Physica A 388 (2009) 732–746

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

A Tsallis' statistics based neural network model for novel word learning

Tarik Hadzibeganovic [a,b], Sergio A. Cannas [c,*,1]

[a] Cognitive Science Section, Department of Psychology, University of Graz, A-8010, Austria

[b] Cognitive Neuroscience Research Unit, Department of Psychiatry & Forensic Medicine, Faculty of Medicine, Hospital del Mar, Universitat Autònoma de Barcelona, 08003 Barcelona, Spain

[c] Facultad de Matemática, Astronomía y Física, Universidad Nacional de Córdoba, Ciudad Universitaria, 5000 Córdoba, Argentina

ARTICLE INFO

Article history:
Received 22 August 2008
Available online 5 November 2008

PACS:
87.19.lj
05.10.-a
87.19.lv
43.71.Hw

Keywords:
Novel word learning
Perceptron
Tsallis entropy
Nonextensive statistical mechanics

ABSTRACT

We invoke the Tsallis entropy formalism, a nonextensive entropy measure, to include some degree of non-locality in a neural network that is used for simulation of novel word learning in adults. A generalization of the gradient descent dynamics, realized via nonextensive cost functions, is used as a learning rule in a simple perceptron. The model is first investigated for general properties, and then tested against the empirical data, gathered from simple memorization experiments involving two populations of linguistically different subjects. Numerical solutions of the model equations corresponded to the measured performance states of human learners. In particular, we found that the memorization tasks were executed with rather small but population-specific amounts of nonextensivity, quantified by the entropic index q. Our findings raise the possibility of using entropic nonextensivity as a means of characterizing the degree of complexity of learning in both natural and artificial systems.

1. Introduction

As shown by Montemurro [1], Zipf–Mandelbrot law satisfies the first-order differential equation of the type $\frac{df}{ds} = -\lambda f^q$, with its solutions asymptotically taking the form of pure power laws with decay exponent $1/(q-1)$. Further modification of the expression into $\frac{df}{ds} = -\mu f^r - (\lambda - \mu) f^q$ (now with a new parameter and a new exponent), allows for the presence of two global regimes [2] characterized by the dominance of either exponent depending on the particular value of f. After this formalism was applied to experimental datasets on re-association in heme proteins [3], within the framework of non-extensive statistical mechanics [4,5], Tsallis suggested its potential usefulness in describing linguistic and neurocognitive phenomena (see e.g. Refs. [1,6]).

Ever since, there has been growing interest within a variety of fields [7–9], including biomedical engineering and computational neuroscience [10–15], in the non-extensive statistical mechanics based on Tsallis' generalized entropy $S_q = k\frac{1-\sum_i p_i^q}{q-1}$ ($\sum_i p_i = 1; q \in \mathcal{R}$), which in the limit of $q \to 1$ (and $k = k_b$) reduces to conventional Boltzmann–Gibbs entropy.

The parameter q that underpins the generalized entropy of Tsallis is linked to the underlying dynamics of the system and measures the amount of its non-extensivity. In statistical mechanics and thermodynamics, systems characterized by the property of nonextensivity are systems for which the entropy of the whole is different from the sum of the entropies of the respective parts. Such are usually the systems with interactions over long distances, with long memories of perturbations, and with very often fractal or multi-fractal structural properties. Since Tsallis' formalism is rooted on a non-extensive

* Corresponding author.
E-mail addresses: ta.hadzibeganovic@uni-graz.at (T. Hadzibeganovic), cannas@famaf.unc.edu.ar (S.A. Cannas).

[1] Member of CONICET, Argentina.

0378-4371/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2008.10.042

entropy, it appears to be a suitable candidate for describing systems with any kind of microscopic interactions (both short- and long-ranged). In other words, the generalized entropy of the whole is *greater* than the sum of the generalized entropies of the parts if $q < 1$ (superextensivity), whereas the generalized entropy of the system is *smaller* than the sum of the generalized entropies of the parts if $q > 1$ (subextensivity).

As noted by Hopfield [16] and then applied to attractor networks by Amit et al. [17], neural network models have direct analogies in statistical physics, where the investigated system consists of a large number of units each contributing individually to the overall, global dynamic behavior of the system. The characteristics of individual units represent the *microscopic* quantities that are usually not directly accessible to the observer. However, there are *macroscopic* quantities, defined by parameters that are fixed from the outside, such as the temperature $T = 1/\beta$ and the mean value of the total energy $\langle E \rangle$. The main aim of statistical physics is to provide a link between the microscopic and the macroscopic levels of an investigated system. An important development in this direction was Boltzmann's finding that the probability of occurrence for a given state $\{x\}$ depends on the energy $E(\{x\})$ of this state through the well-known Boltzmann–Gibbs distribution $P(\{x\}) = \frac{1}{Z} \exp[-\beta E(\{x\})]$, where Z is the normalization constant $Z = \sum_{\{x\}} \exp[-\beta E(\{x\})]$.

In the context of neural networks, statistical physics can be applied to study learning in the sense of a stochastic dynamical process of synaptic modification [18]. In this case, the dynamical variables $\{x\}$ represent synaptic couplings, while the error made by the network (with respect to the learning task for a given set of values of $\{x\}$) plays the role of the energy $E(\{x\})$. The usage of gradient descent dynamics as a synaptic modification procedure leads then to a stationary Boltzmann–Gibbs distribution for the synapses [18]. However, the gradient descent dynamics corresponds to a strictly *local* learning procedure, while non local learning dynamics may lead to a synaptic couplings distribution different from the Boltzmann–Gibbs one [19,20].

In the present study, we employ the *nonextensive* statistics theory of Tsallis to include some degree of non-locality in a two-level perceptron model. This q-generalized artificial neural network is further used to simulate the novel word learning process in two linguistically different populations of subjects. With respect to the computational simulations, our goal has been to investigate whether novel word learning occurs in an extensive or nonextensive manner. The core of the model is represented by a particular kind of *non-extensive cost function* that should induce a *non-local learning rule* in the neural network. In this sense, it is possible to think of non-extensivity as a particular form of globality or non-locality, at least in principle.

Alternatively, an implementation of a cost function that would induce a *local learning rule*, would cause the variation of the synapse between any two neurons at a given time to depend only on the instantaneous post-synaptic potentials (PSP) received by them, and *not* on the PSPs received by the rest of the neurons. It seems, therefore, more reasonable to assume that the full specification of a given neural representation depends on a non-local, distributed pattern of activity, emerging from the interaction of the constituents of whole neuronal ensembles, rather than from the activity in any particular, single neuron [21].

Representing linguistic knowledge by the distributed patterns of activity in neural networks has a long tradition in computational neuroscience [22–26]. More recently developed techniques for recording the simultaneous activity in populations of neuronal cells [27,28] provide substantial evidence for the non-local, distributed patterns hypothesis. Furthermore, there is growing evidence that neuronal populations distributed over distant cortical areas synchronize and work in synergy as *functional webs* during language processing [29].

Through the reciprocal links with the language areas, ventral visual stream, and the hippocampal formation, the anteroventral temporal cortex integrates a variety of aspects of letter-string information during processing such as visual, lexical, semantic and mnestic [30]. Novel word learning, which depends upon the structures in the medial temporal lobe, eventually becomes independent of these structures, relying more on other neocortical areas, such as those in temporal and temporo-parietal regions (see e.g. Ref. [31] for a review). Thus, the representation of the lexical information does not remain strictly limited to a particular area, but instead, it becomes distributed across different brain regions relevant for storing different aspects of information such as word meanings (temporal lobe) and word sounds (temporoparietal regions). For such reasons, it is necessary to investigate the effects of introducing non-local learning rules in neural network models for language learning, especially where they outperform purely local neural dynamics and better fit psychological and neuroscientific phenomenology.

A full understanding of the neural bases of learning also requires an accurate characterization of the learning processes as they occur in behavioral experiments. Learning is generally believed to include a gradual restructuring and strengthening of underlying connections between neural cells [32,33], which is behaviorally manifested in the gradual decrease of error after a series of repeated learning trials. Such asymptotic behavior is usually measured by using the learning curve, which is a plot of the magnitude or frequency of the response accuracy (or error) as a function of the number of learning trials. The agreement that is often found between the investigations of group-averaged learning behavior and the widely accepted neurobiological theories of individual animal learning, has caused many neuroscientists to use population-averaged learning curves for comparing the asymptotic learning behavior between differently treated groups of subjects (e.g. Refs. [34,35]). For a brief review, and an opposite viewpoint on this issue, see Ref. [36].

In the present study, two simple memorization tasks were carried out in two groups of learners with orthographically different native languages [37–39]. Subjects monitored 5×5 and 7×6 nonbinary letter matrices for a fixed number of seconds. Letter sequences in the matrix rows formed novel word items with very low summated type bigram frequencies (STBFs) and sparse orthographic neighborhoods (ONs). Learning was measured following each of the 10 stimulus exposures.