Dynamical properties of strongly interacting Markov chains

by

Nihat Ay, and Thomas Wennekers

Preprint no.: 107 2001

E-mail: (nay,wenneker)@mis.mpg.de

Tel: +49-341-9959-558
Fax: +49-341-9959-555

January 7, 2002

many conceptual approaches to the understanding of first principles for neural organization and learning, where information theory provides an appropriate framework for the formulation and analysis of such principles [15, 16, 6, 20]. A well known measure that quantifies relations of interacting units is a generalized version of the so-called *mutual information* of two units: Consider N binary units $1, 2, \ldots, N$ and a joint probability distribution p on the configuration set $\{0,1\}^N$. Then the Kullback-Leibler divergence [13, 8] of p from the set of factorizable distributions is a natural measure for the "spatial" interdependence of the units:

$$I(p) \;:=\; \inf_{\substack{p_k,\ 1\le k\le N,\\ \text{distributions on } \{0,1\}}} D(p \,\|\, p_1 \otimes \cdots \otimes p_N)\ . \tag{1}$$

The Kullback-Leibler divergence represents the basis of many approaches to neural complexity [19, 9] and has been theoretically studied in [4, 5] from the *information geometric* point of view, where it is referred to as (*stochastic*) *interaction*. In order to capture the intrinsically temporal aspects of interaction, I has been extended in [7] to the dynamical setting of Markov chains. In this case, the factorized or split Markov chains are the ones that can be composed by a family of individual chains. So, there is no interaction of the units with respect to a split Markov chain. In analogy to the approach given by (1), we can consider the divergence of a Markov chain X from being split:

$$I(X) \;:=\; \inf_{Y \text{ split Markov chain}} D(X \,\|\, Y).$$

(This formal definition is specified in Section 3).

ciple for learning in neural networks. This principle has been proposed in the context of spatial interaction in layered networks in [6], where the connection to the *infomax principle* by Linsker [15, 16] is discussed.

In the present paper we consider processes that optimize the dynamical version of interaction. The analytical framework is formulated for general Markov chains without direct reference to particular neural network models. This allows to investigate essential dynamical properties of strongly interacting units unconstrained by architectural or other specific assumptions concerning neural dynamics. This provides a necessary first stage for an investigation of learning processes also in more detailed recurrent neural network models. Thus, the approach presented here optimizes interaction in the full space of Markov chains, which leads to analytical results concerning the most fundamental feature of strongly interacting stochastic systems, i.e., the development of the system dynamics towards determinism. In contrast, a specific neural model would restrict optimization to constrained (and hardly controllable) manifolds. We expect that the essential properties described in the sequel for general Markov chains also carry over to more realistic models, but have to leave the study of such models to future work.

The paper is organized as follows. In Section 2 we briefly introduce basic notations and preliminaries for probability spaces and Markovian transition kernels. In generalization of the usual mutual information for stationary probability distributions, Section 3 provides the main definition of *stochastic interaction* as the divergence of a Markov kernel from the product of its split marginal kernels. The main analytical result for strongly interacting units is also given in Section 3. It states that kernels with optimal stochastic interaction must reveal a drastically reduced entropy and, hence, degree of

2 Preliminaries

2.1 Discrete Probability Spaces

In the following, Ω denotes a non-empty and finite set of *states*. The vector space $\mathbb{R}^\Omega$ of all functions $\Omega \to \mathbb{R}$ carries the natural topology, and we consider subsets as topological subspaces. The closed set of all probability distributions on Ω is given by

$$\bar{\mathcal{P}}(\Omega) \;:=\; \left\{ p = (p(\omega))_{\omega\in\Omega} \in \mathbb{R}^\Omega \;:\; p(\omega) \geq 0 \;\text{ for all }\; \omega \in \Omega, \quad \sum_{\omega\in\Omega} p(\omega) = 1 \right\}.$$

For $p \in \bar{\mathcal{P}}(\Omega)$, $\operatorname{supp} p := \{\omega \in \Omega \;:\; p(\omega) > 0\}$ denotes the *support* of p. The interior $\mathcal{P}(\Omega)$ of $\bar{\mathcal{P}}(\Omega)$ consists of all elements with total support Ω.

The *Shannon entropy* of a distribution $p \in \bar{\mathcal{P}}(\Omega)$ is defined as

$$H(p) \;:=\; -\sum_{\omega\in\operatorname{supp} p} p(\omega)\, \ln p(\omega).$$

It measures the uncertainty about the outcomes of an experiment goverened by p. A related quantity which can be interpreted as a "distance" is the *Kullback-Leibler divergence* or *relative entropy* of distributions $p, q \in \bar{\mathcal{P}}(\Omega)$:

$$D(p \,\|\, q) := \begin{cases} \sum_{\omega\in\operatorname{supp} p} p(\omega)\, \ln \frac{p(\omega)}{q(\omega)}, & \text{if } \operatorname{supp} p \subset \operatorname{supp} q \\ \infty \qquad\qquad , & \text{otherwise} \end{cases}.$$

is called *Markovian transition kernel* if $K(\cdot\,|\,\omega) \in \bar{\mathcal{P}}(\Omega_B)$ for all $\omega \in \Omega_A$, that is

$$\sum_{\omega' \in \Omega_B} K(\omega'\,|\,\omega) \;=\; 1, \qquad \text{for all} \quad \omega \in \Omega_A \;.$$

The set of all such kernels is denoted by $\bar{\mathcal{K}}(\Omega_B\,|\,\Omega_A)$. We write $\mathcal{K}(\Omega_B\,|\,\Omega_A)$ for its interior and $\bar{\mathcal{K}}(\Omega_A)$ respectively $\mathcal{K}(\Omega_A)$ as abbreviation in the case $A = B$. If $A = \emptyset$, then Ω_A consists of exactly one element, namely the empty configuration ϵ. In that case, $\bar{\mathcal{K}}(\Omega_B\,|\,\Omega_\emptyset) = \bar{\mathcal{K}}(\Omega_B\,|\,\epsilon)$ can naturally be identified with $\bar{\mathcal{P}}(\Omega_B)$ by $p(\omega) := K(\omega\,|\,\epsilon)$, $\omega \in \Omega_B$.

Given a probability distribution $p \in \bar{\mathcal{P}}(\Omega_A)$ and a transition kernel $K \in \bar{\mathcal{K}}(\Omega_B\,|\,\Omega_A)$, the *conditional entropy* for the pair (p, K) is defined as

$$H(p, K) \;:=\; \sum_{\omega \in \Omega_A} p(\omega)\, H(K(\cdot\,|\,\omega)) \;. \tag{2}$$

For two random variables X, Y with $\mathrm{Prob}\{X = \omega\} = p(\omega)$ for all $\omega \in \Omega_A$, and $\mathrm{Prob}\{Y = \omega' \,|\, X = \omega\} = K(\omega'\,|\,\omega)$ for all $\omega \in \Omega_A$ with $p(\omega) > 0$ and all $\omega' \in \Omega_B$, we set $H(Y\,|\,X) := H(p, K)$. This measures the uncertainty about Y given X.
The "distance" of two kernels $K, L \in \bar{\mathcal{K}}(\Omega_V)$ with respect to a distribution $p \in \bar{\mathcal{P}}(\Omega_V)$ can be measured by

$$D_p(K \,\|\, L) \;:=\; \sum_{\omega \in \Omega_V} p(\omega)\, D\Big(K(\cdot\,\|\,\omega) \,\|\, L(\cdot\,\|\,\omega)\Big).$$

This extends the Kullback-Leibler divergence to the setting of transition kernels.

$$I(p) \;:=\; \inf_{p' \text{ factorizable}} D(p \,\|\, p') \;=\; \sum_{\nu \in V} H(p_\nu) \;-\; H(p). \tag{3}$$

Here, the p_ν, $\nu \in V$, denote the marginals of p. Ay [7] extended the definition (3) to a dynamical version, where temporal interdependencies among the units are also considered: A kernel $K \in \Omega_V$ is called *split* if there exist kernels $K^{(\nu)} \in \bar{\mathcal{K}}(\Omega_\nu)$, $\nu \in V$, such that

$$K(\omega' \,|\, \omega) \;=\; \prod_{\nu \in V} K^{(\nu)}({\omega'}_\nu \,|\, \omega_\nu), \qquad \text{for all} \quad \omega, \omega' \in \Omega_V.$$

The split transition kernels represent the dynamical version of the factorizable distributions. Thus, in analogy to (3) we define the stochastic interaction of the units with respect to $p \in \bar{\mathcal{P}}(\Omega_V)$ and $K \in \bar{\mathcal{K}}(\Omega_V)$ to be the p-divergence of K from being split [7]:

$$I(p, K) \;:=\; \inf_{K' \text{ split}} D_p(K \,\|\, K').$$

The measure $I(p, K)$ extends the notion of spatial interdependence to the dynamical setting. We also have the following representation ([7], Proposition 3.2):

PROPOSITION 3.1. *Consider a probability distribution $p \in \mathcal{P}(\Omega_V)$, a transition kernel $K \in \mathcal{K}(\Omega_V)$, and the corresponding* marginal kernels *$K_\nu \in \mathcal{K}(\Omega_\nu)$, $\nu \in V$, of K defined by*

$$K_\nu(\omega' \,|\, \omega) \;:=\; \frac{\sum_{\substack{\sigma, \sigma' \in \Omega_V \\ \sigma_\nu = \omega,\, \sigma'_\nu = \omega'}} p(\sigma)\, K(\sigma' \,|\, \sigma)}{\sum_{\substack{\sigma \in \Omega_V \\ \sigma_\nu = \omega}} p(\sigma)}, \qquad \omega, \omega' \in \Omega_\nu. \tag{4}$$

main question we focus on in the present paper is the degree of randomness (uncertainty of the future, given the present) in systems with strongly interacting units. The uncertainty of a kernel K with respect to a distribution p is measured by the entropy of (p, K). It vanishes exactly when the next state ω' is determined by the current state ω for all $\omega \in \operatorname{supp} p$, that is $|\operatorname{supp} K(\cdot \,|\, \omega)| = 1$ for all $\omega \in \operatorname{supp} p$. According to (6), $I(p, K)$ is large if the marginal processes (p_ν, K_ν) have high entropy, but that of the full process (p, K) is low. Thus, heuristically, systems with high interaction I prefer determinism. An implication of the following main theorem of the present work states that the entropy of a system with N strongly interacting binary units is reduced to at most $\ln(N + 1)$. This represents a strong reduction from the maximal value $N \ln 2$ for complete randomness.

THEOREM 3.2. *Consider a probability distribution $p \in \bar{\mathcal{P}}(\Omega_V)$ and a transition kernel $K \in \mathcal{K}(\Omega_V)$. If (p, K) is a local maximizer of I then for all $\omega \in \operatorname{supp} p$ one has*

$$|\operatorname{supp} K(\cdot \,|\, \omega)| \;\leq\; 1 - |V| + \sum_{\nu \in V} |\Omega_\nu|. \tag{7}$$

For the binary case, (7) implies

$$|\operatorname{supp} K(\cdot \,|\, \omega)| \;\leq\; 1 + |V|. \tag{8}$$

The somewhat technical proof of Theorem 3.2 is given in Appendix A. Note, that in (7) and (8) the expression $|\operatorname{supp} K(\cdot \,|\, \omega)|$ counts the number of

where the upper bound reduces to $\ln(1+|V|)$ *if* $|\Omega_\nu| = 2$ *for all* ν.

The upper bound for the entropy given in Corollary 3.3 proofs that randomness is strongly reduced in systems with high interaction.

4 Simulations

4.1 Preliminaries on Markov Chains

Consider a Markov chain $X_t = (X_{\nu,t})_{\nu\in V}$, $t = 0, 1, 2, \ldots$, that is given by an initial distribution $p \in \bar{\mathcal{P}}(\Omega_V)$ and a kernel $K \in \bar{\mathcal{K}}(\Omega_V)$. The probabilistic properties of this stochastic process are determined by the following set of finite marginals:

$$\begin{aligned} & \mathrm{Prob}\{X_0 = \omega_0, X_1 = \omega_1, \ldots, X_t = \omega_t\} \\ = \;& p(\omega_0)\, K(\omega_1 \,|\, \omega_0) \cdots K(\omega_t \,|\, \omega_{t-1}), \qquad t = 0, 1, 2, \ldots \end{aligned}$$

Thus, the set of Markov chains on Ω_V can be identified with

$$\bar{\mathcal{P}}(\Omega_V) \times \bar{\mathcal{K}}(\Omega_V).$$

A probability distribution $p \in \bar{\mathcal{P}}(\Omega_V)$ is called *stationary* with respect to $K \in \bar{\mathcal{K}}(\Omega_V)$ if

$$\sum_{\omega\in\Omega_V} p(\omega)\, K(\omega' \,|\, \omega) \;=\; p(\omega), \qquad \text{for all} \quad \omega' \in \Omega_V.$$

rem 3.2 to the constrained setting. Nevertheless, the simulation results are compatible with this theorem and can be discussed with regard to it. Our main intention is to demonstrate the emergence of determinism in systems with strongly interacting units.

Section 4.2 describes the simulation scheme, that is, the Markov dynamics chosen for the activity update and the optimization method to obtain systems with large interaction by random search. Section 4.3 and 4.4 present simple examples with two and three units, which already reveal all qualitative properties visible in highly interacting systems. Section 4.5 and 4.6 afterwards consider the convergence of the optimization procedure from mere randomness towards global determinism, and Section 4.7 describes the accompanied activation patterns observable during different phases of convergence. Finally, Section 4.8 displays asymptotic entropy distributions obtained from a large number of simulations.

4.2 Description of the Simulations

We employ a parallel update dynamics for the short-time activity of N binary units, i.e. $\Omega_\nu = \{0,1\},\ \nu = 1,\cdots,N$. That is, we start from a set of individual kernels $K^{(\nu)} \in \mathcal{K}\left(\{0,1\} \,|\, \{0,1\}^N\right),\ \nu = 1,\cdots,N$, cf. (9). Transitions $K^{(\nu)}(1\,|\,\omega)$, $\omega \in \{0,1\}^N$, are initialized with independent equally distributed random values in $]0,1[$; transitions $K^{(\nu)}(0\,|\,\omega)$ are then fixed by the normalization condition $K^{(\nu)}(1\,|\,\omega) + K^{(\nu)}(0\,|\,\omega) = 1$ for probabilities and are not stored explicitly. The full Markov kernel K is computed from the $K^{(\nu)}$ by means of (9). Initial states for the single units are also chosen independently

ual kernels $K^{(\nu)}(1\,|\,\omega)$ and perturb it by a small random number, ξ, equally distributed on $[-\epsilon, \epsilon]$, where ϵ is the learning rate ($\epsilon = .025$, if not stated otherwise). Perturbed values are clipped to the range $[0, 1]$, and the $K^{(\nu)}(0\,|\,\omega)$ are again fixed by the normalization condition. That is, for randomly chosen $\nu \in V$ and $\omega \in \{0, 1\}^N$ we set

$$K^{(\nu)}_{t+1}(1\,|\,\omega) = \phi\left(K^{(\nu)}_t(1\,|\,\omega) + \xi\right) \tag{10}$$

$$K^{(\nu)}_{t+1}(0\,|\,\omega) = 1 - K^{(\nu)}_{t+1}(1\,|\,\omega), \tag{11}$$

where the function $\phi(x)$ is zero for $x \leq 0$, one for $x \geq 1$, and $\phi(x) = x$ else. The new full Markov kernel and interaction measure are computed, and if the interaction increases the perturbation is accepted. Otherwise, the optimization is repeated for a new random entry of the parallel kernel. The simulation proceeds to the next time-step if either I can be increased or 5 unsuccessful optimization trials have been performed. We stop simulations after typically several thousand steps. At that time convergence is usually acceptable (cf. Fig. 4 and below).

The above algorithm is numerically quite expensive, since it has exponential complexity in the number of units N. Therefore, we are restricted to small systems, $N < 10$. In the sequel, we represent the resulting strongly interacting systems in different ways: First, we show samples of the time-course of activity as raster plots. Determinism in these plots is expressed as repetetive activation patterns. Second, we display the full Markov kernels after convergence. Deterministic transitions from a state ω to ω' in these kernels have values of 1, and only the element $K(\omega'\,|\,\omega)$ in the ωth column of the full kernel can be different from zero. Third, from $K(\omega'\,|\,\omega)$ we derive

amples the optimization almost reached a deterministic process as can be inferred from the Markov matrices displayed in the upper left portions of the plots. These contain only a single strong transition in every column. In fact, most transitions are zero, and only two transitions in the left and four in the right example reveal a tiny residual probability. Therefore, the entropy of the full kernels is small and would even converge to zero for longer simulation times. The figure also plots the state transition graphs comprising all non-vanishing edges. A generic Markov process in principle allows for transitions between arbitrary states. In Fig. 1, however, only very few edges arise, expressing the almost deterministic nature of the dynamics. Skipping the residual edges with probabilities near zero, the transition graphs show that the dynamics is basically confined to sets of cycles or permutations: a 3-cycle and a 1-cycle (fixed point) in the left, and a 2-cycle and 2 fixed points in the right example. The 3-cycle can also be observed in the sample trajectory on the left side, whereas the example trajectory on the right first is kept in the 00-fixed point and switches to the 2-cycle at step 9 driven by the small residual transition probability from state 00 to state 01.

Finally note that although the dynamics is almost deterministic and the full kernel entropy is low, the marginal kernel entropies are almost $\ln(2) = .6931471$. This implies that entropies computed for single unit activities are maximal. Predictions of the next state of a single unit given nothing but its present state are, hence, impossible. In strong contrast, they are almost perfectly possible, if the whole system state is known, because the full kernel entropy is near zero. This is well in accord with the philosophy behind the interaction measure: As (6) shows, the interaction consists of the summed marginal kernel entropies, which measure the disorder in the individual unit

Figure 1: Two examples of strongly interacting systems comprising $N = 2$ units. Shown in each plot are the Markov kernel $K(\omega' \,|\, \omega)$, $\omega, \omega' \in \{0,1\} \times \{0,1\}$, in the upper left (the area of each circle represents the strength of a particular transition; sums over columns, i.e. ω', must be 1), and its representation as a state transition graph. At the bottom sample trajectories are displayed as raster plots (here, time runs from left to right and unit numbers from top to bottom; each circle represents an activity value of $\omega_\nu = 1$ in the respective time-step). Values for the interaction $I := I(p, K)$, entropy of the full Markov kernel $HK_F := H(p, K)$, and the marginal kernel entropies $HK_\nu := H(p_\nu, K_\nu)$ are also given. The dynamics consist of almost deterministic cycles in both examples. Accordingly, the kernel entropies are low, but note that the marginal kernel entropies almost approach their maximum value $\ln(2)$.

activities, subtracted by the full kernel entropy, which accounts for the disorder in the full system. Maximizing the interaction should, thus, increase the randomness in individual units, but decrease that of the whole state trajectories.

4.4 Branching Nodes and Nested Cycles

Figure 2 shows a somewhat more complex example for $N = 3$ units. The dynamics does not converge to a simple permutation, but reveals additional structure. First, observe that the Markov matrix has columns with more than a single entry. The dynamics is, thus, not fully deterministic. The

node "transient": Once left, activity never flows back to it. Therefore, the probability $p(001)$ is zero. In contrast the other nodes determine the "ergodic components" of the dynamics: They support a positive stationary probability since activity repeatedly cycles back to them. This subdivision into non-empty transient and ergodic components is typical in our simulations. Also nested loops and branching states with more than one but only a few outgoing edges with nonvanishing transition probabilities are typical. Therefore, most strongly interacting systems are only *almost* deterministic: At branching nodes they have the freedom to chose between several possible target states, which, however, are only a small subset of all possible states. The kernel entropy of a system with branching nodes, of course, is always positive – only completely deterministic systems, which result in rare cases in our simulations and cannot reveal any branching nodes, have vanishing entropy.

The branching states have an interesting impact on sample trajectories. Those are periodic for strict n-cycles (cf. Fig. 3), but they consist of repetetive patterns of various length for systems with branching nodes. An example is shown at the bottom of Fig. 2, where the activity switches several times at state 001 between the two possible pattern sequences generated by the two nested loops. Of course, in larger systems one expects even more complicated situations than just two nested cycles (cf., e.g., Fig. 7).

Theorem 3.2 provides an upper bound for the number of outgoing edges for optimal Markov transitions. For N binary units this bound is $N + 1$ for all ω in the ergodic components. No bound is imposed on transient states. This is consistent with the present (as well as all subsequent) simulations, which, however, are restricted to Markov chains in SMC_{par}. In fact, these simulations usually reveal an even smaller number of outgoing transitions

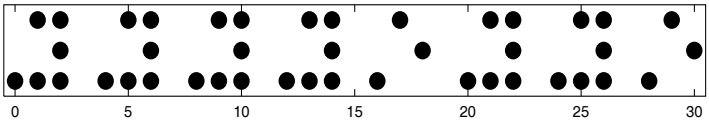

Figure 2: Same as Fig. 1 but for $N = 3$ units. Here, the dynamics consists of a 1-cycle (fixed point), i.e. $\omega = 011$, two nested loops with branching node $\omega = 001$, and a transient node $\omega = 110$ which has transitions to all other nodes. Since the branching node has positive transition probability to nodes in both of the nested loops, the sample trajectory reveals randomly interleaved but repetetive sub-patterns corresponding to eiter of the two loops.

than $N + 1$, with the majority of nodes having just one, and a subset of nodes – the branching nodes – with two outgoing connections. Although they should in principle be possible, we never observed branching nodes with more than 2 outgoing edges in a large number of simulations. In contrast, transient nodes can develop deterministic transitions (see, e.g., node 11000 in Fig. 3) as well as connect to a varying number ot target states up to the full range of all 2^N possible states (e.g., node 110 in Fig. 2).

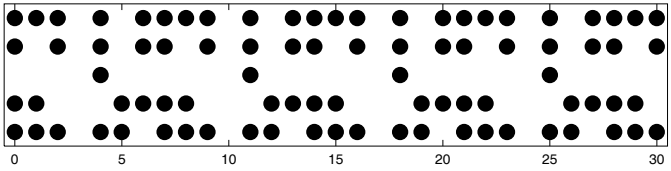

Figure 3: Same as Fig. 1 but for $N = 5$ units. The dynamics separates into several disjoint loops and transients. Only the deterministic transients are shown in the transition graph. Because the loops are disjoint and have no branching nodes the kernel entropy is very low and sample trajectories after transients have died out are strictly periodic.

4.5 The Evolution of Global Determinism and Individual Non-Determinism

The time-course of convergence is depicted in Fig. 4 for two example simulations, one with $N = 4$ and a learning rate of $\epsilon = .05$ on the left hand side, and a second simulation on the right with $N = 7$ and a ten-fold learning rate $\epsilon = .5$. Displayed are the interaction measure (thick solid line), the entropy of the full Markov kernel (dashed line), and the mean number of outgoing edges (thin solid line) averaged over all nodes and divided by 10 in the left figure and by 32 in the right to fit suitably into the plot (counted as outgoing are only transitions with probability larger than .000001). There is no qualita-

the kernel entropy (dashed line), and the average number of outgoing edges (thin solid line; scale has to be taken times 10 on the left and times 32 on the right). Because the kernel entropies approach small values, the corresponding processes become almost deterministic.

tive difference between the curves in both examples, but clearly convergence is slower for larger systems. Therefore, the learning rate was 10 times larger in the $N = 7$ example as compared to $N = 4$. Moreover, the asymptotic interaction measure for $N = 4$ is bounded by $4\ln(2) = 2.7726$, whereas the bound is $7\ln(2) = 4.8520$ for $N = 7$. Obviously, the optimization process almost reaches these upper bounds. Conversely, the kernel entropies converge to small values near zero. Corollary 3.3 in Section 3 provides an upper bound of $\ln(N + 1)$ for the entropy of strongly interacting units. For $N = 4$ this bound is 1.609 and for $N = 7$ it is 2.079, both consistent with Fig. 4. The bound becomes more stringent in larger systems, because $\ln(N+1)/(N\ln 2)$ goes to zero with increasing system size. The small kernel entropies imply that the network dynamics as a whole becomes almost deterministic, but that the marginal kernel entropies must stay large (cf. (6)). So, again, the future can be well predicted from whole states, but single units do not contain much information about it.

4.6 Two Phases in the Evolution of Complex Systems

The optimization curves in Fig. 4 roughly separate into two phases: Initially the interaction I increases approximately linearly but it saturates for large

is initially increased by confining the flow of actvity to restricted pathways in state space. Only some of the remaining transitions, however, are already truly deterministic, that is, have a probability of 1 – most have intermediate values. Therefore, at any time step of the activation dynamics several target states are possible, such that it is not yet possible to predict the future well, even if the current state is known perfectly. In the second part of the optimization the (asymptotically reached) transient and ergodic components of the dynamics separate. On the ergodic component the convergence proceeds as before: The number of outgoing connections is further constrained until many nodes are left with only a single deterministic outgoing transition. Branching nodes are similarly confined to a minimal number of non-vanishing outgoing connections. On the other hand, the probability $p(\omega)$ for (asymptotically) transient nodes decreases to zero. Therefore, these nodes have a decreasingly small influence on the interaction measure, which contains the $p(\omega)$ as weighting coefficients, cf. (6) and (2). Nonetheless, although it is small, changing outgoing connections of transient nodes still has an impact on the interaction measure. During the second phase of the optimization, therefore, most optimization steps refer to transient nodes accompanied by only minor increases in I and a slow convergence. In addition, the increase in the average number of outgoing connections during the second optimization phase shows that many of the transient nodes must again redevelop quite a large number of non-vanishing connections. Hence, as a rule, the number of connections of nodes in the ergodic component is reduced to a minimum, whereas in the transient component this number can become large for at least some nodes (cf., e.g., node 110 in Fig. 2, but also the almost deterministic transient transitions in Fig. 3).

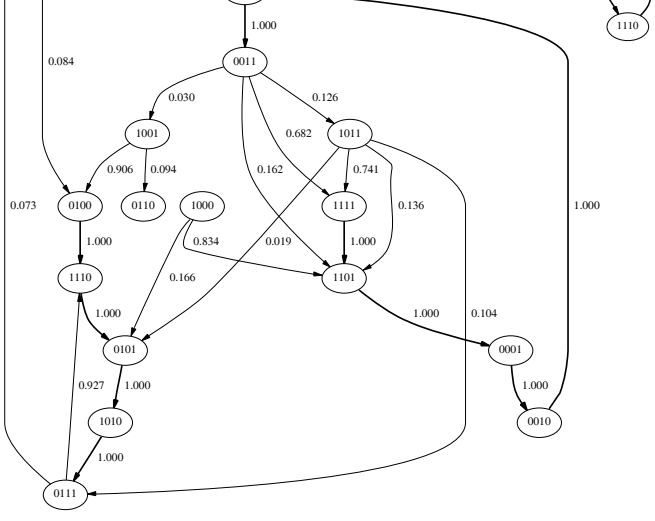

Figure 5: Example for the development of a transition graph for $N = 4$ units. Displayed are graphs corresponding with times $t = 300,\ 1300$ and 2500 of the left simulation in Fig. 4. Only transitions originating from nodes with *at most* 4 outgoing non-zero connections and probabilities larger than .01 are drawn. The optimization first reduces the number of positive connections in the network (t=300). At intermediate times (t=1300) all nodes have less than 5 outgoing transitions and loops have built up; some transitions are already deterministic (bold transitions are $> .95$). After convergence the ergodic component consists mostly of deterministic transitions (t=2500), and most often (though not here) nested loops with branching states, cf. Fig. 2.

would appear more often than others. The corresponding graph, nonetheless, is still broad and flat. At intermediate times, $t = 1300$ *all* 16 states reveal less than 5 outgoing transitions, the average number is about 2, and quite a few transitions are already deterministic. The graph also contains relatively long sequences of almost deterministic activity patterns, some arranged in loops. So, the graph structure deepens, and predictions about subsequent states become increasingly reliable. However, there still are a lot random transitions between the asymptotically ergodic states. Those are removed in the final phase of the optimization process, such that we are finally left with deep and narrow graphs, which allow for predictions of future states but (beside nested loops and branching nodes) are basically free of random transitions between ergodic states. Note that the graph for $t = 2500$ in Fig. 5 still contains two transitions with low probability. These, presumably, die out for very long simulation times leaving just two disjoint cycles. Furthermore, remember that we have stripped off most of the transients from the graphs in Fig. 5, because these have more than 5 outgoing transitions. As Fig. 4 demonstrates the average number of outgoing transition of these nodes must be quite large. Displaying them would have made the figures unnecessary complicated.

4.7 Determinism and Randomness in Activation Patterns

Activation patterns of the four cells resulting from the transition graphs in Fig. 5 are displayed in Fig. 6. At $t = 300$ these are apparently almost random,

in Fig. 5. The activity patterns change over time from almost random to repetitive. Accordingly, determinism of the process and predictability of future states increase gradually.

although a closer inspection shows that some transitions already appear more often than chance level, for instance, the pattern $0011 \to 1111 \to 1101$ occurs twice, and is indeed also part of the eventually emerging repetetive pattern at large times. The pattern approached for $t > 2500$ is perfectly periodic in the displayed time-frame because the transition graph converges to two disjoint cycles, a 4-cycle and a 6-cycle (cf. Fig. 5, $t = 2500$). At intermediate times the 6-cycle is also already visible but occasionally perturbed by non-converged random transitions with low probability.

Figure 7 shows a similar plot than Fig. 6 but for seven units, i.e., $2^7 = 128$ states. This simulation converged to a complicated graph comprising transient states as well as several nested loops of different length. The final Markov kernel and transition graph are not shown since they are utterly complex. Because the converged graph is nested and comprises branching nodes, at first glance all activity traces in Fig. 7 look more or less random even at large times. Nonetheless, the asymptotic entropy of the full Markov kernel was only 0.168 as compared to an interaction of $4.66 \approx 7\ln(2)$ indicating a high degree of determinism. In fact, for instance the first 9 consecutive steps in the bottom frame from 5000 to 5008 appear identically starting at 5033; they are furthermore partly overlapping with a 9-cycle that perfectly repeats three times starting at 5038. A closer investigation of Fig. 7 shows this way, that as in the previous example for $N = 4$, determinism and predictability

Figure 7: Same as Fig. 6 but for $N = 7$ (corresponding with the right plot in Fig. 4). Whereas the asymptotic process approached in Fig. 6 contains only two disjoint cycles (cf. also Fig. 5 for $t = 2500$) the asymptotic transition graph (not shown) in the present example is considerably more complex and consists of several nested loops. Therefore, instead of becoming repetetive as in Fig. 6 (bottom frame) activation patterns now appear much more random even if the network is almost converged.

increase continuously during the time-course of the optimization process.

4.8 Entropy Distributions

Figure 8 displays $H(p)$ and $H(p, K)$ for a large number of simulations and $N = 4, 5, 6$ and 8 units. In all plots ϵ was .5 and individual systems were optimized for 25.000 steps to guarantee convergence. A strict upper bound for the kernel entropies is $N \ln(2)$ corresponding with pure random kernels. Apparently in all cases the kernel entropy $H(p, K)$ converges to much smaller values indicating a high degree of determinism. Some systems at least up to $N = 6$ are even deterministic, that is, have $H(p, K) = 0$. For larger systems simulations become too slow to sample the distributions in Fig. 8 appropriately. So, the case $H(p, K) = 0$ cannot be excluded in such systems.

The value $N \ln(2)$ is also a strict upper bound for $H(p)$. It is reached,

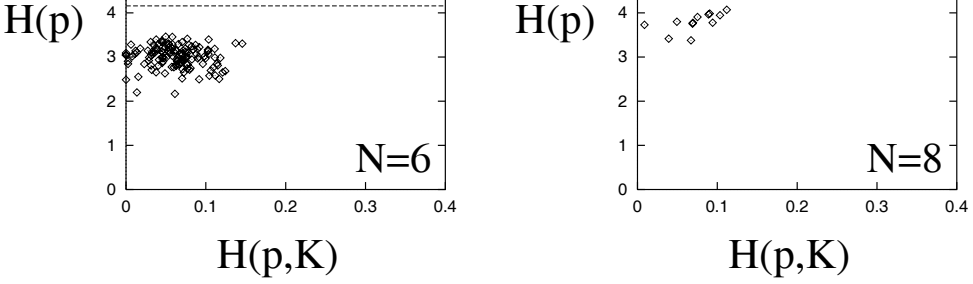

Figure 8: Distribution of $H(p)$ and $H(p,K)$ for $N = 4, 5, 6, 8$. Horizontal lines in each plot depict the theoretical upper bound $N \ln 2$ for $H(p)$ (and $H(p,K)$). Optimized kernel entropies are usually much lower, indicative for a high degree of determinism. $H(p)$ falls below the upper bound with increasing N, because a fraction of all states becomes transient.

if all 2^N states belong to the ergodic component and have equal probability $p(\omega) = 2^{-N}$. For small systems (especially for $N < 4$, not shown) this bound is occasionally obtained, but with increasing systems size transient and branching states become influential in $H(p)$: The first reduce $\operatorname{supp} p$, the second lead to deviations from equal-distributedness. Hence, $H(p)$ falls below the bound $N \ln(2)$ for increasing N. This indicates, that a number of states in a given optimized system should be transient. Nonetheless, $H(p)$ still remains high also for large N. Thus, a considerable fraction of states must belong to the ergodic components.

tivity distributions has been proposed as a fundamental organization principle in real neural systems [21, 15, 16]. As a first step in our dynamic framework, we therefore considered Markovian systems that optimize their stochastic interaction. The restriction to general Markov chains here allows to rigorously prove some strong results concerning complexity in strongly interacting systems. In especially, it was shown that given the current state the number of possible successor states must be linearly bounded in contrast to an exponential number of possible successor states in pure random systems. This way given the present state, future states of the system can be predicted with high reliability. On the other hand, the entropy of the marginal kernels is maximized by the optimization. So, the activity of single elements in a strongly interacting system does not carry much information about the future.

The presented simulations confirm this prediction and reveal that the optimized systems can be represented as almost deterministic transition graphs, consisting of an ergodic component and transient states which eventually direct activity into the ergodic component. The ergodic component carries the asymptotic network activity and in turn comprises nested but almost deterministic cycles of activity patterns, linked by branching nodes, which allow for a low number of transitions between several deterministic cycles.

Simulations in the present work were retricted to small systems. Nevertheless, Corollary 3.3 in Section 3 implies that also in large systems the Markovian matrices of strongly interacting systems and their induced state transition graphs must be sparse. Thus also in networks of reasonable size the phenomena described in the previous sections must appear.

Future work will be directed into two main directions: First, in the present

The same may be expected also in the dynamic framework developed in the present paper. Of course, this needs further evaluation.

Second, sofar we only considered stochastic interaction in somewhat formal Markovian systems. Although some artifical neural networks (Boltzman machines, Hopfield networks) can be reformulated in that context, much work remains to be done to transfer the concept to a mathematical framework for more realistical neural models. However, Linsker's results on the stationary case strongly encourage us to proceed into that direction.

consider an affine subspace V of $\operatorname{aff}\bar{\Delta}$ *that is given by r linear equations:*

$$V = \{x \in \operatorname{aff}\bar{\Delta} \ : \ x \text{ satisfies the } r \text{ given linear equations}\}.$$

If a point $x_0 \in \mathcal{C} := V \cap \bar{\Delta}$ locally maximizes a strictly convex function $f : \mathcal{C} \to \mathbb{R}$, then

$$|\operatorname{supp} x_0| \ \leq \ d + 1 - \dim V \ \leq \ \min\{r, \ d - \dim \mathcal{C}\} + 1. \tag{12}$$

PROOF. Consider the set $\mathcal{S} := V \cap \Delta(\operatorname{supp} x_0)$. We have $x_0 \in \mathcal{S} \subset \mathcal{C}$, where x_0 locally maximizes the strictly convex restriction $f|_{\mathcal{S}}$ of f to $\mathcal{S}$. Thus, x_0 must be an extreme point of $\mathcal{S}$, which is only possible if $\mathcal{S} = \{x_0\}$ ($\mathcal{S}$ is open in $\operatorname{aff}\mathcal{S}$). This implies

$$V \ \cap \ \operatorname{aff}\Delta(\operatorname{supp} x_0) \ = \ \{x_0\}. \tag{13}$$

Now we apply the dimension formula

$$\begin{aligned} d \ &= \ \dim \operatorname{aff}\bar{\Delta} \\ &\geq \ \dim \operatorname{aff}\Big(V \ \cup \ \operatorname{aff}\Delta(\operatorname{supp} x_0)\Big) \\ &= \ \dim V \ + \ |\operatorname{supp} x_0| - 1 \ - \ \underbrace{\dim\left(V \ \cap \ \operatorname{aff}\Delta(\operatorname{supp} x_0)\right)}_{\overset{(13)}{=} 0} \\ &\geq \ \max\{d - r, \ \dim \mathcal{C}\} + |\operatorname{supp} x_0| - 1. \end{aligned}$$

This gives us the estimations (12).

$$\left(\sum_{\substack{\beta\in\Omega_V \\ \beta_\nu=\sigma_\nu}} \qquad \sum_{\substack{\beta\in\Omega_V \\ \beta_\nu=\sigma_\nu}} \right)$$

which can be represented as the intersection of $\bar{\Delta}$ with an affine subspace of $\mathbb{R}^{\Omega_V} \times \mathbb{R}^{\Omega_V \times \Omega_V}$ that is given by $\sum_{\nu\in V} (|\Omega_\nu| - 1)$ equations. In order to apply Lemma A.1, we prove that the interaction I is strictly convex on $\mathcal{C}$:
The entropic representation (6) in Proposition 3.1 immediately implies that the continuous function I can be explicitely written as

$$\begin{aligned} I(p,K) \;&=\; \sum_{\nu\in V} \left\{ \sum_{\sigma_\nu,\sigma'_\nu\in\Omega_\nu} h\left(\sum_{\substack{\alpha,\beta\in\Omega_V \\ \alpha_\nu=\sigma_\nu,\,\beta_\nu=\sigma'_\nu}} p(\alpha)\, K(\beta\,|\,\alpha) \right) \;-\; \sum_{\sigma_\nu\in\Omega_\nu} h\left(\sum_{\substack{\alpha\in\Omega_V \\ \alpha_\nu=\sigma_\nu}} p(\alpha) \right) \right\} \\ &\quad - \sum_{\sigma\in\Omega_V} p(\sigma)\, H(K(\cdot\,|\,\sigma)). \end{aligned} \tag{14}$$

Here, the continuous function $h : [0,1] \to \mathbb{R}$ is defined by

$$x \;\mapsto\; h(x) := \begin{cases} -x\,\ln x, & x \neq 0 \\ 0 \quad , & x = 0 \end{cases} .$$

Consider the restriction of the interaction I to the set $\mathcal{C}$. According to formula (14), for all $(q,L) \in \mathcal{C}$ one has

$$\begin{aligned} I(q,L) \;&=\; \sum_{\nu\in V} \left\{ \sum_{\sigma_\nu,\sigma'_\nu\in\Omega_\nu} h\left(\sum_{\substack{\alpha,\beta\in\Omega_V \\ \alpha_\nu=\sigma_\nu,\,\beta_\nu=\sigma'_\nu}} q(\alpha)\, L(\beta\,|\,\alpha) \right) \;-\; \sum_{\sigma_\nu\in\Omega_\nu} h\left(\sum_{\substack{\alpha\in\Omega_V \\ \alpha_\nu=\sigma_\nu}} q(\alpha) \right) \right\} \\ &\quad - \sum_{\substack{\sigma\in\Omega_V \\ \sigma\neq\omega}} q(\sigma)\, H(L(\cdot\,|\,\sigma)) \;-\; q(\omega)\, H(L(\cdot\,|\,\omega)) \end{aligned}$$

References

[1] Abeles, M. (1991) *Corticonics: Neural circuits of the cerebral cortex*. Cambridge: Cambridge University Press.

[2] Abeles, M., Vaadia, E., Bergman, H., Prut, Y., Headman, I., & Slovin, H. (1993) Dynamics of Neuronal Interactions in the Frontal Cortex of Behaving Monkeys. *Concepts in Neuroscience*, 4, 131–158.

[3] Aertsen, A.M.H.J., Gerstein, G.L., Habib, M.K., & Palm, G. (1989) Dynamics of Neuronal Firing Correlation: Modulation of "Effective Connectivity". *Journal of Neurophysiolgy*, 61, 900–917.

[4] Amari, S.-I. (2001) Information Geometry on Hierarchy of Probability Distributions. *IEEE Transactions on Information Theory*, 47, 1701–1711.

[5] Ay, N. (2001a) An Information Geometric Approach to a Theory of Pragmatic Structuring. *Annals of Probability*, in press.

[6] Ay, N. (2001b) Locality of global stochastic interaction in directed acyclic networks. Submitted.

[7] Ay, N. (2001c) Information Geometry on Complexity and Stochastic Interaction. Submitted.

[12] Gray, C.M. (1994) Synchronous Oscillations in Neuronal Systems: Mechanisms and Functions. *Journal of Computational Neuroscience*, 1, 11–38.

[13] Kullback, S. ,& Leibler, R. A. (1951) On information and sufficiency. *Annals of Mathematical Statistics*, 22, 79–86.

[14] Laughlin, S. (1981) A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung*, 36, 910–912.

[15] Linsker, R. (1986) From Basic Network Principles to Neural Architecture. *Proceedings of the National Academy of Sciences*, USA 83, 7508–7512.

[16] Linsker, R. (1988) Self-organization in a perceptual network. *IEEE Computer*, 21, 105–117.

[17] Malsburg, C. von der (1981) The correlation theory of brain function. *Internal Report 81-2*. Max-Planck Institut für Biophysikalische Chemie, Göttingen.

[18] Martignon, L.; von Hasseln, H.; Grün, S.; Aertsen, A., & Palm, G. (1995) Detecting higher-order interactions among the spiking events in a group of neurons. *Biological Cybernetics*, 73, 69–81.

[19] Martignon, L.; Deco, G.; Laskey, K.; Diamond, M.; Freiwald, W. & Vaadia, E. (2000) Neural Coding: Higher-order temporal patterns in the neurostatistics of cell assemblies. Neural Computation 12, 2621–2653.

555–586.

[24] Tononi, G., Sporns, O., & Edelman, G.M. (1994) A measure for brain complexity: Relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences USA*, 91, 5033–5037.

[25] Wennekers, T., & Palm G. (2000) Cell Assemblies, Associative Memory and Temporal Structure in Brain Signals. In: Miller, R. (ed.) *Time and the Brain. Conceptual Advances in Brain Research*, vol.2, pp. 251–273, Harwood Academic Publishers.