# Max-Planck-Institut
## für Mathematik
## in den Naturwissenschaften
## Leipzig


**Information-theoretic grounding of finite automata in neural systems**

by


*Thomas Wennekers and Nihat Ay*



Preprint no.: 52 2002


# Information-theoretic grounding of finite automata in neural systems

Thomas Wennekers and Nihat Ay

*Max-Planck-Institute for Mathematics in the Sciences,*

*Inselstraße 22–26, D-04103 Leipzig, Germany*

(June 26, 2002)

## Abstract

We introduce a measure "stochastic interaction" that captures spatial *and* temporal signal properties in recurrent systems. The measure quantifies the Kullback-Leibler divergence of a Markov chain from a product of split chains for the single units. Maximization of stochastic interaction, also called "Temporal Infomax", is shown to induce almost deterministic dynamical systems for unconstrained Markov chains. If part of the units are clamped to prescribed stochastic processes providing external input, Temporal Infomax leads to finite automata, either completely deterministic or at most weakly non-deterministic. This way, computational capabilities may arise in neural systems.

84.35.+i, 87.19.La, 02.50.Ga

A fundamental question in computational neuroscience asks for the nature of codes employed by cortical neurons [1]. Information theory provides a framework to study neural coding taking into account the stochasticity of neural spike trains as well as interactions or correlations in the joint activities of populations of cells [2–4]. The concept of *mutual information* has turned out a particularly useful guiding principle in these studies: According to common theorizing about neural function it has been suggested that individual neurons try to maximize the information they convey in their output spike trains about a given stimulus ensemble at their input. For instance, Fairhall et al. [5] have recently shown that single neurons adapt their firing statistics and range of operation to the mean and variance of stimuli in such a way that the mutual information between input and output is maximized.

Linkser has further shown that information maximization in layered feedforward systems leads to the development of spatial filters similar to receptive fields of neurons in primary visual areas, and, in addition, to a self-organization of the network into layered maps with smoothly varying filter properties inside each layer but increasingly complex filters in the hierarchy of layers [6]. These computationally demonstrated phenomena closely reflect the organization of the early visual system as known from neuroanatomy and -physiology [7]. Accordingly, Linsker's principle links two previously unrelated areas of research: Information theory and the primary processing of visual information in real brains.

Linkser's Infomax principle is related to *principle component analysis* [8,9]. That is, the spatial filter developed by a given neuron basically represents the eigenvector corresponding to the largest eigenvalue of the covariance matrix of the neuron's inputs. This again maximizes the output variance and, thus, the mutual information between the probability distributions of the spatial inputs and the neuron's output. In feedforward systems higher order correlations can further be represented in subsequent layers of the model. This way cortical feature hierarchies may arise, eventually yielding distributed "object" representations in "cognitive" brain areas.

Neural systems, however, are in general reccurently connected and non-stationary, both properties not reflected by the classical Infomax principle. In order to capture intrinsically

temporal aspects of dynamic interactions in recurrent networks the concept of information maximization has been extended by Ay [10] to the dynamical setting of Markov processes, where it is referred to as (*stochastic*) *interaction*. Linsker's approach for stationary input-output transformations can be shown to be a special case of this more general framework [9]. In the present paper we consider the optimization of the spatio-temporal stochastic interaction measure in Markov chains and demonstrate that this leads to globally almost deterministic dynamical systems, where nonetheless every single unit generates virtually random activity as characterized by a high entropy. Furthermore, we investigate Markov chains, where a part of the system is clamped to prescribed stochastic processes. Surprisingly, Markov processes that optimize stochastic interaction under such an input constraint turn out to be almost deterministic finite automata, where the internal dynamics is driven by the input through complex, globally almost deterministic state sequences. Therefore, our approach relates spatio-temporal information maximization ("Temporal Infomax") to computing devices.

We consider a set $V = \{1, \ldots, N\}$ of binary units with state sets $\Omega_\nu = \{0, 1\}$, $\nu \in V$. For a subsystem $A \subset V$, $\Omega_A := \{0, 1\}^A$ denotes the set of all configurations restricted to $A$, and $\bar{P}(\Omega_A)$ is the set of probability distributions on $\Omega_A$. Given two subsets $A$ and $B$, $B \neq \emptyset$, $\bar{K}(\Omega_B \,|\, \Omega_A)$ is the set of Markov kernels from $\Omega_A$ to $\Omega_B$. If $A = B$ we also write $\bar{K}(\Omega_A) = \bar{K}(\Omega_A \,|\, \Omega_A)$.

For a probability distribution $p \in \bar{P}(\Omega_A)$ and a Markov kernel $K \in \bar{K}(\Omega_B \,|\, \Omega_A)$ we define a *Markov transition* as the pair $(p, K)$ and the *conditional entropy* of $(p, K)$ as

$$H(p, K) \;=\; -\sum_{\substack{\omega' \in \Omega_B \\ \omega \in \Omega_A}} p(\omega)\, K(\omega' \,|\, \omega)\, \ln K(\omega' \,|\, \omega)\;. \tag{1}$$

$H(p, K)$ is a natural extension of the Shannon-entropy to Markov transitions, since $-\ln K(\omega' \,|\, \omega)$ in (1) is the information content of an individual state transition supposed $\omega$ is known and $K(\omega' \,|\, \omega)p(\omega)$ is the probability for that transition. Thus, $H(p, K)$ measures the average information generated by the Markov transition $(p, K)$.

In correspondence with marginal probability distributions for stationary joint distribu-

tions, we define marginal kernels of $K$ for all $\omega_\nu, \omega'_\nu \in \Omega_\nu$ by

$$K_\nu(\omega'_\nu \,|\, \omega_\nu) \;:=\; \frac{\sum_{\substack{\sigma,\sigma' \in \Omega_V \\ \sigma_\nu = \omega_\nu,\, \sigma'_\nu = \omega'_\nu}} p(\sigma)\, K(\sigma' \,|\, \sigma)}{\sum_{\substack{\sigma \in \Omega_V \\ \sigma_\nu = \omega_\nu}} p(\sigma)}. \tag{2}$$

Equation (2) projects the full kernel $K(\sigma' \,|\, \sigma)$ defined on the whole state space to a kernel $K_\nu(\omega'_\nu \,|\, \omega_\nu)$ for only unit $\nu$. In (2), $p_\nu(\omega_\nu) = \sum_{\substack{\sigma \in \Omega_V \\ \sigma_\nu = \omega_\nu}} p(\sigma)$ are the marginal probabilities for unit $\nu$, such that the pairs $(p_\nu, K_\nu),\ \nu = 1, \cdots, N$ are the marginal Markov transitions of the transition $(p, K)$.

The conditional entropy in (1) enables the definition of a divergence or distance of a given transition from its product of marginal transitions: This *(stochastic) interaction measure* of $K$ with respect to $p$ is defined as

$$I(p, K) \;:=\; \sum_{\nu \in V} H(p_\nu, K_\nu) \;-\; H(p, K)\ . \tag{3}$$

Equation (3) has the form of a Kullback-Leibler divergence and generalizes the usual mutual information to Markov transitions. It measures how much $(p, K)$ deviates from $N$ independent transitions, or, in other words, how strong the units in $(p, K)$ "interact" stochastically. For $N$ binary units $I(p, K) \leq N \ln 2$, because $H(p, K)$ is zero for deterministic systems, and the maximum entropy of a single unit with $|\Omega_\nu| = 2$ states is $\ln |\Omega_\nu| = \ln 2$.

Now, assume that the set of units $V$ is separated into a subset $\partial \subset V$ of "peripheral" or "input" units and "internal" units, $V \setminus \partial$. The dynamics on $\partial$ is given by a fixed Markov transition $(p^\partial, K^\partial)$ independent of the internal units. Then, the Markov kernel $K$ reads

$$K(\omega' \,|\, \omega) = K(z', a' \,|\, z, a) = K'(z' \,|\, z, a) K^\partial(a' \,|\, a)\ , \tag{4}$$

where $\omega, \omega' \in \Omega_V$, $\omega = (z, a),\ \omega' = (z', a'),\ a,\ a' \in \Omega_\partial$, and $z,\ z' \in \Omega_{V \setminus \partial}$. In this setting, $(p^\partial, K^\partial)$ represents some stochastic spatio-temporal process in the "outer world" and in generalization of the classical Infomax principle our aim is to optimize the internal dynamics of the system (given by $K'$ in (4)) such that the stochastic interaction measure, $I(p, K)$, is maximized.

For $K$ as in (4) it is straightforward to show that the conditional entropy satisfies

$$H(p,K) = H(p,K') + H(p^\partial, K^\partial)\ . \tag{5}$$

Thus, the total kernel entropy can be written as a sum of the entropy of the periphery and that of the internal transition $(p, K')$. With (5) $I(p,K)$ in (3) becomes

$$I(p,K) = \sum_{\nu\in V\setminus\partial} H_\nu(p_\nu, K_\nu) - H(p,K') + I(p^\partial, K^\partial). \tag{6}$$

Eqn. (6) reveals, that also the interaction $I(p,K)$ can be written as a sum of terms for the periphery and the internal Markov transition $(p,K')$. Because $I(p^\partial, K^\partial)$ is constant during optimization, maximizing $I(p,K)$ is therefore equivalent to the maximization of $\sum_{\nu\in V\setminus\partial} H_\nu(p_\nu,K_\nu) - H(p,K')$. First, consider $H(p,K')$:

$$H(p,K') = -\sum_{\substack{z'\\ z,a}} p(z,a)K'(z'\,|\,z,a)\ln K'(z'\,|\,z,a) \tag{7}$$

$$= \sum_{z,\ a} p(z,a)\underbrace{\left(-\sum_{z'} K'(z'\,|\,z,a)\ln K'(z'\,|\,z,a)\right)}_{=H(K'(\cdot\,|\,z,a))}\ . \tag{8}$$

The underbraced term in (8) is obviously the Shannon-entropy generated by the internal state transitions induced by $K'$ restricted to the fixed source state $(z,a)$. Clearly, if $K'(\cdot\,|\,z,a)$ is deterministic, i.e., if it is 1 for only a single target state $z'$, the Shannon-entropy $H(K'(\cdot|z,a))$ vanishes. Thus, if all $K'(\cdot\,|\,z,a),\ z\in\Omega_{V\setminus\partial},\ a\in\Omega_\partial$ are deterministic, the total entropy $H(p,K')$ obtains its absolute minimum of 0. Maximizing $-H(p,K)$, therefore, favours determinism in *global* state transitions. However, observe that maximizing $I(p,K')$ requires in addition that the marginal entropies $H_\nu(p_\nu,K_\nu),\ \nu\in V\setminus\partial$ are as large as possible, that is, the *local* single unit activities must be as unpredictable as possible. This prohibits "degenerate" chains where, e.g., some units are constant or otherwise input-independent.

The subsequent simulations show that both seemingly contradicting constraints can be satisfied either perfectly or at least in good approximation. They implement the usual Markov dynamics on a set of $N$ binary units to generate sample trajectories and a random search scheme to optimize $I(p,K)$ of Markov chains with kernels

$$K(z',a'\,|\,z,a) = \left[\prod_{\nu\in V\setminus\partial} K^{(\nu)}(z'_\nu\,|\,z,a)\right] K^\partial(a'\,|\,a)\ . \tag{9}$$

Details of the simulations will be presented elsewhere. In contrast to (4) the kernel $K'$ in square brackets in (9) is of product form reflecting the independent output generation of cells given their inputs in neural modeling. A "parallel kernel" $K'$ as in (9) is deterministic, if all its entries are either 0 or 1. Beside that, results for parallel and general kernels are identical.

Figure 1 displays an optimized system with $N = 3$ units and *no* units clamped, i.e., $\partial = \emptyset$. Figure 1A shows the optimized Markov matrix $K(\omega' \,|\, \omega)$. Most columns reveal only a single possible transition indicating determinism. However, there are two exceptions, states $\omega = 111$ and 010. State 111 is a *transient* state: It has outgoing transitions to some (here, all) other states, but none of the other states projects back to it. Therefore, once left, state 111 is never occupied again. State 010 is what we call a *branching state*. As Fig. 1C shows, which just redisplays the matrix in A as a state transition graph, state 010 is part of two nested loops of states with deterministic transitions between nodes. Only state 010 has two outgoing – and therefore non-deterministic – transitions. As a consequence, activity flows deterministically along consecutive states of the individual loops, but at state 010 it can switch randomly between two possible targets leading back to one or the other deterministic sequence of states. Therefore, sample trajectories of the dynamics are characterized by randomly interleaved sequences of repetetive deterministic firing patterns as shown in Fig. 1B. Individual units, however, reveal largely unpredictable firing.

The example in Fig. 1 is also characteristic for larger systems. Whereas in generic Markov chains transitions between arbitrary states are possible, the dynamics of strongly interacting chains is confined to a core of nested deterministic subsequences of states linked by branching nodes, and augmented by a set of transient states. Accordingly, strongly interacting isolated Markov chains are globally almost deterministic but locally unpredictable.

Figure 2 shows an example system comprising $N = 4$ units, but two units clamped to a Markov chain with equal transition probabilities between peripheral states. Figure 2A displays the respective peripheral kernel $K^{\partial}$ and Fig. 2B the optimized full kernel.

The most prominent difference between the kernels in Fig. 2B and Fig. 1A is that the

columns in Fig. 2B do not reveal just one, but four entries. These entries are grouped into blocks, as indicated in the figure, and all transitions for one global state $(z, a)$ target in exactly one of the blocks. Moreover, the blocks are uniquely characterized by internal states, $z,\ z'$, whereas the peripheral states $a,\ a'$ only indicate the precise location inside each block. Thus, given an internal state $z$ and a peripheral state $a$, the next *internal* target state $z'$ is uniquely defined, that is, $K'(z' \,|\, z, a)$ is deterministic. Nonetheless, because the dynamics on the periphery is random, sample trajectories again do not reveal much determinism.

Computer science [11] defines deterministic finite automata (DFAs) as a quintuple $M = (Z, \Sigma, \delta, z_0, E)$, where $Z = \{z_1, \cdots, z_n\}$ is a finite *set of states* and $\Sigma = \{a_1, \cdots, a_m\}$ a finite *alphabet*. The designated state $z_0 \in Z$ is called the *initial state* and $E \subseteq Z$ the set of *final or accepting states*. Operation of the automaton is defined by the *transition table* $\delta : Z \times \Sigma \to Z$ which maps every pair $(z, a) \in Z \times \Sigma$ to exactly one successor state.

Now, observe that as in a finite automaton, the strongly interacting Markov chain in Fig. 2 provides a total mapping from $\Omega_{V\setminus\partial} \times \Omega_\partial$ to $\Omega_{V\setminus\partial}$. We may, thus, identify the internal state space $\Omega_{V\setminus\partial}$ with the state set $Z$ of a DFA, and the peripheral states $\Omega_\partial$ with the set of symbols $\Sigma$. Then the Markov kernel $K'(z' \,|\, z, a)$ reflects exactly the transition table of a DFA, and can be represented by a labeled state transition graph, see Fig. 3. (For a complete correspondence we also have to designate an initial state, $z_0$, and accepting states, $E$, in our Markov models, but these issues are of secondary importance for the present paper – cf., e.g., [12] for related work).

The example in Fig. 2 again is also typical for larger systems and arbitrary Markov chains on the periphery. However, the optimized internal dynamics must not always be perfectly deterministic, but as for $K$ in Fig. 1 some columns $K'(\cdot \,|\, z, a)$ in optimized constrained chains may occasionally contain more than one positive entry. In that case given an internal state and input two or several internal target states are possible, a situation corresponding with *nondeterministic finite automata* in computer science [11]. In fact, it can be proven mathematically that the number of outgoing edges of any node in the ergodic (i.e., non-transient) component of the dynamics is linearly bounded in system size

$N$. Since $2^N$ transitions are possible, maximization of stochastic interaction in constrained Markov chains therefore leads to systems characterized by deterministic or at most weakly nondeterministic, input-driven, internal state transitions. The development of the internal structure is controled by minimization of $H(p, K')$ which favours global determinism, and maximization of the single unit entropies $H_\nu$, which enforces an "unfolding" of nested loop attractors in avoidance of degenerate (e.g., constant) Markov chains.

We should finally emphasize that we cannot touch on semantic issues in the present work, that is, whether and how the finite automata resulting from our Temporal Infomax principle might be related to "cognitive processes" (e.g., [12]). We rather discussed them merely as intrinsic dynamic modes of activity. With that respect it seems interesting that physiological experiments in higher brain areas indeed give evidence for behavior-related, repetetive, deterministic firing patterns in neural populations [13] accompanied by complex spatio-temporal correlations [14]. Moreover, in behaviorally relevant cortical areas it has been observed that neural activity flips between quasi-stationary states [15], a phenomenon well describable by Markov models [16]. Thus, just as classical Infomax explains properties of sensory systems, it seems possible that Temporal Infomax structures recurrent networks in higher cortical processing stages.

# REFERENCES


[1] L.F. Abbott and T.J. Sejnowski *Neural Codes and Distributed Representations.* (MIT press, Cambridge, MA, 1999).

[2] P. Dayan and L.F. Abbott, *Theoretical Neuroscience.* (MIT Press, Cambridge, MA, 2001).

[3] F. Rieke, D. Warland, R. Ruyter van Steveninck and W. Bialek, *Spikes: Exploring the Neural Code* (MIT Press, Cambridge, MA, 1998).

[4] T.M. Cover and J.A. Thomas, *Elements of Information Theory.* (Wiley, New York, 1991).

[5] A.L. Fairhall, G.D. v. Lewen, W. Bialek and R. Ruyter van Steveninck, Nature **412**, 787 (2001).

[6] R. Linsker, Proc.Natl.Acad.Sci. **83**, 7508 (1986); **83**, 8390 (1986); **83**, 8779 (1986).

[7] J.G. Nicolls, *From Neuron to Brain, 4th ed.*, (Sinaur Assoc., Sunderland, MA, 2001).

[8] J. Hertz, A. Krogh and R.G. Palmer, *Introduction to the Theory of Neural Computations* (Perseus Books, Cambridge, MA, 1991).

[9] N. Ay, Neural Comput., in press.

[10] N. Ay, IEEE Trans.Info.Theory, submitted.

[11] J.E. Hopcroft and J.D. Ullman, *Introduction to Automata Theory, Languages, and Computation.* (Addison-Wesley, Reading, MA, 1979).

[12] T. Wennekers, Technical Report # 98-08, Faculty for Computer Science, University of Ulm, 1998.

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

[14] A.M.H.J. Aertsen, G.L. Gerstein, M.K. Habib and G. Palm, J. Neurophysiol. **61**, 900 (1989).

[15] M. Abeles, H. Bergman, I. Gat, I. Meilijson, E. Seidemann, N. Tishby and E. Vaadia, Proc.Natl.Acad.Sci. **92**, 8616 (1995).

[16] I.Gat, N. Tishby, M. Abeles, Network - Comp. Neural **8**, 297 (1997).

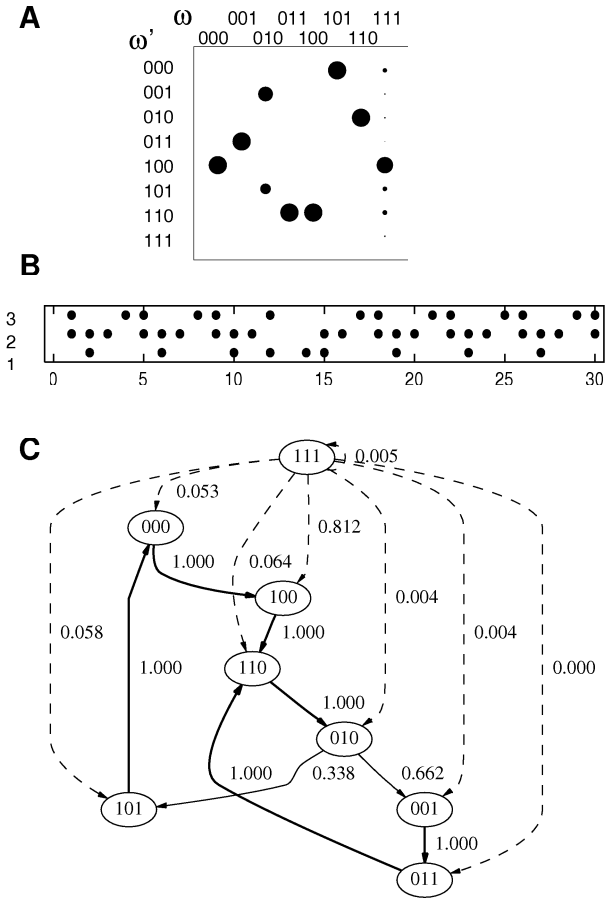


FIG. 1. Example for unconstrained optimization using 3 units. A) Optimized Markov matrix (dot-size indicates transition probability); B) sample trajectory (dots correspond with an output of 1); C) transition graph representing the matrix in A (node labels denote states, edge labels transition probabilities). State 111 is 'transient' and 010 a 'branching state'.

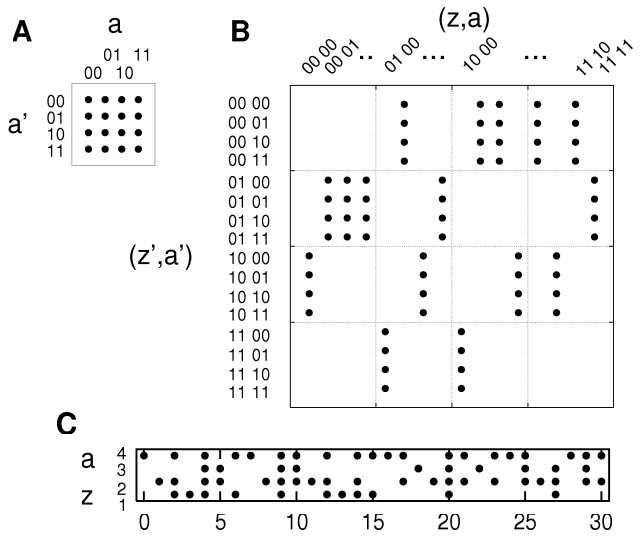


FIG. 2. A strongly interacting Markov chain with $N = 4$ units, $|\partial| = 2$ of which clamped to a peripheral chain with equal transition probabilities (.25) between peripheral states $a,\ a'$. A: the peripheral kernel $K^{\partial}(a' \,|\, a)$; B) full Markov kernel $K(\omega' \,|\, \omega) = K(z', a' \,|\, z, a)$; C) sample trajectory.

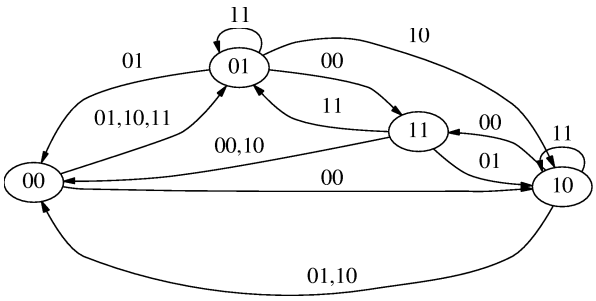


FIG. 3. Deterministic finite automaton corresponding with Fig. 2. Nodes are labeled by internal states $z \in \Omega_{V\setminus\partial}$ and edges by peripheral states $a \in \Omega_{\partial}$.