Grounding Bayesian accounts of numerosity and variability effects in a similarity-based framework: the case of self-organising maps.

Grounding Bayesian Accounts of Numerosity and Variability Effects in a Similarity Based Framework: the case of Self-Organizing Maps

Valentina Gliozzi[(1)] and Kim Plunkett[(2)]
(1) Center for Logic, Language, and Cognition & Dipartimento di Informatica - Università di Torino
C.so Svizzera 185 - 10149 Torino, Italy
e-mail: valentina.gliozzi@unito.it
(2) Department of Experimental Psychology, University of Oxford
South Parks Road, Oxford OX1 3UD, UK
e-mail: kim.plunkett@psy.ox.ac.uk

Abstract

Category generalization is a central function in human cognition. It plays a crucial role in a variety of domains, such as learning, everyday reasoning, specialized reasoning, and decision making. Judging the content of a dish as edible, or a hormone level as healthy, are examples of category generalization. In this paper we propose self-organizing maps as possible candidates to explain the psychological mechanisms underlying category generalization. Self-organizing maps are psychologically and biologically plausible neural network models that can learn after limited exposure to positive category examples, without any need of contrastive information. Just like humans. They reproduce human behavior in category generalization, in particular the Numerosity and Variability effects, which are usually explained with Bayesian tools. Where category generalization is concerned, self-organizing maps deserve attention to bridge the gap between the computational level of analysis in Marr's hierarchy (where Bayesian models are often situated) and the algorithmic level of analysis in which plausible mechanisms are described.

Keywords: category generalization, self-organizing maps, connectionist modeling, Bayesian models

1. Introduction

Category generalization is a central function in human cognition that plays a crucial role in a variety of domains, such as learning, everyday reasoning, specialized reasoning, decision making. Judging the content of a dish as edible, or a hormone level as healthy, are examples of category generalization. In this

paper we propose a psychologically plausible, simple mechanism that can underlie such a central function and offers an account of two phenomena related to category generalization reported in the literature.

More formally, category generalization can be stated as follows: Starting from the observation that an object x belongs to a category C (or has the property P), how do we generalize C (or P) to objects other than x? As Shepard (1987) suggests, 'because any object or situation experienced by an individual is unlikely to recur in exactly the same form and context, psychology's first general law should be a law of generalization' (Shepard, 1987), p.1317.

Both probabilistic (Shepard, 1987; Anderson, 1991; Tenenbaum and Griffiths, 2001), and similarity-based (Tversky, 1977; Medin and Schaffer, 1978; Nosofsky, 1986) accounts of generalization have been proposed: the former suggesting that our generalization judgements are based on probabilistic reasoning, the latter suggesting that we base our generalization judgements on an evaluation of similarity between stimuli. According to the probabilistic account, the doctor's judgement on whether a new hormone level is healthy, similarly to a previously observed one, is based on a set of considerations involving the attribution of a probability to every possible range of values corresponding to healthy hormone levels, which in turn takes into account the likelihood of encountering the observed examples of healthy hormone levels if *that* was indeed the correct range of possible hormone values. In contrast, according to the similarity-based approach, the doctor's judgement on whether a new hormone level is healthy is based on a different set of considerations based on the similarity (for instance Euclidean similarity) of the new hormone level with respect to a previously built representation of the category of healthy hormone levels.

In this paper we show that a similarity-based account, that only uses Euclidean distance between stimuli feature vectors, can explain two effects of category generalization that were thought to be explainable only with Bayesian tools and that were indeed one of the arguments in favor of the superiority of a Bayesian account over similarity-based explanations of category generalization (Tenenbaum and Griffiths, 2001): these are the Numerosity and Variability Effects.

We illustrate this point by considering a specific similarity based kind of model, namely self-organizing maps (Kohonen et al., 2001), which is a particularly plausible neural network model that learns in an unsupervised way, which reproduces many experimental results on category formation (Schyns, 1991; Miikkulainen et al., 2005; Li et al., 2007; Gliozzi et al., 2009; Mayor and Plunkett, 2010), and which reflects basic constraints of plausible brain implementations, as extensively motivated by Miikkulainen et al. (2005). In the specific self-organizing map we consider, learning occurs from a limited set of positive examples (although this is not generally true for all self-organizing maps that may be trained with larger training sets-always in a non supervised way).

Our similarity-based account describes a plausible *mechanism* underlying category generalization. This complements Bayesian models of category generalization, as the one by Tenenbaum and Griffiths (2001), which are formulated at Marr's *computational* level of analysis, and describe what an optimal solu-

tion to the problem of category generalization would be. Approximations of Bayesian models of category generalization have been proposed for instance by Sanborn et al. (2010); Shi et al. (2010); Sanborn and Chater (2016), as well as some possible neural implementations of some of the ingredients of probabilistic inference (R. Zemel, 1998). Although the proposed approximations are certainly very interesting, as we will see, they are based on tools (such as Monte Carlo methods and stored hypothesis) that may imply a level of inferential power and memory that is unrealistic for the human cognizer. As a difference, here we propose a mechanism which is simpler and more readily executable by an individual with *limited memory and reasoning capacities*, and which uses tools, such as self-organizing maps, that are considered to reflect basic constraints of plausible brain implementations in the cortex. By the simplicity of the tools used, our model also differs from other Bayesian algorithms of incremental category formation that deal with more complex categories (e.g. Frermann and Lapata, 2016 that also uses methods like sequential Monte Carlo inference mechanisms). The other side of the coin of the simplicity of of our model is that it deals with simple stimuli, as Tenenbaum and Griffiths (2001). We will discuss this point in the Discussion section.

2. Numerosity and Variability effects in Category Generalization

Two effects have been highlighted that characterize category generalization. These effects are described by Tenenbaum and Griffiths (2001) in their extension of Shepard's (1987) universal law of generalization. Shepard describes an exponential decay in humans' inclination to consider a new stimulus as belonging to the same category as a previously considered one. The decay is proportional to the distance between the new stimulus and the examples already observed. Tenenbaum and Griffiths (2001) extend Shepard's analysis to the case where the examples already observed are multiple. In this case, the tendency to generalize a category to a new stimulus is affected by:

1. A *Numerosity Effect:* the more examples of a category observed within a given range, the lower the generalization outside that range.
2. A *Variability Effect:* the higher the variability in the set of observed examples of a category, the higher the generalization outside the examples' range.

These effects are amongst the primary motivations supporting Bayesian analyses of category generalization, in addition to the strongest argument in favor of Bayesian analyses of cognition, namely that they offer a single, coherent framework for understanding multiple cognitive phenomena, ranging from categorization to causal learning, to perception, to prediction and argumentation (see for instance N. et al. (2001), Griffiths et al., 2010). Bayesian models explain both the Numerosity and the Variability Effect in a very elegant way. However, by offering an explanation at Marr's computational level of analysis, they do not provide a description of a possible underlying mechanism: some possible correlates have been found for basic ingredients of Bayesian inference (R. Zemel,

1998), and some approximations have been proposed (Sanborn et al., 2010; Shi et al., 2010; Sanborn and Chater, 2016). However, as we will discuss below, a simple and plausible mechanism of all the computations involved is still needed.

In contrast to earlier claims (Tenenbaum and Griffiths, 2001), we argue that both Numerosity and Variability effects can be explained at an algorithmic level within a similarity-based framework that generalizes category membership out of few plausible examples, and in so doing exhibits the two effects. We provide a simple and psychologically plausible mechanism of category generalization, based on self-organizing maps, that captures the Numerosity and the Variability effects, thereby meeting the challenge of reconciling different levels of analysis in Marr's hierarchy, when considering the specific problem of category generalization.

Taking simplicity of a mechanism as a metric, we will argue that our similarity-based SOM mechanism (which as already specified is well-founded both from a biological and a psychological point of view) is more parsimonious than other mechanistic approximations of Bayesian analyses that have been proposed in the literature (Sanborn et al., 2010; Shi et al., 2010; Sanborn and Chater, 2016).

3. Bayesian analyses of category generalization

In contrast to more traditional models of cognition, which attempt to describe the psychological processes underlying basic cognitive abilities, Bayesian models of cognition are formulated at Marr's (1982) level of 'computational theory'. In Bayesian models cognitive tasks are described as computational problems posed by the environment. Humans are assumed to find optimal solutions to these problems using inductive, probabilistic inference: given some limited data, humans find the best possible hypothesis compatible with the evidence.

The problem of generalizing a category C to a new object y is formulated in this context as the problem of estimating the *probability* that y belongs to C, after observing X examples of C (Shepard, 1987; Tenenbaum and Griffiths, 2001)[1]. The probability estimation proceeds through two steps. *First,* the *posterior probability* $p(h|X)$ is computed by *Bayes' Rule* for all possible extensions h of C, where a possible extension is any consequential region, also called the *hypothesis (h)*. The Bayes Rule uses both priors and likelihoods: each hypothesis h in the space of hypotheses H has a *prior* probability $p(h)$, independent of any observed example. $p(X|h)$ is the *likelihood* of observing the examples X if the true extension of the category is indeed h. The likelihood obeys the *size principle*: the smaller the size $|h|$ of the consequential region h including all elements of X, the higher the probability of sampling all elements of X as examples of C, and therefore the higher the likelihood. The size principle is expressed as follows:

[1] Shepard (1987) considers the case in which there is one single category example, whereas Tenenbaum & Griffiths (2001) extend the approach to multiple examples. In this paper we refer to Tenenbaum & Griffiths' (2001) theory.

If $X = \{x_i\}$ (with $1 \leq i \leq n$), then

$$p(X|h) = \begin{cases} \frac{1}{|h|^n} & \text{if} \quad x_1 \ldots x_n \in h; \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The priors and likelihood are combined together in order to determine the *posterior* probability of h by Bayes' rule:

$$p(h|X) = \frac{p(X|h)p(h)}{\sum\limits_{h' in H} p(X|h')p(h')} \tag{2}$$

Second, once the posterior probability $p(h|X)$ for all possible extensions h has been computed by Bayes' Rule, the probability that y belongs to C is obtained by summing up the probability of all extensions containing y:

$$p(y \in C|X) = \sum_{h:y \in h} p(h|X) \tag{3}$$

In the hormone level example, if a doctor observes a *healthy* hormone level x, and she has to decide if another close-by hormone level y is still healthy, she first has to infer a probability distribution over the set of possible extensions of the healthy hormone level. Only then can she estimate the probability that the new hormone level y is still healthy.

The computations just defined entail the *Numerosity* and the *Variability Effects* mentioned above. The fact that Bayesian models of category generalization explain Numerosity and Variability effects in such an elegant way is a strong support for Bayesian analyses of category generalization (Tenenbaum and Griffiths, 2001).

Indeed, while the two effects also hold in human categorization, they are not easily captured by alternative theories of categorization such as exemplar and prototype theories, and back-propagation-based neural networks. In the next section we will argue that, on the contrary, a specific kind of psychologically plausible neural network, namely self-organizing maps, can account for these phenomena.

For the moment, it is worth re-emphasizing that Bayesian models of cognition are formulated at Marr's computational level of analysis, and do not describe a mechanism underlying these computations. Bayesian cognitive scientists recognize the need to establish a bridge between Marr's different levels of analysis, and acknowledge the identification of psychologically plausible processes underlying Bayesian inferences as a key challenge in their overall effort to reverse-engineer the human brain.

A possible neural implementation of probabilistic inference is described by R. Zemel (1998).

Furthermore, some proposals have been made to provide a mechanistic account of Bayesian models of generalization (Sanborn et al., 2010; Shi et al., 2010; Sanborn and Chater, 2016). In particular, Sanborn et al. (2010) propose Monte Carlo methods as possible mechanisms underlying Bayesian models. Shi

et al. (2010) propose mechanisms underlying Bayesian models that are based on exemplar models in which the stored exemplars correspond to hypotheses rather than stimuli. These models are certainly informative but imply a level of inferential power and memory that may be unrealistic for the human cognizer: the tools they employ, including both Monte Carlo methods and stored hypotheses, are potentially too complex to pass as psychologically plausible mechanisms.

Relatedly, there are Bayesian algorithms of incremental category formation that deal with more complex categories (e.g. Frermann and Lapata, 2016) that also rely on Monte Carlo inference mechanisms.

With respect to these previous proposals, in this paper, we attempt to formulate mechanisms which are simpler and more readily executable by an individual with *limited* memory and reasoning capacities.

Relations between Bayesian approaches and neural networks have been studied in the past (MacKay, 1995; Neal, 1996; McClelland, 1998). However, the neural network models considered by these researchers suffered from a general criticism raised against neural networks as models on human category formation: in order to properly categorize and generalize, these neural networks must be exposed to a large sample of training examples (rather than the small number of examples required by humans), and with both positive and negative examples (whereas humans can learn from positive examples only). In the following, we describe a neural network model that overcomes these shortcomings and provide a neuro-computationally plausible account of category generalization.

4. A similarity-based account of category generalization based on self-organizing maps

Our objective is to demonstrate that the Numerosity and Variability effects can be accommodated within a simple and psychologically plausible similarity-based account. This contrasts what was previously maintained Tenenbaum and Griffiths (2001). In short, we will show that in the model the numerosity and the variability of the known instances of a category affect the *quality* of the category representation: the numerosity of known examples of the category improves the precision of category representation whereas the variability of these examples diminishes this precision. In turn, the precision of the category representation is the core of our generalization judgements. Roughly speaking, we judge a new stimulus as belonging to a category by comparing the distance of the stimulus from the category representation to the precision of the category representation (see Definition 6 below).

We will use self-organizing maps (SOMs, introduced by Kohonen, 2001) as our instantiation of a similarity-based framework. We leave open the possibility that other similarity-based models can account for the category generalization properties that we consider.

In this paper, we focus on SOMs since these are particularly plausible neural network models that learn in a human-like manner. In particular:

- SOMs learn to organize stimuli into categories in an *unsupervised* way, without the need of a teacher providing a feedback. Even young infants are able to form categories spontaneously in the absence of corrective feedback (e.g., Younger and Cohen, 1986; Eimas and Quinn, 1994);

- SOMs can learn with just a few positive stimuli, without the need for negative examples or contrastive information, as humans do (e.g., Gliozzi et al., 2009, 2013). With these features, SOMs overcome some of the main criticisms raised against neural network approaches, namely that they have to be trained with large amount of data and with explicit contrastive information, in order to mimic characteristics of human category formation;

- SOMs reflect basic constraints of a plausible brain implementation in different areas of the cortex (Miikkulainen et al., 2005), and are therefore biologically plausible models of category formation;

- SOMs have proven to be capable of explaining experimental results. In particular:

  SOMs have been very successful at modelling the architecture of the primary visual cortex (Miikkulainen, Bednar, Choe, & Sirosh, 2005) where neighbouring neurons are responsive to similar orientations of the visual scene (Hubel & Wiesel, 1962);

  SOMs have been successfully used to model conceptual acquisition (Schyns, 1991 was the first such proposal);

  SOMs have been successfully used to simulate aspects of word learning: Miikkulainen (1997) introduced a SOM model of dyslexic and aphasic disorders (DISLEX) that, through selective lesioning procedures, was able to mimic language dysfunction, semantic slips, category-specific aphasic impairments and dyslexic behavior. Li et al.(2004; 2007) proposed two SOMs models of early lexical development. The DevLex models simulated the acquisition of linguistic categories, such as nouns, verbs and adjectives, as well as lexical confusions as a function of word density and semantic similarity. These models accounted for age-of-acquisition effects in the course of learning a lexicon and simulated the challenge infants face when they learn to articulate phonemic sequences of words.

4.1. The architecture of a SOM

SOMs consist of a set of neurons, or units, spatially organized in a grid (Kohonen et al., 2001), as in Figure 1.

Each map unit u is associated with a weight vector w_u of the same dimensionality as the input vectors. At the beginning of training, all weight vectors are initialized to random values, outside the range of values of the input stimuli. During training, the input elements are sequentially presented to all neurons of the map. After each presentation of an input x, the *best-matching unit* (BMU_x)

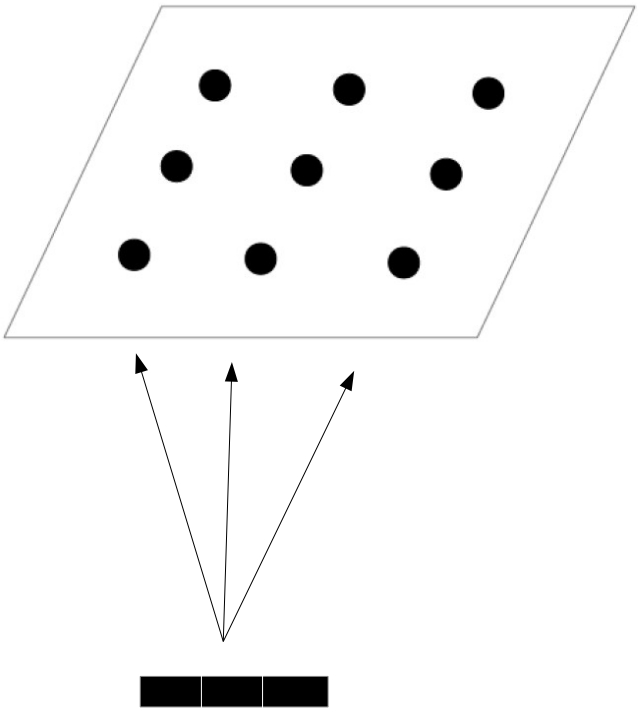

Figure 1: An example of SOM. The set of rectangles stands for the input presented to the SOM (in the example the input is three-dimensional). This is presented to *all* neurons of the SOM (these are the neurons-dots-in the upper grid) in order to find the BMU.

is selected: this is the unit i whose weight vector w_i is closest to the stimulus x (i.e. $i = \arg\min_j \|x - w_j\|$).

The weights of the best matching unit and of its surrounding units are updated in order to maximize the chances that the same unit (or the surrounding units) will be selected as the best matching unit for the same stimulus or for similar stimuli on subsequent presentations. At iteration $n+1$, the weights for neuron j are updated as follows:

$$w_j(n+1) = w_j(n) + \eta(n) h_{BMU_x,j}(n)(x - w_j(n)) \tag{4}$$

where η is the *learning rate*, and $h_{BMU_x,j}$ is the neighborhood function between the best-matching unit BMU_x and j. $h_{BMU_x,j}(n)$ is defined as:

$$h_{BMU_x,j}(n) = exp^{\frac{-d^2_{BMU_x,j}}{2\sigma(n)^2}} \tag{5}$$

where $d_{BMU_x,j}$ is the distance between BMU_x and j on the map's grid, and $\sigma(n)$ is the width of the gaussian.

Both the learning rate η and the width of the gaussian σ decrease with n.

This weight change has a twofold effect:

1. It reduces the distance between the best matching unit (and its surrounding neurons) and the incoming input, so that subsequently the same unit (and the surrounding ones) will most likely be the best matching unit for the same or similar inputs.
2. It organizes the map topologically so that the weights of close-by neurons are updated in a similar direction, and come to react to similar inputs.

The learning process is incremental: after the presentation of each input, the map's representation of the input (and in particular the representation of its best-matching unit) is updated in order to take into account the new incoming stimulus.

Although in standard self-organizing maps Kohonen et al. (2001) this weight update can go on for hundreds of epochs each consisting of the presentation of all the stimuli of the training set, we consider here a specific configuration of parameters (similar to Gliozzi et al. (2009) or Gliozzi et al. (2013)) with a high learning rate η(and a narrow neighborhood function h) that allows the map to learn after a *single* presentation of a stimulus of the training set (details in Section 4.2.1 below). In this way self-organizing maps learn following the same schedule than humans. This can be seen as a plausible mechanism by which humans form categories: starting from a first stimulus, that gives rise to an initial representation, the representation is updated each time a new stimulus is considered, in order to accommodate it. The final representation of the stimuli is the result of this iterative process. At the end of the whole process, the SOM has learned to organize the stimuli in a topologically significant way: similar inputs (with respect to Euclidean distance) are mapped to proximal areas in the map, whereas inputs which are far apart from each other are mapped to distal areas of the map.

Once the SOM has learned to categorize, we assess category generalization. We define the map's disposition to consider a new stimulus y as a member of a known category C as a function of the *distance* of y from the *map's representation* of C. We take a minimalist notion of what is the map's category representation: this is the ensemble of best-matching units corresponding to the known instances of the category. We will use BMU_C to refer to the map's representation of category C.

More precisely, we define category generalization as depending on two elements:

- the distance of the new stimulus y with respect to the category representation
- *compared to* the maximal distance from that representation of all known instances of the category

This captured by the following notion of *relative distance, rd for short,*:

$$rd(y, C) = \frac{min\|y - BMU_C\|}{max_{x \in C}\|x - BMU_x\|} \tag{6}$$

where $min\|y - BMU_C\|$ is the (minimal) Euclidean distance between y and C's category representation, and $max_{x \in C}\|x - BMU_x\|$ expresses the *precision* of category representation[2], and it is the (maximal) Euclidean distance between any known member of the category and the category representation.

With this definition, a given Euclidean distance from y to $C's$ category representation will give rise to a higher *relative distance rd* if the maximal distance between C and its known examples is low (and category representation is precise) than if it is high (and category representation is coarse).

We are now ready to define the *map's Generalization Degree* of category C membership to a new stimulus y. This is a function of the relative distance of Equation 6. The map's Generalization Degree exponentially decreases with the increase of the relative distance as follows [3]:

$$\text{Generalization Degree} = e^{-rd(y,C)} \tag{7}$$

It is worth observing that the above notion of relative distance (Equation 6) requires there to be a memory of some of the known instances of the category being used (this is needed to calculate the denominator in the equation). This

[2] $max_{x \in C}\|x - BMU_x\|$ is indeed inversely related to precision: the higher its value, the higher the maximal distance between the category representation and the worst represented stimulus, hence the lower the precision; and viceversa, the lower the precision, the higher the maximal distance between the category representation and the worst represented stimulus, and the higher the expression's value.

[3] The use of an exponential function of a distance when defining the categorization of a stimulus, or the activation of a unit when receiving a stimulus, is common in the neural network literature (see for instance Westermann and Mareschal, 2004; Mayor and Plunkett, 2010).

gives rise to a sort of hybrid model in which category representation and some exemplars coexist. An alternative way of formulating the same notion of relative distance would be to calculate *online* the distance between known category instance currently examined and the representation of the category being formed. Even in this alternative case, the Numerosity and Variability effects would still hold.

For a simple example of the use of the equations introduced in this section, see the Supplemental Material.

4.2. Modeling Numerosity and Variability effects

Do Numerosity and Variability impact the nature of the representations formed by self-organizing maps, and consequently, the Generalization Degree?

In order to answer these questions, we have run three sets of simulations.

In a first set of simulations, stimuli are points in a continuous metric psychological space (e.g., hormone levels) that vary along one dimension (as in Tenenbaum & Griffiths, 2001). In a second set of simulations, stimuli are points in a continuous metric psychological space that vary along two dimensions (as in Tenenbaum & Griffiths, 2001). Since in these first two simulations we have used a very simplified self-organizing map, that allowed us to make the point by considering a training set uniquely made of the stimuli under observation (rather than a richer training set including those stimuli), we have also run a third set of simulations with a more standard self-organizing map.

Each set of simulations compares the map's Generalization Degree in a Base Condition with respect to a Numerosity Condition (in which the numerosity of the stimuli augments but the range remains the same than in the Base Condition), and then the map's Generalization Degree in a Base Condition with respect to a Variability Condition (in which the number of the stimuli is the same than in the Base Condition but the range of the stimuli augments).

4.2.1. First simulation: Stimuli varying along one dimension.

In our first set of simulations we have used an architecture and a set of parameters similar to those used by (Gliozzi et al., 2009, 2013): a 3*3 hexagonal SOM, initialized to random values[4]. The map was trained with the learning rate initially set to 0.8, with a neighborhood gaussian starting at $\sigma = 0.1$. This low value of the neighborhood parameter results in just a single unit being used to form the representation of the stimuli. With learning rate in the range between 0.7 and 0.9 simulations achieve statistical significance (σ, simplified here, does not vary since we want to consider one best-matching unit at a time; see Section 4.2.3 for a more standard set of SOMs parameters, including σ). This is a rather

[4]Our model was implemented using the SOM Toolbox (http://www.cis.hut.fi/somtoolbox). This is the Matlab toolbox of reference when implementing SOMs. Our SOM's weights were initialized with the somrandinit function, and then slightly modified in order to contain values separate from the input space: to this end we have multiplied by 2 the initial values provided by somrandinit

theoretical and abstract use of SOMs that we chose in order to investigate our starting hypothesis: that Numerosity and Variability impact the way in which the stimuli are categorized and that this in turn can explain the Numerosity and Variability Effects. In the third simulation we will show that the same effects hold for SOMs trained with more standard neighborhood function and parameters.

As in Gliozzi et al. (2009), each input stimulus was presented to the SOM *once* during learning. With this training schedule SOMs learn without extensive training, and from positive examples only.

In this set of simulations, in the *Base Condition*, we trained the map with two stimuli: the points $[50, 0]$ and $[60, 0]$ (see Figure 4.2.1). We then compared the map's Generalization Degree with the Generalization Degree of a map trained with a more numerous training set, by keeping the range of values the same as in our Base Condition. The training stimuli in this *Numerosity Condition* are the points: $[50, 0], [53, 0], [55, 0], [57, 0], [59, 0], [60, 0]$. In the *Variability Condition* we kept the number of stimuli presented to the SOM constant with respect to the Base Condition, while varying the range of values. In this condition, the stimuli considered are $[30, 0]$ and $[60, 0]$.

We then evaluated whether the representation of the stimuli in the different conditions changes. The answer is positive: the Variability and the Numerosity Conditions affect the precision by which the category examples are represented by the SOM. In the Numerosity Condition, the examples are represented more precisely than in the Base Condition: the (maximal) Euclidean distance between the map's category representation and the examples is lower in the Numerosity Condition than in the Base Condition. In contrast, variability leads to a less precise representation of the examples: in the Variability Condition the (maximal) distance between the map's category representation and the category examples is higher than in the Base Condition. This difference is illustrated in Figure 2, where there is one subfigure for the Base Condition, one for the Numerosity Condition, and one for the Variability Condition; each subfigure represents the stimuli used to train the SOM in the corresponding condition, as well as the SOM's units, when training is complete, plotted with respect to the values of their weights (x and y axis). The precision of the SOM's category representation of the examples is figured by the dashed line (the longer the line the lower the precision).

Nor is this pattern of results affected by changing the exact values of the stimuli, provided the stimuli are selected with some care: in the Numerosity condition, stimuli must be uniformly spread in the interval; in the Variability condition stimuli must have the same Numerosity as the Base condition.

This difference in the quality of representation in the two conditions leads to the Numerosity and Variability effects. When the number of category examples within a given range increases, the generalization curve outside that range shrinks, and the SOM is less likely to attribute to the same category a new stimulus y outside that range.

The opposite effect holds when the variability of the category examples increases: in this case the generalization curve widens, and the SOM's disposition

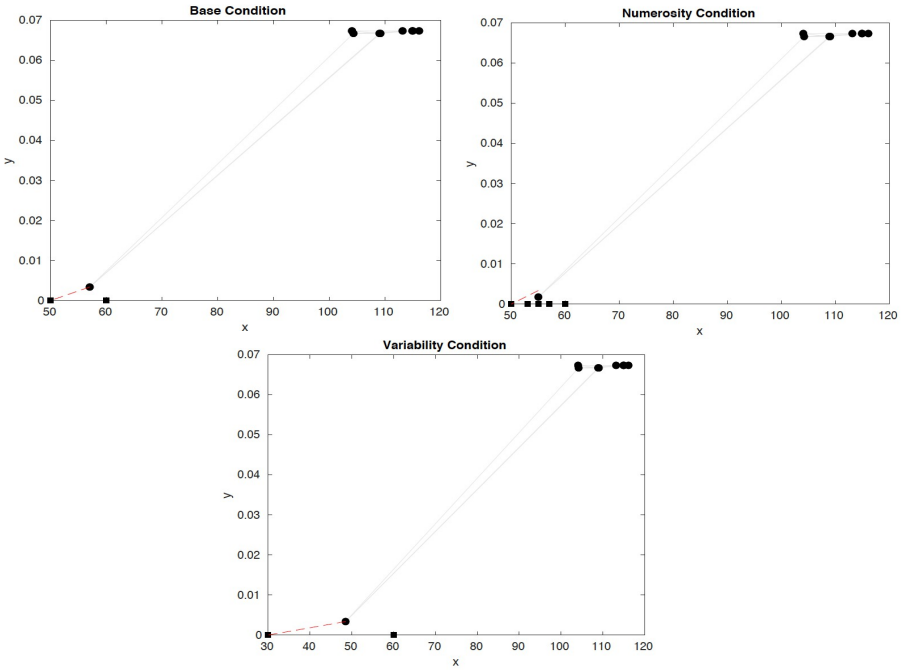

Figure 2: SOM organization in the Base Condition, in the Numerosity Condition, and in the Variability Condition. Each plot has nine black dots, one for each neuron of the SOM (plotted with respect to its weights' values, on the $x-$ and $y-$axis respectively). The grey lines connect adjacent neurons in the map's grid. Each plot represents the neurons' organization after training with the stimuli in the Base, Numerosity, and Variability Condition, respectively. Black boxes represent the stimuli (plotted with respect to their values on the x and y axis respectively) in the three conditions. The maximal Euclidean distance between the stimuli and their best-matching unit is represented by the dashed line: this is longer in the Base Condition than in the Numerosity Condition, whereas it is shorter in the Base Condition than in the Variability Condition.

to attribute a new stimulus y to the same category increases.

The Numerosity and Variability effects are illustrated in Figure 3. The left plot compares the Generalization Degree for the Base Condition versus the Numerosity Condition, and shows that numerosity of the known category examples produces a decrease in generalization. The right plot compares the Generalization Degree for the Base Condition versus the Variability Condition, and shows that variability of the known category examples generates an increase in generalization.

We have run 100 simulations for each condition. When comparing the Generalization Degree in the Base Condition and in the Numerosity condition, a two-tailed t-test[5] revealed that the effect is significant (the mean of the Generalization Degree in the Base Condition at the leftmost point $= 0.008$, $sd = 0.001$, whereas it is $= 0.004, sd = 0.004$ in the Numerosity case, $t(198) = 7.96$, $p < 0.001$; for the rightmost point the mean of the Generalization Degree in the Base Condition $= 0.05$, $sd = 0.006$, whereas it is $= 0.03, sd = 0.03$ in the Numerosity case, $t(198) = 7.86$, $p < 0.001$).

Furthermore, when comparing the Generalization Degree in the Base Condition and in the Variability Condition, a two-tailed t-test revealed that the effect is significant (the mean of the Generalization Degree in the Base Condition at the leftmost point $= 0.008$, $sd = 0.001$, whereas it is $= 0.26, sd = 0.0001$ in the Variability case, $t(198) = -648.5$, $p < 0.001$; for the rightmost point the mean of the Generalization Degree in the Base Condition $= 0.05$, $sd = 0.006$, whereas it is $= 0.28, sd = 0.0008$ in the Variability case, $t(198) = -197$, $p < 0.001$).

4.2.2. Second simulation: Stimuli varying along two dimensions.

In the second set of simulations we have then replicated the simulations above, with the same self-organizing map architecture and the same training schedule and parameters, by considering the other set of stimuli considered by Tenenbaum and Griffiths (2001), namely bi-dimensional stimuli that vary on two dimensions. The stimuli are represented in Figure 4 and 5.

As before, the numerosity of the known category examples lowers the degree of generalization outside the range of the examples, whereas the variability of the known category examples augments the Generalization Degree outside that range[6]. The effect of numerosity is shown in Figure 4, where the thickness of the rectangles represents the Generalization Degree for the stimuli whose values lie on the rectangles. The leftmost plot refers to the Base Condition whereas the rightmost plot refers to the Numerosity Condition, and the thickness of the rectangles, i.e., the Generalization Degree, decreases faster in the second case than in the first. The lower plot explicitly represents the different trend in the decrease of the Generalization Degree. We ran 100 independent simulations.

[5] We applied t-test to the Generalization Degrees for the extreme points, those more distant from the known category examples

[6] The Generalization Degree in this case refers to a rectangle rather than to single points. This is calculated as the maximal Generalization Degree for reference points that define the edges of each rectangle

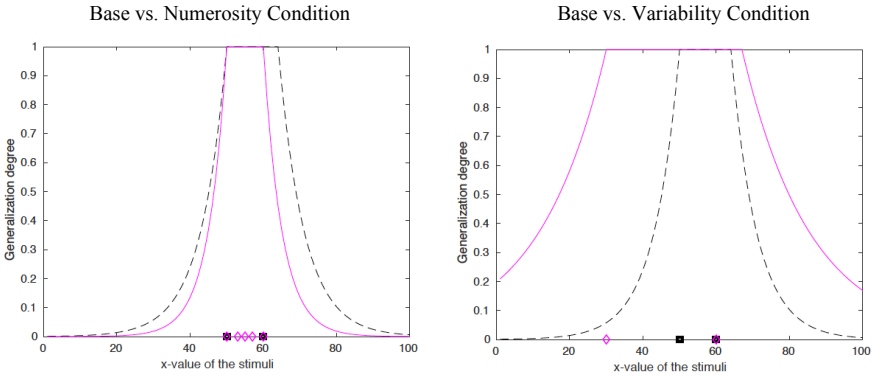

Figure 3: Map's generalization. On the $x-$axis the stimuli varying along that dimension. On the $y-$axis the corresponding *Generalization Degree* from the map's category representation. Left plot: squares are known examples for the Base Condition, diamonds are known examples for the Numerosity Condition; dashed curve Generalization Degree for the Base Condition, continuous line Generalization Degree for the Numerosity Condition. The right plot is as the left one, with the Variability Condition instead of the Numerosity Condition. The plot refers to a single simulation.

A two tailed t-test revealed that the difference in the Generalization Degree in the two conditions is significant (the mean Generalization Degree for the points at a distance corresponding to the outer rectangle in the Base case is 0.29, $SD = 0.11$, whereas it is 0.19, $SD = 0.06$ for the Numerosity condition, $t(198) = 7$, $p < 0.001$).

For the Variability, the effect is shown in Figure 5, with the upper leftmost plot referring to the Base Condition, the right uppermost plot to the Variability Condition (in which the stimuli vary more on the $y-$axis), and the thickness of the rectangles to the strength of the Generalization Degree for stimuli on the rectangles: when comparing the two plots it appears that the variability of the known stimuli makes the decrease of the Generalization Degree slower; this is explicitly represented in the lower plot, where the dashed line refers to the decrease in the Generalization Degree for the Base Condition, the continuous line for the Variability Condition. We ran 100 independent simulations. A two tailed t-test revealed that the difference in the Generalization Degree in the two conditions is significant (the mean Generalization Degree for the points at a distance corresponding to the outer rectangle in the Base case is 0.29, $SD = 0.11$, whereas it is 0.37, $SD = 0.03$ for the Variability condition, $t(198) = -8.22$, $p < 0.001$).

4.2.3. Third simulation: Stimuli varying along two dimensions and more standard training parameters.

In the simulations above, the SOMs are trained on the known category examples only. Furthermore, they are trained in a simplified way, with a high learning rate in order to have a single epoch of training, and with a low neighborhood value so that the overall category representation is simple. This abstract way of training allowed us to make the point of this paper: that the Numerosity and Variability Effects can be the consequence of a change in category representation in the two conditions, and that the two effects can be accounted for within a representation-based algorithm.

Now we report results from a third set of simulations, indicating that the Numerosity and Variability effect also hold in a SOM trained with a richer training set made of different categories, and with standard parameters.

In order to see whether the Numerosity and Variability effect also hold in a SOM trained with a richer training set made of different categories, and with more standard parameters (lower learning rate, higher number of epochs, higher neighborhood function), we have run a third set of simulations in which we have trained a SOM made of 25 units with a training set made of the same stimuli than those considered above augmented with 10 new categories each containing the same number of stimuli. As above, we have compared the map's Generalization Degree when the target category contained more known examples than when it contained less such examples, and when it contained more variable examples than less variable examples. In these simulations, SOMs weights

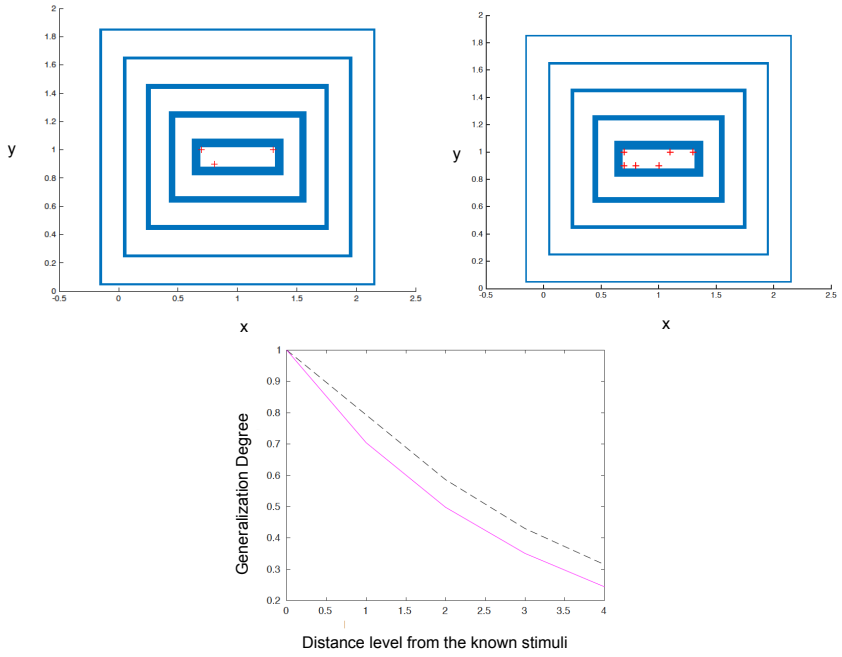

Figure 4: The Generalization Degree (represented by the thickness of the rectangles' edges) decreases faster when the known examples of a given category augment (Numerosity Condition, rightmost plot referring to the stimuli are crosses), compared to when there are less known examples (Base Condition, leftmost plot). In the lower plot the dashed line represents the decrease of the Generalization Degree in the Base Condition, the continuous line in the Numerosity Condition. Results plotted refer to a single simulation.

were initialized to random values [7]; the values for the learning rate and the neighborhood function parameters were the standard ones, as calculated by the som_seqtrain function of the SOM Toolbox[8].

The map was trained for 1000 epochs. Results match those of Figures 4 and 5.

We ran 100 independent simulations. As for Numerosity, a two tailed t-test revealed that the difference in the Generalization Degree in the two conditions is significant: the mean Generalization Degree for the points at a distance corresponding to the outer rectangle in the Base case is $0.33, SD = 0.07$, whereas it is $0.13, SD = 0.01$ for the Numerosity condition, $t(198) = 27.21$, $p < 0.001$. For Variability, the mean Generalization Degree for the points at a distance corresponding to the outer rectangle in the Base case is $0.08, SD = 0.01$, whereas

[7] This was done as in the som_randinit function of the SOM Toolbox

[8] We imposed the width of the neighborhood function to decrease to very low values so that during the last epochs of training, only one single unit was concerned at a time

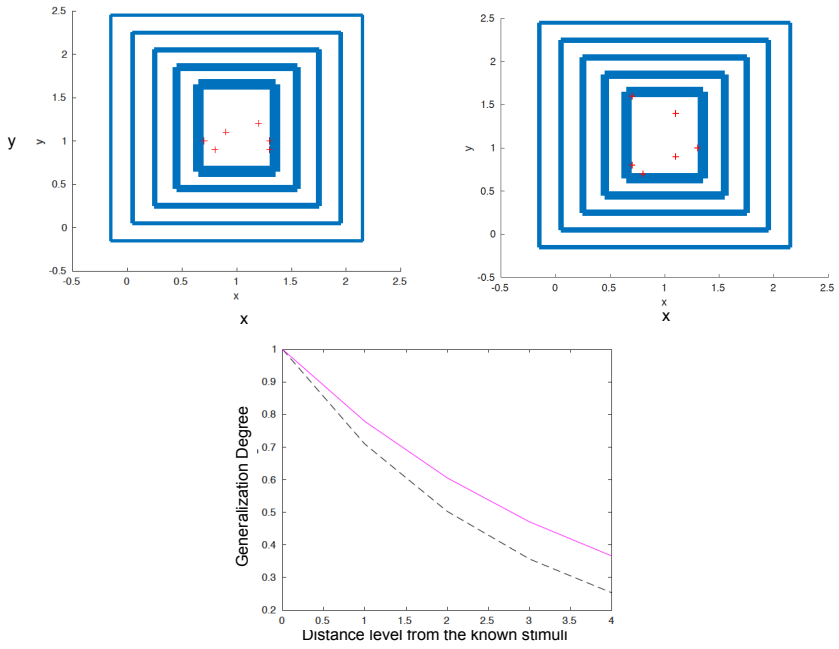

Figure 5: The Generalization Degree (represented by the thickness of the rectangles' edges) decreases slower when the known examples of a given category are more variable, here along the y-dimension (Variability Condition, rightmost plot, stimuli are crosses), compared to when they are less variable (Base Condition, leftmost plot). In the lower plot the dashed line represents the decrease of the Generalization Degree in the Base Condition, the continuous line to the Variability Condition.Results plotted refer to a single simulation.

it is $0.26, SD = 0.01$ for the Variability condition, $t(198) = -185.95$, $p < 0.001$.

5. Discussion

The results of the previous section show that SOMs can provide a mechanistic account of the Numerosity Effect as well as of the Variability Effect observed in human categorization. In this way SOMs can complement the Bayesian account of category generalization which lies at Marr's computational level and does not provide a description of a possible plausible mechanism of category generalization that leads to the two effects.

As a difference with respect to the explanation provided by a Bayesian account, the SOM's explanation of the two effects relies only on the notions of:

- *category representation* and
- of *distance of the new stimulus from the category representation*

As far as we know, this is the first quantified argument demonstrating that the two effects, which characterize human category generalization, can be explained within the similarity-based paradigm. Furthermore, the SOMs we consider exhibit the two effects when exposed only to few positive category examples (as in humans), without the need of extensive or contrastive learning. This is very different from what is usually argued against neural networks in general, namely that they have to be trained in a massive way with huge amount of data and with explicit contrastive information. SOMs' results easily extend to stimuli that vary along more than one dimension, as long as the notion of Euclidean Distance between stimuli can be clearly defined.

These effects cannot be jointly explained within traditional theories of categorization based on similarity, such as the prototype theory (Posner and Keele, 1968) or the exemplar theory (Medin and Schaffer, 1978; Nosofsky, 1986). Indeed, for the prototype theory: the prototype remains the same independently from the number of instances considered, whereas the variability leads to a shift of the prototype position in the direction of the increased variance. This leads to an increase of the generalization curve only in that direction rather than to a general increase, as postulated by the Variability Effect.

For the exemplar theory as well the increased variability leads to an increased generalization only in the direction of increased variance, rather than to a general increase of generalization for category membership, as the one postulated by the Variability Effect. Similar considerations apply to a model as SUSTAIN (Love et al., 2004). It is possible that a mechanism can be added to the above formalisms in order to achieve the Numerosity and Variability effects. But for the moment this does not hold.

Numerosity and Variability effects cannot be explained within neural networks based on backpropagation either, since these networks need a lot of information, including contrastive information in order to achieve a reasonable

categorization of the inputs, whereas in the examples considered here there are only few positive instances of the category.

As already mentioned in Section *Bayesian analyses of category generalization*, previous important proposals have been made to provide a mechanistic account of Bayesian models of category generalization (Sanborn et al., 2010; Shi et al., 2010; Sanborn and Chater, 2016; L. Frermann, 2016). These models are certainly informative but, as discussed in that section, imply a level of inferential power and memory that may be unrealistic for the human cognizer: the tools they employ, including both Monte Carlo methods (Sanborn et al., 2010) and stored hypotheses (Shi et al., 2010) are potentially too complex to pass as psychologically plausible mechanisms.

In this paper, we attempted to formulate a mechanism which is simpler and more readily executable by an individual with limited memory and reasoning capacities. The outlined mechanism lies on a framework, as self-organizing maps, that is considered to reflect basic constraints of plausible brain implementations in the cortex, and that has proven to be capable of explaining experimental results. The other side of the coin of the simplicity of of our model is that it deals with simple stimuli, as Tenenbaum and Griffiths (2001). Although dealing with higher dimensional stimuli is not a problem for self-organizing maps per se, it is not clear how easy would be the specific use we make of the model, for category induction out of few training instances, especially when more categories are treated at the same time and the category structure is intricated. Understanding if in the context of several highly intricated categories something more refined than pure Euclidean distance would be needed is the object of future work.

Before we conclude, let us make some general considerations on the precise relations between self-organizing maps and Bayesian models. This paper shows that SOMs capture *some* aspects of Bayesian analyses. Being psychologically and biologically plausible, this makes SOMs good candidates to bridge the computational, Bayesian level of analysis and the algorithmic level of analysis in the study of category generalization (and purportedly for other cognitive tasks). However, understanding the exact extension of the correspondence between SOMs and Bayesian models requires future research.

For the time being we can say that there are small differences between SOMs and Bayesian analyses predictions. For instance, SOMs are sensitive to the specific position of repeated category examples within a given range, whereas Bayesian analyses are not (at least when the size principle is used as a likelihood estimation, as in Tenenbaum and Griffiths, 2001). Take the two following sets of category examples, whose values vary in the same range but in which the exact values of the instances change:

$$\text{Set } 1 = \{[30, 0], [40, 0], [60, 0]\};$$

$$\text{Set } 2 = \{[30, 0], [50, 0], [60, 0]\}$$

Figure 6 shows that for the Bayesian analysis of categorization in the two conditions the generalization curve will be the same (Tenenbaum and Griffiths,

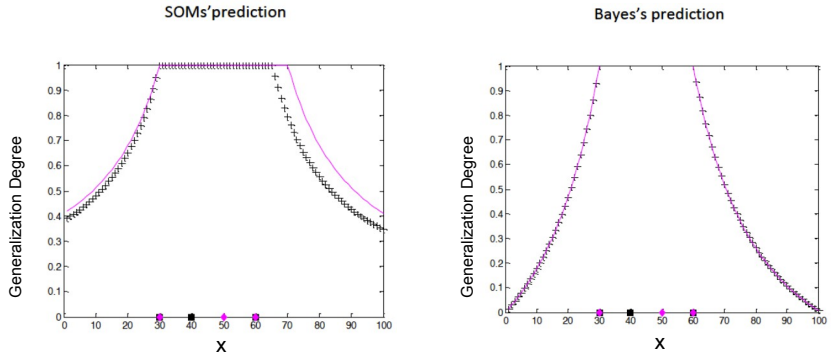

Figure 6: In both plots: on the $x-$axes the stimuli varying along that dimension; boxes on the $x-$axis are the examples of Set 1, whereas diamonds are the examples of Set 2. Left plot: on the $y-$axes the crossed line represents the Generalization Degree for the SOM trained with Set 1, whereas the continuous line refers to the Generalization Degree for the SOM trained with Set 2. Right plot: Probability (Equation 3) that the corresponding stimuli belong to the same category than the observed examples of Set 1 (crossed line) and Set 2 (continuous line), respectively.

2001). This is clear in the rightmost plot, where the crossed curve refers to the trend of the Probability (Equation 3) that the $x-$stimulus belongs to the same category than the category of observed examples of Set 1, whereas the continuous line represents the Probability that the $x-$stimulus belongs to the same category than the category of observed examples of Set 2. The two lines overlap, indicating that the probability in the two cases is the same.

On the contrary, for SOMs there will be a difference in the generalization curve in the two conditions. This is illustrated in the leftmost plot of Figure 6, where the crossed line represents the Generalization Degree for the SOM trained with Set 1, whereas the continuous line refers to the Generalization Degree for the SOM trained with Set 2. The difference between Set 1 and Set 2 produces a difference on the two curves, i.e. on the Generalization Degree in the two cases.

This difference in the predictions is the consequence of the fact that SOMs form a representation of the examples, whose position is shifted depending on the exact values of the category examples. Instead, in Bayesian models there is no representation being formed, therefore no shift, and no consequent effect of the exact values of category examples.

6. Conclusions

In this paper we have shown that a biologically and psychologically plausible neural network architecture can provide a mechanistic account of Numerosity and Variability effects usually explained, at the computational level, with Bayesian tools. SOMs can do so when exposed to limited category examples without any need of contrastive information, thus contradicting the main criticism against neural networks models of category generalization alternative to the Bayesian ones. We leave for future research the investigation of the extension of the correspondence between SOMs and Bayesian models, and whether SOMs can be seen as describing the mechanisms underlying Bayesian analyses in general.

Where category generalization is concerned, the model proposed in this paper allows to bridge the gap between Marr's different levels of analysis, and in particular between Marr's *computational* level of analysis, where we find Bayesian models of category generalization and in particular the one by Tenenbaum and Griffiths (2001) we consider and Marr's *algorithmic* level in which we put the similarity-based mechanism discussed in this paper.

References

Anderson, J., 1991. The adaptive nature of human categorization. Psychological Review 98(3), 409–429.

Eimas, P., Quinn, P., 1994. Studies on the formation of perceptually based basic-level categories in young infants. Child Development 65, 903–917.

Gliozzi, V., Althaus, N., Mayor, J., Plunkett, K., 2013. Primacy/recency effects in infant categorisation. Proc. of the 35th Annual Meeting of the Cognitive Science Society .

Gliozzi, V., Mayor, J., Hu, J., Plunkett, K., 2009. Labels as Features (Not Names) for Infant Categorization: A Neurocomputational Approach. Cognitive Science 33(4), 709–738.

Kohonen, T., Schroeder, M.R., Huang, T.S. (Eds.), 2001. Self-Organizing Maps. 3rd ed., Springer-Verlag New York, Inc., Secaucus, NJ, USA.

L. Frermann, M.L., 2016. Incremental Bayesian Category Learning From Natural Language. Cognitive Science .

Li, P., Zhao, X., Mac Whinney, B., 2007. Dynamic self-organization and lexical development. Cognitive Science 31, 581–612.

Love, B.C., Medin, D.L., Gureckis, T.M., 2004. SUSTAIN: A Network Model of Category Learning. Psychological Review 111, 309–332.

MacKay, D.J.C., 1995. Probable networks and plausible predictions Ñ A review of practical Bayesian methods for supervised neural networks. Network: Computation in Neural Systems .

Mayor, J., Plunkett, K., 2010. A neuro-computational account of taxonomic responding and fast mapping in early word learning. Psychological Review 117(1), 1–31.

McClelland, J.L., 1998. Connectionist models and bayesian inference. M. Oaksford and N. Chater. Rational Models of Cognition. Oxford University Press .

Medin, D.L., Schaffer, M., 1978. Context theory of classification learning. Psychological Review 85, 207–238.

Miikkulainen, R., Bednar, J., Choe, Y., Sirosh, J., 2005. Computational Maps in the Visual Cortex. Springer.

N., C., M., O., U., H., E., H., 2001. Bayesian models of cognition. Wiley Interdisciplinary Reviews: Cognitive Science 1, 811–823.

Neal, R.M., 1996. Bayesian learning for neural networks. Lecture Notes in Statistics 118.

Nosofsky, R., 1986. Attention, similarity, and the identification categorization relationship. Journal of Experimental Psychology: General 115, 39–57.

Posner, M.I., Keele, S.W., 1968. On the genesis of abstract ideas. Journal of Experimental Psychology 77, 241–248.

R. Zemel, P. Dayan, A.P., 1998. Probabilistic Interpretation of Population Codes. Neural Computation, 10(2): 403-430 (1998). 10, 403–430.

Sanborn, A.N., Chater, N., 2016. Bayesian Brains without Probabilities. Trends in Cognitive Science 20(12), 883–893.

Sanborn, A.N., Griffiths, T.L., Navarro, D., 2010. Rational approximations to rational models: alternative algorithms for category learning. Psychol. Rev. 117(4), 1144–1167.

Schyns, P., 1991. A modular neural network model of concept acquisition. Cognitive Science 15, 461–508.

Shepard, R.N., 1987. Toward a Universal Generalization for Psychological Science. Science 237(4820), 1317–1323.

Shi, L., Griffiths, T.L., Feldman, N.H., Sanborn, A.N., 2010. Exemplar models as a mechanism for performing Bayesian inference. Psychonomic Bulletin & Review 17, 443–464.

Tenenbaum, J.B., Griffiths, T.L., 2001. Generalization, similarity, and Bayesian inference. Behavioral and Brain Sciences 24, 629–641.

Tversky, A., 1977. Features of similarity. Psychological Review 84, 327–352.

Westermann, G., Mareschal, D., 2004. From Parts to Wholes: Mechanisms of Development in Infant Visual Object Processing. Infancy 5, 131–151.

Younger, B., Cohen, L., 1986. Developmental change in infants' perception of correlations among attributes. Child Development 57, 803–815.

7. Supplemental Material

7.1. Example of use of the Equations of Section4

We here provide some very simple examples to illustrate the application of the Equations described in Section 4.1.

Equation 4 Consider a SOM only containig 3 neurons, with weight vector respectively $[1, 1]$, $[2, 2]$, $[3, 3]$, receiving at the n^{th} iteration input $x = [0, 1]$. When computing the Euclidean distance between x and all the neurons' weight vectors, it results that the BMU is the first neuron. Suppose the learning rate $\eta(n)$ is 0.5, and suppose the neighborhood as calculated by Equation 5 is $h(n) = 1$ for the BMU and 0.5 for the other surrounding neurons, at iteration $n + 1$, the weight vectors resulting from learning will be respectively $[0.5, 1]$, $[1.5, 1.75]$, $[2.25, 2.5]$.

Equation 6 Suppose now the situation above is the final situation of training, resulting from the presentation to the map of the two known instances of category C, namely $x' = [1, 0.5]$ and $x = [0.1]$. For both x and x' the best matching unit is the first neuron. Now take a new stimulus $y = [0, 3]$. We want to calculate its relative distance from the map's representation of category C, which in this case is given by the only first neuron. Hence $min\|y - BMU_C\| = 2.06$, whereas the denominator $max_{x \in C}\|x - BMU_x\|$ is the maximal Euclidean distance between BMU_C and x or x': in this case it is 0.5. The overall relative distance between y and category C is therefore 4.12 (notice that with higher numerosity presumably the precision of the representation would improve, hence the distance between BMU_C and the known category examples would diminish, and the overall relative distance would increase; the opposite pattern for higher variability for which the precision of representation would diminish. We see in the next point that this has an effect on the Generalization Degree in the two cases).

Equation 7 The relative distance just calculated gives us a way to calculate the Generalization Degree for the novel unknown stimulus y, i.e., our predisposition to attribute y to the same category C than x and x'. In our case the Generalization Degree for y would be 0.02. By the considerations we did for the relative distance, in the Numerosity Condition the Generalization Degree (that decreases when the relative distance augments) would decrease, whereas in the Variability Condition it would increase.

7.2. Higher dimensionality stimuli

In this paper we have considered the same stimuli than those considered by Tenenbaum & Griffiths (2001) in order to compare the two models. These stimuli are very simple and two dimensional. One may wonder if the same pattern of results would hold for higher dimensional stimuli. As an example Figures 7 and 8 below show that the answer is positive for higher dimensionality, reporting patterns of results with 5 dimensional stimuli. However, it is possible that when more categories are considered at the same time, which are more complex and overlapping, (as for instance those considered by Lapata et al., 2016), more sophisticated tools might be necessary, whereas Euclidean distance would no longer be enough.

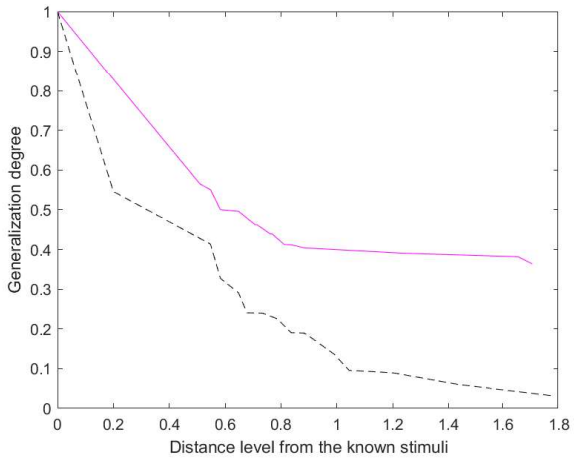

Figure 7: Variability Effect with 5 dimensional stimuli. The plot illustrates the presence of a Variability Effect: when considering stimuli which are more and more distant from the known category instances, generalization increases in the case of higher variability of the known category instances (continuous line) than with lower variability (dashed line). In the plot the generalization curve corresponding to higher variabiity condition is represented with a continuous line, whereas for the lower numerosity with a dashed line

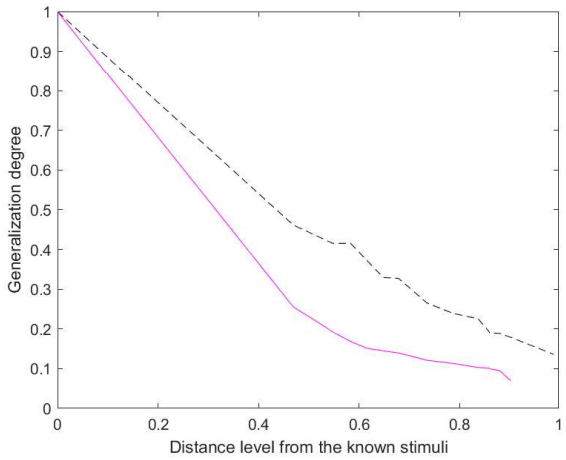

Figure 8: Numerosity Effect with 5 dimensional stimuli. The plot illustrates the presence of a Numerosity Effect: when considering stimuli which are more and more distant from the category instances, generalization decreases in the case of higher numerosity of the known category instances than with lower numerosity. In the plot the generalization curve corresponding to higher numerosity condition is represented with a continuous line, whereas for the lower numerosity with a dashed line.