New Journal of Physics
The open-access journal for physics

Free association transitions in models of cortical latching dynamics

Eleonora Russo[1,3], **Vijay M K Namboodiri**[2,4], **Alessandro Treves**[1,5] **and Emilio Kropff**[1]

[1] SISSA, Cognitive Neuroscience, via Beirut 4, 34014 Trieste, Italy
[2] Department of Physics, IIT Bombay, Powai, Mumbai, India 400076
E-mail: russo@sissa.it, vijay_mkn@iitb.ac.in, ale@sissa.it and kropff@sissa.it

New Journal of Physics **10** (2008) 015008 (19pp)
Received 31 July 2007
Published 31 January 2008
Online at http://www.njp.org/
doi:10.1088/1367-2630/10/1/015008

Abstract. Potts networks, in certain conditions, hop spontaneously from one discrete attractor state to another, a process we have called *latching* dynamics. When continuing indefinitely, latching can serve as a model of infinite recursion, which is nontrivial if the matrix of transition probabilities presents a structure, i.e. a rudimentary *grammar*. We show here, with computer simulations, that latching transitions cluster in a number of distinct classes: effectively random transitions between weakly correlated attractors; structured, history-dependent transitions between attractors with intermediate correlations; and oscillations between pairs of closely overlapping attractors. Each type can be described by a reduced set of equations of motion, which, once numerically integrated, matches simulations results. We propose that the analysis of such equations may offer clues on how to embed meaningful grammatical structures into more realistic models of specific recursive processes.

[3] Principal contributor to section 5.
[4] Principal contributor to section 4.
[5] Author to whom any correspondence should be addressed.

Contents

1. Introduction

Complex thought processes, as well as their uniquely human expression through language, appear to be based on the same cortical machinery that we essentially share with other mammals, despite major variations in absolute size and relatively minor variations in internal organization [1]–[4]. This suggests, in our view, that in order to understand cognitive capacities that are apparently uniquely human one should consider the possibility that they may arise 'spontaneously', through phase transitions induced by quantitative changes in certain parameters of cortical organization. If so, it would not be unreasonable to utilize simple generic models of cortical processing in order to model language processes [5], provided proper consideration is taken of quantitative aspects.

Following a suggestion by Chomsky and colleagues [6], we have focused on the emergence of *latching* dynamics in models of cortical networks, as a simplified model of a recursive process [7]. A computational mechanism for recursion provides, as pointed out in [6], for the generation of an infinite range of expressions of arbitrary length, out of a finite set of elements. Such sequences of elements, if not spanning uniformly the space of all possibilities but rather constrained by a non-trivial structure of transition probabilities at each recursion step, in fact implement a syntax, comparable in principle to those observed in natural language or in reasoning [8]—although in the reduced models that we can simulate such syntactic structure is a rather abstract and apparently pointless statistical characterization of the latching transition matrix.

A transition from finite to infinite recursion may have occurred in the human species tens of thousands of years ago, as a sudden result of a gradually expanding cortical connectivity [7] and it may have later been bootstrapped and refined, of course, by additional complex processes of cultural evolution [8].

In previous reports, we have considered a simplified model of a semantic memory system, implemented as a Hopfield associative network with Potts variables [7, 9, 10]. We have shown how to analyse the storage capacity of the model [11], which characterizes it even in regions of parameter space in which no latching dynamics occurs. We have also provided a first description of the structure of latching transitions [12], which we aim to characterize in more detail in the present study. We refer to these earlier publications for a more extensive introduction on semantic memory, on the representation of concepts through features and on cortical organization, all crucial elements to motivate the analysis of the model. We also refer to a recent paper that discusses the storage of correlated representations, a necessary trigger for latching dynamics to occur [13].

2. The Potts–Hopfield memory model

The model can be regarded as an attractor neural network whose units represent themselves local attractor networks, realized in small patches of cortex, and which can each converge dynamically into one of S local attractors. The activity of local network i can then be described synthetically by an analog 'Potts' unit, i.e. a unit that can be correlated to various degrees with any one of S local attractor states. The state variable of the unit, σ_i, is thus a vector in S dimensions, where each component of the vector measures how well the corresponding feature is being retrieved by the local network. The possibility of no significant retrieval—no convergence and hence no correlation with any local attractor state—can be added through an additional 'zero' state. Because the local state cannot be fully correlated, simultaneously, with all S features and with the zero state, one can use a simple normalization $\sum_{k=0}^{S} \sigma_i^k = 1$. Having introduced such Potts units as models of local network activity, in the following we will use the terms 'local network' and 'unit' as synonyms.

The global network is comprised of N (Potts) units connected to one another through tensor sets of weights, which represent collections of long-range synaptic connections, between distant patches of cortex. The network has stored p Potts activity patterns, as global attractor states that represent concepts in semantic memory. When global pattern ξ^μ is being retrieved, the state of the local network i is in the local attractor state $\sigma_\mathbf{i} \equiv \xi_i^\mu$, retrieving feature ξ_i^μ, a discrete value which ranges form 0 to S (the zero value standing for no contribution of this group of features to concept μ). As shown in [11], such a compositional representation of concepts as sparse constellations of features (with a global sparsity parameter a measuring the average fraction of features active in describing a concept) leads to the desired global attractor states when long range connections have associated weights J_{ij}^{kl}

$$J_{ij}^{kl} = \frac{C_{ij}}{c_\mathrm{M} a(1-(a/S))} \sum_{\mu=1}^{p} \left(\delta_{\xi_i^\mu k} - \frac{a}{S}\right)\left(\delta_{\xi_j^\mu l} - \frac{a}{S}\right)(1-\delta_{k0})(1-\delta_{l0}), \tag{1}$$

which can be interpreted as resulting from Hebbian learning. In this expression, each element of the connection matrix C_{ij} is 1 if there is a connection between units i and j, and 0 otherwise (the diagonal of this matrix is filled with zeros), while c_M stands for the average number of connections arriving to Potts unit i. In this model, the maximum number of patterns, or concepts, which the network can store and retrieve scales roughly like $c_\mathrm{M} S^2/a$. We refer to [11] for an extensive analysis of the storage capacity of the Potts model.

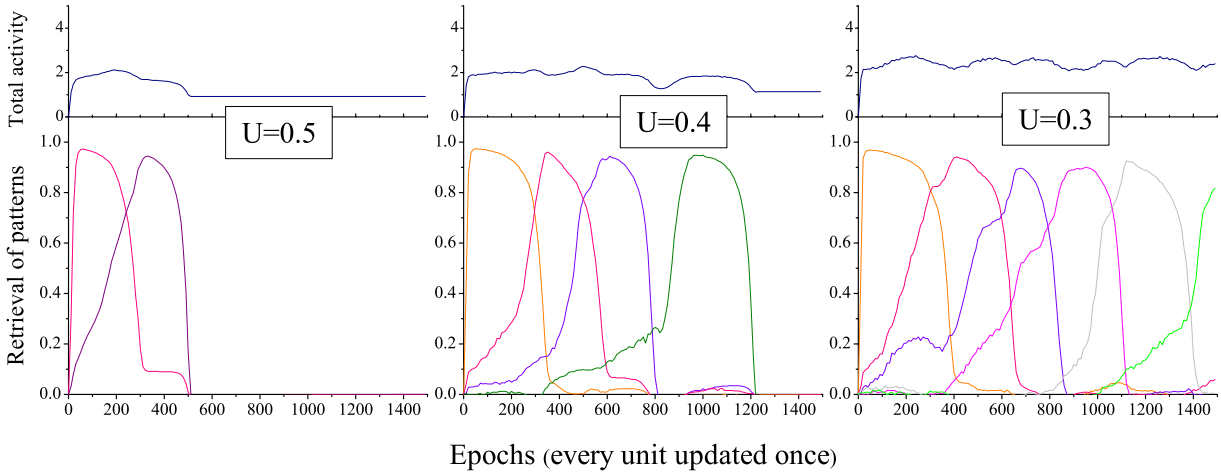

Figure 1. Examples of latching dynamics for the 3 values of U: 0.5, 0.4 and 0.3 (from left to right). Top plots: the evolution of the sum of all the activity in the network. Bottom: overlap of the state with the most relevant patterns. Each colour corresponds to a different pattern.

3. Basic conditions for latching dynamics

Here, we are interested in studying not the storage capacity but rather the dynamics of such a Potts model of a semantic network. Latching dynamics emerges as a consequence of incorporating two additional crucial elements in the Potts model: neuronal adaptation and correlation among attractors. Intuitively, latching may follow from the fact that all neurons active in the successful retrieval of some concept tend to *adapt*, leading to a drop in their activity and a consequent tendency of the corresponding Potts units to drift away from their local attractor state. At the same time, though, the residual activity of several Potts units can act as a cue for the retrieval of patterns *correlated* to the current global attractor. As usual with autoassociative memory networks, however, the retrieval of a given pattern competes, through an effective inhibition mechanism, with the retrieval of other patterns. One can then imagine a scenario in which two conditions are fulfilled simultaneously: the global activity associated with a decaying pattern is weak enough to release in part the inhibition preventing convergence toward other attractors; but, as an effective cue, it is strong enough to trigger the retrieval of a new, sufficiently correlated pattern. In such a regime of operation, after the first, externally cued retrieval, the network state experiences the concatenation in time of successive memory patterns, i.e. it latches from attractor to attractor (see figure 1).

In a previous report [12], we have offered a first description of the *complexity* of latching dynamics, and discussed which parameters control it. Latching transitions were seen to be neither deterministic nor random, nor to depend solely on the correlation between consecutive attractor states. Furthermore, a marked asymmetry was observed in the transition matrix, controlled by a threshold parameter U.

3.1. Introducing a model of adaptation

In retrieval dynamics without adaptation, units are updated with the rule

$$\sigma_i^k = \frac{\exp(\beta h_i^k)}{\sum_{l=0}^{S} \exp(\beta h_i^l)}, \tag{2}$$

under the influence of a tensorial local 'current' signal which sums the weighted inputs from other units, with a fixed threshold U favouring the zero state

$$h_i^k = \sum_{j=1}^{N} \sum_{l=0}^{S} J_{ij}^{kl} \sigma_j^l + U\delta_{k0}. \tag{3}$$

To model firing rate adaptation, however, we introduce a modification in the individual Potts unit dynamics. The update rule

$$\sigma_i^k = \frac{\exp(\beta r_i^k)}{\sum_{l=0}^{S} \exp(\beta r_i^l)} \tag{4}$$

is now mediated, for $k \neq 0$, by the vectors r (the 'fields', or 'local potentials', which integrate the h 'currents') and θ (the dynamic thresholds specific to each state), which are integrated in time

$$r_i^k(t+1) = r_i^k(t) + b_1[h_i^k(t) - \theta_i^k(t) - r_i^k(t)], \tag{5}$$

$$\theta_i^k(t+1) = \theta_i^k(t) + b_2[\sigma_i^k(t) - \theta_i^k(t)], \tag{6}$$

where the fields are assumed to change more rapidly than the thresholds, i.e. with time constants $(b_1)^{-1} < (b_2)^{-1}$. We also include a nonzero local field for the zero state, driven by the (slow) integration of the total activity of unit i in all nonzero directions, $(1 - \sigma_i^0)$.

$$r_i^0(t+1) = r_i^0(t) + b_3[U + 1 - \sigma_i^0(t) - r_i^0(t)], \tag{7}$$

with now $(b_1)^{-1} \ll (b_3)^{-1}$. The local field for the zero state, which is taken to be initially equal to U, eventually increases towards $U + 1$ for active units, down-regulating their activity and thus preventing local 'overheating'—and at the same time destabilizing ordinary fixed-point attractors. Note that a fixed threshold U of order 1 is crucial to ensure a large storage capacity (as shown in [14]) and to enable unambiguous memory retrieval, precisely by stabilizing the fixed-point attractors that here we destabilize over a slower timescale $(b_3)^{-1}$.

A final element we include, partially correcting the effect of the field for the zero state, is an effective self-coupling J_{ii}^{kk}, constant for every i and $k \neq 0$, which adds stability to the local network.

3.2. Generating correlated distributions

A standard mathematical procedure to introduce model correlations in a group of p patterns is through a hierarchical algorithm, which may be parametrically varied from producing independent to highly correlated patterns. Patterns are defined using one or more generations of *parents*, from which they descend, emulating a genetic tree. Since many patterns share the same parents, the generation process introduces correlations among descendant patterns, which are simpler for one-parent families and more complex in the case of multiple parents. We adopt a multi-parent scheme, in particular we allow for up to 200 parents, which we call *factors* [7].

They represent semantic category generators, relating directly the correlation between patterns to categorization in a real semantic system, so as to preserve a possibility to link the correlational statistics of our model to observations in the cognitive neuroscience of semantic memory, which we pursue in a cognate report [13].

These factors are defined simply as distinct random subsets of the entire set of Potts units. In the simulations, each subset includes Nf units out of the total N units, and a total of 200 such factors are generated. The overlaps in the spatial distribution of different factors therefore are purely random, and clustered around their mean value Nf^2.

Next, global patterns are generated from the factors, which have been indexed by n in order of decreasing mean importance. For each global pattern, the specific importance of each factor is given by a coefficient $\gamma_{\mu n}$ obtained by multiplying the overall factor $\exp(-\zeta n)$ by a random number, taken to be 0 with probability $1-a$, and otherwise drawn with a flat distribution between 0 and 1, specifically for pattern μ. A value taken by factor n, σ_n, is randomly drawn among the S 'genuine' attractors, and a contribution $\gamma_{\mu n}$ is added to the field onto each Potts unit over which factor n has been defined, in the direction σ_n. After accumulating contributions from all factors, the direction in which each unit received the largest field is computed, and the Na units receiving the largest maximal fields are assigned the corresponding direction σ_n in pattern μ, while the remaining $N(1-a)$ units are assigned the null state in pattern μ.

With this procedure, pairs of Potts units have uncorrelated activity when averaged across patterns (because the different patterns that both engage the pair will span nearly evenly the different local states). Pairs of patterns, instead, can by highly correlated once averaged across units, particularly if they share one or a few most important factors; and positively correlated if these factors have been assigned the same direction in Potts space. Thus, correlations among patterns will be higher if the importance of different factors decreases rapidly (e.g. in the simulations the value $\zeta = 0.02$ was used, equivalent to assuming of order 50 'important' factors); and they will tend to vanish if all factors are equally important ($\zeta = 0$). When correlations are very high each pattern tends to be significantly correlated with a specific subset of the others, those sharing the main factor that influences them, and positively correlated with a fraction $1/S$ of this subset. In this scheme, the number of memory items significantly overlapping with one recently retrieved, and which can be the target of a non-random transition, scales up as p/S, and does not depend on the connectivity. By contrast, the storage capacity for retrieval can still scale up as in the case of uncorrelated patterns, if a proper learning rule is used [13].

To characterize statistically the correlations among the resulting set of patterns we introduce for each pair of patterns μ and ν the quantities

$$C_0^{\mu\nu} = N(1-a)C_0^{\mu\nu} = \sum_{i=1}^{N} \delta_{\xi_i^\mu \xi_i^\nu} \delta_{\xi_i^\nu 0}, \tag{8}$$

$$C_1^{\mu\nu} = NaC_1^{\mu\nu} = \sum_{i=1}^{N} \delta_{\xi_i^\mu \xi_i^\nu} (1 - \delta_{\xi_i^\nu 0}), \tag{9}$$

$$C_2^{\mu\nu} = NaC_2^{\mu\nu} = \sum_{i=1}^{N} (1 - \delta_{\xi_i^\mu \xi_i^\nu})(1 - \delta_{\xi_i^\mu 0})(1 - \delta_{\xi_i^\nu 0}). \tag{10}$$

For any two patterns $\mu \neq \nu$ (in the following we drop their indices for simplicity) C_0 is obviously the number of inactive units they share, C_1 the number of active units which are shared and

in the *same* state and C_2, on the other hand, the number of shared active units which are in *different* states. C_0, C_1 and C_2 are the corresponding fractions, i.e. normalized to their respective maximum values. In [12] it was shown how to estimate the means and variances of these quantities, given the hierarchical algorithm for generating correlations.

4. The statistics of the latching transitions

We ran a large set of simulations using the dynamics explained in section 3.1. First of all, we created sets of p patterns using the algorithm described in section 3.2. Each simulation started by giving an initial cue to the network (as an additional term in the local field) in order to induce the retrieval of one of the stored patterns. The network was then left free to evolve until one of two stop conditions was reached: either the activity decayed to zero or else each unit was updated 50 000 times—keeping track of latching events. The simulation was run 5 or 100 times for each cued pattern, with different random seeds, and all p patterns were used as the cued pattern. In this way, we collected datasets of latching events, with which we constructed an estimate of the transition probability matrix M. Since we found that all statistical quantities stabilize within the shorter simulations, the longer ones were used merely as control data. For the simulations in this section we have set the parameters $b_1 = 0.1$, $b_2 = 0.005$, $b_3 = 1$ and $J_{ii}^{kk} = 1.8$, where the 'latching' time constant $(b_3)^{-1}$ was made implausibly fast so as to speed up the simulations and collect sufficient statistics.

The probability matrix is a square matrix with $p+1$ rows and columns, the additional one corresponding to the 'null' attractor, with each unit in the zero state. To estimate the transition probability between state μ and ν, we counted the times a latching event between these two attractors appeared in the dataset. We added a transition to the 'null' state whenever global activity decayed to zero and assumed a probability of 1 for the transition from the null state to itself. Finally, given that M_{ij} represents the probability of having a latching transition from global attractors i to j, the sum of matrix elements over each row was normalized to 1.

4.1. The role of the threshold

In [12], we have studied the way latching dynamics depends on the threshold U. We reproduce figure 1, which shows examples of the latching behaviour for $U = 0.3$, 0.4 and 0.5.

In terms of the transition matrix M, we have observed that, as expected, a high threshold U selects a subset of transitions, and the matrix has a few large and many zero or vanishingly small elements. As U decreases, more elements of M are nonzero, and they tend to span more of a continuum of values. We have also found that M is far from symmetric (even though the correlations between patterns are symmetric by definition). As the threshold U decreases and randomness grows, the transition probability matrix was observed to become somewhat more symmetric. The complexity of the transitions was also quantified with Shannon's information measure, computed over each row of M

$$I_\mu = \frac{1}{\log_2(p+1)} \sum_{\nu=1}^{p+1} M_{\mu\nu} \log_2 \left(\frac{1}{M_{\mu\nu}} \right), \tag{11}$$

so that $I_\mu \sim 0$ both if the attractor μ generates no latching (and thus decays to zero) or if it latches to another fixed attractor, deterministically; if instead the process of latching is completely random, $I_\mu = 1$. In terms of such complexity, we have observed that decreasing

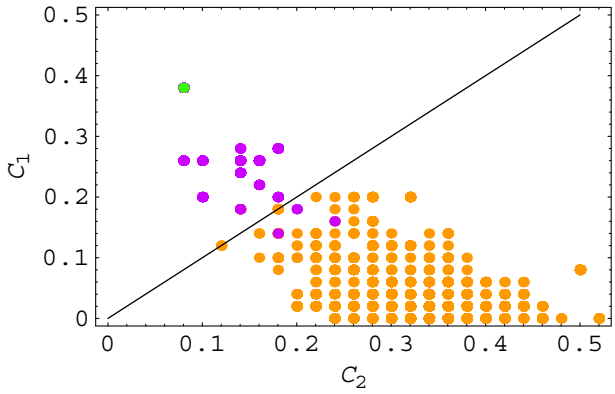

Figure 2. The space spanned by C_1 and C_2 shows three distinct latching regions (in three different colours) which correspond to different latching behaviours. These behaviours are classified based on λ (see text). The line drawn is the one at which $C_1 = C_2$. The parameter values are $N = 200$, $p = 50$, $S = 7$, $a = 0.25$ and $U = 0.4$.

the threshold, from $U = 0.5$ to $U = 0.3$, I_μ increases from a nearly deterministic mean value of $I_\mu \simeq 0.03$ to a largely random mean of $I_\mu \simeq 0.7$, thus suggesting that raising the threshold can effectively span the entire range from random to deterministic.

4.2. Diversity of latching dynamics

Here, we study the distinct types of latching transitions that can be observed even at a fixed value of the threshold U. Since correlations between patterns are obviously a major determinant of the transitions, we considered the distribution of correlations, parametrized for each μ and ν pair by the quantities $\mathcal{C}_0^{\mu\nu}$, $\mathcal{C}_1^{\mu\nu}$ and $\mathcal{C}_2^{\mu\nu}$ defined above. We computed these distributions using (i) the whole set of patterns and (ii) the dataset of latching events. In the first case, each pair of patterns enters the average once and only once. In the second case, only pairs of attractors visited one after another in at least a latching event are considered, with a weight proportional to their frequency of occurrence in the dataset. In [12] we found that the distribution of $\mathcal{C}_0$ values does not vary appreciably between (a) and (b) (to a large extent, quite probably, because its variance is limited). Hence, we focus here on the distribution of $\mathcal{C}_1$ and $\mathcal{C}_2$, which vary significantly across pairs.

We have found three different kinds of latching behaviour in the space formed by C_1 and C_2. They are as shown in figure 2.

We characterize these three regions based on another variable, which is the value of the retrieval overlap at which two consecutive latching patterns cross over each other (see e.g. figure 1). We call this new variable λ.

It is seen that when the value of λ is high (between 0.6 and 0.8), the value of $\mathcal{C}_1$ is generally higher than that of $\mathcal{C}_2$—latching occurs between patterns that are significantly correlated (note that, for uncorrelated patterns, $C_1 \simeq C_2/S$). This region consists of the points shown in violet in figure 2. The points in orange are the ones with λ small (less than 0.2) and these fall, with the exception of a few points, into the region where $\mathcal{C}_1 \leqslant \mathcal{C}_2$. The line $\mathcal{C}_1 = \mathcal{C}_2$ separates the two

Table 1. Second and third largest eigenvalues of M and the corresponding decay times n_{dec}, as defined in equation (13), for the 3 random number generator seeds used in the simulations shown in the figures.

Seed	λ_2	λ_3	n_{λ_3}	n_{λ_3}
1	$\geqslant$0.99	0.88	$\geqslant$230	18
2	$\geqslant$0.98	0.23	$\geqslant$115	1.6
3	$\geqslant$0.98	0.27	$\geqslant$115	1.7

regions. Note that no transitions are observed with intermediate values of λ, which shows that there is no continuous transition between these two regions.

The dichotomy can be intuitively understood because a high value of λ is expected when there is a large overlap between the corresponding patterns of activity (see figure 1) and this implies a high value of $\mathcal{C}_1$, since $\mathcal{C}_1$ is the number of units which are shared between the two patterns, i.e. active and in the same state.

The lone green data point falls into yet another 'region', since in this case the two latching patterns oscillate among themselves in activity and meet at a very high value of λ, of around 0.85 or more.

4.3. Eigenvalue analysis

As M is a transition probability matrix, the eigenvalues of M can be shown to have a modulus lower than or equal to one. Because of the construction of the matrix, the eigenvalue corresponding to the zero pattern, which projects entirely into itself, is $\lambda_0 = 1$. In the general case, when applying the transition matrix n times to an initial pattern μ, the result may as usual be decomposed as

$$M^n \hat{\mathbf{x}}_\mu = A D^n A^{-1} \hat{\mathbf{x}}_\mu = A^{-1}_{0\mu} \hat{\mathbf{x}}_0 + \sum_{k=1}^{p} \lambda_k^n A^{-1}_{k\mu} \mathbf{v}_k, \tag{12}$$

where D is the diagonal matrix with the same eigenvalues as M, A is the basis change matrix with the eigenvectors of M as columns, λ_k is the kth eigenvalue of M, $\mathbf{v}_k$ the corresponding eigenvector and $\hat{\mathbf{x}}_\eta$ is the unitary versor with elements $(\hat{\mathbf{x}}_\eta)_i = \delta_{i\eta}$. Thus, we see that for large values of n activity will eventually decay to the 'null' attractor, unless some non-null eigenvector of M has an eigenvalue of 1. Whenever this is not the case, the decay time is given by the second largest eigenvalue of M. More specifically, for any eigenvalue λ_k, the number of transitions for its eigenspace to decay, for example, to 0.1 of its original amplitude is given by

$$n_{\text{dec}} = \log_{\lambda_k}(0.1). \tag{13}$$

In table 1, we show n_{dec} for the second and the third largest eigenvalues, and for three different random number seeds. For each of the three seeds, the second largest eigenvalue corresponds to modes that do not decay over the entire length of the simulation (the convergence to an attractor and subsequent drift away from it take, with our parameters, between 300 and 500 updates of each unit, which multiplied by $n_{\text{dec}} \geqslant 100$ is of the same order as the 50 000 updates we set as the maximum duration of the simulation). So in each of the three examples, some sort of latching dynamics did occur indefinitely, although in the case of the first seed it was clearly of a

peculiar type. The third largest eigenvalue, when also close to 1, indicates that there are at least two groups of states that dominate the long term behaviour, and are dynamically kept separate for a long time.

In general, the emergence of unitary eigenvalues in the matrix, apart from the one corresponding to the null state, is of great interest, because it indicates the transition from high-order (but finite) recursion to infinite recursion. More analysis is obviously required to understand this phase transition, but it appears from our simulations that quenched disorder (the random seed generator, determining the exact realization of correlated attractors) can bring the system into either phase, even when all parameters take the same values. It remains to be seen whether in a large enough system the variability due to quenched disorder progressively vanishes. The way the probability of observing indefinite latching depends on connectivity parameters, like c_M and S (which is indirectly a connectivity parameter in the local cortical network interpretation of each Potts unit) has been sketched in [7].

4.4. Non-ergodicity appears as distinct latching behaviours

In the case of a particular seed (seed 1), latching was dominated by two patterns which fell in the green region of figure 2. This can be taken to be a somewhat pathological case, determined in part by the high correlation between the two patterns, and in part by the lack of a suitable 'escape route' from the limit cycle they effectively comprise. For the other two seeds, the latching patterns ranged through all the values of C_1 (note that we ran the simulations twice, one for 5 cycles with each external cue and the other for 100 cycles with each external cue, without noting any appreciable difference).

These results are shown in figure 3.

The interesting feature to note in these three examples is that the frequencies of C_1 values over latching pairs, relative to the general distribution, show an initial dip, beyond values of $\mathcal{C}_1$ of 0 and 1: there are fewer transitions between pairs with $\mathcal{C}_1 = 2$ and onward, than with $\mathcal{C}_1 = 0$ or 1. Though this decreased frequency is a small effect, it stands in contrast with the notion, suggested by previous analyses, that the latching transition probability simply increases monotonically with correlation, i.e. with C_1 [12], an effect that is still valid for high values of C_1. The dip is seen from the graphs to be due to the many transitions that have λ very small, i.e. the random transitions, that fall in the orange region of figure 2. Such transitions occur preferentially at very low correlation, $\mathcal{C}_1 = 0$ or 1, and relatively less frequently between pairs of patterns with higher values of C_1. This observation motivates the study of the detailed dynamics of individual transitions, both in the low- and in the intermediate-correlation regimes.

5. Transition dynamics

In order to study the dynamical behaviour of the system during a single latching transition, we may complement the simulation approach with an analytical one. Simulations were performed following the same updating rules defined in section 3.1, but on larger networks ($N = 10\,000$) and with parameters more appropriate to follow the detailed dynamics of individual transitions. The analytical approach considers the field affecting each unit as a sum of 'signal' and 'noise' terms, as described in the following, and derives differential equations that govern the dynamics of each subgroup of units which receive the same signal.

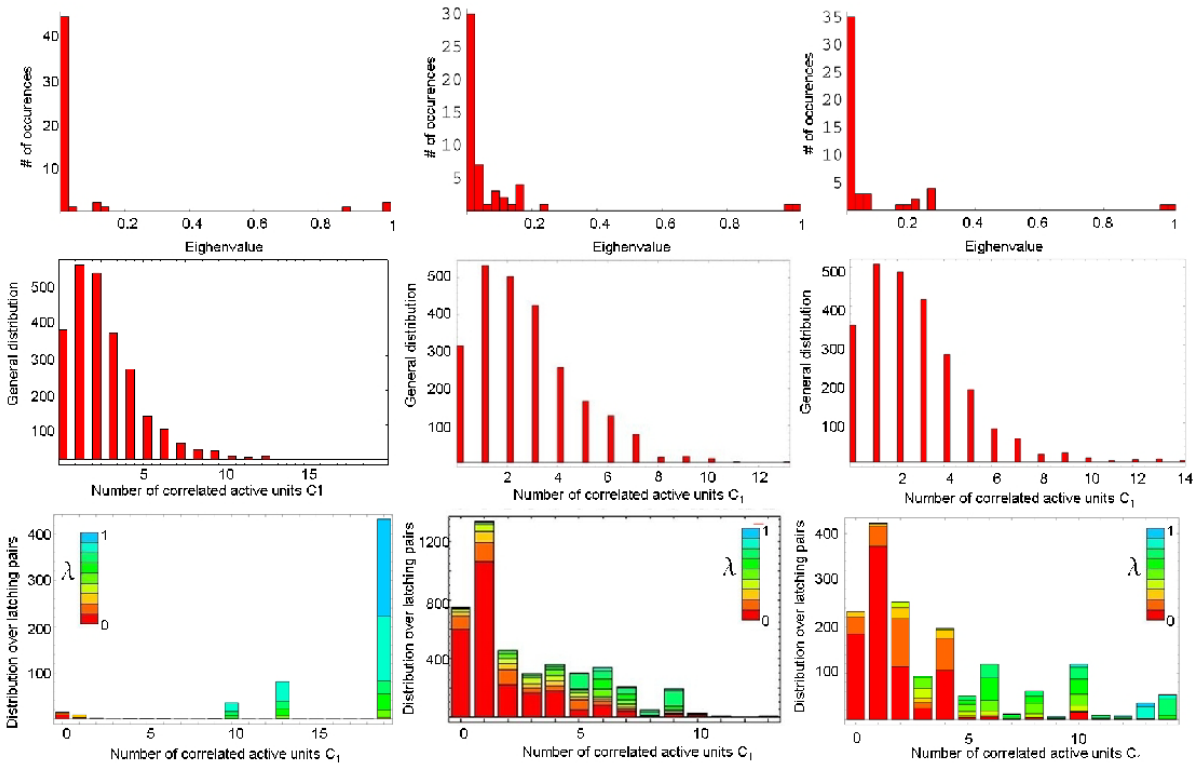

Figure 3. In columns organized from left to right, histograms describing the results of three sets of simulations with different seeds. Inside each of these columns, the top graph shows a histogram of the eigenvalues of the transition matrix, while the centre and bottom graphs show, respectively, the distributions of C_1 for all pairs of patterns and for pairs of patterns participating in latching. In the bottom graphs the bars are subdivided in regions of different colour indicating the λ value at what the corresponding transition occured. For the column in the left, the behaviour is dominated by transitions which occur for a very high value of C_1 (and λ). As indicated by the colour of the bar, it is due to the transitions that fall in the green region of figure 2. This pathological case is not present in the simulations in the centre and right column. The parameter values are, again, $N = 200$, $p = 50$, $S = 7$, $a = 0.25$ and $U = 0.4$.

5.1. Signal-to-noise analysis

We start by writing the expression of the field affecting each nonzero state of unit i, from section 3.1

$$h_i^k = \sum_{j=1}^{N} \sum_{l=0}^{S} J_{ij}^{kl} \sigma_j^l, \tag{14}$$

in terms of the *overlaps* between the state of the system and each pattern μ, defined as

$$m_\mu \equiv \frac{1}{Na(1-(a/S))} \sum_{j\neq i}^{N} \sum_{l\neq 0}^{S} \left(\delta_{\xi_j^\mu l} - \frac{a}{S}\right) \sigma_j^l \tag{15}$$

and neglecting both the self interaction terms and the quenched disorder implied by the sparse connectivity among Potts units; the mean-field expression for the field becomes

$$h_i^k \simeq \sum_{\mu}^{p} m_\mu \left(\delta_{\xi_j^\mu k} - \frac{a}{S}\right) + U\delta_{k0}. \tag{16}$$

The signal-to-noise analysis [11] proceeds by singling out any pattern with macroscopic overlap with the current state, and treating other patterns as contributing only to the noise. For example, when focusing on a latching transition between two patterns $\mu = 1$ and $\nu = 2$, one may write

$$h_i^k \simeq m_1 \left(\delta_{\xi_j^1 k} - \frac{a}{S}\right) + m_2 \left(\delta_{\xi_j^2 k} - \frac{a}{S}\right) + \sqrt{\frac{\alpha a}{S^2}} q\eta + U\delta_{k0}, \tag{17}$$

where the noise amplitude q reflects the global activity of the network, defined as follows

$$q \equiv \frac{1}{Na(1-(a/S))} \sum_{j\neq i}^{N} \left[\sum_{l\neq 0}^{S} \sigma_j^{l\,2} - \frac{a}{S}(1-\sigma_j^0)^2\right], \tag{18}$$

$\alpha = p/c_{\mathrm{M}}$ parametrizes the storage load and the Gaussian variable η has zero mean and unitary variance [11]. Note that the description in terms of equations (17) and (18), which is a reasonable approximation when patterns are uncorrelated, is much more delicate in the presence of correlations. Even when $p \ll c_{\mathrm{M}}$ the effect of the 'noise' term may remain important, due to correlations with other patterns, that make equation (17) inappropriate.

Choosing an asynchronous updating procedure, in which a unit i is randomly selected and all relevant dynamical variables (its own h_i, r_i and θ_i, plus the global m and q) are updated at each micro-step, the update of the network takes of order N single updates. We take this timescale as the unitary (macroscopic) time step, and we focus on the equations detailing the changes occurring within a micro-step of duration $\Delta t = 1/N$.

Considering the definition of m, we can write the overlap value at time $t + \Delta t$ in terms of the old value

$$m_\mu(t+\Delta t) = m_\mu(t) + \frac{1}{Na(1-(a/S))} \sum_{l\neq 0}^{S} \left(\delta_{\xi_j^\mu l} - \frac{a}{S}\right)(\sigma_i^l(t+\Delta t) - \sigma_i^l(t))$$

and from this, we derive the differential equation for the overlaps as

$$\begin{aligned}\frac{\mathrm{d}m_\mu(t)}{\mathrm{d}t} &= \frac{m_\mu(t+\Delta t) - m_\mu(t)}{(1/N)} \\ &= \frac{1}{a(1-(a/S))} \sum_{l\neq 0}^{S} \left(\delta_{\xi_i^\mu l} - \frac{a}{S}\right)(\sigma_i^l(t+\Delta t) - \sigma_i^l(t)),\end{aligned}$$

where unit i was randomly chosen to be updated at time t. Averaging over such random choices we write

$$\begin{aligned}\frac{\mathrm{d}m_\mu(t)}{\mathrm{d}t} &= \frac{1}{Na(1-(a/S))} \sum_{i}^{N} \sum_{l\neq 0}^{S} \left(\delta_{\xi_i^\mu l} - \frac{a}{S}\right)(\sigma_i^l(t+\Delta t) - \sigma_i^l(t)) \\ &\simeq -m_\mu(t) + \frac{1}{a(1-(a/S))} \left\langle \sum_{l\neq 0}^{S} \left(\delta_{\xi^\mu l} - \frac{a}{S}\right)\sigma^l(t^+)\right\rangle_{\text{all units}},\end{aligned} \tag{19}$$

where the notation t^+ means after updating the whole network. A similar procedure can be followed for the variable q to derive the equation

$$\frac{\mathrm{d}q(t)}{\mathrm{d}t} = -q(t) + \frac{1}{a(1-(a/S))}\left\langle\left[\sum_{l\neq 0}^{S}\sigma^{l2}(t^+) - \frac{a}{S}(1-\sigma^0(t^+))^2\right]\right\rangle_{\text{all units}} . \quad (20)$$

Together with equation (19) and the updating of individual units in terms of current values of m and q, the above equation describes completely the dynamics of the system. For example, if we focus on a transition between two patterns with macroscopic overlap at time t, we may write

$$\frac{\mathrm{d}m_1(t)}{\mathrm{d}t} = -m_1(t) + \frac{1}{a(1-(a/S))}\left\langle\sum_{k\neq 0}^{S}\left(\delta_{\xi^1 k} - \frac{a}{S}\right)\frac{\exp(\beta r^k)}{\sum_{l=0}^{S}\exp(\beta r^l)}\right\rangle_{\text{all units}},$$

$$\frac{\mathrm{d}m_2(t)}{\mathrm{d}t} = -m_2(t) + \frac{1}{a(1-(a/S))}\left\langle\sum_{k\neq 0}^{S}\left(\delta_{\xi^2 k} - \frac{a}{S}\right)\frac{\exp(\beta r^k)}{\sum_{l=0}^{S}\exp(\beta r^l)}\right\rangle_{\text{all units}},$$

$$\frac{\mathrm{d}q(t)}{\mathrm{d}t} = -q(t) + \frac{1}{a(1-(a/S))}\left\langle\left[\sum_{k\neq 0}^{S}\left(\frac{\exp(\beta r^k)}{\sum_{l=0}^{S}\exp(\beta r^l)}\right)^2 - \frac{a}{S}\left(1-\frac{\exp(\beta r^0)}{\sum_{l=0}^{S}\exp(\beta r^l)}\right)^2\right]\right\rangle_{\text{all units}},$$

$$\frac{\mathrm{d}r_i^k(t)}{\mathrm{d}t} = b_1[h_i^k(t) - \theta_i^k(t) - r_i^k(t)],$$

$$\frac{\mathrm{d}\theta_i^k(t)}{\mathrm{d}t} = b_2[\sigma_i^k(t) - \theta_i^k(t)],$$

$$\frac{\mathrm{d}r_i^0(t)}{\mathrm{d}t} = b_3[U + 1 - \sigma_i^0(t) - r_i^0(t)].$$

The information we have about the patterns is only statistical, as explained in section 3.2. Thus, the key to using the above equations is to group the last three sets (which comprise $(2S+1)\times N$ individual equations) for units and states that receive the same signal.

5.2. Memory retrieval dynamics

As a simple example, we may consider the retrieval of a single memory pattern by an external cue. In this case, we have to follow separately the dynamics of units which are active or quiescent in the memory pattern to be retrieved, by tracking the fields affecting their quiescent and active states.

In the simulation shown in figure 4, the network is prepared at time 0 in a quiescent activity state, with zero fields and thresholds for the active states and $r^0 = 1.5$. An external cue arrives at time step $t = 200$, providing a signal term pointing at pattern $\mu = 1$ in the field to each unit. We used $p = 5$ patterns, hence a low memory load regime, in which the noise due to interference with other patterns, which were constructed without correlations, may be safely neglected in the analytical treatment; thus we neglect the variable q. Other parameters were $S = 3$, $a = 0.1$, $U = 1.5$, $b_1 = b_2 = 0.01$ and $b_3 = 0.0$ (for simplicity we thus omit also the evolution of the field affecting the zero state).

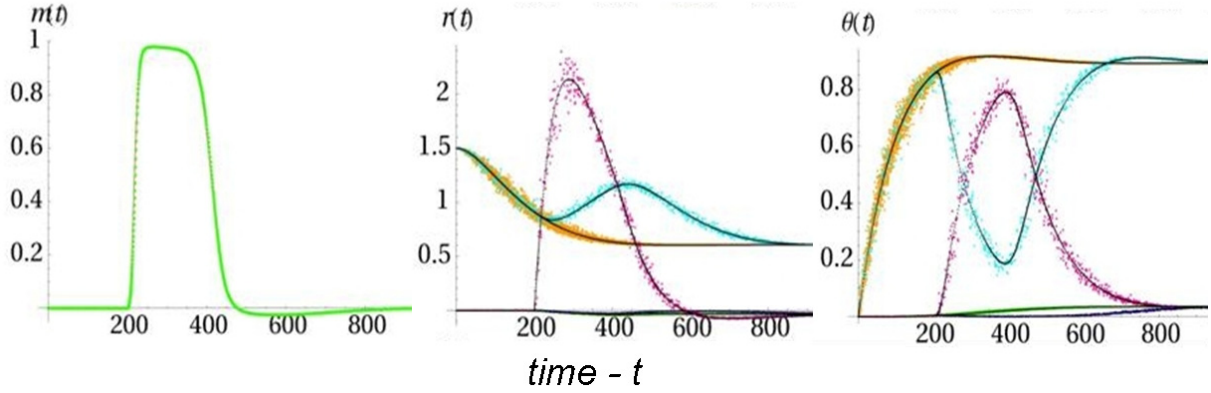

Figure 4. Evolution of the overlap with a single cued pattern (left) and of the fields (middle) and thresholds (right) to five groups of units receiving each the same mean-field signal, as explained in the text. Colour data points are extracted from the simulations, upon the updating of successive units, whereas black curves are obtained by numerical integration of approximate dynamical equations.

Even in this simple situation, we need to distinguish:

1. units that are active in ξ^1 and their corresponding fields in the state $k = \xi^1$ (magenta), in other active states $l \neq \xi^1$ (blue dots, slightly distinguishable) and in the zero state (light blue);
2. units that are inactive in ξ^1 and their corresponding fields in active states (green, not distinguishable) and in the zero state (orange).

In figure 4, the variables that we identify as 'not distinguishable' do not contribute to the dynamics since their values remain close to zero.

With these simplifications, we are left with five differential equations to integrate for the fields, five for the thresholds, and one for the single relevant overlap. Their integration leads to a time evolution of the different quantities (fields and thresholds are shown as black curves in figure 4) in excellent agreement with the simulations (the data points in colour—a data point corresponds to a unit being updated). The latter show some unit-to-unit variability, which can be reduced by taking a moving average over units updated at similar times (not shown). After retrieving the cued pattern, in this simulation the network relapses into the quiescent state.

5.3. Quasi-random transitions

One can extend the analysis above to the more interesting case of latching transitions between pairs of patterns. The grouping of units and states into coherent mean-field ensembles is much more tedious, however. We focus here on the relatively simple case of only two correlated patterns, between which we may observe latching, given $\mathcal{C}_1$ units that share the same active state in both patterns, and we neglect to consider any other pattern. The $\mathcal{C}_1$ units may be expected to be active during the retrieval of the first pattern and to remain active and in the same state

during the retrieval of the second one. In any case, this group of units obviously follows its own dynamics, that differs from the one followed by other groups of units. In total, we need to consider five distinct groups of units, that correspond to:

1. units active in ξ^1 and in ξ^2, and in the same state;
2. units active in ξ^1 and in ξ^2, but in a different state;
3. units active in ξ^1 but not in ξ^2;
4. units inactive in ξ^1 but active in ξ^2;
5. units inactive both in ξ^1 and in ξ^2;

and for these different groups, to distinguish the relevant states (among $S+1$ ones), and consider their evolution separately; including equations for two overlaps and for the variable q, even in this most simplified situation we obtain 28 integrable differential equations, that we do not write down here (see Eleonora Russo, unpublished MSc Thesis, for the full derivation).

In the simulation shown in figure 5(b), we identified an example of a latching transition characterized by a low correlation between the two patterns (in fact, an anticorrelation: they share in the same state only $\mathcal{C}_1 = 25$ out of their 2500 active units, and another $\mathcal{C}_2 = 25$ in different states). At time 0 all network units have the value $\frac{1}{S+1}$ for all the states, with threshold $\theta_i = \sigma_i$, and an external cue arrives at time step $t = 500$, providing a signal term pointing at pattern $\mu = 1$ in the field to each unit. Other parameters were $p = 2$ (hence a single other pattern is present, to simplify the analysis of the noise), $S = 3$, $a = 0.25$, $U = 0.1$, $b_1 = 0.05$, $b_2 = 0.001$ and $b_3 = 0.0005$.

The figure shows the overlaps with the two successive patterns crossing over at a vanishingly small value of λ (slightly negative, in fact), characteristic of a random transition. Remarkably, the fields in the direction of the second pattern build up slowly, in that the decaying first pattern does not provide any useful cue in the direction of the second. Once the fields reach a given effective value, however, a self-regenerating process is initiated and the second overlap rises very rapidly towards 1 (without fully reaching it).

5.4. Transitions between correlated attractors

In another simulation (figure 5(a)), we identified a high-crossover latching transition between two substantially correlated patterns, which were constructed to share $\mathcal{C}_1 = 475$ out of their 2500 active units. Again, network units have at time 0 the value $\frac{1}{S+1}$ for all the states, with threshold $\theta_i = \sigma_i$, and an external cue arrives at time step $t = 500$, providing a signal term pointing at pattern $\mu = 1$ in the field to each unit. Other parameters were $p = 2$ (again, a single other pattern is present), $S = 3$, $a = 0.25$, $U = 0.1$, $b_1 = 0.005$, $b_2 = 0.001$ and $b_3 = 0.0001$.

One observes in figure 6 the latching transition occurring at a fairly high cross-over value $\lambda \simeq 0.68$. In fact, the overlap with the second pattern starts effectively at a nonzero values already imposed by the cue to the first pattern, with which it is significantly correlated. Its overlap then builds up gradually, eventually reaching the self-regeneration threshold. Interestingly, its accrual appears to sustain the overlap with the first patterns, rather than speeding up its demise. The overlap with the second patterns rises relatively slowly even after the self-regeneration process has started, and does not reach particularly high values either. Clearly, much more extensive work is required, in order to confirm the generality of these observations.

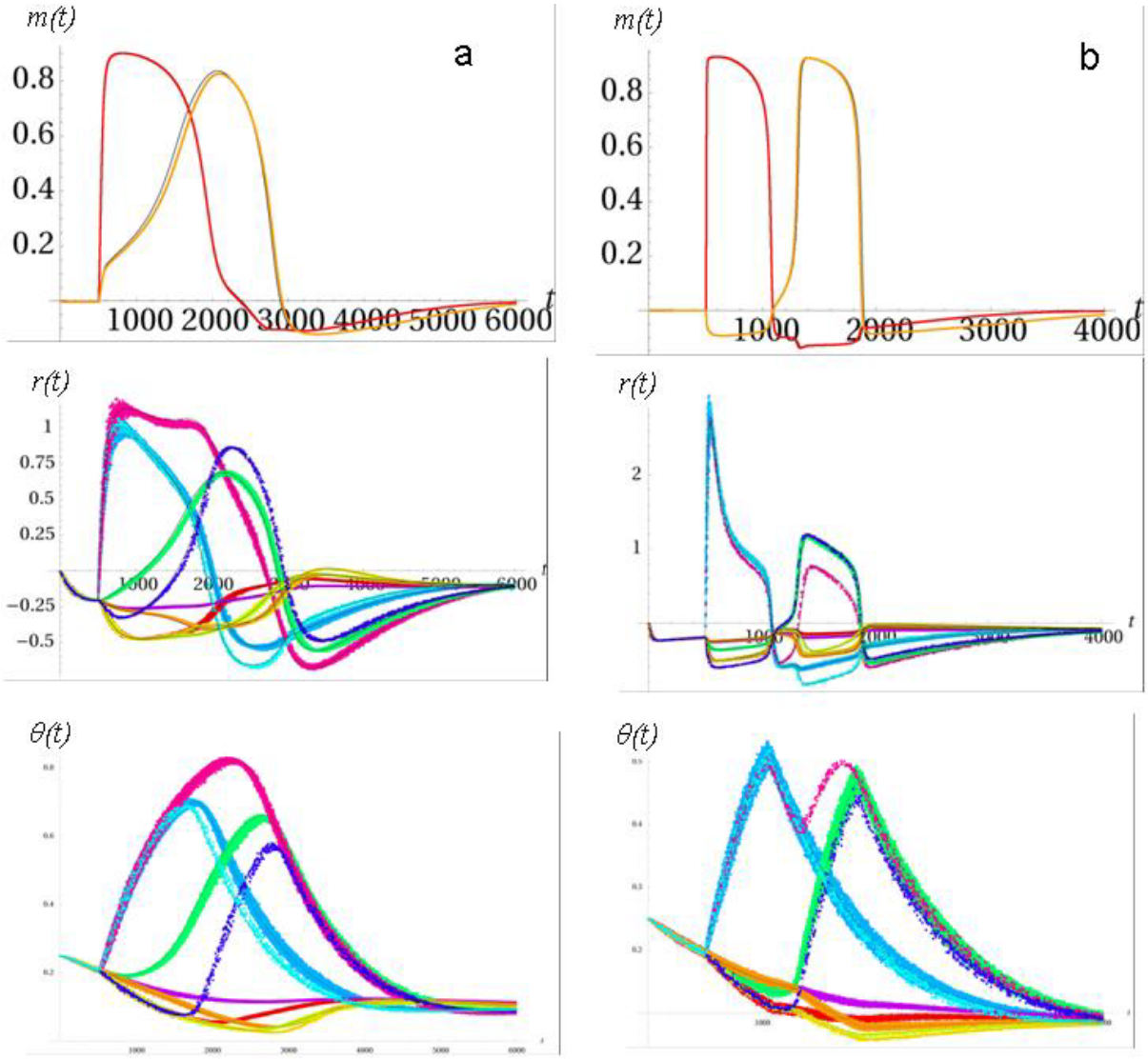

Figure 5. Two examples of simulations that represent the interesting regions shown in figure 2: a latching transition between largely correlated patterns (left column) and a random latching transition between uncorrelated patterns (right column). The three panels show, as in figure 4, the overlaps with the two patterns (top) and the fields (but only on the active states; middle) and thresholds (bottom) of various groups of units, as they are updated.

Finally, to meet the results of section 4, we perform a search of latching dynamics for the solution of the differential equations in the region of parameters spanned by all possible values of $\mathcal{C}_1$ and $\mathcal{C}_2$. To achieve this complete search we chose two combinations of these parameters that span between 0 and 1. Figure 6 shows the regions of latching for $0 \leqslant (\mathcal{C}_1 + \mathcal{C}_2)/(aN) \leqslant 1$ on the y-axis and $0 \leqslant \mathcal{C}_1/(\mathcal{C}_1 + \mathcal{C}_2) \leqslant 1$ on the x-axis.

The region $L1$ shows the latching guided by $\mathcal{C}_1$, or, in other words, shared active units in the same state. It appears when the total amount of shared units $\mathcal{C}_1 + \mathcal{C}_2$ is rather small in comparison with aN and at the same time $\mathcal{C}_1$ is larger than $\mathcal{C}_2$. The pathological $L2$ region is

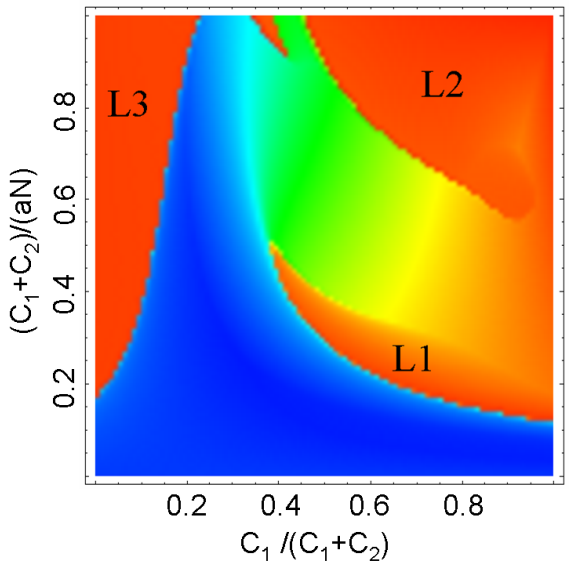

Figure 6. Three latching regions are found, parallel to the ones described in section 4. The axes show the two described combinations of $\mathcal{C}_1$ and $\mathcal{C}_1$, chosen in such a way to have a range spanning between 0 and 1. For each combination of parameters, a numerical integration of the dynamical equations, similar to those shown in figure 5, was performed. The colour at each point indicates the maximum value of m_2 during this integration. Region $L1$ corresponds to the kind of transition shown in the violet points in figure 2, while region $L2$ corresponds to the green point and region $L3$ to the orange points. In each of the marked regions, $m_1 < m_2$ at the time in which m_2 reaches its maximum, but not in the area between $L1$ and $L2$, that could otherwise be regarded as a latching region itself.

associated to $\mathcal{C}_1 \sim (aN)$, i.e. the number of shared units in the same active state is close to its saturation value. Finally, the region $L3$ corresponds to the condition $\mathcal{C}_1 \ll \mathcal{C}_2$ (uncorrelated or anti-correlated patterns). This picture fits exactly what has been described in figure 2 through a completely different approach, suggesting a strategy to follow in future developments. While the dynamical equations can give a mechanistic description of latching in 'noiseless' situations with only two patterns, the observations can therefore be extrapolated to more general simulations for which only a statistical approach is possible, given the high dimensionality.

6. Discussion

The notion of dynamical attractors has recently emerged, in cognitive neuroscience, as having the potential to bridge the gap between the analog processing performed by individual neural elements and the digital operations subsumed in cognitive descriptions. In this context, dynamics which take place largely in the neighbourhood of 'quasi-attractor' states (of states that would be stable attractors were it not for a simple mechanism that destabilizes them, such as firing rate fatigue) offer a model for free associations in semantic space [15], perhaps including the highly constrained trajectories expressed in natural language. This emerging view calls for

a quantitative, first principles modelling of higher order attractor-based processes, that has so far been only partially explored. Here, following up on our previous reports [7, 11, 12], we have begun to analyse the transitions between attractor states demonstrated by a simple Potts associative memory model, in the region of parameter space where it shows latching dynamics.

The model itself is based on the idea that associative memory retrieval operates throughout the cortex at two levels [16], and as a generic functional mechanism rather than as a separate dedicated system [17]. In this spirit, we have earlier suggested a rough description of how attractor dynamics in the network model gives rise to a complex and structured set of transitions, that could be regarded as a model of infinite recursion. This complexity, grounded in the correlation between patterns, was shown to be controlled mainly by the threshold, that also sets the global activity in the network. An appropriate value of the threshold ensures the transient coexistence of decaying and newly emerging attractors at critical points in the retrieval process, when latching between attractors takes place.

Here, we show that even for a given value of the threshold, one observes a considerable diversity of latching transitions. Apart from the extreme case of oscillations between nearly overlapping attractors, latching transitions can be roughly categorized between random ones, and those driven by positive correlations. It appears that the latter are responsible for embedding structure of a potentially usable form into the dynamics. There might, however, be finer structure also in the random transitions, as suggested by the prevalence, among those, of transitions between anticorrelated patterns ($\mathcal{C}_1 = 0$ or 1). Understanding such finer structure is essential, if one aims at embedding real syntactic or semantic constraints in latching dynamics. Features that require continuity between successive elements, like number or gender in the syntax of predicates, or topic in semantic concatenation, have to be engineered to sustain their activation across latching events, while features like subject markers, if any, have to be engineered to be terminated at latching, and maybe to activate complementary markers as distinct local states of the same units. These aspects obviously require massive additional study, preliminary to which is a much more complete analysis of latching dynamics, that we could only begin to sketch here.

References

[1] Barsalou L W 2005 Continuity of the conceptual system across species *Trends Cogn. Sci.* **9** 309–11
[2] Barton R 2007 Evolutionary specialization in mammalian cortical structure *J. Evol. Biol.* **20** 1504–11
[3] Gil-da-Costa R, Braun A, Lopes M, Hauser M D, Carson R E, Herscovitch P and Martin A 2004 Toward an evolutionary perspective on conceptual representation: species-specific calls activate visual and affective processing systems in the macaque *Proc. Natl Acad. Sci. USA* **101** 17516–20
[4] Pillay P and Manger P R 2007 Order-specific quantitative patterns of cortical gyrification *Eur. J. Neurosci.* **25** 2705–12
[5] Garagnani M, Wennekers T and Pulvermüller F 2007 A neuronal model of the language cortex *Neurocomputing* **70** 1914–9
[6] Hauser M, Chomsky N and Fitch W 2002 The faculty of language: what is it, who has it, and how did it evolve? *Science* **298** 1569–79
[7] Treves A 2005 Frontal latching networks: a possible neural basis for infinite recursion *Cogn. Neuropsychol.* **6** 276–91
[8] Amati D and Shallice T 2007 On the emergence of modern humans *Cognition* **103** 358–85
[9] Bolle D, Cools R, Dupont P and Huyghebaert J 1993 Mean-field theory for the Q-state Potts-glass neural network with biased patterns *J. Phys. A: Math. Gen.* **26** 549–62
[10] Kanter I 1988 Potts-glass models of neural networks *Phys. Rev.* A **37** 2739–42

[11] Kropff E and Treves A 2005 The storage capacity of Potts models for semantic memory retrieval *J. Stat. Mech.* **2005** P08010
[12] Kropff E and Treves A 2007 The complexity of latching transitions in large scale cortical networks *Nat. Comput.* **6** 169–85
[13] Kropff E and Treves A 2007 Uninformative memories will prevail: effective storage of correlated representations and its consequences *HFSP J.* **1** 249–62
[14] Tsodyks M V and Feigel Man M V 1988 The enhanced storage capacity in neural networks with low activity level *Europhys. Lett.* **6** 101–5
[15] Herrmann M, Ruppin E and Usher M 1993 A neural model of the dynamic activation of memory *Biol. Cybern.* **68** 453–63
[16] Braitenberg V and Schuz A 1991 *Anatomy of the Cortex: Statistics and Geometry* (Berlin: Springer)
[17] Fuster J M 1999 *Memory in the Cerebral Cortex: An Empirical Approach to Neural Networks in the Human and Nonhuman Primate* (Cambridge: MIT Press)