x}^i \in \mathbb{R}^n$ be M samples that we want to cluster without further a priori knowledge about them. This lack of further information or target function is which tells apart unsupervised learning from supervised learning. The clustering process consists of assigning a label to each of the samples according to some similarity measure [**27**, 4], see Figure 7 for an example.

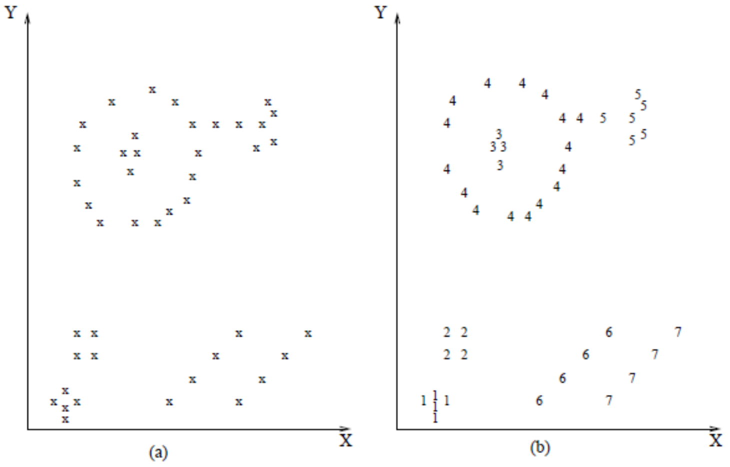


Figure 7. Example of data clustering. Source: [**27**, Figure 1].

Clustering algorithms may be split into hierarchical and partitional algorithms; see [**27**, Figure 7] for more details on this. In addition, among partitional algorithms, we have square error type algorithms and other types as graph theoretic, mixture resolving and mode seeking. In square error algorithms, the missing target function is replaced by the square error criterion [**27**, 5.2.1]. This means that the algorithm search for an appropriate number m of clusters and for the centroid of each cluster $\mathbf{w}^1, \ldots, \mathbf{w}^m$, in such a way that the square error function below is close to a local minimum,

$$(3) \qquad SE = \sum_{k=1}^{m}\sum_{i \in C_k} ||\mathbf{x}^i - \mathbf{w}^k||^2.$$

Here, $C_k \subseteq \{1,\dots,M\}$ consists of the samples belonging to the k-th cluster C_k. Hence, $C_k \cap C_{k'} = \emptyset$ if $k \neq k'$ and $\cup_{k=1}^{m} C_k = \{1,\dots,M\}$. Some well-known square error algorithms are k-means algorithm, winner-takes-all strategy and self-organizing map (SOM). Below we describe the two first approaches, and, in the next section, we start describing SOM, which is the main object of study of this work.

- K-means algorithm: choose and fix a number m of clusters and initialize the centroids $\mathbf{w}^1,\dots,\mathbf{w}^m$ to m random samples. Then repeat the following steps until the clusters stabilize or the decrease in the function (3) is below a threshold,
  (1) Update the C_k's: Assign each sample $\mathbf{x}^i$ to the cluster C_k whose centroid $\mathbf{w}^k$ is the closest to $\mathbf{x}^i$.
  (2) Update the $\mathbf{w}^k$'s: Define $\mathbf{w}^k$ as the centroid (barycenter) of the samples in the cluster C_k.
- Winner-takes-all strategy: this is the particular case of the SOM algorithm in which the metric D among neurons (see Definition 4.3.1) is given by $D(i,i) = 0$ and $D(i,j) = \infty$ if $i \neq j$. After choosing the number m of clusters and randomly initializing the centroids $\mathbf{w}^1,\dots,\mathbf{w}^m$, the step below is repeated for each one of the samples $\mathbf{x}^1,\dots,\mathbf{x}^M$,
  (1) For the sample $\mathbf{x}^i$, update the centroid $\mathbf{w}^{k_0}$ which is closest to $\mathbf{x}^i$. Move $\mathbf{w}^{k_0}$ towards $\mathbf{x}^i$ an amount between 0 and 1. This amount decreases with the index i.

  The final clusters are given as before, i.e., assigning each sample $\mathbf{x}^i$ to the cluster C_k whose centroid $\mathbf{w}^k$ is the closest to $\mathbf{x}^i$.

4.3. Structure of the SOM

Self-organizing maps (SOMs) can be understood as a type of ANN with unsupervised learning. They were introduced by Kohonen in 1982 [**34**]. They have no hidden layers and, as main structural difference, they have connections between neurons of the output layer. Depending on the layout of these connections, an internal dimension d can be associated with a SOM. For instance, one-dimensional SOMs ($d = 1$) can be depicted as follows:

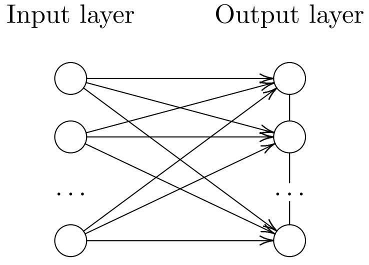


FIGURE 8. SOM of dimension one.

Here, a connection exists between each pair of consecutive neurons in the output layer. A two-dimensional SOM ($d = 2$) is represented as follows:

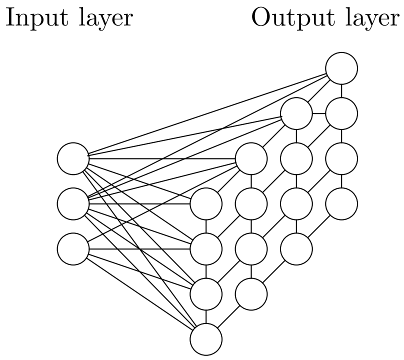


FIGURE 9. SOM of dimension two.

In this case, each neuron in the output layer is connected to four neighbourhoods (except if the neuron is in the boundary of the grid). Moreover, in the one-dimensional case, the top most neuron can be connected to the bottom-most neuron giving rise to a “circle” configuration. In the two-dimensional case, the left-most neurons can be connected to the right-most neurons, producing a “cylinder” shape. In addition, if the top most neurons are linked to the bottom-most neurons a “toroidal” arrangement is obtained. Similar networks can be defined in any dimension d, and there are exist many variations for the layout of the connections, some of which will be explored in Subsection 5.2.1 and Section 6.1.

Besides, to define the SOM, each pair of neurons j and k in the output layer must have associated a non-negative number $D(j,k)$ that indicates the distance between these neurons. Therefore, the distance of the SOM can be seen as a function

$$D\colon \{1,\dots,m\}\times\{1,\dots,m\}\to\mathbb{R}_+=\{x\in\mathbb{R}|x\geq 0\}. \tag{4}$$

This function D is expected to be a *metric* on the set $\{1,\dots,m\}$, i.e., it must satisfy, the following conditions, see also Definition A.0.9,

- $D(j,k)=0\Leftrightarrow j=k$ for all j, k.
- $D(j,k)=D(k,j)$ for all j, k.
- $D(j,l)\leq D(j,k)+D(k,l)$ for all j, k, l.

We will stick to the following notion of SOM.

DEFINITION 4.3.1. A SOM $\mathcal{S}$ is a quintuple $\mathcal{S}=(n,m,d,\mathcal{G},D)$, where $n\in\mathbb{Z}_+$ is the number of input neurons, $m\in\mathbb{Z}_+$ is the number output neurons, $d\in\mathbb{Z}_+$ is the internal dimension of the SOM, $\mathcal{G}$ is an undirected graph on m vertices and D is a metric on the set $\{1,\dots,m\}$.

REMARK 4.3.2. Note that the topology induced by the metric $\mathcal{D}$ on the finite set $\{1,\dots,m\}$ is the discrete topology, see Example A.0.2.

This is not the usual way of describing SOMs. We have introduced the abstract Definition 4.3.1 as it isolates the different elements of the SOM, and hence permits to discuss them separately. In the next example, we explain the standard way of obtaining a SOM via a *grid.* Unless stated otherwise, all SOMs throughout this work are constructed as in this example.

Example 4.3.3. Fix the number n as the number of input neurons and m as the number of output neurons. Then we use a *grid* with m points in the Euclidean space $\mathbb{R}^d$ for some dimension $d > 0$ to define the graph $\mathcal{G}$ and the distance function D. This *grid* consist of m points $\mathbf{p}^1, \dots, \mathbf{p}^m$ in $\mathbb{R}^d$, the point $\mathbf{p}^j$ corresponding to output neuron j, and a collection of straight segments from $\mathbf{p}^j$ to $\mathbf{p}^k$ for different neurons j and k. We assume that these segments either do not intersect or intersect in a single point of the chosen ones.

Then we define the corresponding graph $\mathcal{G}$ on m vertices with an edge for each selected segment and set the distance function D via

$$D(j,k) = ||\mathbf{p}^j - \mathbf{p}^k||. \tag{5}$$

Thus, we set the distance to be equal to the Euclidean norm in $\mathbb{R}^d$ between the corresponding points. Then we have that D is a metric as expected and that $(n, m, d, \mathcal{G}, D)$ is a SOM.

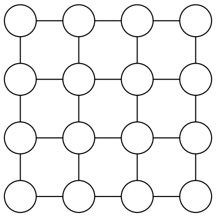
Remark 4.3.4. The grid is the subspace $\mathcal{S}_d$ of $\mathbb{R}^d$ given as the union of the chosen points and segments with the topology inherited from $\mathbb{R}^d$.

Remark 4.3.5. To consider the right dimension d, one should impose the condition that the points $\{\mathbf{p}^1, \dots, \mathbf{p}^m\}$ span $\mathbb{R}^d$ as an affine space, i.e., that the set of vectors $\{\mathbf{p}^2 - \mathbf{p}^1, \dots, \mathbf{p}^m - \mathbf{p}^1\}$ contains d linearly independent vectors.

Example 4.3.6. Consider the grid 16 points of $\mathbb{R}^2$ given by

$$\{0,1,2,3\} \times \{0,1,2,3\} = \{(0,0),(1,0),(2,0),\dots,(2,3),(3,3)\},$$

and the segments/connections between the points (i,j) and $(i+1,j)$ and the points (i,j) and $(i,j+1)$. This gives a SOM with the following underlying undirected graph that contains 16 vertices or output neurons,

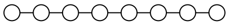


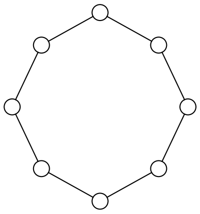
The distance between neuron j with $\mathbf{p}^j = (a,b)$ and neuron k with $\mathbf{p}^k = (a',b')$ for this SOM is defined by

$$D(j,k) = ||(a,b) - (a',b')|| = \sqrt{(a-a')^2 + (b-b')^2}.$$

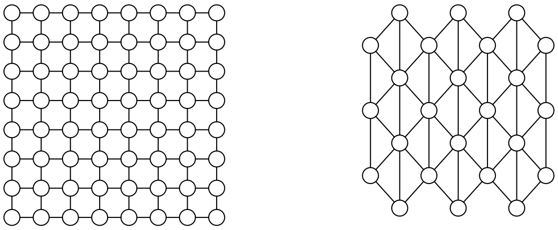
For instance, if the neurons j and k are connected in the graph, we have $D(j,k) = 1$. Note that this is the same configuration as that of Figure 9.

Example 4.3.7. Typical 1-dimensional grids ($d = 1$) are the "linear" and "circle" grids. Below we show an 8×1 linear grid and an 8×1 circle grid. .



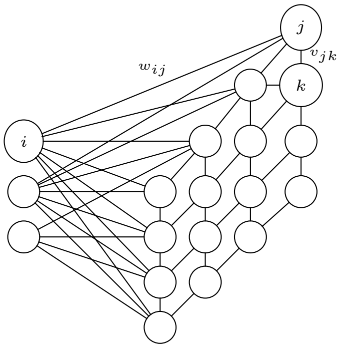


Typical 2-dimensional grids ($d = 2$) are the "square" grid as that of Example 4.3.6 and the "hexagonal" grid. Below we present an 8×8 square grid and a 7×7 hexagonal grid.



4.4. SOM as ANN

Next, we explain how the general behaviour of ANNs gives rise to the SOM algorithm. So consider a SOM $\mathcal{S} = (n, m, d, \mathcal{G}, D)$, let $x_1, \ldots, x_n$ the values at the input neurons and let $y_1, \ldots, y_m$ the values at the output neurons. Denote by w_{ij} the weight of the connection between input neuron i and output neuron j, and denote by v_{jk} the strength of the link between output neurons j and k. (Note that v_{jk} is different from the distance $D(j,k)$.) The following picture represents these variables in the dimension two case.



Mimicking biological hypotheses (which are detailed in [**34**]), the strength v_{jk} is positive for neurons k close to j (small $D(j,k)$) and negative for neurons k far from j (large $D(j,k)$). This assumption and other simplifications permit to deduce that, in the equilibrium state, i.e., for output values $y_1, \ldots, y_m$ satisfying

$$y_j = \sigma(\sum_{i=1}^{n} w_{ij} x_i + \sum_{k=1}^{m} v_{jk} y_k - \theta), \tag{6}$$

these values reach a maximum value y_{j_0} at some output neuron j_0, and that y_k decreases with the distance $D(j_0, k)$. See [**53**, Section 4.1] for a deduction of this fact in the one-dimensional case. In the formula above, σ is an activation function and the threshold θ is common to all output neurons.

The neuron j_0 where the maximum value y_{j_0} is attached may be computed, in principle, solving Equation 6. This is a difficult task and Kohonen suggests instead employing only the external signal, i.e., determining j_0 as follows,

$$\sum_{i=1}^{n} w_{ij_0} x_i = \max_j \sum_{i=1}^{n} w_{ij} x_i.$$

This means that we seek to maximize the dot product of $\mathbf{w}^j$ with $\mathbf{x}$, where $\mathbf{x} = (x_1, \dots, x_n)$ and we have introduced the *weight vector* or *prototype* $\mathbf{w}^j$ associated to the output neuron j,

$$\mathbf{w}^j = (w_{1j}, \dots, w_{nj}). \tag{7}$$

Consider the following equation involving the dot product,

$$||\mathbf{w}^j - \mathbf{x}||^2 = (\mathbf{w}^j - \mathbf{x}) \cdot (\mathbf{w}^j - \mathbf{x}) = ||\mathbf{w}^j||^2 + ||\mathbf{x}||^2 - 2\ \mathbf{w}^j \cdot \mathbf{x}.$$

From this equation, and under the hypothesis that $||\mathbf{w}^j||$ does not depend on j, it turns out that maximize $\mathbf{w}^j \cdot \mathbf{x}$ is equivalent to minimize $||\mathbf{w}^j - \mathbf{x}||^2$, and hence also to minimize $||\mathbf{w}^j - \mathbf{x}||$. In fact, this is the rule that, even without the assumption above on the norms $||\mathbf{w}^j||$'s, governs the choice of j_0 in the SOM algorithm:

$$||\mathbf{w}^{j_0} - \mathbf{x}|| = \min_j ||\mathbf{w}^j - \mathbf{x}||. \tag{8}$$

Thus, Equation 8 provides unsupervised competitive learning: The winning neuron is the neuron j_0 whose prototype $\mathbf{w}^{j_0}$ is closest to the input $\mathbf{x}$ in the input space $\mathbb{R}^n$. These considerations allow us to ignore the strengths v_{jk} among output neurons and focus only on $\mathbf{x}$, $\mathbf{y}$ and the prototypes $\mathbf{w}^j$. Starting from this point, we explain in the next section the SOM algorithm.

4.5. SOM algorithm

Consider a SOM $\mathcal{S} = (n, m, d, \mathcal{G}, D)$ and denote by $\mathbf{x} \in \mathbb{R}^n$ an input vector, by $\mathbf{y}$ the output vector and by $\mathbf{w}^j$ the prototypes of the output neurons for $j = 1, \dots, m$. Notice that, from Equation 7, $\mathbf{w}^j = (w_{1j}, \dots, w_{nj})$ and hence $\mathbf{w}^j \in \mathbb{R}^n$. Thus,

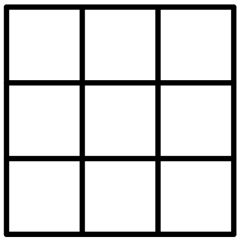
the prototype of an output neuron belongs to the input space.

This implies that we can associate to the abstract graph $\mathcal{G}$ of $\mathcal{S}$ two spaces:

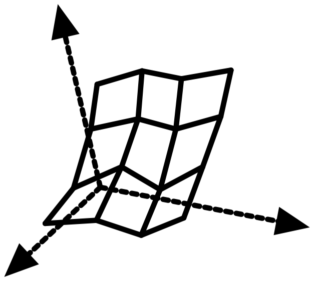
- The subspace $\mathcal{S}_d$ of $\mathbb{R}^d$ if the SOM is constructed from a grid, see Example 4.3.3 and Remark 4.3.4.
- The subspace $\mathcal{S}_n$ of $\mathbb{R}^n$ consisting of the union of the points $\{\mathbf{w}^j\}_{j=1,\dots,m}$ and the segments among them in $\mathbb{R}^n$ corresponding to edges in the graph $\mathcal{G}$.

Note that the space $\mathcal{S}_d$ is defined in such a way that its edges do not intersect. On the other hand, the edges in the space $\mathcal{S}_n$ may intersect.

Example 4.5.1. The SOM defined in Example 4.3.6 for $d = 2$ and $n = 3$ has the following subspaces associated, where we have chosen the prototypes arbitrarily,



Subspace $\mathcal{S}_d$ of $\mathbb{R}^d = \mathbb{R}^2$.



Subspace $\mathcal{S}_n$ of $\mathbb{R}^n = \mathbb{R}^3$.

The (unsupervised) learning phase for a SOM consists of the following two steps. Afterwards, the SOM works in recall phase as explained in Section 4.6.

- Initialization of prototypes of the SOM.
- Adjustment of prototypes for each input vector presented to the SOM.

Two standard initialization procedures are random initialization and principal component analysis initialization. We do not further discuss this topic here. In [**1**], the reader may find a comparison of these two approaches for one-dimensional SOMs.

Let $\mathbf{x}(t) \in \mathbb{R}^n$ be the input presented to the SOM at time t, where we let t takes the values $t = 1, 2, 3, \ldots, N$. So $\mathbf{x}(1), \ldots, \mathbf{x}(N)$ are the training samples the SOM will use in the learning phase. Assume the weight vectors or prototypes have the values $\mathbf{w}^j(t)$ at time t. Next we explain how the updated values $\mathbf{w}^j(t+1)$ are computed. This process has two phases:

(a) Determine the neuron j_0 whose prototype is closest to the input vector $\mathbf{x}(t)$ in the input space $\mathbb{R}^n$ according to Equation 8. Thus, we look for j_0 such that

$$||\mathbf{w}^{j_0}(t) - \mathbf{x}(t)|| = \min_j ||\mathbf{w}^j(t) - \mathbf{x}(t)||. \tag{9}$$

(b) Set the new prototype $\mathbf{w}^j(t+1)$ to be equal to a convex combination of $\mathbf{x}(t)$ and $\mathbf{w}_j(t)$, i.e., to a point in the segment from $\mathbf{w}_j(t)$ to $\mathbf{x}(t)$.

The neuron j_0 in point (a) above is called *best matching unit* (*BMU*). We will abuse notation and call BMU to the neuron j_0, to its associated point in $\mathbf{p}^{j_0} \in \mathbb{R}^d$ and to its associated prototype $\mathbf{w}^{j_0} \in \mathbb{R}^n$. Which one we mean will be clear from the context. We write $BMU(t)$ if we want to emphasize the instant t.

Now we explain the role that the winning neuron j_0 or BMU has in point (b) above. The mentioned convex combination will be of the form,

$$\mathbf{w}^j(t+1) = \mathbf{w}^j(t) + \gamma(t, j, j_0)(\mathbf{x}(t) - \mathbf{w}^j(t)), \tag{10}$$

where the function $\gamma(t, j, j_0)$ satisfies $0 \leq \gamma(t, j, j_0) \leq 1$ and decreases with time and with the distance $D(j, j_0)$. Note that this function depends only on t, j and j_0. See Figure 13.

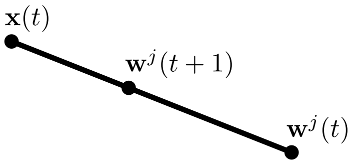
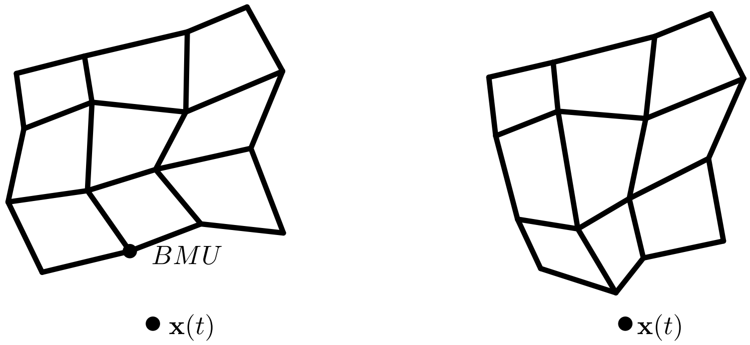


FIGURE 13. $\mathbf{w}^j(t+1)$ lies in the segment from $\mathbf{w}^j(t)$ to $\mathbf{x}(t)$.

The function γ is decomposed as follows,

$$\gamma(t, j, j_0) = \theta(t, j, j_0)\eta(t),$$

where $\theta(t, j, j_0)$ is called the *neighbourhood* function and $\eta(t)$ is called the *learning* function. These two functions are described below in Subsection 4.5.1 and 4.5.2. Together, they produce that, on the learning step t, only the neurons closer to the BMU learn, and the amount of learning decreases with the distance to the BMU (distance D or, analogously, distance in the space $\mathcal{S}_d$) and with time t. See the next figure for a representation of these facts.



(a) Space $\mathcal{S}_n$, input sample $\mathbf{x}(t)$ and BMU before learning step.

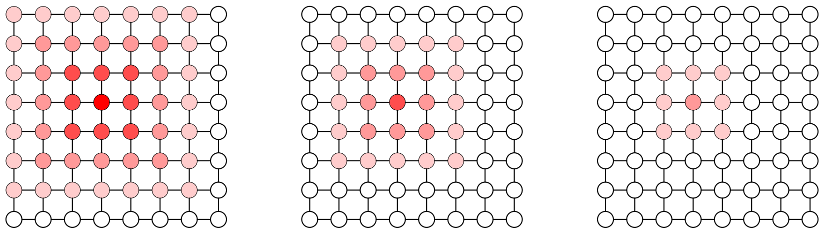
(b) Space $\mathcal{S}_n$ and input sample $\mathbf{x}(t)$ after learning step.

4.5.1. The neighbourhood function. A standard neighbourhood function is the (pseudo) normal distribution

$$\theta(t, j, j_0) = e^{-\frac{D^2(j,j_0)}{2\sigma(t)^2}},$$

where $D(j, j_0)$ is the distance between neurons j and j_0 in the SOM, and the variance $\sigma(t)$ decreases with time. The role of $\sigma(t)$ is measuring the influence of the BMU on its neighbours. Recall that approximately 68% of the distribution θ is taken by neurons j which are at a distance from j_0, $D(j, j_0)$, smaller than $\sigma(t)$.

EXAMPLE 4.5.2. For $d = 2$ and an 8×8 grid, the next picture shows the decreasing neighbourhoods defined by $\sigma(t)$ around a fixed neuron j_0:



A standard definition of the variance $\sigma(t)$ is

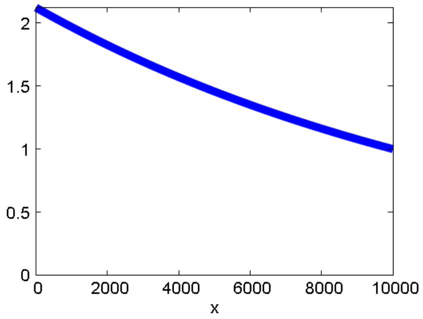
$$\sigma(t) = \sigma_0 e^{\frac{-t}{\lambda}},$$

where σ_0 and λ are appropriately chosen constants. They can be determined by optimization of parameters or they may be chosen heuristically as follows,

$$\sigma_0 = \frac{max_{j,k} D(k,j)}{2} \text{ and } \lambda \qquad = \frac{N}{\ln(\sigma_0)},$$

where N is the number of times that the learning step will be executed by the SOM and σ_0 is half the diameter of the space $\{1, ..., m\}$, i.e, its radius. This implies that $\sigma(N) = 1$ and thus, if the function D is normalized in such a way that $D(j,k) = 1$ whenever j and k are linked, in the last step the influence of the winning BMU j_0 is just itself.

Example 4.5.3. For the SOM of Example 4.3.6, we have that the radius is $\sigma_0 = 3\sqrt{(2)} \approx 2.12$. Setting $N = 10,000$ we get $\lambda = \frac{N}{\log(\sigma_0)} \approx 13297.18$. The corresponding variance function is as follows

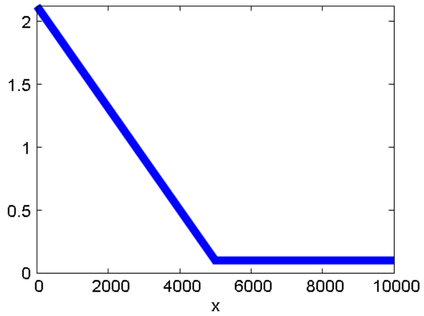


Another common form for the variance is a linear decay followed by a residual value,

$$\sigma(t) = \begin{cases} \sigma_0 - \frac{\sigma_0 - \sigma_1}{\lambda N} x & \text{if } 0 \leq t \leq \lambda N, \\ \sigma_1 & \text{if } \lambda N \leq t \leq N, \end{cases} \tag{11}$$

where again N is the number of times that the learning step will be executed by the SOM, σ_0 is the starting variance, σ_1 is the residual value of the variance and $0 \leq \lambda \leq 1$. A common value is $\lambda = \frac{1}{2}$ and the other parameters can be determined by optimization.

Example 4.5.4. For the SOM of Example 4.3.6, we have that the radius is $\sigma_0 = 3\sqrt(2) \approx 2.12$. Set $N = 10,000$, $\lambda = 0.5$ and $\sigma_1 = 0.1$. The corresponding piecewise linear variance function (11) is as follows



4.5.2. The learning function. The learning function $\eta(t)$ is usually modelled as an exponential decay function,

$$\eta(t) = \eta_0 e^{\frac{-t}{\lambda}},$$

or as a linear decay as before,

$$\eta(t) = \begin{cases} \eta_0 - \frac{\eta_0 - \eta_1}{\lambda N} x & \text{if } 0 \leq t \leq \lambda N, \\ \eta_1 & \text{if } \lambda N \leq t \leq N. \end{cases} \tag{12}$$

Here N is the total number of iterations of the SOM algorithm, $0 \leq \lambda \leq 1$ and a usual value is $\lambda = \frac{1}{2}$. The parameter η_0 is the initial learning value and η_1 is the residual learning value. Both parameters may be determined by optimization.

Remark 4.5.5. As the reader may have noticed, for the SOM $\mathcal{S} = (n, m, d, \mathcal{G}, \mathcal{D})$, the graph $\mathcal{G}$ does not play a role in the learning phase of the SOM. In fact, the graph $\mathcal{G}$ will be relevant when assessing the topological preservation of the SOM as discussed in Section 5.1.

4.6. SOM in recall phase

The learning phase of the SOM algorithm finishes after N iterations in which the training samples $\mathbf{x}(1), \ldots, \mathbf{x}(N)$ have been presented to the SOM. During this process, the weights $\mathbf{w}^j = (w_{1j}, \ldots, w_{nj})$ (see Equation (7)) are adjusted until their final values are reached.

In the recall phase, a single input $\mathbf{x} = (x_1, \ldots, x_n)$ is presented to the SOM algorithm and it produces an output vector $\mathbf{y} = (y_1, \ldots, y_m)$. In fact, denoting by $j_0 \in \{1, \ldots, m\}$ the *BMU*, i.e., the neuron whose prototype $\mathbf{w}^j$ is closest to the input $\mathbf{x}$ (see Equation (8)), the output vector is defined by

$$y_j = \delta(j, j_0) = \begin{cases} 1 & \text{if } j = j_0, \\ 0 & \text{otherwise.} \end{cases}$$

This formula describes the SOM ANN as a function from the input space $\mathbb{R}^n$ to the output space $\mathbb{R}^m$, in the way it was introduced in Section 4.2,

$$\begin{aligned} \chi \colon \mathbb{R}^n &\to \mathbb{R}^m \\ \mathbf{x} = (x_1, \ldots, x_n) &\mapsto \chi(\mathbf{x}) = \mathbf{y} = (y_1, \ldots, y_m) \text{ with } y_j = \delta(j, j_0) \end{aligned} \tag{13}$$

This function takes as values the vectors

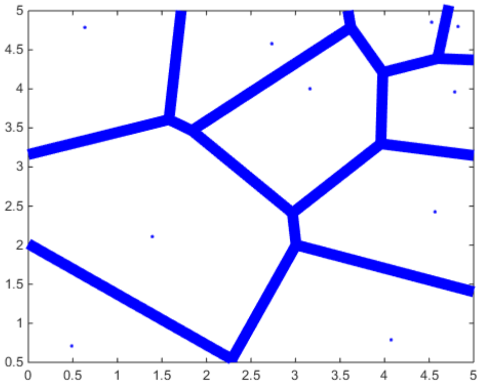
$$(1,0,\dots,0,0),(0,1,\dots,0,0),\dots,(0,0,\dots,1,0),(0,0,\dots,0,1)$$

and the preimage of the j_0-th vector in this list is exactly the following subset of $\mathbb{R}^n$,

$$V_{j_0} = \{\mathbf{x} \in \mathbb{R}^n | \; ||\mathbf{w}^{j_0} - \mathbf{x}|| = \min_j ||\mathbf{w}^j - \mathbf{x}||\}. \tag{14}$$

So the set V_{j_0} consists of the points of $\mathbb{R}^n$ which are closer to $\mathbf{w}^{j_0}$ than to any other prototype $\mathbf{w}^j$. Thus, V_{j_0} is exactly the Voronoi cell of the point $\mathbf{w}^{j_0}$ corresponding to the Voronoi diagram corresponding to the points $\mathbf{w}^1,\dots,\mathbf{w}^m$ of $\mathbb{R}^n$.

Example 4.6.1. The following picture shows 10 points within the subset $[0,5]\times[0,5] \subseteq \mathbb{R}^2$ together with their corresponding Voronoi cells.



If, in the recall phase, we have a set of M test samples, $\mathbf{x}^1,\dots,\mathbf{x}^M$, a common measure of the performance of the SOM is the Mean Squared Error. In order to introduce it, denote by $j_i = BMU_i$ the neuron whose prototype is closest to the input test sample $\mathbf{x}^i$, for $i = 1,\dots,M$. Then we define,

$$MSE = \frac{1}{M}\sum_{i=1}^{M} ||\mathbf{x}^i - \mathbf{w}^{j_i}||^2, \tag{15}$$

which can be also written as

$$MSE = \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{m} P(i,j)||\mathbf{x}^i - \mathbf{w}^j||^2,$$

where

$$P(i,j) = \begin{cases} 1 & \text{if } \mathbf{x}^i \in V_j, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

Note that Equation (15) is the mean of Equation (3), which corresponds to the squared error criterion. As it will be explained in Section 5.1, there exist other performance measures of the SOM that focus on the so-called topology preservation, a concept closely related to the mathematical notion of continuity. Please see Appendix A for basic notions on mathematical continuity. For instance, the map χ

defined above (13) is not continuous. In preparation for the discussion of topological preservation in Section 5.1, we define in the next subsection some other maps and study their continuity.

4.6.1. Continuity of some maps. So let $\mathcal{S} = (n, m, d, \mathcal{G}, \mathcal{D})$ the SOM under study. Then we can consider, besides χ, two other maps, one map Φ_1 from the input space $(\mathbb{R}^n, ||\cdot||)$ to the metric space $(\{1, \ldots, m\}, \mathcal{D})$, and one map Φ_2 from the metric space $(\{1, \ldots, m\}, \mathcal{D})$ to the input space $(\mathbb{R}^n, ||\cdot||)$. We have emphasized that we consider the euclidean norm in $\mathbb{R}^n$ by writing $(\mathbb{R}^n, ||\cdot||)$,

$$\begin{tikzcd}(\mathbb{R}^n, ||\cdot||) \arrow[r, shift left, "\Phi_1"] & (\{1, \ldots, m\}, \mathcal{D}) \arrow[l, shift left, "\Phi_2"]\end{tikzcd}$$

The map $\Phi_1 \colon (\mathbb{R}^n, ||\cdot||) \to (\{1, \ldots, m\}, \mathcal{D})$ is defined by

$$\mathbf{x} \mapsto j_0,$$

where j_0 is again the neuron whose prototype is closest to $\mathbf{x}$ within $\mathbb{R}^n$. Note that $\mathbf{x}$ may belong to the intersection of two or more Voronoi cells (14), i.e., there could be more than one neuron that minimizes the distance to $\mathbf{x}$. In this case, we arbitrarily define $\Phi_1(x)$ as one of these neurons. In an implementation of the SOM, the selected neuron could be the first neuron in the natural order $1 < 2 < \ldots < m$. In different setups, the problem of multiple minimizing points is solved differently, for instance, considering the center of the smallest ball containing all minimizing points [**48**, p. 73].

The map $\Phi_2 \colon (\{1, \ldots, m\}, \mathcal{D}) \to (\mathbb{R}^n, ||\cdot||)$ is defined by

$$j \mapsto \mathbf{w}^j.$$

If Φ_2 is injective[1], we have that $\Phi_1 \circ \Phi_2 = 1_{\{1,\ldots,m\}}$. Regarding continuity, we have the following result.

LEMMA 4.6.2. *Let $\mathcal{S} = (n, m, d, \mathcal{G}, \mathcal{D})$ be a SOM. Then:*

(1) The map $\Phi_1 \colon (\mathbb{R}^n, ||\cdot||) \to (\{1, \ldots, m\}, \mathcal{D})$ is not continuous[2].
(2) The map $\Phi_2 \colon (\{1, \ldots, m\}, \mathcal{D}) \to (\mathbb{R}^n, ||\cdot||)$ is continuous.

PROOF. For the first part, take a point $\mathbf{x}$ in the intersection of at least two Voronoi cells, $\mathbf{x} \in V_{j_1} \cap V_{j_2}$, with $j_1 = \Phi_1(\mathbf{x})$. Then there is a sequence of points $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots$ contained in $int(V_{j_2}) \setminus V_{j_1}$ and converging to $\mathbf{x}$. Then $\Phi_1(\mathbf{x}_i) = j_2$ and $\Phi_1(\mathbf{x}) = j_1$. Thus Φ_1 is not continuous, see A.0.12. For the second part, by Remark 4.3.2, the topology of $\{1, \ldots, m\}$ is the discrete one, and hence every map with this domain is continuous. □

If the SOM $\mathcal{S}$ was constructed from a grid (see Example 4.3.3), we consider the spaces $\mathcal{S}_d$ and $\mathcal{S}_n$ described at the beginning of Section 4.5, see also Example 4.5.1. Moreover, in this situation, we have another two maps,

$$\begin{tikzcd}(\mathbb{R}^n, ||\cdot||) \arrow[r, shift left, "\Psi_1"] & (\mathcal{S}_d, ||\cdot||), \arrow[l, shift left, "\Psi_2"]\end{tikzcd}$$

[1] this is equivalent to the points $\{\mathbf{w}^1, \ldots, \mathbf{w}^m\}$ being all distinct.
[2] if there are at least two different points in $\{\mathbf{w}^1, \ldots, \mathbf{w}^m\}$

where recall that the topology in $\mathcal{S}_d$ is given by the restriction of the euclidean metric $||\cdot||$ of $\mathbb{R}^d$. The map $\Psi_1\colon(\mathbb{R}^n,||\cdot||)\to(\mathcal{S}_d,||\cdot||)$ is defined by

$$\mathbf{x}\mapsto\mathbf{p}^{j_0},$$

where $\mathbf{p}^{j_0}$ is the point of the grid corresponding to the *BMU* j_0. The map $\Psi_2\colon(\mathcal{S}_d,||\cdot||)\to(\mathbb{R}^n,||\cdot||)$ is the unique linear extension of the map that sends

$$j\mapsto\mathbf{w}^j.$$

Note that the image of this map is exactly the space $\mathcal{S}_n\subseteq\mathbb{R}^n$. In this case, even if Ψ_2 is injective, we do not have $\Psi_1\circ\Psi_2=1_{\mathcal{S}_d}$. The reason is that the image of Ψ_1 is just the set $\{\mathbf{p}^1,\dots,\mathbf{p}^m\}\subsetneq\mathcal{S}_d$. The continuity of these maps is as follows.

LEMMA 4.6.3. *Let $\mathcal{S}=(n,m,d,\mathcal{G},\mathcal{D})$ be a SOM arising from a grid. Then:*

(1) The map $\Psi_1\colon(\mathbb{R}^n,||\cdot||)\to(\mathcal{S}_d,||\cdot||)$ is not continuous.

(2) The map $\Psi_2\colon(\mathcal{S}_d,||\cdot||)\to(\mathbb{R}^n,||\cdot||)$ is continuous.

PROOF. The first part follows as in Lemma 4.6.2. The second part follows from the continuity of piecewise linear maps. □

REMARK 4.6.4. The graph $\mathcal{G}$ does not play a role in the learning phase, see Remark 4.5.5, nor in the recall phase. It will become relevant to determine the level of topological preservation of the SOM as discussed in Section 5.1.

CHAPTER 5

SOM Analysis

In this chapter, we comment on some aspects of SOMs. We start discussing topological preservation in Section 5.1. We focus on the formal definition and its comparison to mathematical continuity, see A.0.4, as well as on some quantitative measures of topological preservation. Then we briefly elaborate on variants of SOMs having to do with the distance in the input space and the distance in the grid that gives rise to the SOM, paying close attention to this last topic, see Section 5.2.

5.1. Topological preservation in SOMs

One of the main ideas underlying SOMs is that they are topological preserving ANNs. In this section, we discuss this notion and we relate it to the mathematical concept of continuity, see Appendix A and A.0.4. The term topographic preservation is used in the literature as a synonym for topological preservation. We do the same here.

Some biological studies indicate that, in some animals, signals from adjacent receptors are conducted to adjacent neurons in the brain, cf. A.0.7. For instance, for the sense of touch in monkeys, the somatosensory cortex is described in [**28**], see Figure 16.

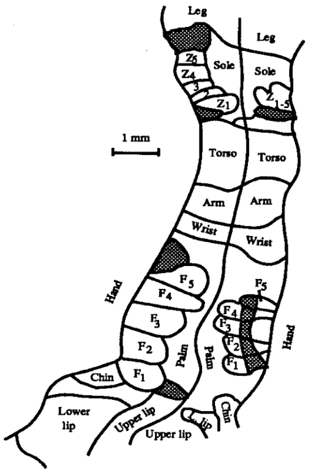


FIGURE 16. Part of the somatosensory cortex of a monkey with the layout of the mapped body parts. Letters F_i correspond to fingers. Source: [**28**, Fig. 1], [**53**, p. 17].

In that figure, it is observed that many neighboorhoud relations in the body are preserved after they are mapped to the cortex. For a further discussion on the biological grounds, we refer the reader to [**53**, Chapter 2]. In the literature, these biological findings are emulated via different ways of measuring how a trained SOM preserves the topology, and some of these ideas are related to the mathematical notion of continuity. For instance, from [**6**, p. 660, l. 6],

> In a colloquial manner, topography of a map means the mapping of similar data points to close locations in the map layer.

And later, in the same paper [**6**, p. 660, l. 14],

> Then topography means topology preservation and is equivalent to continuity of a mapping between the input and output set.

Also, in [**11**, 4.2.1], we have,

> Besides quantization, the second main goal of the SOM is the so-called topology preservation, which means that close data in the input space will be quantized by either the same centroid, either two centroids that are close from one another on a predefined string or grid.

In [**61**, p. 279], we find the following,

> Loosely spoken, topology preservation means that a continuous change of a parameter of the input data leads to a continuous change of the position of a localized excitation in the neural map.

So all these notions refer to the continuity of the map that goes from the input space to the neurons and resemble the characterization of mathematical continuity using the closure operator A.0.7. We have modelled the map from the input space to the neurons in two different ways via the maps Φ_1 and Ψ_1 that employ the BMU neuron, see Subsection 4.6. Nevertheless, these maps are not continuous with the topologies given there, see Lemma 4.6.2(i) and Lemma 4.6.3(i). In other works, this map is required to be even a homeomorphism [**21**, Section 4]. In the rest of this section, we do focus on this map from the input space to the neurons. In fact, in view of the results Lemma 4.6.2(ii) and Lemma 4.6.3(ii) and of the biological background commented above, the map from the neurons to the input space is less interesting.

Some authors have introduced different topologies or alternative input spaces to deal with the continuity issue. For instance, in the work [**61**] as well as in [**62**], different topologies termed "discrete" are defined in the set of neurons $\{1, \ldots, m\}$. In fact, in the mathematical sense, any given set has a unique discrete topology. Namely, the topology where all subsets are open, see A.0.2 and A.0.11. Moreover, the "system of open sets" defining these topologies in [**61**, Corollary 5] do not satisfy that union of open sets is open. Therefore, they are not topologies in the mathematical sense, see A.0.1.

These ideas arose from the constructions in [**43**, Definitions 3,4,5], that we now succinctly describe: let $M \subseteq \mathbb{R}^n$ be a subset of $\mathbb{R}^n$. This set M should be thought of as a topological manifold of some positive dimension that contains the point cloud of inputs. Let $(n, m, d, \mathcal{G}, \mathcal{D})$ be a SOM in recall phase such that $\mathbf{w}^j \in M$ for all $1 \leq j \leq m$. Then this SOM is said to be *topology preserving* if, for all $1 \leq i, j \leq m$, we have

$$\{i, j\} \in \mathcal{G} \Leftrightarrow V_i \cap V_j \cap M \neq \emptyset.$$

Thus, we ask for the vertices i and j to be connected in the graph $\mathcal{G}$ if and only if the corresponding Voronoi cells (14) intersect non-trivially with M. In [**62**, Definition 1], and for SOMs arising from a grid 4.3.3, the authors introduce a variation of this idea by substituting the neighbourhood relation in the graph $\mathcal{G}$ by the neighbourhood relation induced by the maximum norm $|| \cdot ||_\infty$ in the space $\mathbb{R}^d$ where the grid is immersed. See also the definition of Topographic Function below 5.1.2.

This concept resembles the mathematical notion of continuity, except that, if we define the "topological" neighbourhoods of i to be the neighbourhoods of i in the graph $\mathcal{G}$, we do not get a topology because intersection of open sets would not be open in general.

So far in this section, we have faced several times the underlying problem of endowing the finite set of neurons $\{1, \dots, m\}$ with a topology. In fact, topologies on a finite set are given by pre-orders, i.e., by relative transitive relations on the finite set. This is explained in [**5**] for instance, where the author also develops the notions of algebraic topology in this context. As an example, the discrete topology on a finite set corresponds to the trivial pre-order on that set, i.e., each element is related only to itself. It is remarkable that Kohonen's original paper [**34**, p. 60] mentions orders,

> Assume that the events A_k can be ordered in some metric or topological way such that $A_1 R A_2 R A_3 \dots$.. where R stands for a general ordering relation which is transitive (the above implies, e.g., that $A_1 R A_3$).

Therefore, it seems worth investigating the possible connections between topologies on the finite set of neurons, i.e., pre-orders, the biological roots of SOMs as commented above and the very original definition of SOM by Kohonen.

In the rest of this section, we discuss some other ways of measuring the topological preservation in SOMs. These are of quantitative nature instead of qualitative nature. A survey of these notions can be found in [**44**].

5.1.1. Topographic error. This is possibly the simplest qualitative measure of continuity, although it is not the earliest one, see [**32**]. To introduce it, given a set of test samples $\mathbf{x}^1, \dots, \mathbf{x}^M$, denote by $j_i = BMU_i$ the neuron whose prototype is closest to $\mathbf{x}^i$ in $\mathbb{R}^n$, and denote by j'_i the second closest prototype to $\mathbf{x}^i$. Then the topographic error is defined as follows,

$$TE = \frac{1}{M} \sum_{i=1}^{M} TE_i, \tag{17}$$

where TE_i is 1 if the best and second-best matching units (j_i and j'_i) are not connected in the graph $\mathcal{G}$ and 0 otherwise. In symbols,

$$TE_i = \begin{cases} 0 & \text{if } \{j_i, j'_i\} \in \mathcal{G}, \\ 1 & \text{otherwise.} \end{cases}$$

So the topographic error $TE \in [0, 1]$ gives the average of local topographic errors along the test samples.

5.1.2. Topographic function. The topographic function [**62**] is an evolution of the topographic product, which was the first quantitative measure of topographic preservation and goes back to 1992 [**7**]. The topographic function TF has domain the positive integers $k = 1, 2, 3, \ldots$ and codomain the interval $[0, m]$. Let $\mathcal{S} = (n, m, d, \mathcal{G}, \mathcal{D})$ a SOM constructed from a grid with points $\mathbf{p}^1, \ldots, \mathbf{p}^m$ within $\mathbb{R}^d$, see 4.3.3. Assume that $M \subseteq \mathbb{R}^n$ is a topological manifold of positive dimension containing the prototypes $\mathbf{w}^1, \ldots, \mathbf{w}^m$. Then the topographic function is defined as follows, for any $k \geq 1$,

$$TF(k) = \frac{1}{m} \sum_{i=1}^{m} |\{j \in \{1, \ldots, m\} \text{ such that } V_i \cap V_j \cap M \neq \emptyset \text{ and } ||\mathbf{p}^i - \mathbf{p}^j||_\infty > k\}|.$$

Recall that $|A|$ denotes the cardinal of the set A and that V_i is the Voronoi cell corresponding to the prototype $\mathbf{w}^i \in \mathbb{R}^n$. This definition is extracted from [**62**, Equations (9) and (10)]. The value $TF(k)$ gives the average number of prototypes that are adjacent in the Delaunay triangulation in $\mathbb{R}^n$ and whose respective points in the grid of $\mathbb{R}^d$ are at a distance larger than k in the maximum norm $|| \cdot ||_\infty$.

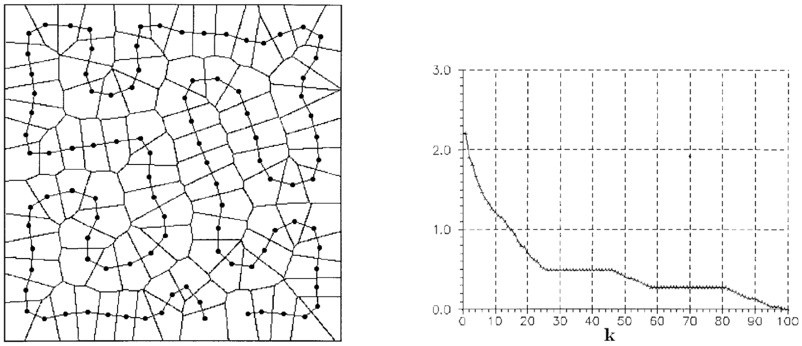


FIGURE 17. Prototypes of a 100×1 linear one-dimensional SOM on a square (left) and the graph of the corresponding Topographic function TF (right). Source: adapted from [**62**, Figures 5 and 6].

REMARK 5.1.1. From the mathematical viewpoint, continuity is defined as an absolute notion, i.e., a function is either continuous or not continuous, see A.0.4. Mimicking what has been described here, it would be interesting to consider quantitative measures that assign a value of "continuity" to a given function in the mathematical setup.

5.2. Variations of SOMs

Let $\mathcal{S} = (n, m, d, \mathcal{G}, \mathcal{D})$ be a SOM arising from a grid of points $\mathbf{p}^1, \ldots, \mathbf{p}^m$ in $\mathbb{R}^d$, see Example (4.3.3). Varying some of the elements employed in the definition of the SOM and/or in the SOM algorithm has given rise to an overwhelming number of alterations of the original SOM, see [**35**, Chapter 5] for instance. We focus on the following two structural elements and discuss some of them in Chapter 6 and some later in this section.

(1) the distance function D is the restriction to the set $\{1, \ldots, m\}$ of the Euclidean distance in $\mathbb{R}^d$, see Equation (5).
(2) the best matching unit BMU is computed minimizing the Euclidean distance in $\mathbb{R}^n$, see Equation (8).

Regarding the first point above, one can consider the inner dimension of the input data and compare it to the dimension d of the grid. More precisely, assume that the training samples $\mathbf{x}(1), \ldots, \mathbf{x}(N)$ are picked from a topological manifold $M \subseteq \mathbb{R}^n$ of some positive dimension d'. Then, an investigation of the performance of the SOM in terms of the difference $|d - d'|$ is relevant. Partial results in this direction have been obtained and are described in Section 6.2.

We could focus instead on the local structure of the input data and the grid: If the input data presents a non-Euclidean local structure, then a non-Euclidean SOM may be better suited for the learning task. This has been considered in [**54**] and we discuss this point of view in Subsection 5.2.1, together with some additional reflections. Other variations may be obtained by keeping the Euclidean distance in $\mathbb{R}^d$ and considering alternative distributions of the points of the grid. For instance, one can take into account regular tessellations of the plane $\mathbb{R}^2$. Results have been obtained in this direction and they are described in Section 6.1.

Still with relation to point (1), it is natural to allow the number of neurons m and the distance function D to change along the learning process. These ideas have given rise, for instance, to the Growing SOM (GSOM) and to the Self-Organizing Dynamic Graphs (SODG), which are discussed in Subsection 5.2.2.

Regarding point (2) above, one could substitute the Euclidean straight segment between two points in the input space $\mathbb{R}^n$ by the geodesic in $\mathbb{R}^n$, where we endow $\mathbb{R}^n$ with a curvature inversely proportional to the density of input samples. This was the objective of the work [**17**] and it will not be discussed here. Another modification consists of substituting the Euclidean straight segment of $\mathbb{R}^n$ by a piecewise linear path that avoids predefined forbidden regions. Results have been obtained in this direction and they are discussed in Section 6.3.

5.2.1. Non-euclidean spaces and SOMs.

Consider the differential manifold of dimension d' and of constant curvature $k = 0, 1$ or -1, given by the Euclidean space $\mathbb{R}^{d'}$, the sphere $\mathbb{S}^{d'}$, and the hyperbolic space $\mathbb{H}^{d'}$ respectively. Then the volume $V_k(r)$ of a metric ball of radius r is given as follows, see [**56**, Chapter IV, problem 1],

$$\begin{aligned} V_0(r) =& \lambda_0 r^{d'}, \\ V_1(r) =& \lambda_1 \int_0^r sin^{d'-1}(t)dt \ (0 < r < \pi), \\ V_{-1}(r) =& \lambda_{-1} \int_0^r sinh^{d'-1}(t)dt, \end{aligned}$$

where λ_0, λ_1, and λ_{-1} are constants. For instance, for $d' = 2$, we obtain,

$$\begin{aligned} V_0(r) =& \pi r^2, \\ V_1(r) =& 2\pi(1 - cos(r)) \ (0 \leq r \leq \pi), \\ V_{-1}(r) =& 2\pi(cosh(r) - 1). \end{aligned}$$

In Figure 18, we can see the graph of these functions for $0 \leq r \leq \pi$. The function $V_{-1}(r)$ has exponential growth and the function $V_0(r)$ has polynomial growth.

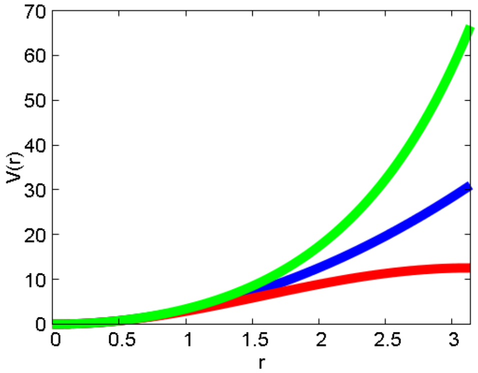
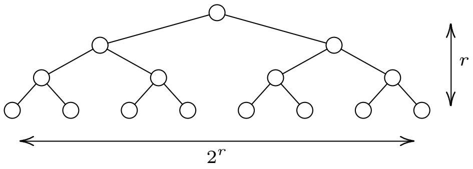


FIGURE 18. Volume of a ball in $\mathbb{R}^2$ (blue), $\mathbb{S}^2$ (red) and $\mathbb{H}^2$ (green) in terms of radius r.

In hierarchical structures such as trees, the number of nodes at distance r or smaller of a given node is an exponential function of r,



Then, and as discussed in [**54**], it could be more appropriate, for hierarchically structured data, to use a SOM arising from a grid whose points are the vertices of a tessellation of some hyperbolic space. This brings up the theme of tessellations of spaces, a notion that is useful and fruitful to construct variants of SOM. In Appendix C we have gathered some basic facts and definitions about tessellations. Below we give a succinct account of tessellations for the spaces $\mathbb{R}^2$, $\mathbb{S}^2$, and $\mathbb{H}^2$. The vertices of these tessellations may be used to generate a SOM as in Example 4.3.3. Further insights into tessellations of the plane $\mathbb{R}^2$ are explained in Section 6.1.

We consider regular tessellations, i.e., edge-to-edge monohedral vertex transitive tesselations with tile a fixed n-gon, see Appendix C for the appropriate definitions. For the plane $\mathbb{R}^2$, there exist three regular tessellations,

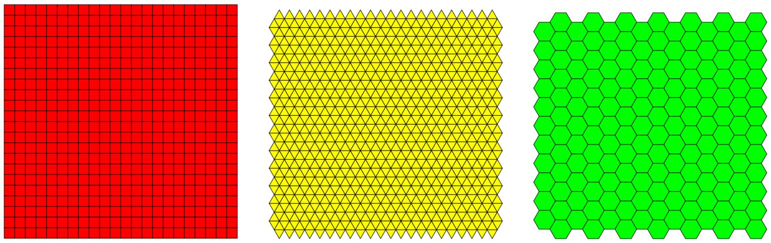


FIGURE 19. Regular tessellations of the plane. Source: Wikipedia

For the 2-sphere, the five platonic solids, namely, tetrahedron, cube, octahedron, dodecahedron, and icosahedron, give rise to five regular tessellations, see Figure 20. Moreover, subdivisions of these may be considered in order to add more points to the grid, at the price of losing some symmetry of course.

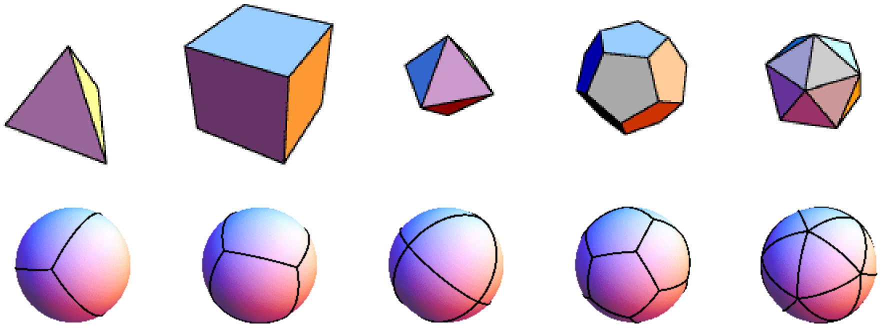


FIGURE 20. The platonic solids and their projection onto a concentric sphere. Source: Math & the Art of MC Escher, Wikipedia

For the hyperbolic space, there are infinitely many tessellations, see [**13**]. In Figure 21, there are three examples of these tessellations.

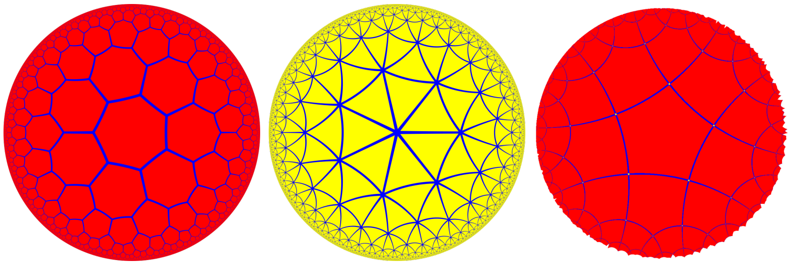


FIGURE 21. Some tessellations of the hyperbolic plane. Source: Wikipedia

Relaxing some of the symmetry constraints of the tessellations, a huge variety of such appear. For instance, instead of monohedral tessellations, one may allow different n-gons, keeping the edge-to-edge condition and the vertex transitivity condition. These are called Archimidean or semi-regular tesselations, and there are 8 of these besides the three regular tessellations of the plane, see Figure 19. The improved performance of a SOM on such a tessellation is studied in Section 6.1.

In general, one may consider many other variations, as polytopes that are not n-gons, more than one orbit on the vertices and, as discussed above and in [**54**], non-Euclidean spaces. Apparently, there is no much literature on SOMs in this direction. In my opinion, this is a topic worth studying.

5.2.2. Growing and dynamic SOMs. The growing self-organizing map or GSOM was introduced in 2000 in the work [**2**]. The main feature of this SOM is that the grid of neurons grows through the learning process. There exist many other growing-type SOMs, as for instance [**42**]. Moreover, there exist variants of SOM, as Self-Organizing Dynamic Graphs [**41**], in which the metric D itself (see Definition 4.3.1) varies along the learning process. Properly describing both growing and dynamical SOMs was one of the reasons that motivated us to introduce the abstract notion of SOM given in Definition 4.3.1.

The idea behind GSOM and other growing-type SOMs is to create new neurons near existing neurons that accumulate a "big enough error" through the learning process. We describe next the basic mechanism underlying GSOM as an instance of this type of SOMs. So let $\mathcal{S} = (n, 4, 2, \mathcal{G}, D)$ be a GSOM at initial state. Hence, GSOM is two dimensional and, in fact, the graph $\mathcal{G}$ and the metric D are built as in Example 4.3.3 for a unit square in $\mathbb{R}^2$, i.e., for the following points,

$$\mathbf{p}^1 = (0,0)\ ,\ \mathbf{p}^2 = (0,1)\ ,\ \mathbf{p}^3 = (1,1) \text{ and } \mathbf{p}^4 = (1,0),$$

and segments,

$$\overline{\mathbf{p}^1\mathbf{p}^2}\ ,\ \overline{\mathbf{p}^2\mathbf{p}^3}\ ,\ \overline{\mathbf{p}^3\mathbf{p}^4} \text{ and } \overline{\mathbf{p}^4\mathbf{p}^1}.$$

If there are m neurons at some instant t $(1 \leq t \leq N)$, the accumulated error for neuron j $(1 \leq j \leq m)$ is given recursively by [**2**, Equation (2)],

$$E_j(t+1) = E_j(t) + ||\mathbf{x}(t) - \mathbf{w}^j(t)||,$$

where $\mathbf{x}(t)$ is the input presented at time t and $\mathbf{w}^j$ is the weight vector or prototype of neuron j (see Equation (7)) at that instant. If this value exceeds a threshold, new nodes are created at the positions $\{\mathbf{p}^j + (1,0), \mathbf{p}^j - (1,0), \mathbf{p}^j + (0,1), \mathbf{p}^j - (0,1)\}$ that are not yet occupied by a neuron. Here, $\mathbf{p}^j$ is the point of $\mathbb{R}^2$ corresponding to neuron j. See Figure 22.

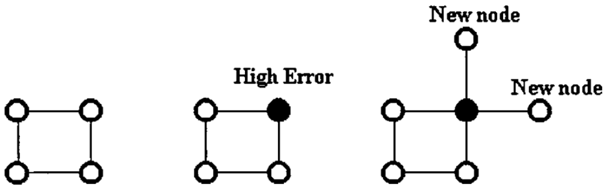


FIGURE 22. New neurons generation in GSOM. Source [**2**, Figure 2]

The weights of the new neurons are initialized to a linear combination of the weights of neurons in its vicinity according to some detailed casuistic, see [**2**, III.B.2]. In this way, GSOM at instant t is described by a SOM $\mathcal{S}(t) = (n, m(t), 2, \mathcal{G}(t), D(t))$, where $m(t) \geq m(0) = 4$ and $\mathcal{G}(t)$ and $D(t)$ are obtained by some planar grid as in Example 4.3.3. It is also interesting that the function $\gamma(t, j, j_0)$ appearing as the coefficient in the affine combination rule (10), i.e., in

$$\mathbf{w}^j(t+1) = \mathbf{w}^j(t) + \gamma(t, j, j_0)(\mathbf{x}(t) - \mathbf{w}^j(t)),$$

is modified in GSOM: it depends very explicitly and strongly on the number of neurons at each instant $m(t)$, see [**2**, IV].

In SODGs or Self-Organizing Dynamic Graphs [**41**], the structure at learning time t is described by a SOM $\mathcal{S}(t) = (n, m, d, \mathcal{G}, D(t))$, where the number of neurons

m and the graph $\mathcal{G}$ are fixed but the metric $D(t)$ depends on the instant t. In fact, the collections of values, $D(t) = \{D(i,j)(t)\}_{i,j=1,\dots,m}$, where $D(i,j)(t)$ is the "distance" between neurons i and j at time t, will not be, in general, a metric in the sense A.0.9. For instance, we may have $D(i,j)(t) \neq D(j,i)(t)$ [**41**, p.101]

Nevertheless, we have $D(i,i)(t) = 0$ for all i and t [**41**, 3] and the alternate condition [**41**, 13],

$$\sum_{j=1,\dots,m} D(i,j)(t) = 1, \tag{18}$$

for all $1 \leq i \leq m$. In more detail, the update formula for $D(t)$ is given in two steps [**41**, (14),(15)]. First, set

$$D(j,i)(t+1) = D(i,j)(t) + \frac{||\mathbf{x}(t) - \mathbf{w}^i(t)||}{||\mathbf{x}(t) - \mathbf{w}^j(t)||}\delta(t),$$

where $\mathbf{x}(t)$ is the input sample at time t, $\mathbf{w}^i(t)$ and $\mathbf{w}^j(t)$ are prototypes at time t and $\delta(t)$ is the "metric" learning rate, which follows a linear decay analogous to Equation 11 and Example 4.5.4. Then normalize $D(t)$ so it satisfies Equation (18), i.e., divide each entry in a row by the sum of that row.

According to [**41**, p.95], these evolution equations for $D(t)$ produce "...a selection of the most vigorously growing synapses at the expense of the others...". In addition, SODG seems to faithfully represent two-dimensional shapes that the standard SOM algorithm cannot properly grasp. As an example of this fact, see Figure 23, where, for SODG, the lines correspond to the three strongest adjacencies for each neuron, i.e., three highest values $D(i,j)$ for fixed i.

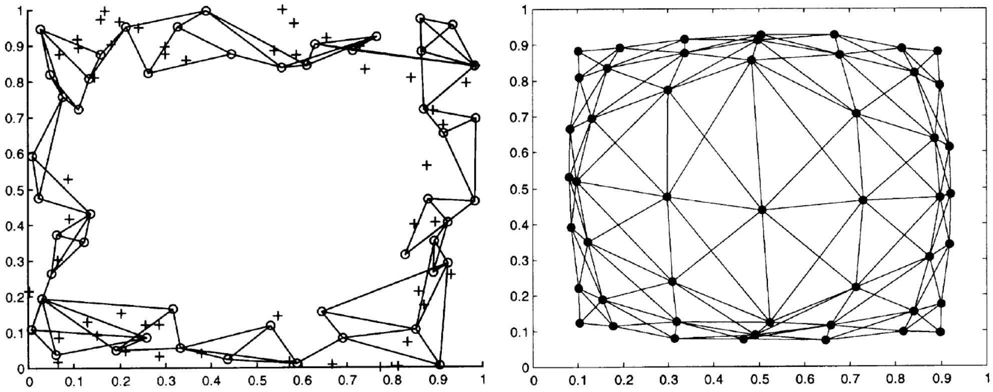


FIGURE 23. Hollow square learnt by SODG (left) and SOM (right). Source [**2**, Figure 8]

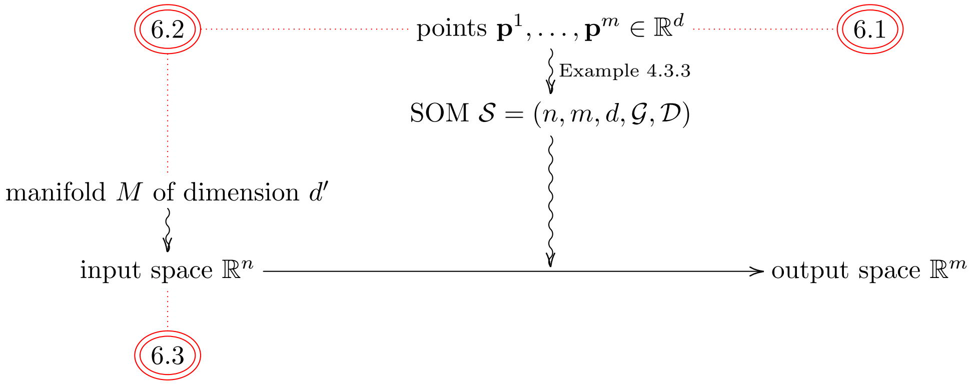
CHAPTER 6

Summary of works

In this chapter, we review three works undertaken by the Ph.D. student. These works have been published in [**38**], [**18**] and [**19**] and may be succinctly described as follows,

(1) In [**38**], we study the performance of SOMs arising from alternative tessellations, see also Subsection 5.2.1 and Appendix C. Details are given in Section 6.1.
(2) In [**18**], we investigate the relation between the intrinsic dimension of the data d' and the inner dimension d of the SOM. More insight is provided in Section 6.2.
(3) In [**19**], we introduce a new type of SOM, the so-called Forbidden Region SOM or FRSOM: displacement of the prototypes in the input space avoids predefined forbidden regions, see 6.3.

Now we clarify how these works fit in the setup of SOMs introduced in earlier chapters. To that aim, in the following diagram we represent the main elements that take part in the SOM, and we indicate the components that are directly related to each of the three works,


6.2
points $\mathbf{p}^1, \dots, \mathbf{p}^m \in \mathbb{R}^d$
6.1
Example 4.3.3
SOM $\mathcal{S} = (n, m, d, \mathcal{G}, \mathcal{D})$
manifold M of dimension d'
input space $\mathbb{R}^n$
output space $\mathbb{R}^m$
6.3

6.1. Grid metric

In Subsection 5.2.1, edge-to-edge monohedral vertex transitive tesselations were considered, see also Appendix C. As discussed there, we may allow n-gons with different n's and still impose the edge-to-edge condition and the vertex transitivity condition. This way one obtains the Archimidean or semi-regular tesselations. Recall that there exist three regular tessellations of the plane, denoted 4^4, 3^6 and 6^3, see Figure 19 or Figure 24 below. Besides these three regular tessellations, there are 8 semi-regular tessellations of the plane, denoted $3^4.6$, $3^3.4^2$, $3^2.4.3.4$, $3.4.6.4$, $3.6.3.6$, 3.12^2, $4.6.12$ and 4.8^2. They are shown in Figure 25 below.

	Sqr	Hex	Tri	Prism	Cairo
MSE	0	0	4	0	**6**
MTR	2	0	**6**	3	3

Table 4. Statistically significant victory counts for machine learning datasets experiments. Best results are marked in **bold**.

	Sqr	Hex	Tri	Prism	Cairo
MSE	42	52	**31**	54	**31**
MTR	51	53	**25**	39	42

Table 5. Rank sums for machine learning datasets experiments. Best results are marked in **bold**.

6.2. Lattice dimensionality

Let $\mathcal{S} = (n, m, d, \mathcal{G}, \mathcal{D})$ be a SOM of intrinsic dimension d and consider input samples lying in a topological manifold $M \subseteq \mathbb{R}^n$ of dimension d'. In fact, $d = 2$, i.e., plane grid, is by far the most used in practice. One-dimensional and three-dimensional SOMs are less used because they are less oriented towards visualization. In the work [**18**], the Ph.D. student analysed the relative merits of the cases $d = 1$, $d = 2$ and $d = 3$ for the distinct values of the intrinsic dimension of the samples d'. In this section, we present a summary of this work. All SOMs in this section are built as in Example 4.3.3 employing a linear ($d = 1$), square ($d = 2$) or cubic ($d = 3$) grid.

6.2.1. Lattice dimensionality: Theory. Four different theoretical analyses were carried out:

(1) Global analysis: consider samples $\mathbf{x}^1, \ldots, \mathbf{x}^M$ and prototypes $\mathbf{w}^1, \ldots, \mathbf{w}^m$. As explained in [**18**, 2.2, 3.1], the SOM algorithm seeks to minimize the following energy function,

$$\mathcal{E} = \mathcal{E}(\mathbf{w}^1, \ldots, \mathbf{w}^m) = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{m} P(i,j) \sum_{k=1}^{m} \theta(j,k) ||\mathbf{x}^i - \mathbf{w}^k||^2, \tag{19}$$

where $P(i,j)$ was defined in Equation (16) and $\theta(j,k) = e^{-D^2(j,j_0)}$ is a time-independent neighbourhood function. In turn, this formula may be decomposed, depending on the dimension d of the SOM, as follows,

$$\mathcal{E} = E_0 + E_1 + \ldots + E_d,$$

where $E_0 = MSE$ is given by Equation (15) and each summand E_i for $i = 1, \ldots, d$ is non-negative and corresponds to the distance between neurons when rearranged for intrinsic dimension i. So $E_0 = MSE$ correspond to the discrete "metric" ρ on the neurons $\{1, \ldots, m\}$ and to a purely competitive network, where

$$\rho(x,y) = \begin{cases} 0 & \textit{if } x = y, \\ \infty & \textit{if } x \neq y, \end{cases}$$

be chosen if possible, see Subsection 6.2.1(3). Emergence of complex patterns in the neuron density function of two-dimensional SOM has been discovered, see Figure [**18**, Figure 4].

- Experiments with real data indicate that competitive learning and one-dimensional topologies are the best SOMs overall, not only with respect to MSE but also to topological map quality. These results agree with the theory developed in Section 6.2.1.

From the above considerations, it can be said that one and three-dimensional SOMs are heavily underutilized, since they clearly outperform the standard two-dimensional SOM in many respects.

6.3. Forbidden regions

Recall that in the standard SOM, the value of the prototypes in the input space are updated following straight segments, see Equation (10). In the work [**19**], the Ph.D. student introduced an alternative SOM for which the prototypes are updated according to a forbidden region procedure. This means, in particular, that the straight segment is substituted by a piecewise linear path. In this section, we give a brief account of this work, and we refer the reader to the work [**19**] for full details.

6.3.1. Forbidden Region Avoidance. Let $\mathcal{B}$ be a collection of pairwise disjoint convex polyhedral sets $B_1, \dots, B_N$ that lie in the space $\mathbb{R}^n$. Each set $B_i \subset \mathbb{R}^n$ is termed a *barrier* and it is the convex hull of a set of points in $\mathbb{R}^n$,

$$B_i = \mathrm{Conv}(\mathbf{p}_{i,1}, \dots, \mathbf{p}_{i,n_i}) = \Big\{ \sum_{j=1}^{n_i} \alpha_j \mathbf{p}_{i,j} | \forall j, \alpha_j \geq 0, \text{ and } \sum_{j=1}^{n_i} \alpha_j = 1 \Big\},$$

where Conv stands for the convex hull of a set of points in $\mathbb{R}^n$.

In this setting, and given two points $\mathbf{a}$ and $\mathbf{b}$ of $\mathbb{R}^n$ that lie outside the barriers $\mathcal{B}$, the aim is to compute the shortest path that joins $\mathbf{a}$ to $\mathbf{b}$ and that avoids all barriers. Under the further condition that $n = 2$, i.e., that the ambient space is two dimensional, it is well known that the solution is a piece-wise linear path with breaking points only in extreme points of barriers [**14**, Lemma 15.1]. Moreover, this path may be found via the visibility graph and Dijsktra's algorithm as we explain below [**14**, Section 15.1].

The vertices of the visibility graph $\mathcal{V}$ are the extreme points of all barriers, $\bigcup_{i=1}^{N} \{\mathbf{p}_{i,1}, \dots, \mathbf{p}_{i,n_i}\}$, and $\mathcal{V}$ has an edge between two vertices if they are visible to each other, i.e., if the straight segment between them does not collide with the barriers. Here, moving along the boundaries of the barriers is allowed and the weight of the edge is the Euclidean distance between the vertices. Once the graph is constructed, Dijsktra's algorithm provides the shortest distance $d_{\mathcal{V}}(\mathbf{p}, \mathbf{q})$ between any pair of vertices $\mathbf{p}$ and $\mathbf{q}$ of $\mathcal{V}$. Thus, given the points $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$, the optimum path is recovered as follows:

(i) Determine the vertices $\{\mathbf{p}_1, \dots, \mathbf{p}_{n_a}\}$ of $\mathcal{G}$ that are visible from $\mathbf{a}$.
(ii) Determine the vertices $\{\mathbf{q}_1, \dots, \mathbf{q}_{n_b}\}$ of $\mathcal{G}$ that are visible from $\mathbf{b}$.
(iii) The shortest path and the shortest distance $d_{\mathcal{G}}(\mathbf{a}, \mathbf{b})$ are found by optimizing

$$\min_{i,j}(d(a, \mathbf{p}_i) + d_{\mathcal{V}}(\mathbf{p}_i, \mathbf{q}_j) + d(\mathbf{q}_j, b)), \tag{20}$$

where d is the Euclidean distance in $\mathbb{R}^2$.

CHAPTER 7

Conclusions

This work is a summary of the works developed by the Ph.D. student during the last five years. The results have been published in the form of three papers ([**38**], [**18**], [**19**]) and all of them have to do with a type of artificial neuronal network (ANN) known as self-organizing map (SOM). The focus has been placed on mathematical analysis of SOMs and on mathematically based variants of these networks. Experiments and statistical analysis has been carried out to tell apart significant improvements among the models tested.

In the report, a short account of these works may be found in Chapter 6. Before that, a general discussion about ANNs and an extended description of SOMs forms the contents of Chapter 4. As motivation and prelude to Chapter 6, in Chapter 5, some aspects of SOMs related to mathematical continuity, mathematical tessellations and the distance function were discussed. We divide this chapter into a compilation of the conclusions of each work and global conclusions.

7.1. Individual conclusions

The particular conclusion for each of the works are reproduced verbatim below. From [**38**, Conclusions]:

> Three alternative grid topologies for self-organizing maps have been proposed. Their choice has been guided by symmetry and vector quantization performance considerations rooted in the geometrical theory of tessellations. A theory of the grid energy of a self-organizing map topology and a parametric measure of its quality have been developed to account for the performance differences among topologies. Experiments have been carried out over unsupervised clustering and classification applications to compare them to the classical topologies used in the literature. Several quantitative performance measures have been obtained, and the statistical significance of the results has been computed. The experimental results are in accordance with the developed theory, which indicates that the proposed topologies are suitable in a wide range of situations. In particular, it has been found that Cairo and triangular topologies allow for a closer adaptation to the input dataset, which makes them adequate for many applications. The other alternatives are preferable in the cases where maintaining the map in an ordered state is more important than quantization error minimization.

From [**18**, Conclusions]:

> Three alternative grid topologies for self-organizing maps have been examined. A theoretical study of them has been carried out from several points of view. Experiments have been carried out over synthetic and real data to compare them. Several quantitative performance measures have been chosen to this end, and the statistical significance of the results has been computed. The results and the further discussion indicate that the 1D and 3D topologies are well suited to many datasets. This indicates that there is room to improve SOM-based systems by employing these relatively uncommon topologies.

From [**19**, Conclusion]

> In this paper, we have proposed FRSOM, a variant of SOM whose prototypes avoid prescribed forbidden regions. To the best of our knowledge, the FRSOM is the first SOM which is designed to avoid prespecified forbidden regions. The proposed learning algorithm for the FRSOM is guaranteed to keep all prototypes outside of the forbidden regions, while it still learns a topological map of the input distribution like the SOM. Therefore, all prototypes lie in meaningful regions of the input space. Unsupervised clustering experiments have been carried out on two kinds of data for SOM, FRSOM, SOMN, and ViSOM. The results show that, with statistical significance, FRSOM is more reliable and suitable in many situations, especially when the forbidden regions have a strong impact on the MSE. The advantage of the FRSOM is not only observed in the vector quantization performance, as measured by the MSE, but also in the quality of the learned topological maps, since the FRSOM produces less topological errors. Furthermore, FRSOM is essential in applications where the prototypes are required to lie outside some prescribed regions. The novelty of FRSOM opens several lines of future research, including application of the forbidden region approach to other well-known variants of SOM as the growing neural gas or the growing hierarchical SOM, further experiments in order to sharpen suitable situations for FRSOM, closer analysis of the relationship between forbidden regions and MSE as well as the development of an algorithm to automatically generate barriers for a given data set, and generalization of the forbidden region problem and of FRSOM to higher dimensions. The same region avoidance procedures that are presented in this paper can be employed for unsupervised competitive learning neural networks. In particular, any unsupervised learning network whose prototypes are updated to a linear combination of the old prototype and the current input vector, is suitable to employ the region avoidance scheme presented here. These extensions are left as future works.

7.2. Global conclusions

Assume we are studying samples that lie on a topological manifold $M \subseteq \mathbb{R}^n$ of some unknown dimension d'. Some of the aims of machine learning algorithms that examine these data are:

(a) Appropriate clustering of the data.
(b) Discovery of the unknown dimension d'.

Let $\mathcal{S} = (n, m, d, \mathcal{G}, D)$ be a SOM as introduced in Definition 4.3.1. If employing such a SOM to analyze the given data, we may conclude from the works described here, [**38**], [**18**], [**19**], that:

(A) Regarding (a), recall that the quality of the clustering is usually assessed either by the Mean Squared Error or some measure of Topology Preservation. Such quality may be improved by either using alternative grids to the standard ones or by modifying the SOM algorithm as to avoid (forbidden) regions that do not contribute significantly to the MSE.

(B) As expected, the internal dimension d that obtains the best results in terms of MSE is equal to the dimension of the data, $d = d'$. In general, low internal dimension d will produce better MSE results and high internal dimension d will yield better grasping of the data structure. This fact is partially due to that lower values for d generate stronger bonds to local minima of MSE.

Regarding topology preservation, we may deduce from earlier chapters that:

(C) Qualitatively, topology preservation in SOMs seems not to be settled, i.e., there is no accepted precise definition of what topology preservation should mean. On the contrary, from a quantitative point of view, there are several good measures of topology preservation for SOMs

From a more general perspective, we draw the following conclusions:

(D) By modifying the elements d, $\mathcal{G}$ and D of the SOM $\mathcal{S}$ as well as the SOM algorithm, there is still room to obtain improvements in terms of MSE and topology preservation.

(E) There is not enough understanding of the behaviour of the SOM concerning the input distribution of samples, either locally or globally, in terms, for instance, of the input density of the samples and the output density of the neurons.

Next, we enumerate some topics that emanate from this document and that would be worth it pursuing in the future:

- Given a data cloud and a fixed number of neurons, compare the global minimum for MSE with the typical values of MSE produced by SOM algorithms (or other algorithms). Given the magnitude of the space of solutions, it is not clear whether the global minimum can be calculated except for highly symmetrical layouts or small size problems.
- Starting from the closed formula for the final positions of the neurons, study the behaviour of neuron densities in a theoretical way and compare it to the existing results in the bibliography. Several variables calculus could be utilized to find a theoretical answer for the optimal parameters, which should also be compared to prior results.
- As commented in 35, it would be exciting to study the topology preservation issue under the framework of (finite) topologies on the finite set of neurons, i.e., pre-orders.
- Regarding tessellations (see 5.2.1,6.1 and C), there is a huge territory to explore, as tessellations of hyperbolic spaces or spheres, non-orientable spaces as the projective plane, etc. As we have already seen, switching

to a non-standard grid may result in an improvement of the SOM efficiency, and it would stimulating to consider grids immersed in alternative topological spaces.
- Regarding topology preservation, try to export the qualitative measures of topology preservation in SOMs to the formal setup of mathematics.

The developments contained in this report are a contribution to some mathematical aspects and mathematical alternatives to the classical SOM algorithm. As aforementioned, there exist an enormous amount of papers on this subject and the trend seems to keep going. Besides, the coming of Big Data and its need for analysis through neural networks will likely maintain SOM research at a high level. It is interesting to wonder whether there exist an upper limit for the efficiency of SOMs in terms, for instance, of MSE, and how to attain such limit through modifications of the algorithm. On the other hand, it is also intriguing to be unaware of the next "generation" of neuronal networks and in which way will they be or not based on particular aspects of biological systems.

Bibliography

[1] A.A. Akinduko, E.M. Mirkes, A.N. Gorban, *SOM: Stochastic initialization versus principal components*, Information Sciences, Volumes 364–365, 2016, Pages 213-221.

[2] D. Alahakoon, S. Halgamuge, B. Srinivasan, *Dynamic self-organizing maps with controlled growth for knowledge discovery*, IEEE Transactions on Neural Networks, vol. 11, 2000, pages 601–614.

[3] A. Asgary, A.S. Naini, J. Levy, *Modeling the risk of structural fire incidents using a self-organizing map*, Fire Safety Journal, 2012, 49, 1–9.

[4] A. Azcarraga, S. Manalili, *Design of a structured 3D SOM as a music archive*, In: Jorma Laaksonen and Timo Honkela, editors, Advances in Self-Organizing Maps, volume 6731 of Lecture Notes in Computer Science, pp. 188–197. Springer Berlin Heidelberg, 2011.

[5] J. Barmak, *Algebraic topology of finite topological spaces and applications*, Lecture Notes in Mathematics, 2032. Springer, Heidelberg, 2011.

[6] H.U. Bauer, M. Herrmann, T. Villmann, *Neural maps and topographic vector quantization*, Neural Networks, Volume 12, Issues 4–5, 1999, pages 659-676.

[7] H.U. Bauer, K. Pawelzik, *Quantifying the neighborhood preservation of Self-Organizing Feature Maps*, IEEE Trans. on Neural Networks 3, 1992, 570–579.

[8] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[9] Y. Borisovich, N. Bliznyakov, Y. Izrailevich, T. Fomenko, *Introduction to topology*, Translated from the Russian by Oleg Efimov. "Mir", Moscow, 1985.

[10] P.G. Ciarlet, *Introduction to numerical linear algebra and optimisation, With the assistance of Bernadette Miara and Jean-Marie Thomas*, Translated from the French by A. Buttigieg. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 1989.

[11] M. Cottrell, E. de Bodt, M. Verleysen, *A Statistical Tool to Assess the Reliability of Self-Organizing Maps*, 2001, In: Advances in Self-Organising Maps. Springer, London

[12] M. Cottrell, J.C. Fort, G. Pagés,*Theoretical aspects of the SOM algorithm*, Neurocomputing, 21(1-3), pp. 119–138, 1998.

[13] H.S.M. Coxeter, *The Non-Euclidean Symmetry of Escher's Picture 'Circle Limit III'*, Leonardo, vol. 12, no. 1, 1979, pp. 19–25.

[14] M. de Berg, M. van Kreveld, M. Overmars, O. Schwarzkopf,*Computational geometry*, Algorithms and applications. Springer-Verlag, Berlin, 1997.

[15] J. Demšar, *Statistical comparisons of classifiers over multiple data sets*, Journal of Machine Learning Research, 2006, 7(1), p. 1-30.

[16] O.J. Dunn, *Multiple comparisons among means*, Journal of the American Statistical Association, 1961, 56, p. 52-64.

[17] A. Díaz Ramos, *Redes neuronales no supervisadas con topología dinámica para la segmentación de imágenes en color*, Proyecto de Fin de Carrera (Ingeniero Técnico en Informática de Sistemas), 2011.

[18] A. Díaz Ramos, E. López-Rubio and E. J. Palomo, *The role of the lattice dimensionality in the self-organizing map*, Neural Network World, Volume 28, 2018, p. 57–86.

[19] A. Díaz Ramos, E. López-Rubio and E. J. Palomo, *The Forbidden Region Self-Organizing Map Neural Network*, IEEE Transactions on Neural Networks and Learning Systems, 2019 March 18, p. 1-11.

[20] J.A. Flanagan, *Self-organized criticality and the self-organizing map*, Physical Review E - Statistical, Nonlinear, and Soft Matter Physics, 63(3 II), pp. 361301–361306, 2001.

[21] J.C. Fort, *SOM's mathematics*, Neural Networks, Volume 19, Issues 6–7, 2006, Pages 812-816.

[22] M. Friedman, *The use of ranks to avoid the assumption of normality implicit in the analysis of variance*, Journal of the American Statistical Association, 1937, 32, p. 675-701.

[23] M. Friedman, *A comparison of alternative tests of significance for the problem of m rankings*, Annals of Mathematical Statistics, 1940, 11, p. 86-92.

[24] B. Grünbaum, G. C. Shephard, *Tilings by regular polygons*, Mathematics Magazine, Vol. 50, No. 5, Nov., 1977, pp. 227–247.

[25] B. Grünbaum, G. C. Shephard, *Tilings and patterns*, W. H. Freeman and Company, New York, 1987.

[26] D. Hebb, *The Organization of Behavior*, 1949, New York: Wiley & Sons.

[27] A.K. Jain, M.N. Murty, P.J. Flynn, *Data Clustering: A Review*, ACM Comput. Surv., Sept. 1999, 31, 3, 264–323.

[28] J.H. Kaas, R.J. Nelson,M. Sur M, C.S. Lin, M.N. Merzenich MM, *Multiple Representations of the Body within the Primary Somatosensory Cortex of Primates*, Science, 1979, 204:521-523.

[29] A. Kaever, T. Lingner, K. Göbel, I. Feussner, P. Meinicke, *MarVis: A tool for clustering and visualization of metabolic biomarkers*, BMC Bioinformatics, 10, 2009.

[30] S. Kaski, J. Kangas, T. Kohonen, *Bibliography of Self-Organizing Map (SOM) Papers: 1981-1997*, 1998, Neural Computing Surveys, 1, 102-350.

[31] D. Kaye, I. Ivrissimtzis, *Implicit surface reconstruction and feature detection with a learning algorithm*, In: John Collomosse and Ian Grimstead, editors, Theory and Practice of Computer Graphics, pp. 127–130. European Association for Computer Graphics, 2010.

[32] K. Kiviluoto, *Topology preservation in self-organizing maps*, In: Proceedings IEEE International Conference on Neural Networks, Bruges, June 3–6, 1996, pp. 294–299.

[33] T. Kohonen, *Automatic formation of topological maps of patterns in a self-organizing system*, in Erkki Oja and Olli Simula, editors, Proc. 2SCIA, Scand. Conf. on Image Analysis, pages 214-220, Helsinki, Finland, 1981, Suomen Hahmontunnistustutkimuksen Seura r. y.

[34] T. Kohonen, *Self-organized formation of topologically correct feature maps*, Biological Cybernetics, 1982, 43 , 59–69.

[35] T. Kohonen, *Self-Organizing Maps*, 2001, Springer-Verlag Berlin Heidelberg.

[36] W.H. Kruskal, W. A. Wallis, *Use of Ranks in One-Criterion Variance Analysis*, Journal of the American Statistical Association, vol. 47, no. 260, 1952, pp. 583–621.

[37] D. Liu, Z. Ruirui, J. Liang, Z.J. Juan, *Application of selforganizing feature maps neural network on hydrographic zones partitioning*, Int. J. Advancements Comput. Technol., vol. 4, no. 8, pp. 232–239, 2012.

[38] E. López-Rubio, A. Díaz Ramos, *Grid topologies for the self-organizing map*, Neural Networks, Volume 56, August 2014, Pages 35–48.

[39] E. López-Rubio, R.M. Luque-Baena, E. Domínguez, *Foreground detection in video sequences with probabilistic self-organizing maps*, International Journal of Neural Systems, 21(3), pp. 225–246, 2011.

[40] E. López-Rubio, J. Muñoz-Pérez, J.A. Gómez-Ruiz, *Invariant pattern identification by self-organising networks*, Pattern Recognit. Lett., vol. 22, no. 9, pp. 983–990, 2001.

[41] E. López-Rubio, J. Muñoz-Pérez, J.A. Gómez-Ruiz, *Self-Organizing Dynamic Graphs*, Neural Process. Lett. 16, 2, Oct. 2002, 93–109.

[42] E. Lopez-Rubio, E.J. Palomo, *Growing Hierarchical Probabilistic Self-Organizing Graphs*, IEEE Transactions on Neural Networks, vol. 22, no. 7, July 2011, pp. 997–1008.

[43] T. Martinetz, K. Schulten, *Topology representing networks*, Neural Networks, Volume 7, Issue 3, 1994, Pages 507-522.

[44] E. Merényi, K. Tasdemir, L. Zhang, *Learning Highly Structured Manifolds: Harnessing the Power of SOMs*, In: Biehl M., Hammer B., Verleysen M., Villmann T. (eds) *Similarity-Based Clustering*, Lecture Notes in Computer Science, 2009, vol 5400. Springer, Berlin, Heidelberg

[45] J.R. Munkres, *Topology: a first course*, Second Edition, Prentice-Hall, Inc., Upper Saddle River, NJ, 2000.

[46] M. Oja, S. Kaski, T. Kohonen, *Bibliography of Self-Organizing Map (SOM) Papers: 1998-2001 Addendum*, Neural Computing Surveys, 2003, 3, 1-56.

[47] S. Okajima, Y. Okada, *Treecube+3D-ViSOM: Combinational visualization tool for browsing 3D multimedia data*, In: 11th International Conference on Information Visualization, pp. 40–45, 2007.

[48] S.Y. Oudot, *Persistence Theory: From Quiver Representations to Data Analysis*, American Mathematical Society, 2015, Mathematical Surveys and Monographs.

[49] E.J. Palomo, D. Elizondo, G. Brunschwig, *Land usage classification: A hierarchical neural network approach*, J. Agricult. Sci., vol. 152, no. 5, pp. 817–828, 2014.

[50] Y.S. Park, R. Céréghino, A. Compin, S. Lek,*Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters*, Ecological Model., vol. 160, no. 3, pp. 265–280, 2003.

[51] V.M. Panaretos, *Statistics for mathematicians. A rigorous first course.*, Compact Textbooks in Mathematics. Birkhäuser/Springer, 2016.

[52] M. Pöllä, T. Honkela, T. Kohonen, *Bibliography of Self-Organizing Map (SOM) Papers: 2002-2005 Addendum. TKK Reports in Information and Computer Science*, Helsinki University of Technology, Report TKK-ICSR24, 2009.

[53] H. Ritter, T. Martinetz, and K. Schulten, *Neural Computation and Self-Organizing Maps; An Introduction*, 1992 (1st ed.), Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[54] H. Ritter, *Self-Organizing Maps on non-euclidean Spaces*, In: Editor(s): Erkki Oja, Samuel Kaski, *Kohonen Maps*, Elsevier Science B.V., 1999, Pages 97-109,

[55] S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall Series in Artificial Intelligence, Third Edition, 2010.

[56] T. Sakai, *Riemannian geometry*, Translated from the 1992 Japanese original by the author. Translations of Mathematical Monographs, 149. American Mathematical Society, Providence, RI, 1996.

[57] J. Schmidhuber, *Deep learning in neural networks: An overview*, Neural Networks, Volume 61, January 2015, Pages 85–117.

[58] D.J. Sheskin, *Handbook of parametric and nonparametric statistical procedures*, 3rd ed., Boca Raton, FL, Chapman & Hall/CRC.

[59] A. Skupin, R. Hagelman, *Visualizing demographic trajectories with self-organizing maps*, GeoInformatica, vol. 9, no. 2, pp. 159–179, 2005.

[60] M. Van Der Voort, M. Dougherty, S. Watson, *Combining Kohonen maps with ARIMA time series models to forecast traffic flow*, Transportation Research Part C: Emerging Technologies, 1996, 4(5), 307–318.

[61] T. Villmann, *Topology Preservation in Self-Organizing Maps*, In: Editors Erkki Oja, Samuel Kaski, *Kohonen Maps*, Elsevier Science B.V., 1999, Pages 279-292.

[62] T. Villmann, R. Der, M. Herrmann, and T. M. Martinetz, *Topology preservation in self-organizing feature maps: exact definition and measurement*, IEEE Transactions on Neural Networks, vol. 8, no. 2, pp. 256-266, March 1997.

[63] D. Williams, *Weighing the odds*, A course in probability and statistics. Cambridge University Press, Cambridge, 2001.

[64] H. Yin, *Visom - a novel method for multivariate data projection and structure visualization*, IEEE Transactions on Neural Networks , vol. 13, no. 1, pp. 237–243, Jan 2002.

[65] H. Yin, N.M. Allinson, *Self-organizing mixture networks for probability density estimation*, IEEE Transactions on Neural Networks, vol. 12, no. 2, pp. 405-411, 2001.

[66] Artificial intelligence: Go master Lee Se-dol wins against AlphaGo program, BBC News Online. 13 March 2016.

[67] AlphaZero AI beats champion chess program after teaching itself in four hours, The Guardian Online, 7 December 2017.

APPENDIX A

Topology

Here we give a brief account on basic notions about topology, for more details see [**9**] and [**45**].

Definition A.0.1. A *topological space* is a pair (X, τ) such that X is a set and τ is a collection of subsets of X satisfying the following properties:

(1) $\emptyset, X \in \tau$,
(2) arbitrary union of elements from τ belongs to τ, and
(3) finite intersection of elements from τ belongs to τ.

The subsets in the collection τ are said to be *open* (in X) and τ is called a *topology* on X.

Example A.0.2. Let X be a set and let τ be the collection of *all* subsets of X, then (X, τ) is a topological space and τ is termed the *discrete* topology on X.

Note that if (X, τ) is a topological space and the singleton $\{x\}$ belongs to τ for all $x \in X$, then τ is the discrete topology by (ii) in Definition A.0.1. If $Y \subset X$ is a subset of X and τ is a topology on X, we may define a topology on Y as follows.

Definition A.0.3. Let (X, τ) be a topological space and let $Y \subseteq X$. The *subspace* or *inherited* topology τ_Y on Y is given by the following collection of subsets of Y,

$$\tau_Y = \{U \cap Y,\, U \in \tau\}.$$

Next, we introduce the notion of continuous map between topological spaces.

Definition A.0.4. Let (X, τ) and (Y, σ) be topological spaces. A map $f\colon X \to Y$ is said to be *continuous* if

$$f^{-1}(V) \in \tau \text{ for all } V \in \sigma.$$

Below we define closed subsets, and then characterize continuity in terms of this kind of subsets.

Definition A.0.5. Let τ be a topology on the set X. Then $U \subseteq X$ is *closed* if $X \setminus U$ is open. If $A \subseteq X$ is an arbitrary subset of X, then:

(1) its *closure*, denoted $cl(A)$, is the smallest closed subset of X containing A.
(2) its *interior*, denoted $int(A)$, is the largest subset of A which is open in X.

Proposition A.0.6. *Let (X, τ) and (Y, σ) be topological spaces and let $f\colon X \to Y$ be a map. Then f is continuous if and only if*

$$f^{-1}(C) \subseteq X \textit{ is closed in } X \textit{ for all } C \subseteq Y \textit{ closed in } Y.$$

Proposition A.0.7. *Let (X, τ) and (Y, σ) be topological spaces and let $f\colon X \to Y$ be a map. Then f is continuous if and only if*

$$f(cl_X(A)) \subseteq cl_Y(f(A)) \text{ for all } A \subseteq X,$$

where cl_X and cl_Y denote the closure operators in X and Y respectively.

Definition A.0.8. Let (X, τ) and (Y, σ) be topological spaces and let $f\colon X \to Y$ be a map. We say that f is a *homeomorphism*, and that (X, τ) and (Y, σ) are *homeomorphic*, if f is a bijection and both f and f^{-1} are continuous.

Metric spaces are a particular case of topological spaces.

Definition A.0.9. A *metric space* is a set X endowed with a map

$$\rho\colon : X \times X \to \mathbb{R} \cup \{\infty\}$$

satisfying the following properties,

(1) $\rho(x, y) \geq 0$ for all $x, y \in X$,
(2) $\rho(x, y) = 0$ if and only if $x = y$, for all $x, y \in X$,
(3) $\rho(x, y) = \rho(y, x)$ for all $x, y \in X$, and
(4) $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$ for all $x, y, z, \in X$.

The map ρ is called a *metric* on X, and the non-negative real number $\rho(x, y)$ is called the *distance* between the points x and y of X.

Example A.0.10. Let X be a set. The *discrete* metric on X is the metric ρ defined by

$$\rho(x, y) = \begin{cases} 0 & \textit{if } x = y, \\ 1 & \textit{if } x \neq y. \end{cases}$$

If (X, ρ) is a metric space, then the collection τ_ρ of all subsets which are union of sets of the form

$$B_\epsilon(x) = \{y \in X | \rho(x, y) < \epsilon\},$$

where $x \in X$, $\epsilon > 0$, is a topology on X, called the *metric topology.* The subset $B_\epsilon(x)$ is the *open ball* of radius ϵ and centre x.

Remark A.0.11. It is easy to see that the metric topology on any finite set X is the discrete topology.

Continuity between metric spaces can be characterized in terms of limits of converging sequences.

Proposition A.0.12. *Let (X, ρ_X) and (Y, ρ_Y) be metric spaces and let $f\colon X \to Y$ be a map. Then f is continuous for the metric topologies if and only if*

$$\{x_n\}_{n\in\mathbb{N}} \text{ converges to } x_0 \in X \Rightarrow \{f(x_n)\}_{n\in\mathbb{N}} \text{ converges to } f(x_0) \in Y.$$

To finish this section, we introduce isometries, which are, roughly speaking, homeomorphisms between metric spaces that preserve the distances.

Definition A.0.13. Let (X, ρ_X) and (Y, ρ_Y) be metric spaces and let $f\colon X \to Y$ be a map. We say that f is an *isometry* if

$$\rho_X(x, x') = \rho_Y(f(x), f(x')) \text{ for all } x, x' \in X.$$

We say that (X, ρ_X) and (Y, ρ_Y) are *isometric* if there exists a bijective isometry between them.

It is easy to prove that isometric metric spaces are homeomorphic spaces.

APPENDIX B

Probability and Statistics

Further details on the concepts introduced in this chapter may be found in the books [**63**] and [**51**].

Definition B.0.1. A *probability space* is a triple (Ω, Σ, μ) such that Ω is a set (the set of "events"), Σ is a collection of subsets of Ω and $\mu\colon \Sigma \to \mathbb{R}$ is a map satisfying the following properties,

(1) Σ is a *σ-algebra*, i.e., it satisfies,
   - (a) Ω in Σ,
   - (b) Σ is closed under complements, i.e., $U \in \Sigma \Rightarrow \Omega \setminus U \in \Sigma$, and
   - (c) Σ is closed under countable unions, i.e.,

$$U_n \in \Sigma \text{ for } n \in \mathbb{N} \Rightarrow \bigcup_{n\in\mathbb{N}} U_n \in \Sigma.$$

(2) μ is a *probability measure*, i.e., it satisfies the following properties,
   - (a) $\mu(U) \geq 0$ for all $U \in \Sigma$,
   - (b) $\mu(\Omega) = 1$,
   - (c) μ is countable additive, i.e., if $\{U_n\}_{n\in\mathbb{N}}$ are mutually disjoint subsets of Σ, then

$$\mu(\bigcup_{n\in\mathbb{N}} U_n) = \sum_{n=1}^{\infty} \mu(U_n).$$

The pair (Ω, Σ) is called a *measurable space*. To define random variable, we restrict ourselves to real-valued random variables, i.e., we choose as target measurable space the pair $(\mathbb{R}, \Sigma_{\mathbb{R}})$, where $\Sigma_{\mathbb{R}}$ is the Borel σ-algebra on $\mathbb{R}$.

Definition B.0.2. Let (Ω, Σ, μ) be a probability space. A *random variable* is a *measurable map* $X : \Omega \to \mathbb{R}$, i.e., a map that satisfies that

$$X^{-1}((-\infty, x]) = \{w \in \Omega | X(w) \leq x\} \in \Omega \text{ for all } x \in \mathbb{R}.$$

The *distribution function* of the random variable X is the function $F_X : \mathbb{R} \to [0, 1]$ given by

$$F_X(x) = \mu(X^{-1}((-\infty, x]))$$

and the *probability measure* of the random variable X is

$$P_X(A) = \mu(X^{-1}(A)) \text{ for any Borel set } A \subseteq \mathbb{R}.$$

Note that, by Proposition A.0.6, continuous maps are random variables. We will omit the subindex X and write $F(x)$ and $P(A)$ instead of $F_X(x)$ and $P_X(A)$ whenever the random variable under study is clear from the context.

B.1. Test of Hypotheses

Along this section, we consider a fixed probability space (Ω, Σ, μ). In the application to SOMs, Ω consists of all possible outcomes (either final positions of neurons or MSEs) of a concrete SOM algorithm with fixed parameters acting on a fixed dataset and μ models the probability of the different outcomes.

Definition B.1.1. For a collection of independent identically distributed random variables $X_1, \ldots, X_n$, $X_i \colon \Omega \to \mathbb{R}$, with common distribution $F = F_{X_i}$, we write *iid random variables.* Typically, this distribution will depend on one or more parameters, $F = F_\theta = F(x;\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^p$.

Note that $X_1 \times \ldots \times X_n : \Omega^n \to \mathbb{R}^n$ is a measurable map. We denote the induced probability measure on $\mathbb{R}^n$ by $P_{X_1,\ldots,X_n}$ and recall that the (joint) distribution on $\mathbb{R}^n$ is

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F(x_1) \cdot \ldots \cdot F(x_n).$$

So the value $P_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ is the probability of obtaining the *random sample* $(x_1, \ldots, x_n)$. Next, we briefly discuss one population Test of Hypotheses. Later we will list common hypotheses tests for one and various populations/samples. So let $X_1, \ldots, X_n$ be iid random variables with common distribution $F = F_\theta$, $\theta \in \Theta$. Then consider a random sample $(x_1, \ldots, x_n)$ obtained for a particular value of the parameter $\theta \in \Theta$. We use Test of Hypotheses to ascertain whether the value of the parameter belongs to a particular subset of Θ. So write $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \emptyset$ and the two hypotheses:

$$H_0 \text{ (null hypothesis) } : \theta \in \Theta_0,$$
$$H_1 \text{ (alternative hypothesis) } : \theta \in \Theta_1.$$

Definition B.1.2. A *test statistic* is a random variable $T \colon \mathbb{R}^n \to \mathbb{R}$ and a *critical region* is a subset $C \subseteq \mathbb{R}$.

We *reject* the hypothesis H_0 if $T(x_1, \ldots, x_n) \in C$ and we *do not reject* the hypothesis H_0 if $T(x_1, \ldots, x_n) \notin C$. A Type I error consists of rejecting the hypothesis H_0 when it is true, i.e., when $\theta \in \Theta_0$. A Type II error consists of not rejecting the null hypothesis when it is false, i.e., when $\theta \in \Theta_1$. Typically, critical regions are parametrized by the so-called significance level.

Definition B.1.3. We call a value $\alpha \in (0, 1)$ *significance level*, and we assume that there exists a critical region C_α for each such value. Moreover, these subsets must satisfy the following,

$$\alpha \leq \alpha' \Rightarrow C_\alpha \subseteq C_{\alpha'}.$$

We only consider test statistics that comply with the Neyman-Pearson framework [**51**, Definition 4.9], i.e., with Equation (23) below. Notice that, in that Equation and in the definition of power in Definition B.1.4, the dependence on θ is through the random variable T.

Definition B.1.4. Let T be a test statistic and $\{C_\alpha\}_{\alpha \in (0,1)}$ a family of critical regions. We assume that T satisfies the following condition

$$\sup_{\theta \in \Theta_0} P_T(C_\alpha) \leq \alpha. \tag{23}$$

For such a test statistic T, we define its *power* as the following function on $\theta \in \Theta_1$,

$$\text{power}_T(\theta) = P_T(C_\alpha),\ \theta \in \Theta_1.$$

This means, on the one hand, that the probability of Type I Error is smaller than α for any $\alpha \in (0,1)$. On the other hand, $1 - \text{power}(\theta)$ gives the probability of Type II Error, so we want to maximize the power. Recall that, in the earlier definition, we may unfold P_T as follows,

$$P_T(C_\alpha) = P_{X_1,\ldots,X_n}(\{(x_1,\ldots,x_n) \in \mathbb{R}^n | T(x_1,\ldots,x_n) \in C_\alpha\}).$$

Next, we define the p-value of a sample.

Definition B.1.5. Let $(x_1,\ldots,x_n)$ a random sample of the iid random variables $X_1,\ldots,X_n$ with common distribution function $F = F_\theta$. The *p-value* of the sample is defined as

$$p(x_1,\ldots,x_n) = \inf\{\alpha \in (0,1) | T(x_1,\ldots,x_n) \in C_\alpha\}.$$

So the p-value of the sample if the smallest significance level for which, given that sample, we would reject the null hypothesis H_0. In particular, because of Equation 23, if the p-value is z, the probability of making a Type I Error is at most z. Figure 36 contains a summary of the notions introduced in this section. Common ways of defining critical regions for significance level α are

$$C_\alpha = \{t \in \mathbb{R} | t \geq t_{1-\alpha}\} \text{ or } C_\alpha = \{t \in \mathbb{R} | t \leq t_\alpha\},$$

where t_β is the β-quantile for the distribution F_T of T, i.e.,

$$F_T(t_\beta) = P_T((-\infty, t_\beta]) = \beta.$$

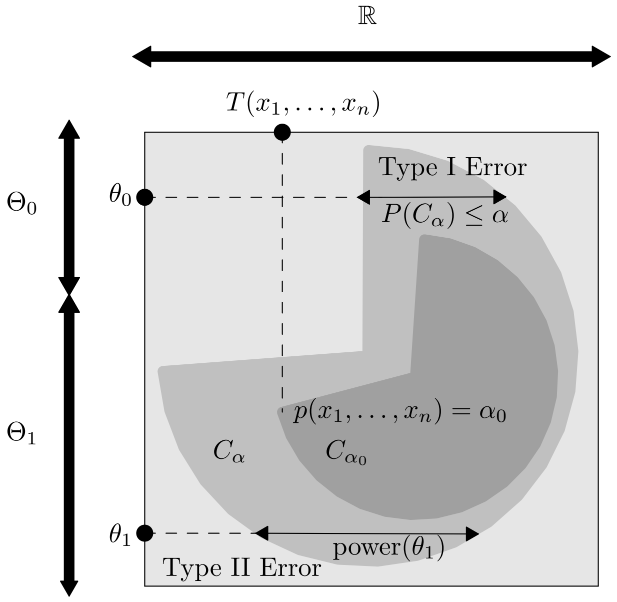


Figure 36. Diagram summarizing Test of Hypotheses.

B.2. Common test statistics

In this section, we sum up the test statistics employed in Chapter 6, see [**58**] for more details. Some common procedures for post-hoc analysis after one of these tests are given in Subsection B.2.5.

Test	Parametric	Populations	Type of samples
1-way ANOVA,B.2.1	Yes (normal distribution)	≥ 2	Independent
2-way ANOVA,B.2.2	Yes (normal distribution)	≥ 2	Dependent
Kruskall-Wallis,B.2.3	No	≥ 2	Independent
Friedman,B.2.4	No	≥ 2	Dependent

B.2.1. One-way ANOVA test, [**58**, Test 21]**.** Hypothesis evaluated with test: In a set of k independent samples (where $k \geq 2$), do at least two of the samples represent populations with different *mean* values? If we designate by θ_i the mean of the i-th population, then the hypotheses are,

$$\begin{aligned} H_0 \text{ (null hypothesis) } &: \theta_1 = \theta_2 = \ldots = \theta_k, \\ H_1 \text{ (alternative hypothesis) } &: \exists 1 \leq i, j \leq k | \theta_i \neq \theta_j. \end{aligned}$$

Procedure: let $\{x_{ij}\}_{1 \leq i \leq n, 1 \leq j \leq k}$ be the data,

(1) Compute $x_j = \frac{\sum_{i=1}^n x_{ij}}{n}$ for each $1 \leq j \leq k$, and $x_T = \frac{\sum_{j=1}^n x_j}{nk}$.
(2) Compute $SS_{BG} = n \sum_{j=1}^k (x_j - x_T)^2$ and $MM_{BG} = \frac{SS_{BG}}{k-1}$.
(3) Compute $SS_{WG} = \sum_{j=1}^k \sum_{i=1}^n (x_{ij} - x_j)^2$ and $MM_{WG} = \frac{SS_{WG}}{nk-k}$.
(4) The following test statistic T follows a Fischer distribution with parameters $k-1$ and $nk-k$, $F = F(k-1, nk-k)$,

$$T = \frac{MM_{BG}}{MM_{WG}}.$$

We reject H_0 at level of significance α if

$$F(\alpha) \leq T(x_{ij}),$$

where $F(\alpha)$ is the quantile defined by $P(F \geq F(\alpha)) = \alpha$. The p-value is given by

$$p(x_{ij}) = P(F \geq T(x_{ij})).$$

In the next figure, we represent some Fischer distributions.

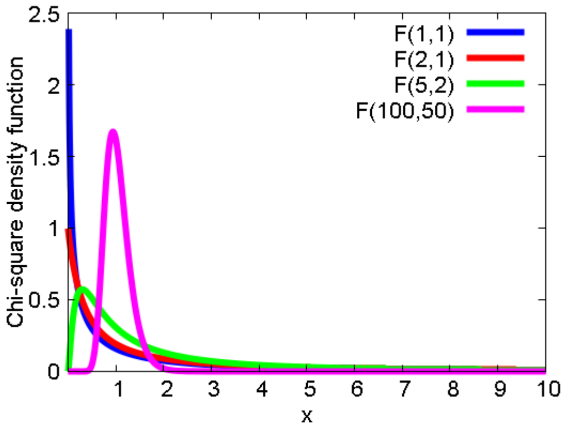


FIGURE 37. Fischer probability density function for several parameters.

B.2.2. Two-way ANOVA test, [**58**, Test 24]**.** Hypothesis evaluated with test: In a set of k dependent samples (where $k \geq 2$) from normal distributions, do at least two of the samples represent populations with different *mean* values? If we designate by θ_i the mean of the i-th population, then the hypotheses are,

$$H_0 \text{ (null hypothesis) } : \theta_1 = \theta_2 = \ldots = \theta_k,$$
$$H_1 \text{ (alternative hypothesis) } : \exists 1 \leq i, j \leq k | \theta_i \neq \theta_j.$$

Procedure: let $\{x_{ij}\}_{1\leq i\leq n, 1\leq j\leq k}$ be the data,

(1) Compute $x_j = \frac{\sum_{i=1}^n x_{ij}}{n}$, $x_i = \frac{\sum_{j=1}^k x_{ij}}{k}$ and $x_T = \frac{\sum_{j=1}^n x_j}{nk}$.

(2) Compute $SS_{BC} = n \sum_{j=1}^k (x_j - x_T)^2$ and $MM_{BC} = \frac{SS_{BC}}{k-1}$.

(3) Compute $SS_{res} = \sum_{j=1}^k \sum_{i=1}^n \big((x_{ij} - x_T) - (x_i - x_T) - (x_j - x_T)\big)^2$ and $MM_{res} = \frac{SS_{res}}{(n-1)(k-1)}$.

(4) The following test statistic T follows a Fischer distribution with parameters $k-1$ and $(n-1)(k-1)$, $F = F(k-1, (n-1)(k-1))$,

$$T = \frac{MM_{BC}}{MM_{res}}.$$

We reject H_0 at level of significance α if

$$F(\alpha) \leq T(x_{ij}),$$

where $F(\alpha)$ is the quantile defined by $P(F \geq F(\alpha)) = \alpha$. The p-value is given by

$$p(x_{ij}) = P(F \geq T(x_{ij})).$$

B.2.3. Kruskal-Wallis test, [**36**], [**58**, Test 22]**.** Hypothesis evaluated with test: In a set of k independent samples (where $k \geq 2$) from normal distributions, do at least two of the samples represent populations with different *median* values? If we designate by θ_i the median of the i-th population, then the hypotheses are,

$$H_0 \text{ (null hypothesis) } : \theta_1 = \theta_2 = \ldots = \theta_k,$$
$$H_1 \text{ (alternative hypothesis) } : \exists 1 \leq i, j \leq k | \theta_i \neq \theta_j.$$

Procedure: let $\{x_{ij}\}_{1\leq i\leq n, 1\leq j\leq k}$ be the data,

(1) Compute the matrix $\{r_{ij}\}_{1i\leq n, 1\leq j\leq k}$, where r_{ij} is the rank of x_{ij} within all data, i.e, within $\{x_{ij}\}_{1\leq i\leq n, 1\leq j\leq k}$.
(2) Compute $r_j = \frac{\sum_{i=1}^{n} r_{ij}}{n}$ for $1 \leq j \leq k$.
(3) Compute the Kruskal-Wallis test statistic:

$$T(x_{ij}) = \frac{12}{nk(nk+1)} \sum_{j=1}^{k} n r_j^2 - 3(nk+1).$$

(4) The test statistic T is approximately a chi-square distribution with $k-1$ degrees of freedom (for $n > 5$). We reject H_0 at level of significance α if

$$\chi^2_{k-1}(\alpha) \leq T(x_{ij}),$$

where $\chi^2_{k-1}(\alpha)$ is the quantile defined by $P(\chi^2_{k-1} \geq \chi^2_{k-1}(\alpha)) = \alpha$. The p-value is given by

$$p(x_{ij}) = P(\chi^2_{k-1} \geq T(x_{ij})).$$

The next figure represents some χ^2 distributions.

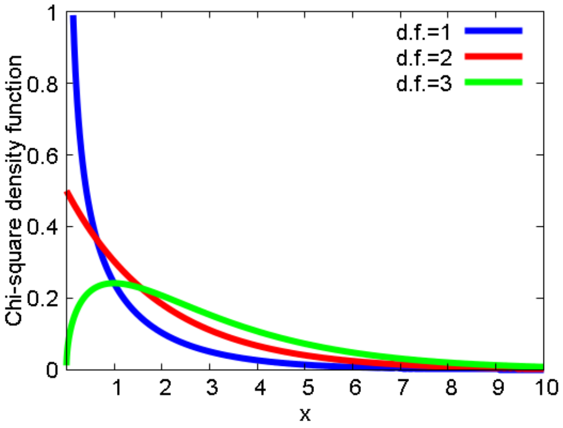


FIGURE 38. Chi-square probability density function for 1, 2 and 3 degrees of freedom.

B.2.4. Friedman test, [22], [58, Test 25]. Hypothesis evaluated with test: In a set of k dependent samples (where $k \geq 2$), do at least two of the samples represent populations with different *median* values? If we designate by θ_i the median of the i-th population, then the hypotheses are,

$$H_0 \text{ (null hypothesis) } : \theta_1 = \theta_2 = \ldots = \theta_k,$$
$$H_1 \text{ (alternative hypothesis) } : \exists 1 \leq i, j \leq k | \theta_i \neq \theta_j.$$

Procedure: let $\{x_{ij}\}_{1\leq i\leq n, 1\leq j\leq k}$ be the data,

(1) Compute the matrix $\{r_{ij}\}_{1i\leq n, 1\leq j\leq k}$, where r_{ij} is the rank of x_{ij} within $\{x_{ij}\}_{1\leq j\leq k}$.
(2) Compute $r_j = \frac{\sum_{i=1}^{n} r_{ij}}{n}$ for $1 \leq j \leq k$.

(3) Compute the Friedman test statistic:

$$T(x_{ij}) = \frac{12}{nk(k+1)} \sum_{j=1}^{k} r_j^2 - 3n(k+1).$$

(4) The test statistic T is approximately a chi-square distribution with $k-1$ degrees of freedom (for $n > 15$ or $k > 4$). We reject H_0 at level of significance α if

$$\chi^2_{k-1}(\alpha) \leq T(x_{ij}),$$

where $\chi^2_{k-1}(\alpha)$ is the quantile defined by $P(\chi^2_{k-1} \geq \chi^2_{k-1}(\alpha)) = \alpha$. The p-value is given by

$$p(x_{ij}) = P(\chi^2_{k-1} \geq T(x_{ij})).$$

B.2.5. Post-hoc analysis. For any of the tests above, if the null hypothesis is rejected, we infer that not all the means (medians) are equal. To find out which population's means (medians) are different, one can proceed comparing each pair of populations. This increases the probability of producing a Type I Error. More precisely, denote by α_O the overall probability of producing a Type I Error and by α the probability of producing a Type I Error in one of the c pairwise comparisons. Then we have,

$$\alpha_O = 1 - (1-\alpha)^c.$$

To cope with this disparity between α_O and α there exists several procedures. Below we briefly comment on the Dunn-Sidak and the Bonferroni-Dunn tests, see [**58**, Test 21.VI] for more details on post-hoc tests.

The Dunn-Sidak correction consists of solving α in the above equation, given the overall wished level of significance α_O,

$$\alpha = 1 - (1-\alpha_O)^{\frac{1}{c}}.$$

According to the Bonferroni-Dunn test, and because of the approximation

$$\alpha_O = 1 - (1-\alpha)^c \approx c\alpha,$$

we take $\alpha = \alpha_O/c$, if α_O is the overall wished level of significance and c is the number of pairwise comparisons to be carried out. The test statistic employed in each pairwise comparison is also modified. For more details, see [**58**, Test 21b] for one-way ANOVA test, [**58**, Test 24b] for two-way ANOVA test, [**58**, Test 22.VI.2] for Kruskall-Wallis test and [**58**, Test 25.VI.2] for Friedman test.

APPENDIX C

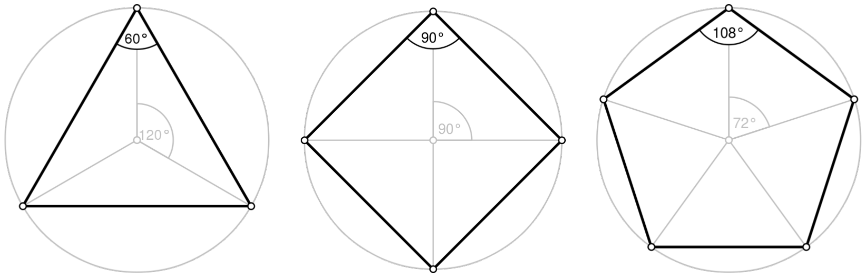
Tessellations

In this section, we succinctly describe the basic notions about tessellations or tilings, for more details see [**24**] or [**25**]. We define these notions for a general metric space (X, ρ), see Definition A.0.9.

Definition C.0.1. Let (X, ρ) be a metric space. A *tessellation* of (X, ρ) is a countable family of closed sets $\{T_1, T_2, \dots\}$, called tiles, which cover X without gaps or overlaps, more precisely:

(1) $X = \cup_{i \in \mathbb{N}} T_i$, and
(2) $int(T_i) \cap int(T_j) = \emptyset$ for $i \neq j$.

Recall that *int* stands for the interior of a subset, see Definition A.0.5. Next, we restrict to two-dimensional tessellations and, in addition, we restrict the kind of tiles that are allowed. By *n-gon*, we mean a regular convex polygon in the plane that has n edges, see Figure 39.



(a) 3-gon (b) 4-gon (c) 5-gon

Figure 39. Some n-gons. Source: Wikipedia

Definition C.0.2. Let $\{T_i\}_{i \in \mathbb{N}}$ a tessellation of the metric space (X, ρ). Assume each tile is isometric to one of finitely many fixed n-gons. We say that the tessellations if *edge to edge* is for every pair of tiles, T_i, T_j, with $i \neq j$, one of the following holds:

(1) $T_i \cap T_j = \emptyset$, or
(2) $T_i \cap T_j$ is a common vertex of T_i and T_j, or
(3) $T_i \cap T_j$ is a common edge of T_i and T_j.

The vertices in (2) and the edges in (3) are called vertices and edges of the tessellation.

Recall that the definition of isometric spaces was given in Definition A.0.13. The simplest kind of tesselation has only one isometry type of tile, as defined below.

Definition C.0.3. Let $\{T_i\}_{i\in\mathbb{N}}$ an edge to edge tessellation of the metric space (X,ρ). We say that the tessellation is *monohedral* if every tile is isometric to a fixed n-gon of the plane $\mathbb{R}^2$.

For edge to edge tessellations, we may define symmetries of the tessellation.

Definition C.0.4. Let $\{T_i\}_{i\in\mathbb{N}}$ an edge to edge tessellation of the metric space (X,ρ). A *symmetry* of the tessellation is an isometry of (X,ρ) that takes vertices to vertices and edges to edges.

Another point of view to classify tessellations is to look at the orbits of the vertices of the tessellations under the action of its symmetries. The simplest case is one orbit, as defined next.

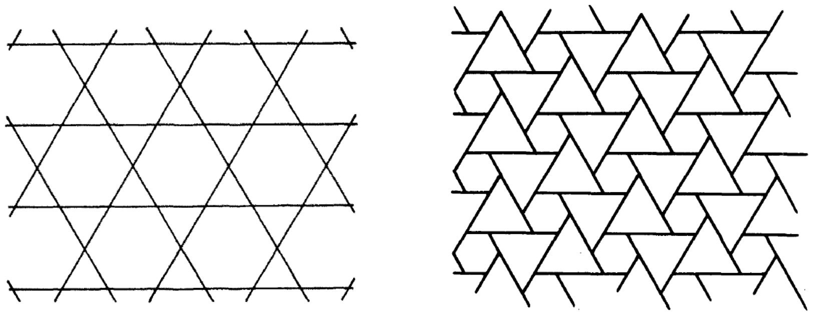
Definition C.0.5. Let $\{T_i\}_{i\in\mathbb{N}}$ an edge to edge tessellation of the metric space (X,ρ). We say that the tessellation is *vertex transitive* if for every two vertices of the tessellation, there exists a symmetry of the tessellation that takes one vertex to the other.

A vertex transitive tessellation may be described by writing down the number of edges of the polygons around any fixed vertex, in consecutive order. For instance, 4.4.4.4 is a tessellation with four squares around every vertex. For brevity, we shorten this notation when possible, as in 4^4. In earlier chapters, we only discuss edge to edge, vertex transitive tesselations, and below we give a name to this kind of tessellations.

Definition C.0.6. Let $\{T_i\}_{i\in\mathbb{N}}$ a tessellation of the metric space (X,ρ). We say that the tessellations is,

(1) *regular* if it is edge to edge, vertex transitive and monohedral.
(2) *semi-regular or Archimedean* if it is edge to edge and vertex transitive.

For the plane $\mathbb{R}^2$ (with the Euclidean distance), the regular tesselations are shown in Figure 24, and the semi-regular tessellations in Figure 25. For completeness, in Figure 40, we show non edge to edge and non vertex transitive tessellations of the plane.



(a) An edge to edge non vertex transitive tessellation.

(b) A non edge to edge tessellation.

Figure 40. Some tessellations that are not regular nor semi-regular. Source: [**24**, Figures 4 and 11].

Work 1

Grid topologies for the self-organizing map.

This chapter is a copy of the paper [**38**]:

E. López-Rubio, A. Díaz Ramos, *Grid topologies for the self-organizing map*, Neural Networks, Volume 56, August 2014, Pages 35-48.

https://doi.org/10.1016/j.neunet.2014.05.001

ABSTRACT

The original Self-Organizing Feature Map (SOFM) has been extended in many ways to suit different goals and application domains. However, the topologies of the map lattice that we can found in literature are nearly always square or, more rarely, hexagonal. In this paper we study alternative grid topologies, which are derived from the geometrical theory of tessellations. Experimental results are presented for unsupervised clustering, color image segmentation and classification tasks, which show that the differences among the topologies are statistically significant in most cases, and that the optimal topology depends on the problem at hand. A theoretical interpretation of these results is also developed.

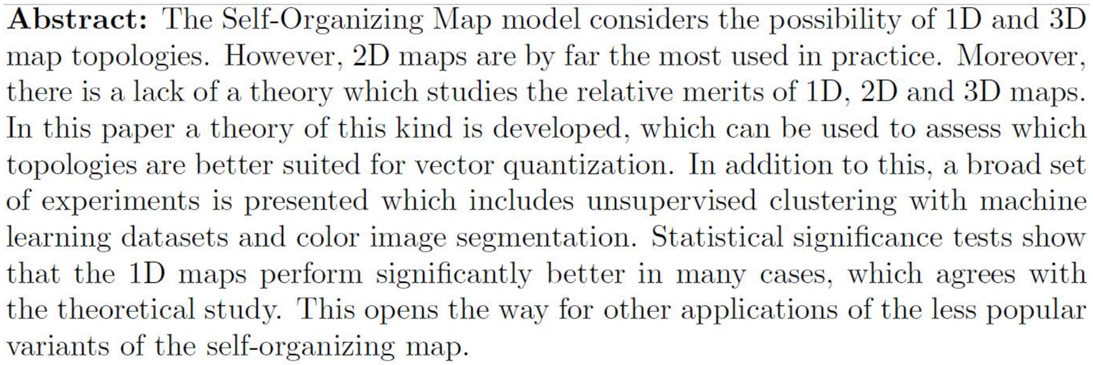
Work 2

The role of the lattice dimensionality in the self-organizing map.

This chapter is a copy of the paper [**18**]:

A. Díaz Ramos, E. López-Rubio and E. J. Palomo, *The role of the lattice dimensionality in the self-organizing map*, Neural Network World, Volume 28, 2018, p. 57–86.

DOI: 10.14311/NNW.2018.28.004



Abstract: The Self-Organizing Map model considers the possibility of 1D and 3D map topologies. However, 2D maps are by far the most used in practice. Moreover, there is a lack of a theory which studies the relative merits of 1D, 2D and 3D maps. In this paper a theory of this kind is developed, which can be used to assess which topologies are better suited for vector quantization. In addition to this, a broad set of experiments is presented which includes unsupervised clustering with machine learning datasets and color image segmentation. Statistical significance tests show that the 1D maps perform significantly better in many cases, which agrees with the theoretical study. This opens the way for other applications of the less popular variants of the self-organizing map.

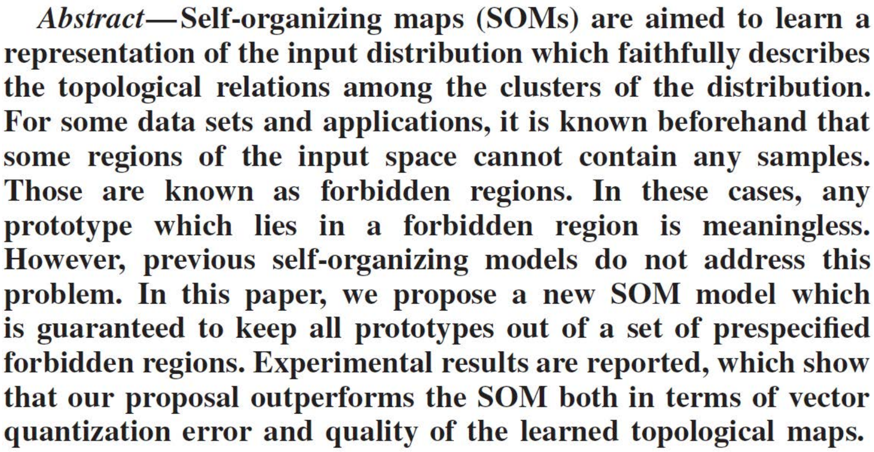
Work 3

The Forbidden Region Self-Organizing Map Neural Network

This chapter is a copy of the paper [**19**]:

A. Díaz Ramos, E. López-Rubio and E. J. Palomo, *The Forbidden Region Self-Organizing Map Neural Network*, IEEE Transactions on Neural Networks and Learning Systems, 2019 March 18, p. 1-11.

DOI: 10.1109/TNNLS.2019.2900091



Abstract—**Self-organizing maps (SOMs) are aimed to learn a representation of the input distribution which faithfully describes the topological relations among the clusters of the distribution. For some data sets and applications, it is known beforehand that some regions of the input space cannot contain any samples. Those are known as forbidden regions. In these cases, any prototype which lies in a forbidden region is meaningless. However, previous self-organizing models do not address this problem. In this paper, we propose a new SOM model which is guaranteed to keep all prototypes out of a set of prespecified forbidden regions. Experimental results are reported, which show that our proposal outperforms the SOM both in terms of vector quantization error and quality of the learned topological maps.**