Journal of Computational Science 2 (2011) 316–323

Contents lists available at SciVerse ScienceDirect

Journal of Computational Science

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

A fast no-rejection algorithm for the Category Game

Francesca Tria [a,*], Animesh Mukherjee [a], Andrea Baronchelli [b], Andrea Puglisi [c,d], Vittorio Loreto [c,a,d]

[a] Institute for Scientific Interchange (ISI), Viale Settimio Severo 65, 10133 Torino, Italy
[b] Departament de Fisica i Enginyeria Nuclear, Universitat Politecnica de Catalunya, Campus Nord B4, 08034 Barcelona, Spain
[c] Dipartimento di Fisica, Sapienza Università di Roma, Piazzale Aldo Moro 5, 00185 Roma, Italy
[d] CNR-ISC Piazzale Aldo Moro 5, 00185 Roma, Italy

ARTICLE INFO

Article history:
Received 21 December 2010
Received in revised form
20 September 2011
Accepted 3 October 2011
Available online 12 October 2011

Keywords:
Category game
Metastable states
No-rejection algorithms
Agent-based simulation

ABSTRACT

The Category Game is a multi-agent model that accounts for the emergence of shared categorization patterns in a population of interacting individuals. In the framework of the model, linguistic categories appear as long lived consensus states that are constantly reshaped and re-negotiated by the communicating individuals. It is therefore crucial to investigate the long time behavior to gain a clear understanding of the dynamics. However, it turns out that the evolution of the emerging category system is so slow, already for small populations, that such an analysis has remained so far impossible. Here, we introduce a fast no-rejection algorithm for the Category Game that disentangles the physical simulation time from the CPU time, thus opening the way for thorough analysis of the model. We verify that the new algorithm is equivalent to the old one in terms of the emerging phenomenology and we quantify the CPU performances of the two algorithms, pointing out the neat advantages offered by the no-rejection one. This technical advance has already opened the way to new investigations of the model, thus helping to shed light on the fundamental issue of categorization.

1. Introduction

The Category Game (CG) is a computational model in which a population of individuals co-evolve their own system of symbols and meanings by playing elementary language games [1]. It has been introduced to investigate how categorization can emerge from scratch in a group of individuals who interact in a pairwise way without any central coordination. The reference problem is color categorization, which is a central issue both in linguistics [2] and in cognitive science [3–5]. Color naming represents in fact a fundamental access point to human cognition, and at the same time provides important clues on language evolution. The evolution of English color categories constitutes an excellent example. English color terms exhibited a gradual semantic shift from largely brightness color concepts (Old English) to almost exclusively hue concepts (Middle English) [6]. The World Color Survey, moreover, showed that color systems across language are not random [7,8], but rather exhibit certain statistical regularities, thus opening the way to a revolution in cognitive science [5,9].

The main point of interest of the CG is that it is able to reproduce qualitatively and, most remarkably, quantitatively the empirical data gathered in the World Color Survey [10]. Kay and Berlin [7] ran a first survey on 20 languages in 1969. From 1976 to 1980, this unique and extensively studied database was enlarged by the same researchers along with W. Merrifield and the data was made publicly available since 2003 on the website http://www.icsi.berkeley.edu/wcs. These data concern the basic color categories of 110 languages without written forms and spoken in small-scale, non-industrialized societies. On average, 24 native speakers from each of this language were interviewed by expert field linguists. Every informant was tasked to name each of 330 color chips produced by the Munsell Color Company that represent 40 gradations of hue and maximal saturation, plus 10 neutral color chips (black-gray-white) at 10 levels of value. The chips were presented in a predefined, fixed random order, to the informant who had to name each of these colors with primarily a "basic color name" from her language (in English, "basic color names" would correspond to words like "red", "green", "yellow", "blue", etc. for more details see [7]).

The Category Game model differs from the other models defined to address similar issues [11–18] in that it accounts for the categorization of a genuinely continuous perceptual channel and it describes a categorization pattern as a continuously evolving metastable state on which the population shares a temporary consensus [1,19]. The latter characteristic is intriguing and underlies the existence of a new framework to address the puzzling problem of language change, which turns out to be at the same time propelled by the interaction among the speakers and impeded by the need of these speakers to understand each other [19]. The presence of this sort of frustration renders the dynamics of the model so slow

* Corresponding author.
E-mail address: fra_trig@yahoo.it (F. Tria).

1877-7503/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jocs.2011.10.002

that it has been so far impossible to investigate and quantify properly the details of the long time behavior, even for small population sizes.

Here we present a fast algorithm suitable for studying the CG dynamics over large timescales and for moderately large population sizes. The algorithm is, in spirit, similar to those suggested for accelerating Monte Carlo simulations (see [20,21] and also [22] for other examples), where the key ingredient is to avoid rejection steps (hence the "no-rejection" tag in the title). Of course, however, the dynamics we are referring to is substantially different, so new methods had to be developed in order to tackle the specific aspects of the model under consideration. In general, this area of research has gained immense popularity over the past few years and as a consequence the new field of computational science came into existence [23–25]. The rest of the paper is organized as follows. In Section 2, we outline a detailed description of the CG model. The next section presents the fast algorithm suitable for investigating the long-time CG dynamics. Section 4 compares the outcomes of the fast algorithm with the original one, showing excellent qualitative as well as quantitative agreement. In this context, a detailed investigation of a set of relevant observables is also performed, shedding new light on the "microscopic" origin of the "macroscopic" behavior of the system. In Section 5 we show the computational complexity of the proposed fast algorithm, compared with the one of the original model. We conclude in Section 6 by summarizing our contributions and pointing to possible future applications of this method.

2. The Category Game model

The Category Game model [1] is crafted to examine how a population of interacting individuals can develop through a series of language games [26] a shared form-meaning repertoire from scratch without any pre-existing categorization. We consider a population of N artificial agents each of them having, without any loss of generality, a one-dimensional continuous perceptual space spanning the $[0, 1)$ interval. A categorization can be identified as a partition of this space into discrete sub-intervals which we shall denote from now onwards as perceptual categories. Each agent has a dynamical inventory of associations linking the perceptual categories (meanings) to words (forms) that are used to name each of these categories. The perceptual categories as well as the words associated to them co-evolve over time through a series of simple communication interactions among the agents (or "games").

In each game, two individuals are randomly selected from the population and one of them is assigned the role of a speaker while the other the role of a hearer. Both the speaker and hearer are presented with a scene of $M \geq 2$[1] stimuli (objects) where each stimulus corresponds to a real number in the $[0, 1)$ interval. By definition, no two stimuli appearing in the same scene can be at a distance closer than $d_{\min}$ which is the only parameter of the model encoding the finite resolution power of any perception, for instance, the human Just Noticeable Difference (JND). In psychophysics, JND is defined as the minimum amount by which the stimulus intensity must be changed in order to produce a noticeable variation in sensory experience [27].

One of the objects is the *topic* of the communication. The task of the speaker is to communicate this to the hearer using the following prescription. The speaker utters a word associated with the topic while the hearer tries to guess its meaning from the word she "listened". The speaker always checks whether the topic is the unique stimulus to lie in a specific perceptual category among all the presented stimuli. If it is not, i.e., if the two stimuli collide on the same perceptual category, then a new boundary is created in the perceptual space at a location corresponding to the middle of the segment connecting the two stimuli. A new word is invented for each of the resultant two new categories. In addition, both of them inherit all the words corresponding to the old category. This process is termed as *discrimination*. Subsequently, the speaker utters the "most relevant" name for the category corresponding to the topic where the most relevant name is either the one used in a previous successful communication or the newly invented name in case the category has just been created due to a discrimination. For the hearer, there can be the following possibilities: (a) the hearer does not have any category containing an object and associated with the name, in which case the game is a failure, (b) there are one or more categories associated with this name and containing an object in the hearer's inventory. In this case, the hearer randomly chooses one of them. If the category chosen corresponds to that of the topic, the game is a success, otherwise it is a failure.

Depending on the outcome of the game one or both the agents update their repertoires. In case of a failure, the hearer adds the word in her repertoire linked to the category corresponding to the topic (eventually discriminating it). In case of success, this word becomes the most relevant name for the category corresponding to the topic for both agents and they remove all other competing words from their respective repertoires linked with this category. Note that if both the speaker and the hearer already have only the successful word in the corresponding category, the inventory of both of them remains unaltered after the game. This situation, as already mentioned in the introduction, corresponds to a rejection step of the Monte Carlo algorithm. The time t of the dynamics is simply measured as the number of games played by the agents.

All the agents start without perceptual categories and without category labels. During the dynamics, discrimination creates finer and finer boundaries ("perceptual categories") and produce more and more category labels, unsuccessful games spread such labels through the population, and successful games reduce the dictionaries associated to the categories to single winning words shared by at least two players. Quite rapidly the N agents develop a fine structure of perceptual categories represented by one or two words, shared by the population. But there is also another fundamental phenomenon which occurs unexpectedly in the dynamics: nearby perceptual categories (after a lucky sequence of games) may share the same category label. In this way, a spontaneous competition of labels arises, to *conquer* larger and larger groups of adjacent perceptual categories. These groups are what we call "linguistic categories".

In summary, the CG dynamics results in the emergence of a hierarchical category structure comprising two distinct levels: a basic layer, responsible for the fine discrimination of the perceptual space (i.e., the perceptual categories), and a second shared linguistic layer that groups together perceptions having the same name to guarantee communicative success (linguistic categories). Note that while the number of perceptual categories is tuned by a parameter of the model and can be very large (of the order of $1/d_{\min}$), the number of linguistic categories turns out to be finite and small, as observed in natural languages (see Fig. 2). On an extremely much larger timescale, however, tiny variations of categories and labels may occur, and eventually these linguistic categories may merge. This long time dynamics is something very interesting but extremely slow to be investigated.

3. A fast algorithm for the Category Game

The primary goal of developing a fast algorithm is to study the long time dynamics of the model. In the original algorithm

[1] Without any loss of generality in all our simulations we shall use $M = 2$.