LETTER **Communicated by Peter Tino**

Finite State Automata Resulting from Temporal Information Maximization and a Temporal Learning Rule

Thomas Wennekers
Thomas.Wennekers@plymouth.ac.uk
Centre for Theoretical and Computational Neuroscience, University of Plymouth, Plymouth PL4 8AA, U.K.; Institute for Neuroinformatics, Ruhruniversity Bochum, 44780 Bochum, Germany; and Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany

Nihat Ay
ay@mi.uni-erlangen.de
Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany; Mathematics Institute, Friedrich Alexander University Erlangen-Nuremberg, 91054 Erlangen, Germany; and Santa Fe Institute, Santa Fe, NM 87501, U.S.A.

We extend Linkser's Infomax principle for feedforward neural networks to a measure for stochastic interdependence that captures spatial *and* temporal signal properties in recurrent systems. This measure, *stochastic interaction*, quantifies the Kullback-Leibler divergence of a Markov chain from a product of split chains for the single unit processes. For unconstrained Markov chains, the maximization of stochastic interaction, also called *Temporal Infomax*, has been previously shown to result in almost deterministic dynamics. This letter considers Temporal Infomax on constrained Markov chains, where some of the units are clamped to prescribed stochastic processes providing input to the system. Temporal Infomax in that case leads to finite state automata, either completely deterministic or weakly nondeterministic. Transitions between internal states of these systems are almost perfectly predictable given the complete current state and the input, but the activity of each single unit alone is virtually random. The results are demonstrated by means of computer simulations and confirmed analytically. It is furthermore shown numerically that Temporal Infomax leads to a high information flow from the input to internal units and that a simple temporal learning rule can approximately achieve the optimization of temporal interaction. We relate these results to experimental data concerning the correlation dynamics and functional connectivities observed in multiple electrode recordings.

1 Introduction

A fundamental question in computational neuroscience asks for the nature of codes employed by cortical neurons (Dayan & Abbott, 2001; Abbott & Sejnowski, 1999). Experiments suggest a considerable interaction of neurons already on the level of spikes, for instance, expressed by spatiotemporal correlations in multiple unit recordings (Abeles, Bergman, et al., 1993a; Eckhorn, 1999; Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1998; Singer & Gray, 1995; Vinje & Gallant, 2000). Such correlations have been a matter of intensive theoretical and conceptual research, (cf., e.g., Abeles, 1991; Abeles, Vaadia, et al., 1993b; Aertsen, 1993; Gerstein, Bedenbaugh, & Aertsen, 1989; Palm & Aertsen (1986); Wennekers, Sommer, & Aertsen, 2003).

A well-known measure that quantifies spatial relations of interacting units is the *multi-information* (see Studený & Vejnarova, 1998) shared among the units, which is called *mutual information* in the case of two units (see Cover & Thomas, 1991). Multi-information can be expressed in terms of the Kullback-Leibler divergence as

$$I(p) := D(p \parallel p_1 \otimes \cdots \otimes p_N) = \sum_{\nu=1}^{N} H(p_\nu) - H(p). \tag{1.1}$$

In equation 1.1, $H(\cdot)$ denotes the usual Shannon entropy and p_ν the νth marginal of p. $I(p)$ measures the distance of p from the factorized distribution $p_1 \otimes \cdots \otimes p_N$. It is a natural measure for spatial interdependence of N stochastic units and a starting point of many approaches to neural coding and complexity, e.g., Martignon, von Hasseln, Grün, Aertsen, & Palm, 1995; Martignon et al., 2000; Nakahara & Amari, 2002; Rieke et al., 1998; Sporns, Tononi, & Edelman, 2000; Tononi, Sporns, & Edelman, 1994; Ay, 2002; Wennekers & Ay, 2003.

Linsker (1986a, 1986b, 1986c), for instance, considered layered feedforward neural systems that maximize the mutual information between (stationary) input and output probability distributions. This approach is closely related to maximizing equation 1.1 (cf. Ay, 2002). Linsker's work revealed surprising relations between information maximization and the spatial structure of receptive fields in the visual system. Recent experiments in that direction suggest that individual neurons may even adapt dynamically to maximize their information transfer with respect to a given stimulus ensemble (Fairhall, Lewen, Bialek, & de Ruyter van Steveninck, 2001).

As a fundamental concept regarding neural coding, the Infomax principle is still being discussed and developed further toward confined models for the development and functional significance of receptive fields (see, e.g., Abbott & Sejnowski, 1999; Barlow, 2001; Bell & Sejnowski, 1995; Li & Atick, 1994; Penev & Atick, 1996). Additional sparseness constraints have been shown to lead to more realistic receptive field types than pure Infomax

(Simoncelli & Olshausen, 2001). In addition, theoretical work revealed a close relation between information maximization, on one hand, and principal or independent component analysis, on the other. This puts Linsker's Infomax in line with projection techniques for high-dimensional data analysis (Cutler & Breiman, 1994; Hertz, Krogh, & Palmer, 1991; Bell & Sejnowski, 1995; Lee, Girolami, Bell, & Sejnowski, 2000). For these reasons, information-based maximization principles can be seen an important guiding principle in computational neuroscience and neuroinformatics.

Infomax methods reduce redundancy in feedforward neural networks for artificial and real data. This appears useful for primary and secondary sensory systems, which then encode stimuli into independent channels, a concept that is quite widely believed in neuroscience. A second common belief, however, is that real neural systems are highly interacting and reveal complex spatiotemporal correlations. In fact, those correlations are usually assigned to recurrent synaptic pathways, which are neglected in the classical feedforward frameworks. It might well be that early sensory processing makes heavy use of decorrelation and independent channels, but on a higher level, information must be reintegrated into functional ensembles and combined depending on the context that a set of features appears in. Hebbian learning in recurrent synapses has been proposed to provide this (Hebb, 1949; Wennekers et al., 2003). It can be seen as amplifying synapses connecting related cells in response to correlations in input patterns. Also, brain processes are dynamic and to a high degree independent of sensory stimulation. Accordingly, correlations between cells can also build up without sensory input, and the respective processes are most naturally considered in the spatiotemporal domain, not just for stationary stimulus ensembles with spatial correlations only. It is therefore of interest to study information maximization in more general settings than classical Infomax. In order to capture intrinsically temporal aspects of dynamic interactions in recurrent networks, the measure I in equation 1.1 has been extended by Ay (2001) to the dynamical setting of Markov processes, where it is referred to as (*stochastic*) *interaction* and can be seen as measuring the amount of correlations in a Markov process. In a previous paper (Ay & Wennekers, 2003), we have shown that the maximization of stochastic interaction in Markov chains leads to globally almost deterministic dynamical systems, where, nonetheless, every unit generates virtually random activity as characterized by a high entropy. That work neglected external input into the systems under study. This letter therefore investigates the more interesting case of Markov chains, where a part of the system is clamped to prescribed stochastic processes, but only the internal dynamic is optimized toward large stochastic interaction. Importantly, Markov processes optimized under this input constraint turn out to be finite state automata, where the internal dynamic is driven by the external input through complex, almost deterministic global state sequences, but again, single unit activity is virtually random.

To demonstrate and explain these phenomena, the article is organized as follows. Section 2 introduces the basic formalism of constrained Markov chains and generalizes Shannon entropy and mutual information to the spatiotemporal dynamics of Markov chains. Section 3 presents detailed simulations of small example systems with numerically maximized stochastic interaction under the constraint that parts of the systems follow prescribed stochastic processes. Section 4 explains the basic features of strongly interacting systems observed in the simulations on the base of mathematical properties of $I(p, K)$ and derives strict upper bounds for the amount of order and entropy in optimized systems. Proofs of the stated theorems are sketched in the appendix. The letter closes with a discussion that relates our results to experiments concerning spatiotemporal correlations in biological neural ensembles.

2 Temporal Infomax on Constrained Markov Chains

Consider a set $V = \{1, \ldots, N\}$ of binary units with state sets $\Omega_v = \{0, 1\}$, $v \in V$. For a subsystem $A \subset V$, $\Omega_A := \{0, 1\}^A$ denotes the set of all configurations restricted to A, and $\bar{\mathcal{P}}(\Omega_A)$ is the set of probability distributions on Ω_A. Given two subsets A and B, where B is nonempty, $\bar{\mathcal{K}}(\Omega_B \mid \Omega_A)$ is the set of all Markov kernels from Ω_A to Ω_B. In the case $A = B$, we use the abbreviation $\bar{\mathcal{K}}(\Omega_A) = \bar{\mathcal{K}}(\Omega_A \mid \Omega_A)$. Informally, a Markov kernel transforms probability distributions over Ω_A into probability distributions over Ω_B: $p_B(\omega') = \sum_{\omega \in \Omega_A} K(\omega' | \omega) p_A(\omega)$, where $p_A \in \bar{\mathcal{P}}(\Omega_A)$, $p_B \in \bar{\mathcal{P}}(\Omega_B)$, and $K \in \bar{\mathcal{K}}(\Omega_B \mid \Omega_A)$. If $p_A(\omega) = \delta_{\omega,\omega''}$, that is, if p_A is localized in a single state ω'', then $p_B(\omega') = K(\omega' | \omega'')$ for all $\omega' \in \Omega_B$. The $K(\omega' | \omega'')$ can therefore be interpreted as transition probabilities, and because probability distributions are normalized, $\sum_{\omega' \in \Omega_B} p_B(\omega') = 1$, so must be Markov kernels in their first argument, $\sum_{\omega' \in \Omega_B} K(\omega' | \omega'') = 1$ for all $\omega'' \in \Omega_A$.

A Markov chain over $A \subset V$ is an infinite sequence of random variables $X_n = (X_{v,n})_{v \in A}$, $n = 0, 1, 2, \ldots$, where X_{n+1} depends on only X_n. n is called *time* or *step*. A stationary Markov chain can be specified by an initial distribution $p_0 = P(X_0) \in \bar{\mathcal{P}}(\Omega_A)$ and a Markov kernel $K \in \bar{\mathcal{K}}(\Omega_A)$. The probability distributions for X_n then evolve like $P(X_{n+1}) = K\,P(X_n)$, that is, $P(X_n) = K^n P(X_0) = K^n p_0$. In nonstationary Markov chains, K depends on n, such that transition probabilities can change over time.

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

$$H(p, K) = - \sum_{\omega \in \Omega_A,\, \omega' \in \Omega_B} p(\omega)\, K(\omega' \mid \omega)\, \ln K(\omega' \mid \omega). \tag{2.1}$$

$H(p, K)$ defined this way is a natural extension of the Shannon entropy to Markov transitions, because $-\ln K(\omega' \mid \omega)$ in equation 2.1 is the

information content of an individual state transition supposing ω is known and $K(\omega' \mid \omega)p(\omega)$ is the probability for that transition. Thus, equation 2.1 measures the average information generated by the Markov transition (p, K) just as the Shannon entropy measures the average information contained in a stationary probability distribution p: $H(p) = -\sum_{\omega} p(\omega) \ln p(\omega)$.

Note that a Markov transition is not the same as a Markov chain. It describes just a single transformation step between states in two not necessarily equal spaces Ω_A and Ω_B. However, if $\Omega_A = \Omega_B$ and p is a stationary probability distribution of K, then a Markov transition induces a stationary Markov chain in a natural way.

As probability distributions, Markov transitions can be marginalized. We define the marginal kernels $K_\nu \in \mathcal{K}(\Omega_\nu)$, $\nu \in V$, of a kernel $K \in \mathcal{K}(\Omega_V)$ by

$$K_\nu(\omega'_\nu \mid \omega_\nu) := \frac{\sum_{\substack{\sigma,\sigma' \in \Omega_V \\ \sigma_\nu = \omega_\nu,\, \sigma'_\nu = \omega'_\nu}} p(\sigma)\, K(\sigma' \mid \sigma)}{\sum_{\substack{\sigma \in \Omega_V \\ \sigma_\nu = \omega_\nu}} p(\sigma)}, \qquad \omega_\nu, \omega'_\nu \in \Omega_\nu. \tag{2.2}$$

Equation 2.2 projects the full kernel $K(\sigma' \mid \sigma)$ defined on the whole state space to a kernel $K_\nu(\omega'_\nu \mid \omega_\nu)$ for only unit ν. Clearly, in equation 2.3, the expression $p(\sigma)\, K(\sigma' \mid \sigma)$ is the probability that the system is in state σ and transits to σ'. Thus, summing over all states σ, σ' with unit ν clamped to ω_ν and ω'_ν, respectively, gives the total probability for transitions where unit ν is in state ω_ν before and in state ω'_ν after the transition, regardless of the rest of the system. The normalization by $p_\nu(\omega_\nu) := \sum_{\substack{\sigma \in \Omega_V \\ \sigma_\nu = \omega_\nu}} p(\sigma)$ in equation 2.2 ensures that K_ν is a proper Markov kernel, that is, $\sum_{\omega'_\nu \in \Omega_\nu} K_\nu(\omega'_\nu \mid \omega_\nu) = 1$ for all $\omega_\nu \in \Omega_\nu$. In fact, p_ν is the marginal probability distribution for unit ν. Further, the pairs (p_ν, K_ν), $\nu = 1, \ldots, N$ are the marginal Markov transitions of the transition (p, K). Note that the marginal transition kernels are defined in equation 2.2 only for kernels $K \in \mathcal{K}(\Omega_V)$, and not for more general kernels in $\mathcal{K}(\Omega_B \mid \Omega_A)$. A generalization to different source and target spaces is straightforward. Marginalization with respect to sets $A' \subset A$, $B' \subset B$ is defined analogous to equation 2.3 with sums restricted to the complement sets of A' and B'.

The stochastic interaction measure of K with respect to p is defined as

$$I(p, K) := \sum_{\nu \in V} H(p_\nu, K_\nu) - H(p, K), \tag{2.3}$$

with values in the range $\left[0, \sum_{\nu \in V} \ln |\Omega_\nu|\right]$, because the minimum of $H(p, K)$ is zero for deterministic systems and the maximum entropy of a single unit with $|\Omega_\nu|$ states is $\ln |\Omega_\nu|$. Evidently, for N binary units, the maximal interaction is $N \ln 2$ (or N bits if we would use dual logarithms). Comparison with equation 1.1 shows that equation 2.3 has the form of a Kullback-Leibler divergence and generalizes the usual mutual information to Markov

transitions. It measures the divergence of (p, K) from the product of its marginal transitions, thereby indicating how much (p, K) deviates from a product of independent single unit transitions or, in other words, how strong the units in (p, K) "interact" stochastically. Observe that $I(p, K)$ is particularly large if the marginal transitions have high entropy, but that of the full transition is low. Then, supposing the current state $\omega \in \Omega_V$ is known, the next global state is predictable with high confidence, but, conversely, not much information is gained from knowledge about single units, ω_v. We call such systems *strongly interacting* and study some of their properties in the sequel. Although single unit entropies are high in strongly interacting systems, this does not mean that stochastic interaction measures basically entropy (i.e., disorder). A set of independent processes has high marginal entropies, but the joint entropy is as high as the sum of the marginals, such that the interaction of independent processes is always zero (remember that $I(p, K)$ is a KL distance). For probability distributions, Amari (2001) has shown that $I(p)$ in equation 1.1 can be represented in terms of a series of correlations of all orders. A similar representation likely exists for $I(p, K)$ (but is unproved). In this sense, $I(p, K)$ measures total correlation rather than entropy.

In order to study strongly interacting systems, we consider Markov chains $X_n = (X_{v,n})_{v \in V}$, $n = 0, 1, 2, \ldots$, given by an initial distribution $p_0 \in \bar{\mathcal{P}}(\Omega_V)$ and a kernel $K \in \bar{\mathcal{K}}(\Omega_V)$. We further separate V into two sets: a set of units ∂ called *periphery* of the system and the set $V \setminus \partial$, called the *interior* or set of *internal units*. The peripheral units represent the environment of the internal units; they provide input for the rest of the system. Given this distinction, we restrict attention to Markov kernels of the form

$$K(\omega' \mid \omega) = K(z', a' \mid z, a) = K'(z' \mid z, a)\, K^{\partial}(a' \mid a) \tag{2.4}$$

$$= \left[\prod_{v \in V \setminus \partial} K^{(v)}(\omega'_v \mid z, a) \right] K^{\partial}(a' \mid a), \tag{2.5}$$

where $\omega, \omega' \in \Omega_V$ are global states, $\omega_v \in \Omega_v$ are states of single units, $z, z' \in \Omega_{V \setminus \partial}$ are internal states, and $a, a' \in \Omega_\partial$ are states on the periphery. For frequent later use, we abbreviated the Markov kernel in square brackets in equation 2.5 as $K'(z' \mid z, a)$ in equation 2.4. From equation 2.4, it is evident that the Markov chain on the periphery behaves independent of the internal units. This can be relaxed—in principle, we could let the periphery be influenced by activity in the interior, $K^{\partial}(a' \mid z, a)$, but we do not consider this case in this work.

K^{∂} in equation 2.5 is a general Markov kernel reflecting that the environment may contain arbitrary spatiotemporal stochastic one-step dependencies. In contrast, K' has the form of a so-called parallel kernel. Given the current global state ω, each kernel $K^{(v)}$, $v = 1, \ldots, V \setminus \partial$ in equation 2.5

determines the next state of only a single unit ν independent of transitions in other units. Therefore, the kernels can be termed *local*, and the global transition is of product form similar to that of independent probability distributions. General kernels represent mappings between arbitrary global states, such that a source state ω can specifically target arbitrary subsets of other global states ω'. The state transition of a certain unit then would depend on the simultaneous transitions of other units, that is, on "nonlocal" information. Parallel Markov chains are a more natural assumption in neural modeling than general Markov chains, because the activity of a neuron is determined only by its own input and internal dynamic, not by the simultaneous activity of other cells.

An even more realistic model of transition kernels is given by parallel transitions that are adapted to a network structure. More precisely, with a graph (V, E) where $E \subseteq V \times V$ denotes the set of edges, we model the transition of each unit ν by a kernel $K^{(\nu)} \in \bar{K}(\Omega_\nu | \Omega_{\mathrm{pa}(\nu)})$. Here, $\mathrm{pa}(\nu)$ denotes the parent set of the unit ν. The global transition K is then defined by a product corresponding to equation 2.5. For a Markov chain $(X_{\nu,n})_{\nu \in V}$, $n = 0, 1, 2, \ldots$, given by such a transition kernel K and an initial distribution p stationary with respect to K, we have the following representation of $I(p, K)$ (Ay, 2002):

$$I(p, K) = \sum_{\nu \in V} I(X_{\nu,n+1} : X_{\mathrm{pa}(\nu),n} | X_{\nu,n}). \tag{2.6}$$

In equation 2.6, $I(X_{\nu,n+1} : X_{\mathrm{pa}(\nu),n} | X_{\nu,n})$ is the conditional mutual information of $X_{\nu,n+1}$ and $X_{\mathrm{pa}(\nu),n}$ given $X_{\nu,n}$, which can be interpreted as the information flow from the parent set of ν to ν. In this sense, our measure of stochastic interaction is nothing but the sum of local information flows. Thus, the simultaneous maximization of the local information flows implies the maximization of the global measure of stochastic interaction. It turns out that the implication in the opposite direction is also true within the setting of natural gradient maximization, (see Ay, 2002). Maximization of information flow for biologically realistic single neurons is subject of current research and supported by close relations to biological findings (Bi & Poo, 2001; Froemke & Dan, 2002; Fairhall et al., 2001; Gütig, Aharonov, Rotter, & Sompolinsky 2003; Chechik, 2003; Bell & Parra, 2005). Our contribution here is to provide a translation of the concept of local information flow maximization to the global setting.

We previously considered parallel Markov chains where $I(p, K)$ was numerically maximized under no further constraint regarding K, that is, there was no input, $\partial = \emptyset$ (Ay & Wennekers, 2003). We call this approach *Temporal Information Maximization*. The optimized, strongly interacting Markov chains were shown to be globally almost deterministic, but the firing of individual units was largely random and unpredictable. Strongly interacting Markov chains turned out to be representable in state-space Ω_V by complex

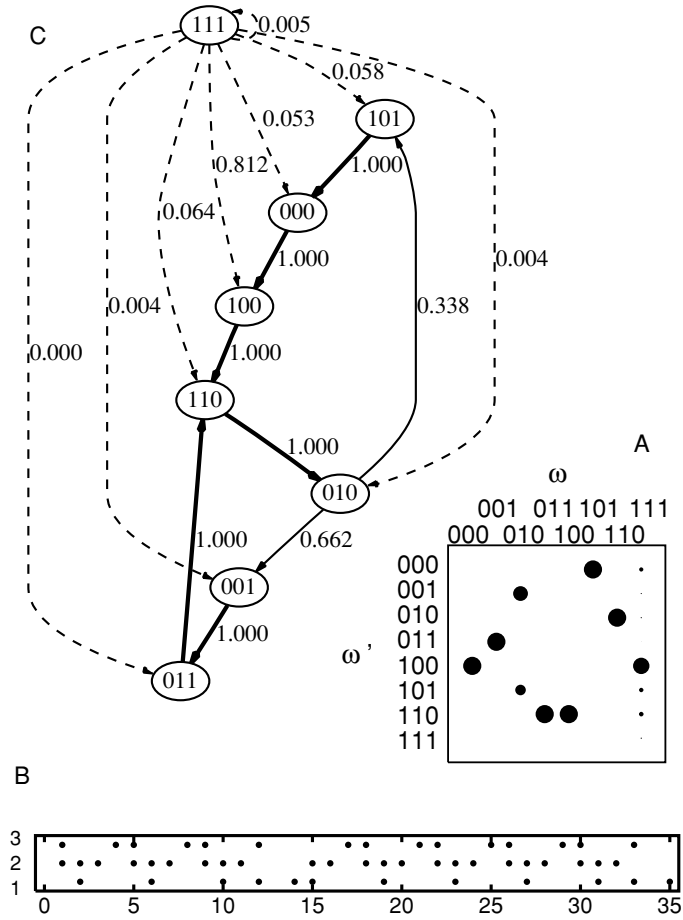

Figure 1: An example for unconstrained optimization using $N = 3$ units. (A) The optimized Markov matrix, where dot size indicates transition probability. (B) A sample trajectory as a raster plot over time; dots correspond with an output of 1. (C) The state transition graph representing the matrix in A. Node labels denote states, ω, and edge labels transition probabilities. Observe the almost deterministic asymptotic state transitions (bold edges). State 111 is transient and 010 a branching state.

but systematically structured transition graphs consisting of a fraction of transient trajectories as well as sets of attracting nested loops that correspond with almost deterministic repetitive firing patterns of various lengths (see Figure 1). The observed globally deterministic but locally random activity patterns can be envisaged as intrinsic modes of activity in strongly

interacting systems. This work considers Markov chains that maximize $I(p, K)$ under the additional constraint that the transition kernel has the form 2.4 where K^{∂} is fixed for a subset $\partial \subset V$ of units during optimization. This corresponds to Linsker's original setting, where mutual information between output and input was maximized in a one-layer feedforward neural network for a stationary stimulus ensemble. Here, we maximize local flows in a recurrent setting. A second, perhaps more direct generalization of Linsker's Infomax principle in the recurrent case would be to maximize the global information flow into the system (cf. equation 3.3). However, simulations in section 3.4 show that the maximization of stochastic interaction also leads to a quite high global flow. The relation between interaction maximization, local flows, and global flow is further worked out for the nontemporal case in feedforward systems in Ay (2002).

3 Examples for Strongly Interacting Systems

This section presents simulations of strongly interacting Markov chains with various peripheries. The simulations numerically optimize the stochastic interaction measure $I(p, K)$ in equation 2.3 for kernels of the form 2.5.

3.1 Simulation Procedures

3.1.1 Sample Trajectories. The simulations implement the usual Markov dynamics on a set of N binary units to generate sample trajectories: Starting from a random initial state, the next state is chosen iteratively according to the transition probabilities in the Markov kernel corresponding to the present state. Sample trajectories shown in subsequent figures serve only for visualization purposes; the optimization of $I(p, K)$ is performed directly on the Markov kernel and does not need them.

3.1.2 Kernel Initialization. Entries of the parallel Markov kernel for internal units ν in equation 2.5 are initialized with independent equally distributed random values in the range $[0, 1]$. The peripheral kernels $K^{\partial}(a' \mid a)$ are chosen independent of the internal unit kernels and are kept fixed during optimization. We consider different choices for the peripheral kernels, such as kernels with all entries equal, randomly initialized kernels, or special deterministic kernels. These kernels will be defined where used.

3.1.3 Numerical Optimization of $I(p, K)$. A random search scheme is implemented to optimize stochastic interaction of the Markov chains. In every time step, the interaction $I(p, K)$ is computed with respect to an induced stationary probability distribution p of the present kernel K obtained by solving the equation $Kp = p$. (Usually we start from ergodic Markov chains, where p is unique. If, during the optimization, a Markov chain becomes nonergodic, we select an arbitrary one of the solutions of $Kp = p$.) The kernel is

then perturbed such that I increases. We randomly choose a single entry in one of the individual kernels of the internal units and perturb it by a small random number ξ equally distributed on $[-r, r]$, where r is the learning rate ($r = 0.05$, if not stated otherwise). Perturbed values are clipped to the range $[0, 1]$. Formally, for randomly chosen $\nu \in V \setminus \partial$ and $\omega \in \{0, 1\}^N$, we set

$$K^{(\nu)}_{t+1}(1 \mid \omega) = \phi\left(K^{(\nu)}_t(1 \mid \omega) + \xi\right) \tag{3.1}$$

$$K^{(\nu)}_{t+1}(0 \mid \omega) = 1 - K^{(\nu)}_{t+1}(1 \mid \omega), \tag{3.2}$$

where the clipping function $\phi(x)$ is zero for $x \leq 0$, one for $x \geq 1$, and $\phi(x) = x$ otherwise. The new interaction measure is computed, and if it increases, the perturbation is accepted. Otherwise the optimization step is discarded and the procedure repeated. The simulation proceeds to the next time step if either I can be increased or five unsuccessful optimization trials have been performed.

3.1.4 Convergence. We stop optimization, output optimized values, and reinitialize the next simulation run after a fixed number of steps T (between 2000 and 12,000 depending on N, $|\partial|$, and r). This does not guarantee perfect convergence to a local maximizer, but for big enough T, the interaction becomes virtually constant under the described update scheme. Convergence issues have been studied in detail in Ay and Wennekers (2003), which considers unconstrained optimization ($\partial = \{\}$) and compares different optimization procedures. Results in that work show that a restriction to the special random search scheme described above and a fixed T stopping condition is not crucial.

After convergence, state transition graphs are constructed from the resulting full Markov kernels and plotted using the public domain software *dot*.[1] Simulation programs were implemented using the simulation environment Felix written by one of the authors (T.W.) and run on various Linux platforms. Because the optimization of I is algorithmically complex, simulations are restricted to small N.

3.2 Strongly Interacting Markov Chains and Automata. Figure 2 shows an example system comprising $N = 4$ units, where two of the units have been clamped to a Markov chain with equal transition probabilities between peripheral states. Figure 2A displays the peripheral kernel K^{∂} and Figure 2B the full Markov matrix $K(\omega' \mid \omega) = K(z', a' \mid z, a)$. Here, as well as throughout the article, units are counted from left to right in binary representations of states z, a, and $\omega = (z, a)$.

[1] The program *dot* is part of the Graph Drawing Package *graphviz* from AT&T and Lucent Bell Labs available online at http://www.research.att.com/sw/tools/graphviz.

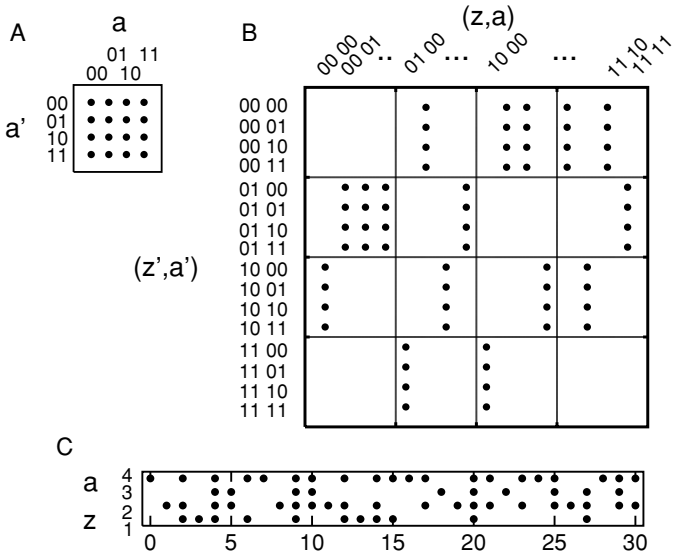

Figure 2: Optimized Markov chain with $N = 4$ units, $|\partial| = 2$ of which clamped to a peripheral chain with equal transition probabilities (0.25) between peripheral states a, a'. Interaction of the system is $I(p, K) = 1.35773 \lessapprox 2 \ln 2$. (A) Peripheral kernel $K^{\partial}(a' \mid a)$. (B) Full Markov kernel $K(\omega' \mid \omega) = K(z', a' \mid z, a)$. (C) Sample trajectory.

The most prominent difference between the kernel in Figure 2B and the unconstrained kernel in Figure 1A is that the columns in Figure 2B reveal not just one but four entries (with probabilities summing up to 1, since K is a Markov kernel). These entries are grouped into blocks as indicated in the figure such that all transitions for one source state (z, a) target exactly one of the blocks. The blocks are uniquely characterized by the internal states, z, z', whereas the peripheral states a, a' indicate only the precise location inside each block (cf., Figure 1B). Thus, given a source state z and a peripheral state a, the next internal target state z' is uniquely defined such that the internal state transition kernel $K'(z' \mid z, a)$ as defined in equation 2.4 is deterministic. Only the next peripheral state is random, because by assumption, it is completely governed by $K^{\partial}(a' \mid a)$ and the current peripheral state a. Thus, starting from some internal initial state, the peripheral Markov dynamic, viewed as input, drives the internal subsystem through deterministic state sequences. Nonetheless, because the dynamic on the periphery is random, sample trajectories do not reveal much determinism at a rough look (see Figure 2C).

Computer science (Hopcroft & Ullman, 1979) defines deterministic finite state automata (DFAs) as a quintuple $M = (Z, \Sigma, \delta, z_0, E)$, where $Z =$

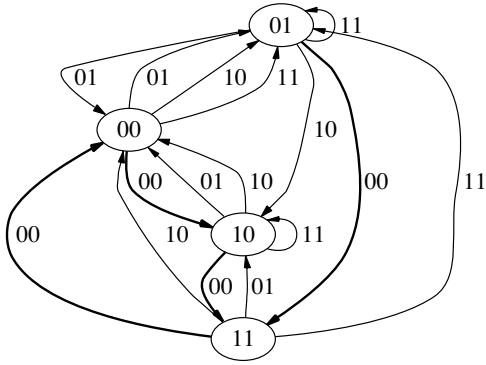

Figure 3: Deterministic finite state automaton corresponding to Figure 2. Nodes are labeled by internal states $z \in \Omega_{V\setminus\partial}$ and edges by peripheral states $a \in \Omega_\partial$. Given the current internal state z and peripheral state a, the automaton predicts the next internal, but not the next peripheral, state. Bold edges are for constant input 00.

$\{z_1, \cdots, z_n\}$ 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* δ, $Z \times \Sigma \to Z$, which maps every pair $(z, a) \in Z \times \Sigma$ to exactly one successor state. A *word* (over Σ) is any finite sequence (or string) consisting of symbols in Σ. A word x is said to be *accepted* by a DFA M iff, reading the word symbol by symbol starting from state z_0, application of the respective transition rules leads to a final state in E when the word is read completely. The set of all accepted words is a regular language.

Now observe that as a finite state automaton, the strongly interacting Markov chain in Figure 2 provides a total mapping from $\Omega_{V\setminus\partial} \times \Omega_\partial$ to $\Omega_{V\setminus\partial}$. We may therefore identify the internal state space $\Omega_{V\setminus\partial}$ with the state set Z of a DFA and the peripheral state space Ω_∂ with a set of symbols Σ. The Markov kernel $K'(z' \mid z, a)$ then corresponds to the transition table of that DFA, and the Markov chain can be represented by a labeled state transition graph as in Figure 3.

Clearly, for a complete correspondence, we would also have to designate an initial state, z_0, and accepting states, E. However, z_0 and E merely specify how a finite state automaton (FSA) actually accepts or rejects a certain input string. We could add equivalent constructs also in our Markov models by selecting initial and accepting states and presenting segmented input (finite words) (cf., e.g., Wennekers, 1998). Strongly interacting systems could in this way serve to recognize temporal patterns in the input stream. These issues are of secondary importance for this article. The main point here is that the maximization of stochastic interaction in constrained Markov chains

leads to systems characterized by deterministic, input-driven internal state transitions.

So far we have simplified things. Although most columns in the unconstrained Markov chain in Figure 1 are deterministic, the example also reveals nondeterministic transitions: transient and branching states have several possible targets. In constrained Markov chains, corresponding phenomena occur, in which case more than one block in one or more rows of the full Markov matrix reveals nonvanishing entries (simulations not shown). Transient global states are nonergodic; once left, they are never occupied again whatever the input sequence is. Corresponding columns may have an arbitrary number of filled blocks comparable to column 111 in Figure 1A. Branching states are ergodic with more than a single target state given the present global state (z, a) similar to column 010 in Figure 1A. In larger systems, a certain number of nondeterministic state transitions become increasingly likely because the number of possible state transitions grows exponentially with system size. Simulations nonetheless indicate that their relative number is always small in strongly interacting systems. Of the 2^N entries per matrix column, a fraction at most linear in N are nonvanishing. For Markov chains slightly different from those used in the simulations, this can actually be proved rigorously (cf. section 4.2). Accordingly, the internal state transitions of strongly interacting Markov chains with constrained periphery are always weakly deterministic. Asymptotically, in system size, the relative fraction of nondeterministic transitions goes to zero.

In the context of automata theory, this has the following consequence. Only completely deterministic Markov chains correspond to deterministic FSAs. However, automata theory also defines nondeterministic finite state automata (NFAs), which differ from DFAs basically by the fact that given the same state and input symbol, several successor states are possible. (As a second less important difference, they also can have more than one initial state; see Hopcroft & Ullman, 1979.) This is what we also observe in strongly interacting constrained Markov chains. Therefore, optimized constrained Markov chains correspond in general to NFAs. Nonetheless, because the relative fraction of nondeterministic transitions is small, the resulting NFAs are only weakly nondeterministic;—they are "almost" DFAs.

3.3 Impact of Periphery on Final FSAs. In section 4.1, we prove that maximizing stochastic interaction in systems of product form 2.5 is equivalent to maximizing the interaction of just K', if K^{∂} is fixed. The kernels K' for binary state variables have $d := 2^N(N - |\partial|)$ independent parameters, $K^{(\nu)}(z_\nu = 1 \mid \omega)$, $\nu \in V \setminus \partial$, $\omega \in \{0, 1\}^N$, because there are $N - |\partial|$ internal units and 2^N possible source states. K' is deterministic if all the $K^{(\nu)}(z_\nu \mid z, a)$ are either 0 or 1, such that unit ν either fires or remains silent with probability 1 given any source state (z, a). For given N and $|\partial|$, the number of deterministic parallel Markov chains therefore is $N_{DFA} := 2^d$, a number

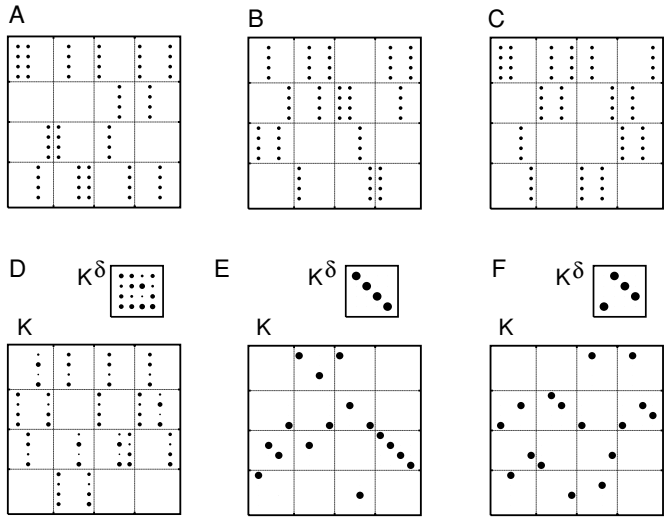

Figure 4: (A–C) Different Markov kernels $K(\omega' \mid \omega)$ resulting for $N = 4$, $|\partial| = 2$ and peripheral transitions all of equal probability as in Figure 2. (Bottom row) Optimized kernels for different types of periphery as indicated by K^{∂} in each frame. (D) Random kernel. (E) Identity. (F) Deterministic cyclic state sequence on ∂.

that grows unimaginably fast with N. For instance, for one peripheral unit $|\partial| = 1$, N_{DFA} equals 16, 65,536, and 2^{48} for $N = 2, 3, 4$, respectively.

Although not all of these deterministic Markov chains are local maximizers of $I(p, K)$, as a consequence of the huge number of possible DFAs, optimized chains appearing in simulations for given N and K^{∂} are highly nonunique. Figures 4A to 4C, for instance, display three (out of $N_{DFA} = 2^{32}$) deterministic chains optimized under the same conditions as the one in Figure 2, that is, $N = 4$, $|\partial| = 2$, and $K^{\partial}(a' \mid a) = 0.25$ for all $a, a' \in \Omega_{\partial}$. Interaction in these systems is $I(p, K) = 1.323508$, 1.35529, and 1.331308 for Figures 4A, 4B, and 4C, respectively. Apparently, all these kernels are deterministic, but the corresponding automata (not shown) are certainly not equivalent. Because different numbers of source states converge onto the internal states in these examples, there cannot be a renumbering of internal states that maps one system onto the other.

Figures 4D to 4F display comparable simulations for various choices of K^{∂}. In Figure 4D, K^{∂} has random entries in $[0, 1]$; in Figures 4E and 4F, the kernels are "deterministic," although in order to yield unique stationary probability distributions, off-diagonal elements of K^{∂} in Figure 4E (and similar in Figure 4F) were set to small, positive values. As before, individual simulation runs for each of these kernels converged to a variety of

different systems as in Figures 4A to 4C, with realized transitions scattered throughout the whole Markov matrices. Note also that the full kernels K reflect the Markov chain on the periphery. If a $K'(z' \mid z, a)$ is 1, the whole ath column of $K^{\partial}(a' \mid a)$ is copied into block z' of the (z, a)th column of K. This is a consequence of the product form of $K(z', a' \mid z, a) = K'(z' \mid z, a) K^{\partial}(a' \mid a)$ (cf. section 4.1).

The common structural features of strongly interacting Markov chains as shown in Figure 4 are that transitions are confined to just one or a few internal target states per source state (z, a) and that realized transitions are merely randomly scattered througout the blocks and columns of K. These features will be explained in section 4 on the base of mathematical properties of the interaction measure $I(p, K)$.

The question remains whether the peripheral Markov chain has an influence on the resulting FSAs at all. Figure 5 displays interaction values for all

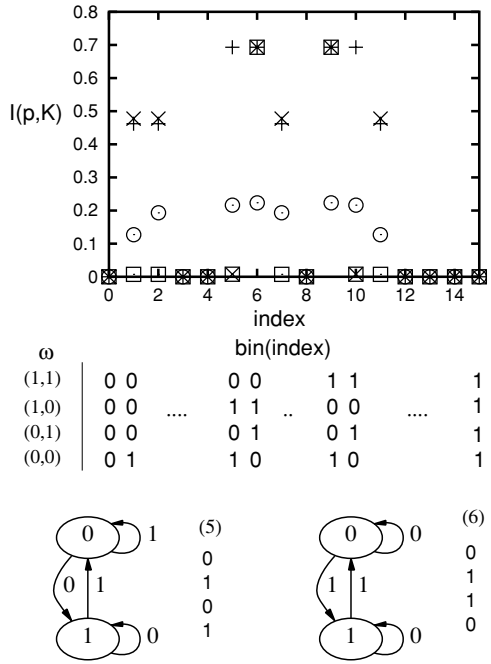

Figure 5: $I(p, K)$ for $N = 2$, $|\partial| = 1$ and all 16 possible parallel deterministic Markov kernels K'. Periphery K^{∂} has been clamped to different choices (cf. also Figure 4): Plus signs: all entries equal; crosses: identity; squares: cyclic sequence $01010\cdots$; circles: a special process (see text). At the bottom, binary representations of the deterministic Markov kernels K' are plotted—$K'(z' = 1 \mid \omega) = bin(index)[\omega]$. Two transition diagrams for Markov chains $5 = 0101$ and $6 = 0110$ are also shown.

16 possible deterministic Markov kernels K' for $N = 2, |\partial| = 1$, and different kernels K^{∂}. Kernels used for that figure did not result from an optimization but were explicitly constructed. Plus signs denote results for a peripheral kernel with all entries equal; that is, the peripheral unit is a Bernoulli process with $prob[z = 1] = 0.5$. Crosses indicate results for an identity kernel where transitions $0 \to 0$ and $1 \to 1$ are almost 1, but small off-diagonal probabilities of 0.001 ensure that the stationary probability distribution $p^{\partial} = (0.5, 0.5)$ is unique. Trajectories of the peripheral unit here consist of long sequences of either zeros or ones, where the small off-diagonal entries occasionally switch between both states. Squares in Figure 5 denote a deterministic cyclic process on the periphery—trajectories $\cdots 01010101 \cdots$. Circles in Figure 5 represent interaction values for a special peripheral process with $K^{(1)}(a' = 1 \mid a) = (0.05, 0.2)$, which is almost a Bernoulli process with rate ≈ 0.05 but an increased probability that the output stays 1 if it was so in the previous state. Because we have only one peripheral unit, the interaction of the periphery is always zero for the above choices, but the conditional entropies of the peripheries are ln 2 (i.e., as large as possible) for the Bernoulli process with rate 0.5, 0.216273 for the special peripheral process, and zero for the identity and cyclic kernels. At the bottom in Figure 5, the indexes of the DFAs used in the upper part of the figure are represented binary. The internal Markov kernels K'_i for DFA number i are chosen such that $K'_i(z' = 1 \mid \omega) = bin(i)[\omega]$ where $bin(i)[\omega]$ is the ωth position in the binary representation of i.

Quite a few of the possible DFAs in Figure 5 have zero interaction for all tested peripheries: DFAs 0, 3, 4, 8, and 12 to 15. These automata are degenerate: For arbitrary input sequences, and thus for all peripheral Markov chains, they run into attractors, where the next internal state is perfectly predictable from its current state alone, regardless of the input. This is most obvious for automata 0 and 15, which map every state (z, a) to internal state 0 or 1, respectively. Asymptotic sequences $\ldots$ 010101 $\ldots$ independent of the input as for DFA 3 have zero conditional entropies as well. The internal unit then has zero entropy, such that the interaction vanishes.

Some DFAs have high interaction for certain peripheries but zero interaction for others. Also, vanishing interaction implies that the ergodic internal state sequences are predictable without knowing the input. Consider DFAs 5 and 6 in Figure 5. For the equal-probability Bernoulli process on the periphery, they lead to Bernoulli processes on the internal state. Consequently, the interaction is large in both cases. But for the identity on the periphery, the internal unit of DFA 5 either outputs $\ldots 00000 \ldots$ or $\ldots 11111 \ldots$ after transients have died out, such that it is predictable. In contrast, for DFA 6, it can output $00000 \ldots$ and $11111 \ldots$ for input $00000 \ldots$ and $010101 \ldots$ or $10101 \ldots$ for input $1111 \ldots$ All internal transitions $z \to z'$ where $z, z' \in \{0, 1\}$ are equally likely, such that the interaction is maximal. Similarly, for the cyclic process on the periphery, DFA 5 leads to the deterministic sequence $\ldots 0101010 \ldots$ with vanishing interaction, but 6 gives maximum interaction

for outputs ... 00110011 ..., that is, all marginal one-step transition probabilities for the internal unit are 0.5, such that its entropy is ln 2.

In summary, the general form of $I(p, K)$ favors optimized Markov chains, where minimizing the conditional entropy $H(p, K)$ induces determinism, but maximizing the marginal single unit entropies leads to an "unfolding" of the chains, such that degenerate input-independent Markov chains are avoided, and in every internal state, many successor states are possible. A large number of systems can reveal these general structural features. The precise periphery selects Markov chains that still lead to unpredictable firing of internal single units for the special input sequences generated by the peripheral dynamics.

3.4 Information Flow. For sets $A, B \in V$, define the information flow from A to B as

$$F(A, B) = H(X_{B,n+1}|X_{B,n}) - H(X_{B,n+1}|X_{A,n}, X_{B,n}). \tag{3.3}$$

In equation 3.3, $H(X_{B,n+1}|X_{B,n}) =: H(B|B)$ is the conditional entropy about the next state in B given the present state in B only. $H(X_{B,n+1}|X_{A,n}, X_{B,n}) =: H(B|A, B)$ is a noiseentropy: it measures the entropy of the next state in B that is left of $H(B|B)$ if we know both the present state of B and that of A. So if we subtract it from $H(B|B)$, we get the information A shares with the next state in B. This cannot be larger than the (Shannon) information of A and must also be smaller than $|B| \ln 2$, the maximum information possible in B, if units in B are binary. Equation 3.3 is a conditional mutual information applied to successive steps of a given Markov chain—hence, the term *information flow*.

Most interesting is the case where $A = \partial$ and $B = V \setminus \partial$. The noise entropy $H(B|A, B)$ then is obviously $H(p, K')$, which is small for strongly interacting systems. $H(B|B)$ is the internal entropy if the input is not known. The flow F in that case is the average information represented in the next state about the current peripheral state. Corollary 1 in section 4.2 provides an upper bound, $H(B|A, B) \leq \log(N + 1)$, such that under the conditions of the corollary, $F(A, B) \geq H(B|B) - \log(N + 1)$. Thus, using equation 3.3, $H(B|B) - \log(N + 1) \leq F(A, B) \leq H(B|B)$, providing lower and upper bounds for F. If the internal entropy increases linearly in system size, $H(B|B) \sim \mathcal{O}(N)$, the term $\log(N + 1)$ becomes small for large N relative to $H(B|B)$, that is, $F(A, B)/H(B|B) \to 1$ for $N \to \infty$. In informal terms, most of the internally observable entropy becomes a reflection of external entropy, which flows (almost) deterministically into the system. This does not imply that repetition of the same stimulus sequence necessarily leads to equal responses. In each repetition, the response sequence depends on the precise internal initial state as well as the relatively small amount of nondeterminism.

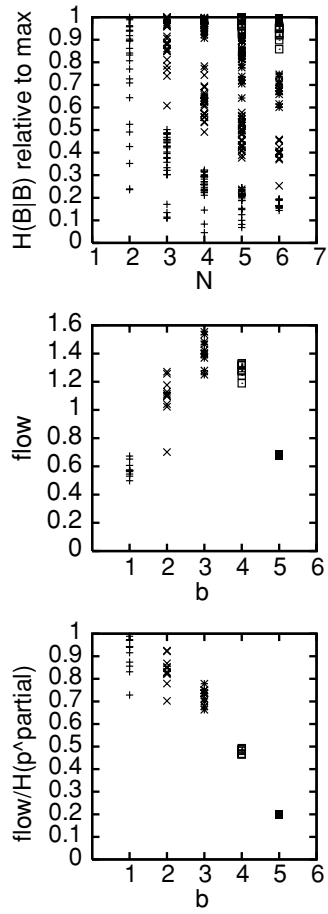

Figure 6: (Top) Internal entropies for various N (x-axis) and $b = 1, \ldots, N-1$ (different symbols) relative to maximally possible value $(N-b)\ln 2$. Each single symbol represents one optimized system with random Markov chains for K^{∂} (see text). (Middle) Flow $F(\partial, V \setminus \partial)$ for $N = 6$. (Bottom) Flow relative to Shannon entropy on periphery, $H(p^{\partial})$.

Figure 6 displays values of $H(B|B)$ and F for optimized Markov chains with N up to 6 and varying $b := |\partial|$. Markov chains on the periphery had independent and identically distributed (i.i.d.) entries equally distributed in the range $[0, 1]$. The parallel Markov chains for the internal units were also initialized with random values in the range $[0, 1]$. The systems were then optimized as before.

Figure 6 (top) shows $H(B|B)$ relative to its maximally possible value $(N-b)\ln 2$ over N. Different symbols denote different numbers of peripheral units $b = 1, \ldots, N-1$ ($+ = 1$, $\times = 2$, $* = 3$, $\square = 4$, ...). Apparently $H(B|B)$ scatters over the full range, so $H(B|B)$ is not maximized to its highest possible value $(N-b)\ln 2$ equivalent to a random walk on the internal states in B. One could expect this because a random walk on the internal states would also imply large marginal entropies. Because $H(B|A, B)$ is small in optimized systems (smaller than $\approx .2$ in most systems shown, and most often zero), $H(B|B)$ is actually similar to the flow $F(A, B)$. Figure 6 (top) therefore implies that the flow is only a small fraction of the maximally possible conditional entropy in the interior if the size of the periphery b is small compared to $N-b$. However, in the opposite case, the internal units explore most of their global state space, that is, $H(B|B)$ and the flow approach the theoretical maximum of $H(B|B)$.

Figure 6 (middle) shows the absolute flow for $n = 6$ in dependence of b. As long as $b \leq N-b$, the flow increases with b, such that an increasing amount of information on the periphery is represented in the next state of the internal units. In contrast, for $b > N-b$, the interior is smaller than the periphery and no longer able to represent the peripheral states, that is, the maximal internal entropy becomes progressively smaller than that on the periphery if b grows. Therefore, the curve in Figure 6 (middle) is unimodal with a maximum near $b = N-b$. Figure 6 (bottom) redisplays the flow relative to the Shannon entropy on the periphery. As long as $b < N-b$, the flow is a high fraction of the peripheral entropy. For high b—as Figure 6 (top) reveals more directly—it approaches the absolute information bound of the internal units, which decreases as $N-b$. This shows that maximizing $I(p, K)$ does not strictly maximize the flow from the periphery to the interior but that the flow nonetheless reaches a significant fraction of its maximally possible value, bounded by either the Shannon information on the periphery or the capacity of the internal units.

3.5 A Relation to Temporal Learning Rules. It is known that a Hebbian-type coincidence learning rule maximizes mutual information in simple feedforward neural networks for stationary input-output relationships (Hertz et al., 1991). We may therefore ask whether temporal stochastic interaction can be optimized on the base of sample trajectories of a given Markov dynamics. Note in that context that this work does not refer to concrete networks but centers on Markov chains. So there is no obvious equivalent of "synaptic" learning. Instead, we implemented the following temporal learning rule based on consecutive firing patterns in subsequent steps. For given N and periphery ∂, transitions on the periphery and the parallel kernels of the internal units were initialized with i.i.d. equi-distributed random values in $[0, 1]$. Sample trajectories of the Markov dynamics were simulated in the usual way as described in section 3.1. Update of the dynamics now did not make use of computations of $I(p, K)$ directly, but only of simulated

trajectories. If in step t the current state was $\omega = (z, a)$ and in step $t + 1$ it was $\omega' = (z', a')$, for each internal unit ν, the kernel $K^{(\nu)}$ was updated according to

$$K^{(\nu)}_{t+1}(1|\omega) = \phi\left[K^{(\nu)}_t(1|\omega) + r \cdot (2z'_\nu - 1)\right], \tag{3.4}$$

subject to normalization, $K^{(\nu)}(1|\omega) + K^{(\nu)}(0|\omega) = 1$. In equation 3.4, r is the learning rate ($r = 0.1$) and ϕ is the same clipping function as in equation 3.1, which keeps transition probabilities in the range $[0, 1]$. The rule increases componentwise the probability for the same firing pattern z', if the state ω appears again, by increasing single unit firing probabilities for units active in z' given ω, and decreasing them otherwise.

Simulations using rule-based update of the transition kernels showed that for all network sizes and peripheries, the rule indeed decreases $H(p, K)$ and increases $I(p, K)$ from their initial values. The decrease (or increase) is not always monotonic. Due to the random nature of the activity dynamic, it can slightly fluctuate, and it also reveals occasional jumps because sometimes the attractor structure reorganizes such that the number of states in the ergodic component changes (not shown). Asymptotically, I and H become virtually constant, because K' gets deterministic. Interaction and entropy approach values in a comparable range as the random search scheme, with final interaction on average somewhat smaller than compared to random search optimization. This is demonstrated in Figure 7 for networks of size $N = 6$ having two peripheral units. Resulting Markov kernels have the same structure as for random search update and are barely distinguishable from those in Figure 4. Systems that were optimized using random optimization in most cases were not further modified if we switched to rule-based updates. Such changes occurred, however, if the optimized system had branching points; rule-based update seems to avoid those. Systems

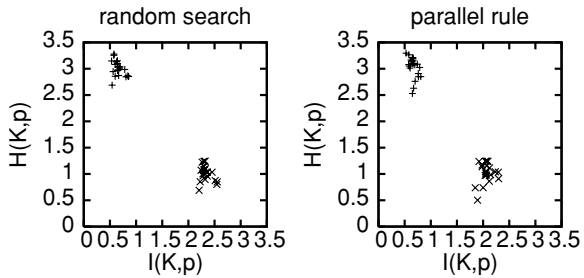

Figure 7: Entropies $H(p, K)$ and interaction $I(p, K)$ for randomly initialized Markov chains ($N = 6$, $|\partial| = 2$). + indicates values before and × after optimization using the random search scheme (left) and a temporal learning rule (right) described in section 3.5.

resulting from rule-based update were in turn also often not further optimized when switching to the random search scheme if they were already deterministic. Thus, the rule-based and random-search methods are largely consistent, though not completely equivalent. Neural systems may therefore employ a local temporal learning rule implementing effectively equation 3.4 to reach high stochastic interaction.

4 Analytical Results

In this section, we consider the optimization of strongly interacting systems mathematically. Section 4.1 explains the most prominent features of constrained optimized Markov chains on the base of analytical properties of the conditional entropy, $H(p, K)$, and the interaction measure, $I(p, K)$. Section 4.2 proves upper bounds for the number of nonvanishing transitions and the maximum entropy of optimized Markov chains.

4.1 Properties of H and I. We start with some instructive calculations assuming a Markov kernel of product form,

$$K(\omega' \mid \omega) = K(z', a' \mid z, a) = K'(z' \mid z, a) K^{\partial}(a' \mid a), \tag{4.1}$$

where $\omega = (z, a)$, $\omega' = (z', a') \in \Omega_V$, a, $a' \in \Omega_\partial$, and z, $z' \in \Omega_{V \setminus \partial}$ as usual. Then the conditional entropy of (p, K) can be written as

$$\begin{aligned}
H(p, K) &= -\sum_{\omega,\, \omega'} p(\omega) K(\omega' \mid \omega) \ln K(\omega' \mid \omega) && (4.2)\\
&= -\sum_{z',\, \underline{a'},\, z,\, a} p(z, a) K'(z' \mid z, a) K^{\partial}(\underline{a'} \mid a) \ln K'(z' \mid z, a) && (4.3)\\
&\quad -\sum_{\underline{z'},\, a',\, z,\, a} p(z, a) K'(\underline{z'} \mid z, a) K^{\partial}(a' \mid a) \ln K^{\partial}(a' \mid a) && (4.4)\\
&= -\sum_{z',\, z,\, a} p(z, a) K'(z' \mid z, a) \ln K'(z' \mid z, a) && (4.5)\\
&\quad -\sum_{a',\, a} \underbrace{\left(\sum_{z} p(z, a) \right)}_{= p^{\partial}(a)} K^{\partial}(a' \mid a) \ln K^{\partial}(a' \mid a) && (4.6)\\
&= H(p, K') + H(p^{\partial}, K^{\partial}). && (4.7)
\end{aligned}$$

In equations 4.5 and 4.6, we used the normalization of the Markov kernels K^{∂} and K' in their first argument. In equation 4.6, $\sum_z p(z, a) = p^{\partial}(a)$, where p^{∂} is the probability distribution on the periphery. p^{∂} is stationary, $K^{\partial} p^{\partial} = p^{\partial}$, because we maximize with respect to stationary distributions

p. Equation 4.7 shows that the total kernel entropy can be written as a sum of the conditional entropy of the periphery and that of the transition (p, K'). Because the entropy of the periphery is a fixed quantity, optimization of the interaction $I(p, K)$ can only influence $H(p, K')$ in equation 4.7.

With equation 4.7, the interaction measure reads

$$I(p, K) = \sum_{\nu \in V} H_\nu(p_\nu, K_\nu) - H(p, K) \tag{4.8}$$

$$= \sum_{\nu \in V \setminus \partial} H_\nu(p_\nu, K_\nu) - H(p, K') + \sum_{\nu \in \partial} H_\nu(p_\nu, K_\nu) - H(p^\partial, K^\partial) \tag{4.9}$$

$$= I(p, K') + I(p^\partial, K^\partial). \tag{4.10}$$

Equation 4.10 reveals also that the interaction $I(p, K)$ can be written as a sum of the interaction of the periphery and that of the Markov transition (p, K'). Again, $I(p^\partial, K^\partial)$ is constant during optimization, such that the maximization of $I(p, K)$ is actually equivalent to the maximization of $I(p, K') = \sum_{\nu \in V \setminus \partial} H_\nu(p_\nu, K_\nu) - H(p, K')$. Thus, for a given periphery, the internal units should maximize their marginal entropies, but global transitions $(z, a) \to z'$ should become as deterministic as possible. Consider $H(p, K')$:

$$H(p, K') = -\sum_{z',\, z,\, a} p(z, a) K'(z' \mid z, a) \ln K'(z' \mid z, a) \tag{4.11}$$

$$= \sum_{z,\, a} p(z, a) \underbrace{\left(-\sum_{z'} K'(z' \mid z, a) \ln K'(z' \mid z, a) \right)}_{= H(K'(\cdot \mid z, a))}. \tag{4.12}$$

The underbraced term in equation 4.12 is the Shannon entropy of the internal state transitions induced by K' restricted to the fixed source state (z, a). $H(p, K')$ then is the weighted average over these entropies, where the weights are the probabilities that the system is indeed in state (z, a) before the transition. To maximize $I(p, K')$, we should make $H(p, K')$ small. Non ergodic states have $p(z, a) = 0$, so their conditional entropy does not count in equation 4.12. Accordingly, they can have an arbitrary number of internal target states, with any distribution of transition probabilities $K'(\cdot|z, a)$. In contrast, for ergodic states (z, a) with $p(z, a) > 0$, equation 4.12 requires minimizing $H(K'(\cdot|z, a))$, but $H(K'(\cdot|z, a))$ is zero if $K'(\cdot \mid z, a)$ is deterministic, that is, if there is only a single target state z' with $K'(z' \mid z, a) = 1$. Thus, if all $K'(\cdot \mid z, a)$, $z \in \Omega_{V \setminus \partial}$, $a \in \Omega_\partial$ are deterministic, the total entropy $H(p, K')$ obtains its absolute minimum of 0, such that every global state has a single internal successor. $K'(z' \mid z, a)$ can then be interpreted as the transition table of a deterministic FSA.

Note, however, that we used parallel Markov kernels for the internal states in the simulations, which constrain $K'(z' \mid z, a)$ to be of the form

$$K'(z' \mid z, a) = \prod_{\nu \in V \setminus \partial} K^{(\nu)}(z'_\nu \mid z, a). \tag{4.13}$$

For any $K'(z'|z, a)$ to be deterministic, it suffices that all $K^{(\nu)}(z'_\nu \mid z, a)$ are either 0 or 1. Under this condition, the $K^{(\nu)}(z'_\nu | z, a)$ in equation 4.13 can be considered binary variables for fixed (z, a) and $z'_\nu \in \{0, 1\}$, $\nu \in V \setminus \partial$. Because a product of binary variables is equivalent to logical AND-ing these variables, for each pair (z, a) there exists exactly one z' with $K'(z' \mid z, a) = 1$. On the other hand, if for a given (z, a) one of the $K^{(\nu)}(z'_\nu = 1 | z, a)$ is not zero or one, there must be at least two possible target states z'_0, z'_1, where unit ν is active in z'_1 and inactive in z'_0 and the other units have the same values in both states. Accordingly, $H(p, K')$ is zero for parallel kernels iff they consist of only transitions with probability 0 or 1 for ergodic states. Maximizing $I(p, K)$ should therefore try to force entries in the parallel internal kernels $K^{(\nu)}$ toward 0 or 1, which implies the same for K'. As a consequence, the product form of K in equation 4.1 then implies for the full kernels that the ath column inside block z, z' is just a copy of the ath column of K^∂. This has been observed in section 3.3.

Maximizing $I(p, K)$ requires not only $H(p, K')$ to be small, but in addition that the marginal entropies $H_\nu(p_\nu, K_\nu)$, $\nu \in V \setminus \partial$ are large. Without this second condition, all deterministic Markov chains would be maximizers, including many degenerate chains like those that map every state (z, a) to a constant target state z'. The marginal entropies of those units would be zero and the interaction of K' therefore vanishing. In section 3.3, we have given further examples for degenerate chains and showed that Markov chains with high interaction are to some degree selective relative to the special periphery (see Figure 5). In fact, if for each state (z, a), the next internal state is random, the sequences of bits of an internal unit will likely be unpredictable too. This explains the random scattering of filled blocks in Figure 4. For specific peripheries, some states and transitions appear more often than others. For the frequent peripheral states, the corresponding columns in the full Markov matrix should be as diverse as possible. This makes the optimized systems selective with respect to the clamped peripheral chains.

If we would maximize the internal marginal entropies alone, the optimum would be independent Bernoulli processes with probability of firing equal to 0.5. These could then be coupled to a more or less degree to the periphery; that is, there could be some more or less high information flow from the peripheral units to the interior. However, by simultaneously minimizing $H(p, K')$, we enforce a low noise entropy and thereby usually an increased coupling between periphery and interior. If instead of the marginal entropies alone we would maximize just $H(B|B)$, where B is the interior

$V \setminus \partial$, a random walk on the internal states z would be the optimum, again depending to an unconstrained degree on states on the periphery (which are marginalized out when going from the full kernel to that of B). Maximizing the flow $F(\partial, V \setminus \partial) = H(V \setminus \partial | V \setminus \partial) - H(p, K')$ would then further reduce the noise entropy, such that the internal transitions become confined to an entropy bounded by the entropy on the periphery (or the capacity of $V \setminus \partial$ itself). Apparently, a high internal entropy would also lead to high internal marginal entropies, because if the state sequence is random, so would be the single bits. Therefore, one expects that maximizing the flow would also lead to high interaction values. In section 3.4, we have shown the opposite. However, if the internal marginal entropies are high—the single units are unpredictable—so too would likely be the full internal states, the more because in the simulations, we started from random initial conditions for the internal transitions and updated randomly selected entries in the kernels in each step. Therefore, the full optimized Markov kernels reveal entries in randomly scattered blocks, which in turn lead to random sequences of states when driven by the periphery. From that point of view, it is plausible that optimizing interaction also leads to high values for the flow into the systems.

Finally, consider the rule-based update in section 3.5. Transitions that appear repeatedly, due to either the initialization that favors them or by chance in a sample realization, will be enforced until they become deterministic. Thus, the rule decreases $H(p, K')$. On the other hand, which transitions become deterministic is largely random, though constrained by the periphery. Therefore, internal states and single unit sequences stay largely random under the update scheme, corresponding to a high interaction and flow. In fact, in the simulations, $\sum_{v \in V \setminus \partial} H_v$ started from high values due to the random initialization, but decreased in the majority of cases only moderately if the Markov kernels got confined to deterministic global transitions. The final change in $H(p, K')$ outperformed the slight loss in the marginal entropies (simulations not shown) such that $I(p, K)$ increased (cf. Figure 7).

The above arguments explain informally the special Markov chains observed in the simulations: first, the convergence toward almost deterministic systems, because those maximize $H(p, K')$, such that the chains become automata-like, and second, the fact that realized internal transitions are apparently randomly scattered throughout the optimized Markov kernels. This increases the number of possible pathways in the state transition graphs, and therefore the unpredictability of activity of individual units $H_v(p_v, K_v)$ as well as the internal entropy. Consequently, maximizing stochastic interaction also increases information flow into the system. Rule-based update decreases $H(p, K')$ but keeps internal state sequences random, such that it behaves similarly rather than directly optimizing $I(p, K)$. The peripheral Markov chain has the further influence of weighting some of the possible pathways in the state transition graphs more strongly than others, according to the probabilities with which input

sequences appear. This makes some of the possible chains more likely, and others less, but still leaves many locally optimal solutions with the described features.

4.2 Theorems. In this section, we state theorems showing that in fact all strongly interacting systems are weakly nondeterministic. As defined earlier, a *strongly interacting system* is a local maximizer of I. We consider a Markov transition with constrained periphery as *weakly nondeterministic*, if for every global state ω with $p(\omega) > 0$ the number of possible internal target states is bounded by $\eta(V/\partial) + 1$, where for a subset $A \subset V$ we define

$$\eta(A) := \sum_{v \in A} \big(|\Omega_v| - 1\big).$$

For binary neurons, that is, $|\Omega_v| = 2$ for all $v \in V$, one has $\eta(A) = |A|$, such that weak nondeterminism imposes a bound linear in the number of internal units on the number of possible internal target states of the Markov kernel given any fixed global state in the ergodic component ($p(\omega) > 0$). For strictly deterministic systems, this number is exactly 1. General Markov chains, to the other extreme, can have exponentially many target states. In fact, for the unconstrained optimization of temporal interaction, $\partial = \emptyset$, we already proved the following theorem (Ay & Wennekers, 2003).

Theorem 1. *Consider a probability distribution $p \in \bar{\mathcal{P}}(\Omega_V)$ and a transition kernel $K \in \bar{\mathcal{K}}(\Omega_V)$. If (p, K) is a local maximizer of I, then for all $\omega \in \operatorname{supp} p$ the following bound on the support of $K(\cdot \mid \omega)$ holds:*

$$|supp\, K(\cdot\,|\omega)| \leq 1 + \eta(V). \tag{4.14}$$

For binary units, the estimate 4.14 implies the linear bound $|\operatorname{supp} K(\cdot \mid \omega)| \leq 1 + |V|$. Instead of the unconstrained optimization, we now consider a driven system. Let ∂ be a subset of the set V of neurons, the periphery of the system. We assume that the process on the periphery is given by the environment of the system, which is fixed. We model this extrinsic process by a probability distribution $p^\partial \in \bar{\mathcal{P}}(\Omega_\partial)$ and a transition kernel $K^\partial \in \bar{\mathcal{K}}(\Omega_\partial)$. As already considered in equation 2.4, the intrinsic information processing is modeled by a transition kernel K' from Ω_V to $\Omega_{V\setminus\partial}$. Thus, we investigate the optimization of I restricted to the set of transition kernels K from Ω_V to Ω_V that have the product structure

$$K(z', a' \mid z, a) \;=\; K'(z' \mid z, a)\, K^\partial(a' \mid a). \tag{4.15}$$

This constrained optimization leads to the following generalization of theorem 1.

Theorem 2. *Let (p, K) be a local maximizer of the restriction of I to the set of transition kernels with product structure 4.15 and a fixed peripheral transition kernel K^{∂}, and let $\omega = (z, a)$ be an element of* supp p. *Then for the intrinsic kernel K' of K, we have*

$$|supp\, K'(z' \mid z, a)| \leq 1 + \eta(V \setminus \partial). \tag{4.16}$$

Note that we recover the estimate 4.14 if we set $\partial := \emptyset$ in equation 4.16. Then, formally, $K^{\emptyset}$ maps the empty state ϵ onto the empty state, such that $K' = K$. Theorem 2 implies the following corollary on the entropy generated by a strongly interacting system:

Corollary 1. *In the situation of theorem 2, the conditional entropy of the next internal state given the current global state satisfies*

$$H_{(p,K)}(X'_{V\setminus\partial} \mid X) \;\leq\; \ln\big(1 + \eta(V \setminus \partial)\big).$$

The proofs of theorem 2 and corollary 1 are given in the appendix. Informally, they imply that all strongly interacting systems of the form 4.15, where information flows only into the system from some periphery, must be weakly nondeterministic. Given any internal state and input, the number of internal target states is small (linear in system size) as compared to the possible number of states (exponential). Accordingly, the internally generated entropy grows at most logarithmically in system size ($|V \setminus \partial|$).

5 Summary and Discussion

We have defined a measure of stochastic interaction including spatial and temporal properties of stochastic processes as the divergence of a Markov chain from its product of marginal chains. We have shown numerically and analytically that the optimization of stochastic interaction in Markov chains with clamped periphery leads to deterministic, or at most weakly nondeterministic, FSAs. This can be envisaged as a consequence of locally maximizing conditional mutual information 2.6, which makes all single unit states in the next step (near) deterministic given the current global state. In such systems, the dynamics prescribed on a set of input units drives the internal dynamics through (almost) deterministic state transitions. Nonetheless, the internal single unit activities in strongly interacting systems are largely unpredictable given their current activity alone. These features are explained by the property of stochastic interaction to combine two goals. On one hand, it minimizes the conditional entropy for global state transitions, but simultaneously it maximizes the single unit entropies. As a consequence, the resulting internal Markov chains are confined from arbitrary Markov kernels toward deterministic kernels, but they unfold from degenerate chains,

such that in every internal state, as many different target states as possible can be approached in dependence of the present input activity. This way, the recurrent internal dynamics of strongly interacting systems reveals complex internal structure, in contrast to pure feedforward networks.

From a dynamical systems viewpoint, strongly interacting systems can be seen as driven or nonautonomous systems with rich internal dynamics. If the input is held constant for some time, activity flows into attractors specific for the particular input, (see Figure 3, where bold lines are for fixed input 00). Similar attractor structures can be constructed for inputs 01, 10, and 11. They correspond directly to the ergodic nested-loop attractors for unconstrained optimization (cf. Figure 1). Accordingly, peripheral activity constant over a certain time can select intrinsic modes of activity, and peripheral state transitions can further switch between such internal dynamic modes. This provides the simplest way of neural computations, because information about the history of the system can be represented and processed. In other words, the optimized DFAs represent spatiotemporal features of the input signals. Interestingly, some experimental evidence indeed suggests the existence of brief intrinsic modes or states in cortical neural activity. As Abeles et al. (1995) have demonstrated, cortical activity in prefrontal areas of monkeys flips among quasi-stationary states of several ten to hundred milliseconds duration defined by short-time firing rate patterns of simultaneously recorded neurons. Similar phenomena appear in our network for slowly varying input patterns, if the intrinsic dynamics is forced into different modes over time. Gat, Tishby, and Abeles (1997) have further shown that the state flips can be well segmented by hidden Markov models, suggesting that intrinsic modes can be switched on a fast timescale, although they persist themselves on a longer scale. It would be interesting to determine the stochastic interaction comprised by these experimentally observed Markovian systems and compare it with complete randomness or order under various behavioral conditions.

Aertsen, Gerstein, Habib, and Palm (1989) defined *functional connectivity* of a set of neurons with reference to short-time correlations in their mutual firing patterns. In experimental data, these correlations were shown to change rapidly over time, with the interpretation that neurons dynamically form varying subgroups of interacting cells, also referred to as *functional cell assemblies*. Interestingly, in our optimized systems, correlations would change similarly if the network activity is driven through different intrinsic modes, but they are constant, determined by the transiently approached attractor, in each particular mode. In this way, our approach may provide an explanatory base for the complex correlation dynamics found in experiments.

A further aspect of cortical neural activity seems important at that point. This is the presence of repetiting firing patterns with interspike intervals up to the order of tens to hundreds of milliseconds. Abeles, Bergman, et al. (1993a) have shown that such long-lasting synfire patterns appear

reliable and behavior dependent in multiple electrode recordings of monkeys performing simple behavioral tasks. In the light of the largely stochastic firing of single units, this observation is highly surprising (Abeles, 1991). The classical synfire chain model explains these long-time correlations by volleys of synchronized activity that propagate repeatedly along the same (i.e., deterministic) neural pathways (Abeles, 1991; Abeles, Vaadia, et al., 1993b). Activity in our optimized Markov chains reveals quite similar properties: Single-unit activity is virtually random, but the state transitions are largely deterministic and proceed along nested repetitive loops. So a network dynamic can be globally deterministic even if every single neuron's activity looks virtually random. In fact, on the background of neural assembly and associative memory theories in Wennekers (1998), we have demonstrated that the classical synfire chain model can be extended in a simple and straightforward way to implement arbitrary deterministic and nondeterministic FSAs. Wennekers and Ay (2003) furthermore argue that synfire chain–type activation patterns appear naturally under the assumption that the brain maximizes temporal interaction. Attractor models of brain function, on the other hand, reveal only small stochastic interaction.

We further mention a relation of our work to a series of papers by Tononi, Sporns, and Edelman (Sporns et al., 2000; Tononi et al., 1994; Tononi, Sporns, & Edelman, 1999). They considered the segregation and integration of neurons into functional ensembles based on several different measures for complexity: Shannon entropy, spatial interaction (also termed *integration* in Tononi et al.'s work), and two further measures that account for information flow between partitions of a set of units. The "integration" measure is equivalent to our stochastic interaction restricted to stationary probability distributions. Tononi et al. (1994, 1999) and Sporns et al. (2000) compared structural features of systems that optimize one or the other complexity measure. As a main result, they found that in particular their partition-based measures lead to networks with distinct structural characteristics such as clustered connectivity and a short wiring length (cf. also Murre & Sturdy, 1995, for an interesting complementary approach), whence the neurons organize into mutually segregated subgroups, with strong internal interactions. The spatial "integration" measure, however, usually leads to systems where most cells are bound into a single strongly interacting cluster. An important difference between our work and that of Tononi et al. (1994, 1999) and Sporns et al. (2000) is that stochastic interaction as used in this work is based on spatial *and* temporal interactions. Our example systems therefore show rich internal and input-dependent dynamics and are better described in space-time—for example, in terms of the intrinsic modes for constant input—rather than in space alone. We leave the mathematical conceptualizations of these issues to future work.

Our principle of temporal information maximization complements Linkser's Infomax principle for stationary input-output relations in layered

feedforward systems. Linkser's work (1986a–1986b) pointed out a surprising link between two previously unrelated and even distant areas of research: Information maximization and the structure of visual receptive fields. The principle of Temporal Infomax as developed in this work presents a reasonable extension of Linsker's classical Infomax to the spatiotemporal domain. Again, it turns out here also that an information-theoretic maximization principle can suddenly be linked to a previously completely unrelated area of research: the theory of computing machines. The possibility of grounding the development of computational structures in neural systems on information-theoretic optimization principles seems as appealing as Linsker's observation that such principles may guide the organization of sensory hierarchies. Both principles, of course, need further theoretical and experimental evaluation.

Appendix: Proofs

Now we come to the proof of theorem 2. It is based on the following lemma, which we proved in Ay and Wennekers (2003):

Lemma 1. *Let $\bar{\Delta}$ be a d-dimensional closed simplex in a real vector space and ext its set of extreme points. For each subset $ext' \subset ext$, $\Delta(ext')$ denotes the open face of $\bar{\Delta}$ with the extreme points ext', and we have the stratification*

$$\bar{\Delta} = \biguplus_{\emptyset \neq ext' \subset ext} \Delta(ext').$$

For a point $x \in \bar{\Delta}$, $\operatorname{supp} x$ is the subset of ext defined by $x \in \Delta(\operatorname{supp} x)$. Now 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

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

Proof of Theorem 2. We fix the local maximizer (p, K) of the interaction I^V, an $\omega \in \operatorname{supp} p$, and define the simplex

$$\begin{aligned}\bar{\Delta} &:= \bar{\Delta}(p, K^\partial, K', \omega) \\ &:= \{(p, K^\partial, L') \in \bar{\mathcal{P}}(\Omega_V) \times \bar{\mathcal{K}}(\Omega_\partial) \times \bar{\mathcal{K}}(\Omega_{V\setminus\partial} \mid \Omega_V) :\end{aligned}$$

$$L'(\cdot \mid \sigma) = K'(\cdot \mid \sigma) \text{ for all } \sigma \in \Omega_V, \sigma \neq \omega\}$$
$$\subset \mathbb{R}^{\Omega_V} \times \mathbb{R}^{\Omega_\partial \times \Omega_\partial} \times \mathbb{R}^{\Omega_V \times \Omega_{V \setminus \partial}}.$$

This set can naturally be identified with $\bar{\mathcal{P}}(\Omega_{V \setminus \partial})$ by the map $\bar{\Delta} \to \bar{\mathcal{P}}(\Omega_{V \setminus \partial})$, $(p, K^\partial, L') \mapsto L'(\cdot \mid \omega)$. Now we define the convex subset,

$$\mathcal{C} := \{(p, K^\partial, L') \in \bar{\Delta} \ : \quad L'_v = K'_v \ \text{ for all } v \in V \setminus \partial\},$$

which can be represented as the intersection of $\bar{\Delta}$ with an affine subspace of $\mathbb{R}^{\Omega_V} \times \mathbb{R}^{\Omega_\partial \times \Omega_\partial} \times \mathbb{R}^{\Omega_V \times \Omega_{V \setminus \partial}}$ that is given by $\eta(V \setminus \partial)$ equations. In order to apply lemma 1, we have to prove that the interaction I^V is strictly convex on $\mathcal{C}$. This part of the proof follows exactly the lines in Ay and Wennekers (2003) for the unconstrained case and is therefore not repeated here. Lemma 1 then implies

$$|\text{supp } K'(\cdot \mid \omega)| \ \leq \ 1 + \eta(V \setminus \partial).$$

Proof of Corollary 1. This follows directly from theorem 2.

Acknowledgments

N. A. thanks the Santa Fe Institute for hosting him during the final work on this letter. We also thank two anonymous referees for their valuable comments, which helped to improve the letter significantly.

References

Abbott, L., & Sejnowski, T. J. (Eds.). (1999). *Neural codes and distributed representations*. Cambridge, MA: MIT Press.

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

Abeles, M., Bergman, H., Gat, I., Meilijson, I., Seidemann, E., Tishby, N., & Vaadia, E. (1995). Cortical activity flips among quasi stationary states. *Proc. Natl. Acad. Sci. (USA)*, *92*, 8616–8620.

Abeles, M., Bergman, H., Margalit, E., & Vaadia, E. (1993a). Spatio-temporal firing patterns in frontal cortex of behaving monkeys. *J. Neurophysiol.*, *70*, 1629–1643.

Abeles, M., Vaadia, E., Bergman, H., Prut, Y., Headman, I., & Slovin, H. (1993b). Dynamics of neuronal interactions in the frontal cortex of behaving monkeys. *Concepts in Neuroscience*, *4*, 131–158.

Aertsen, A. (Ed.). (1993). *Brain theory*. Amsterdam: Elsevier.

Aertsen, A. M. H. J., Gerstein, G. L., Habib, M. K., & Palm, G. (1989). Dynamics of neuronal firing correlation: Modulation of "effective connectivity." *J. Neurophysiol*, *61*, 900–917.

Amari, S. (2001). Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, *47*, 1701–1711.

Ay, N. (2001). *Information geometry on complexity and stochastic interaction* (Preprint 95/2001). Leipzig: Max Planck Institute for Mathematics in the Sciences.

Ay, N. (2002). Locality of global stochastic interaction in directed acyclic networks. *Neural Computation*, *14*(12), 2959–2980.

Ay, N., & Wennekers, T. (2003). Dynamical properties of strongly interacting Markov chains. *Neural Networks*, *16*, 1483–1497.

Barlow, H. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems*, *12*, 241–253.

Bell, A. J., & Parra, L. C. (2005). Maximizing sensitivity in a spiking network. In L. K. Saul, Y. Weiss, & l. Bottou (Eds.), *Advances in neural information processing systems*, *17*. Cambridge MA: MIT Press.

Bell, A. J., & Sejnowski, T. J. (1995). An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*, 1129–1159.

Bi, G., & Poo, M. (2001). Synaptic modification of correlated activity: Hebb's postulate revisited. *Annual Review of Neuroscience*, *24*, 139–166.

Chechik, G. (2003). Spike-timing-dependent plasticity and relevant mutual information maximization. *Neural Computation*, *15*, 1481–1510.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.

Cutler, A., & Breiman, L. (1994). Archetypical analysis. *Technometrics*, *36*, 338–347.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience*. Cambridge, MA: MIT Press.

Eckhorn, R. (1999). Neural mechanisms of scene segmentation: Recordings from the visual cortex suggest basic circuits for linking field models. *IEEE Transactions on Neural Networks*, *10*, 464–479.

Fairhall, A. L., Lewen, G. D., Bialek, W., & de Ruyter van Steveninck, R. (2001). Efficiency and ambiguity in an adaptive neural code. *Nature*, *412*, 787–792.

Froemke, R., & Dan, Y. (2002). Spike-timing depending plasticity induced by natural spike trains. *Nature*, *416*, 433–438.

Gat, I., Tishby, N., & Abeles, M. (1997). Hidden Markov modelling of simultaneously recorded cells in the associative cortex of behaving monkeys. *Network—Computation in Neural Systems*, *8*, 297–322.

Gerstein, G. L., Bedenbaugh, P., & Aertsen, A. M. H. J. (1989). Neuronal assemblies. *IEEE Transactions on Biomedical Engineering*, *36*, 4–14.

Gütig, R. G., Aharonov, R., Rotter, S., & Sompolinsky, H. (2003) Learning input correlations through non-linear temporally asymmetric Hebbian plasticity. *Journal of Neuroscience*, *23*, 3697–3714.

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*. Reading, MA: Addison-Wesley.

Hopcroft, J. E., & Ullman, J. D. (1979). *Introduction to automata theory, languages, and computation*. Reading, MA: Addison-Wesley.

Lee, T. W., Girolami, M., Bell, A. J., & Sejnowski, T. J. (2000). A unifying information-theoretic framework for independent component analysis. *Comput. Math. Appl.*, *39*, 1–21.

Li, Z., & Arick, J. J. (1994). Toward a theory of the striate cortex. *Neural Computation*, *6*, 127–146.

Linsker, R. (1986a). From basic network principles to neural architecture: Emergence of spatial-opponent cells. *Proceedings of the National Academy of Sciences (USA)*, *83*, 7508–7512.

Linsker, R. (1986b). From basic network principles to neural architecture: Emergence of orientation-selective cells. *Proceedings of the National Academy of Sciences (USA)*, *83*, 8390–8394.

Linsker, R. (1986c). From basic network principles to neural architecture: Emergence of orientation columns. *Proceedings of the National Academy of Sciences (USA)*, *83*, 8779–8783.

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.

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.

Murre, J. M. J., & Sturdy, D. P. F. (1995). The connectivity of the brain: Multi-level quantitative analysis. *Biological Cybernetics*, *73*, 529–545.

Nakahara, H., & Amari, S. (2002). Information geometric measure for neural spike trains. *Neural Comput.*, *14*, 2269–2316.

Palm, G., & Aertsen, A. (Eds.). (1986). *Brain theory*. Berlin: Springer.

Penev, P. S., & Atick, J. J. (1996). Local feature analysis: A general statistical theory for object representation. *Network: Computation in Neural Systems*, *7*, 477–500.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., & Bialek W. (1998). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*, 1193–1216.

Singer, W., & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypotheses. *Annual Review of Neuroscience*, *18*, 555–586.

Sporns, O., Tononi, G., & Edelman, G. M. (2000). Connectivity and complexity: The relationship between neuroanatomy and brain dynamics. *Neural Networks*, *13*, 909–922.

Studený, M., & Vejnarova, J. (1998). The multiinformation function as a tool for measuring stochastic dependence. In M. I. Jordan (Ed.). (1998). *Learning in graphical models*. Dordrecht: Kluwer, 1998.

Tononi, G., Sporns, O., & Edelman, G. M. (1994). A measure for brain complexity: Relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci. (USA)*, 91, 5033–5037.

Tononi, G., Sporns, O., & Edelman, G. M. (1999). Measures of redundancy and degeneracy in biological networks. *Proc. Natl. Acad. Sci. (USA)*, *96*, 3257–3262.

Vinje, W. E., & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, *287*, 1273–1276.

Wennekers, T. (1998). *Synfire graphs: From spike patterns to automata of spiking neurons* (Tech. Rep. N. 98-08). Ulm: Faculty for Computer Science, University of Ulm.

Wennekers, T., & Ay, N. (2003). Spatial and temporal stochastic interaction in neuronal assemblies. *Theory Biosci.*, *122*, 5–18.

Wennekers, T., Sommer, F., & Aertsen, A. (Eds.). (2003). Neural assemblies. Special Issue of *Theory in Biosciences*, *112*, 1–104.

Received January 21, 2004; accepted March 9, 2005.