Maximizing local information transfer in Boolean networks

T Haruna and K Nakajima

Abstract
We study a Boolean network model such that rules governing the time evolution of states are not given a priori but emerge from the maximization process of local information transfer and are stabilized if possible. We mathematically derive the class of rules that can be stabilized. With the presence of small noise, those stabilized are such that their output depends on a unique input. We confirm the prediction of the theory by numerical simulation. We argue that the stabilized rules have generic properties of real-world gene regulatory networks, being both critical and highly canalized.

1. Introduction
Quantifying information flow between agents in a complex system is an important step in understanding its structure and behavior. For this purpose, information-theoretic quantities such as mutual information [1] and transfer entropy [2] have been applied to analyzing living, cognitive, economical, and social systems [3]. The relation between the optimal information processing capability of a system and its functional advantages has also been studied in model living and cognitive systems [4, 5].

The relation between the dynamics of a system and the averaged information-theoretic quantities over the whole system has been investigated in the literature. In particular, many authors have asked whether information-theoretic quantities are maximized at criticality. For example, Luque and Ferrera [6] showed that the averaged mutual information between the past and future states of each agent in a random Boolean network is maximized at criticality, using mean field theory. Ribeiro et al [7] proved that the pairwise mutual information averaged over all the pairs of agents in a random Boolean network is maximized at criticality in the thermodynamic limit. However, Barnett et al [8] showed that the global transfer entropy measuring the average information transfer from all agents in the system to an agent has a peak in the disordered phase in a two-dimensional Ising model. Lizier et al [9] found that the balance between information transfer and information storage is optimized close to criticality, by numerical simulation. Lizier et al [10–12] further proposed a framework to localize information-theoretic measures at each spatio-temporal point in the dynamics of complex systems. In particular, they revealed that traveling coherent structures in the spatio-temporal dynamics of cellular automata carry most of the transmitted information by using local transfer entropy [10].

In this paper, we ask the converse question: What kind of dynamics will emerge when local information transfer to an agent in a system is maximized at each time step? We formulate this question in Boolean networks, which are simple models of gene regulatory networks [13, 14]. Such models have also been used as a test bed for discussing the relation between information and dynamics, as we have reviewed above. One of the authors recently proposed an adaptive Boolean network model in which rewiring of links based on local information transfer leads to near critical dynamics [15]. The basic idea is to use the fact that local information-theoretic quantities can change their sign. Each agent attempts to balance the stability of information processing (1) by eliminating links with negative local mutual information, which carry misleading information transfer [10] and (2) by acquiring new links randomly when there are no such links. However, in this model, the ensemble of rules is fixed and the dynamical property of the whole system is varied via changes of links across a longer time scale.
than that of state dynamics. Here, we consider a Boolean network model in which the rules themselves can change in the same time scale as that of state dynamics.

Some authors have used transfer entropy as a fitness function for the evolution of systems’ dynamics. Lizier et al [16] found that coherent traveling waves of information transfer are observed when motion of snake-like modular robots is evolved to maximize the average transfer entropy along components from the tail to the head. Obst et al [17] showed that the learning performance of recurrent neural networks is improved when the parameters of their dynamics are adjusted so that the transfer entropy is optimized at the individual unit level. In contrast with these previous works, local information transfer directly acts on the state dynamics in our approach.

It may appear that the question we pose here is similar to the one discussed by Informax in artificial neural networks [18]. Informax seeks the optimal function between its input and output by maximizing the average mutual information subject to given external constraints. In contrast with that, the local information transfer maximization (LITM) proposed in this paper has no external constraints. It turns out that this latter construction admits a Bayesian interpretation, with a likelihood function depending on the state dynamics that can be seen as an internal constraint.

This paper is organized as follows. In section 2, we present the Boolean network model in which the rule of each agent at each time step is determined by LITM. In section 3, we mathematically derive the stabilized rules and confirm them by numerical simulation. In section 4, we discuss properties of the stabilized rules and links to real-world gene regulatory networks. Finally, in section 5, we summarize the result of this paper and indicate its relation to existing approaches.

2. Model

Let us consider $N$ agents that can each take values 0 and 1 for their state. Each agent $i$ receives $K$ inputs from other agents $j_1, j_2, \ldots, j_K$, which are chosen randomly. The state of the agent evolves according to these inputs. Unlike in usual Boolean networks [19], here the time evolution rule of state for each agent is not given a priori. Rather, we hypothesize that the change of state is based on the ability of an agent to maximize local information transfer toward it from its inputs. Thus, local information transfer has an influence on the change of state in an optimal way in our model. Let $x_{i,t}$ be the state of agent $i$ at time step $t$ and let $x_{i:j_k,t} = (x_{j_1,t}, \ldots, x_{j_K,t})$. Then, $x_{i,t+1}$ is determined as the state $y$ which maximizes the sum of the local mutual information between the states of $i$ and $j_k$. Namely,

$$x_{i,t+1} = \begin{cases} 1 & \text{if LIT}(1; x_{i:j_k,t}) > \text{LIT}(0; x_{i:j_k,t}), \\ 0 & \text{otherwise}, \end{cases}$$

(1)

where

$$\text{LIT}(y; x_{i:j_k,t}) = \sum_{k=1}^{K} \log_{2} \frac{p_{i|j_k}(y|x_{i,j_k,t})}{p_i(y)},$$

(2)

Here, $p_i(y)$ is the probability that the state of $i$ at time step $t + 1$ is $y$ and $p_{i|j_k}(y|x_{i,j_k,t})$ is the conditional probability that the state of $i$ at time step $t + 1$ is $y$, given the state of $j_k$ at time step $t$ is $x_{j_k,t}$. These probabilities at a specific time step are calculated from the frequencies of states and pairs of states in the past $W$ steps in our numerical simulation below. The states in the first $W$ time steps ($t = - (W - 1), \ldots, -1, 0$) are prepared as follows: The initial states and the time evolution rules of all agents are given at random. Then, the states are updated by the rules until $t = 0$. For $t > 0$, the state of $i$ is updated according to equation (1).

We could use transfer entropy instead of mutual information. Indeed, transfer entropy has been accepted as a more appropriate measure to evaluate the magnitude of the impact of one dynamical variable on another one [3]. However, we here use mutual information as an approximation for measuring information transfer to make mathematical argument feasible.

We also note that the underlying network become locally tree-like when $N$ is large since the links between the agents are formed at random. This implies that the inputs to an agent are almost uncorrelated. The agreement between the theory developed below and numerical simulation depends largely on this assumption.

Interestingly, (1) has a Bayesian interpretation. Specifically, (2) can be expressed as

$$\text{LIT}(y; x_{i:j_k,t}) = \log_{2} \frac{\prod_{k=1}^{K} p_{i|j_k}(x_{i,j_k,t}y)}{\prod_{k=1}^{K} p_{i}(x_{i,j_k,t})},$$

(3)

If we introduce $L(x_{i:j_k,t}y) = \prod_{k=1}^{K} p_{j_k}(x_{i,j_k,t}y)$, then LIT$(1; x_{i:j_k,t}) > \text{LIT}(0; x_{i:j_k,t})$ is equivalent to $L(x_{i:j_k,t}1) > L(x_{i:j_k,t}0)$ under the assumption that $\prod_{k=1}^{K} p_{j_k}(x_{i,j_k,t}) > 0$, using the convention $\log_{2} 0 = -\infty$.
for $a \geq 0$. Thus, in our model, each agent performs a maximum likelihood estimation with its inputs as data, $L$ as a likelihood function, and an indeterminate future state as a hypothesis. The time evolution of states can be interpreted as transformation of states-as-hypothesis to states-as-data. We also note that the likelihood function depends on past states and hence is subject to change over time, which is reminiscent of inverse Bayesian inference [20, 21].

If we fix $p_i$ and $p_{ij|k}$ and vary inputs $x_{j_k}$ in (1) and (2), we can act as though a rule outputting $x_{i,t+1}$ is given at each time step $t$. In general, it changes for each time step. However, some rules could persist. Here, we examine which rules can be stabilized. We introduce the following condition for that purpose.

We say that a rule $f$: $\{0, 1\}^K \rightarrow \{0, 1\}$ satisfies the LITM condition when

$$f(x_{j_k}) = 1 \iff \text{LIT}(1; x_{j_k}) > \text{LIT}(0; x_{j_k})$$

holds for all $0 < q_{j_k} < 1$ and $x_{j_k} \in \{0, 1\}^K$. Here, $q_{j_k}$ is the probability that the state of input $j_k$ is 1. The values of $p_i$ and $p_{ij|k}$ are given by

$$p_i(1) = \sum_{x_{j_k} \in \{0, 1\}^K} \prod_{k=1}^K q_{j_k}^x(1 - q_{j_k})^{1-x} f(x_{j_k})$$

and

$$p_{ij|k}(1|z) = \sum_{x_{j_k} \in \{0, 1\}^K} \prod_{k=1, k \neq l}^K q_{j_k}^x(1 - q_{j_k})^{1-x} f(x_{j_k}),$$

respectively. We have used the notation $q_{j_k}$ to make it explicit that the LITM condition is a condition for an isolated single rule and does not depend on the outputs of the other agents.

3. Result

We find that a necessary and sufficient condition for a rule $f$ to satisfy the LITM condition is as follows (theorem 1 in appendix):

There exists an integer $0 \leq l \leq K$, $j_k, \ldots, j_k \in \{j_1, \ldots, j_K\}$, $(z_{j_k}, \ldots, z_{j_k}) \in \{0, 1\}^l$ and $y \in \{0, 1\}$ such that if $(x_{j_k}, \ldots, x_{j_k}) = (z_{j_k}, \ldots, z_{j_k})$, then $f(x_{j_k}) = y$, otherwise $f(x_{j_k}) = 1 - y$.

A rule $f$ satisfying this condition is said to belong to class $l$. For a class $l$ rule $f$, there exist $l$ specific inputs $j_k, \ldots, j_k$ that control the output. If their states are equal to specific values $(z_{j_k}, \ldots, z_{j_k})$, then the output of $f$ is a specific value $y$ and is otherwise $1 - y$, regardless of the values of the other inputs. If we express a rule as a $K$-dimensional Ising hypercube [22], then a rule belonging to class $l$ corresponds to the one such that all the vertices in a single $(K - l)$-dimensional face are colored with the same color and the remaining vertices are colored with the other color (figure 1). One finds the number of class $l$ rules is

![Figure 1. Examples of class $l$ rules for $K = 3$ represented as Ising hypercubes. The labels on vertices are inputs and their color represents output (white and black correspond to 0 and 1, respectively).](image)
by an elementary combinatorial argument. Table 1 shows the value of $C_{K,l}$ for $K \leq 4$. Since the total number of rules with $K$ inputs is $2^{2K}$, the proportion of class $l$ rules approaches 0 as $K$ becomes large.

We examined whether class $l$ rules can indeed be stabilized by numerical simulation. We applied noise by flipping output with probability $\epsilon > 0$. The motivation for introducing noise is to force agents to explore the space of input probabilities $(q_{1}, \ldots, q_{K})$ and thus to make the LITM condition effective. Another reason for including this is that it is natural to consider noise when treating Boolean networks as a model of gene regulatory networks [13, 14] since it is known that gene expression is stochastic [23].

![Figure 2](image)

**Figure 2.** Time evolution of the proportion of class $l$ rules ($0 \leq l \leq 3$) for $K = 3$. ‘Others’ indicates rules that do not satisfy the LITM condition. The values of the other parameters are $N = 1000$, $W = 100$, and $\epsilon = 0.05$.

Table 1. The number of class $l$ rules $C_{K,l}$ for $1 \leq K \leq 4$.

| $l$ | 0 | 1 | 2 | 3 | 4 |
|-----|---|---|---|---|---|
| $C_{1,l}$ | 2 | 2 | 22 | 24 | 8 |
| $C_{2,l}$ | 2 | 4 | 8 | 24 | 16 |
| $C_{3,l}$ | 2 | 6 | 24 | 16 | 32 |

$$ C_{K,l} = \begin{cases} 2K & \text{if } l = 1, \\ 2^{l+1}K & \text{otherwise} \end{cases} $$

4 In the absence of noise, each agent can explore only a small part of the space of input probabilities before its rule is stabilized. Thus, the stabilized rules do not necessarily satisfy the LITM condition and may vary from trial to trial. On the other hand, when there is too much noise, the conditional probability fluctuates largely and thus rules are never stabilized.
4. Discussion

Class \( l \) rules have a high degree of canalization. Canalization is the property that specific values of specific inputs determine the output of a rule regardless of the values of the other inputs. In the context of ontogeny, it enables organisms to explore beneficial combinations of variations while maintaining a robust phenotype by buffering variations in the developmental process [24]. It has been suggested that real-world eukaryotic gene regulatory networks use rules with a high degree of canalization [25, 26]. The degree of canalization of a rule with \( K \) inputs can be quantified as the ratio of the number of \( k \)-tuple inputs that determine the output regardless of the other inputs to the number of all \( k \)-tuple inputs \( p_k \) \((k = 0, 1, \ldots, K - 1)\) [27]. Table 2 shows \( p_k \) of class \( l \) rules for \( K = 3 \). For \( K = 2, 3, p_k \) of rules that do not belong to any class is smaller than or equal to \( p_k \) of class \( l \) rules. For \( K = 4 \), there are 96 rules that do not belong to any class whose \( p_k \) for some \( k \) is greater than that of class 1 and 2 rules. Thus, it is not necessarily true that class \( l \) rules always have a higher degree of canalization than those that do not belong to any class. Class 1 rules are the least canalizing rules among class \( l \) rules. However, we can say that they have a high degree of canalization among all rules. For example, when \( K = 3 \), there are 18 rules whose \( p_k \) is greater than that of class \( l \) rules for some \( k \) among all 256 rules. These 18 rules belong to either class \( 0 \) or class \( 3 \) (Class 2 rules have the same \( p_k \) as class 1 rules. This is also the case for \( K = 4 \)). When \( K = 4 \), there are 194 rules whose \( p_k \) is greater than that of class 1 rules for some \( k \) among all 65,536 rules. These 194 rules consist of rules belonging to classes \( 0, 3, \) and 4 along with the above-mentioned 96 rules that do not belong to any class.

In recent years, it has been suggested that the dynamics of real-world gene regulatory networks are near critical [28]. It has been argued that criticality of dynamics is related to a balance between robustness and evolvability [29] and an optimal information processing ability [30]. The dynamics of a Boolean network in which each agent uses a class 1 rule is expected to be critical. Indeed, a random Boolean network with class 1 rules is equivalent to one with \( K = 1 \) without constant functions. It is known that the dynamics of the latter is critical [31]. Indeed, the stability of Boolean networks can be determined by measuring the rate of perturbation propagation \( \lambda \) [32]. Let \( d_t \) be the size of perturbation at time step \( t \) (the fraction of agents whose states are flipped). By plotting \( d_{t+1} \) against \( d_t \) and computing the slope of the graph (the Derrida plot) at the origin, we can obtain \( \lambda \). The states \( \lambda < 1, \lambda = 1 \) and \( \lambda = 1 \) correspond to stable, unstable and critical dynamics of Boolean networks. Let \( f \) be a class 1 rule. There exists a unique input such that the output of \( f \) changes when the input is flipped. The other inputs are irrelevant. Thus, if each input of \( f \) is flipped with probability \( d_t \), then its output will

![Figure 3](image-url)
also flip with probability $d_i$. If every agent uses class 1 rules, then $d_i$ will be conserved on average and thus $\lambda$ is expected to be close to 1. We numerically checked the stability of the dynamics of Boolean networks obtained in figure 2 by constructing the Derrida plots. Figure 4 shows the Derrida plot of a Boolean network after 10 000 time steps from figure 2. The graph is indistinguishable from the diagonal line $d_{t+1} = d_t$ and indicates a critical dynamics, as we expect.

5. Conclusion

In summary, we showed that maximization of local information transfer in Boolean networks under the presence of noise leads to the emergence of class 1 rules that are both highly canalized and critical. Bassler et al [27] reported that critical rules with a high degree of canalization emerge from a minority game on Boolean networks. Although class 1 rules appear with high frequency in their model, other rules also appear with positive probability. Thus, the criticality dynamic is different from ours. Their model is an example of extremal dynamics since the most losing agent changes its rule randomly [33]. Hence, there is a global selection of rules that is performed from outside. Here, on the other hand, rules are determined locally from within [34, 35]. Our approach suggests an alternative picture in which the determination process of rules based on LITM yields dynamics with criticality and a high degree of canalization, which are generic properties of real-world gene regulatory networks.

Future work can proceed in two directions. One is to investigate what kind of spatio-temporal structures emerges when LITM is applied to each agent in models having the spatial dimension such as cellular automata. This direction is now ongoing and will be reported elsewhere [36]. The other is a generalization of the framework based on the Bayesian interpretation pointed out in section 2. We could consider likelihood functions other than equation (3) and study what kind of dynamics emerges when each agent performs a maximum likelihood estimation similar to the one discussed in this paper for a particular likelihood function. These extensions would make the limits and applicability of our approach clearer.

Acknowledgments

The authors are grateful to the anonymous reviewers for their constructive comments and suggestions. TH was partially supported by JSPS KAKENHI Grant Number 18K03423. KN was supported by JST PRESTO Grant Number JPMJPR15E7, Japan, and JSPS KAKENHI Grant Numbers 15K16076 and 26880010.

Appendix

Let $f : \{0, 1\}^K \rightarrow \{0, 1\}$ be a Boolean function assigned to an agent $i$. For convenience, we renumber $N$ agents so that inputs to $i$ are 1, 2, \ldots, $K$ and $K < i$. $f$ is said to satisfy the LITM condition when
\[ f(x_{1:K}) = 1 \iff \text{LIT}(1; x_{1:K}) > \text{LIT}(0; x_{1:K}) \] (A.1)

for all \(0 < q_1, \ldots, q_K < 1\) and \(x_{1:K} = (x_1, \ldots, x_K) \in \{0, 1\}^K\), where \(q_k\) is the probability that \(x_k = 1\) and

\[ \text{LIT}(y; x_{1:K}) = \log \frac{\prod_{k=1}^{K} p_{ijk}(y|x_k)}{p_i(y)^K}. \] (A.2)

The values of \(p_i\) and \(p_{ijk}\) are given by

\[ p_i(1) = \sum_{x_i \in \{0, 1\}} \prod_{k=1}^{K} q_{ik}^{x_i}(1 - q_k)^{1-x_i} f(x_{1:K}) \] (A.3)

and \(p_i(0) = 1 - p_i(1)\) and

\[ p_{ijk}(1|z) = \sum_{x_i \in \{0, 1\}} \prod_{k=1, k \neq j}^{K} q_{ik}^{x_i}(1 - q_k)^{1-x_i} f(x_{1:K}) \] (A.4)

and \(p_{ijk}(0|z) = 1 - p_{ijk}(1|z)\).

**Theorem 1.** A rule \(f : \{0, 1\}^K \rightarrow \{0, 1\}\) satisfies the LITM condition if and only if there exists an integer \(0 \leq l \leq K\), as well as \(k_1, \ldots, k_l \in \{1, \ldots, K\}\), \((z_{k_1}, \ldots, z_{k_l}) \in \{0, 1\}^l\), and \(y \in \{0, 1\}\) such that if \((x_{k_1}, \ldots, x_{k_l}) = (z_{k_1}, \ldots, z_{k_l})\) then \(f(x_{1:K}) = y\), and otherwise \(f(x_{1:K}) = 1 - y\). A rule satisfying this condition is said to belong to class \(l\).

**Proof.** It is straightforward to check that class 0 rules satisfy the condition stated in the theorem by using the convention mentioned in the main text: let \(\log \frac{1}{a} = -\infty\) for \(a \geq 0\). In the following, we assume that \(f\) does not belong to class 0, namely, that \(f\) is not a constant function. Let \(f\) be a class 1 rule for some \(1 \leq l \leq K\). Without loss of generality, we can assume that \(f(x_{1:K}) = 1\) if and only if \(x_{1:l} = (1, \ldots, 1)\) for \(x_{1:K} \in \{0, 1\}^K\). We have

\[ p_i(1) = \prod_{k=1}^{l} q_k = q, \] (A.5)

\[ p_{ijk}(1|1) = \begin{cases} q p_i^j & \text{for } 1 \leq j \leq l, \\ q & \text{for } l + 1 \leq j \leq K \end{cases} \] (A.6)

and

\[ p_{ijk}(1|0) = \begin{cases} 0 & \text{for } 1 \leq j \leq l, \\ q & \text{for } l + 1 \leq j \leq K. \end{cases} \] (A.7)

To prove that the LITM condition holds for \(f\), we have to show

\[ \text{if } f(x_{1:K}) = 1 \text{ then } \frac{\prod_{i=1}^{l} p_{i}^{j}(1|x_i)}{p_i(1)^K} > \frac{\prod_{i=1}^{l} p_{i}^{j}(0|x_i)}{p_i(0)^K} \] (A.8)

and

\[ \text{if } f(x_{1:K}) = 0 \text{ then } \frac{\prod_{i=1}^{l} p_{i}^{j}(1|x_i)}{p_i(1)^K} \leq \frac{\prod_{i=1}^{l} p_{i}^{j}(0|x_i)}{p_i(0)^K} \] (A.9)

for all \(1 < q_1, \ldots, q_K < 1\).

When \(f(x_{1:K}) = 1\), we have \(x_{1:l} = (1, \ldots, 1)\) by the definition of \(f\). From (A.5)–(A.7), we have

\[ \frac{\prod_{k=1}^{K} p_{ijk}(1|x_k)}{p_i(1)^K} = \prod_{k=1}^{l} q_i^{j} \prod_{k=l+1}^{K} q_i^{k} = \frac{1}{q} \] (A.10)

and

\[ \frac{\prod_{k=1}^{K} p_{ijk}(0|x_k)}{p_i(0)^K} = \prod_{k=1}^{l} \left(1 - \frac{q_i^{j}}{q_i^{k}}\right) \prod_{k=l+1}^{K} (1 - q_i^{k}) = \frac{\prod_{k=1}^{l} (q_i^{j} - q_i^{k})}{q_i(1 - q_i^{j})}. \] (A.11)

Since \((1 - q_i^{j}) > \prod_{k=1}^{l} (q_i^{j} - q_i^{k})\) holds for all \(0 < q_1, \ldots, q_K < 1\), the relation (A.8) holds for all \(0 < q_1, \ldots, q_K < 1\).
When \( f(x_{1:K}) = 0 \), there exists some \( 1 \leq j \leq l \) such that \( x_j = 0 \). Since \( p_{1:j}(1|0) = 0 \) by (A.7), we have
\[
\frac{\prod_{j=1}^{l} p_{1:j}(1|x_j)}{p_{1:l}(1|x)} = 0. \quad \text{(A.9)}
\]
Thus, (A.9) holds for all \( 0 < q_0, \ldots, q_K < 1 \).

Conversely, let \( f \) be a rule that does not belong to any class \( I \leq l \leq K \). We have to show that there exist \( (q_1', \ldots, q_K') \) and \( x_{1:K} \) such that \( (A.1) \) does not hold. Indeed, we will show that \( (A.1) \) does not hold for a positive volume in the space of probabilities \( (q_1', \ldots, q_K') \) for some \( x_{1:K} \). Let \( F \) be the minimum face of the \( K \)-dimensional Ising hypercube corresponding to \( f \) such that \( F \) includes all vertices \( x_{1:K} \) satisfying \( f(x_{1:K}) = 1 \). Since \( f \) does not belong to any class, \( F \) must include at least one vertex \( x_{1:K} \) such that \( f(x_{1:K}) = 0 \). Without loss of generality, we can assume that \( F = \{x_{1:K} \mid x_1 = \ldots = x_K = 1\} \) for some \( 1 \leq l \leq K \). In particular, we have \( z_1 = \ldots = z_K = 1 \) and \( f(x_{1:K}) = 0 \) if \( x_{1:K} \notin F \). We also note that there is at least one \( x_{1:K} \in F \) such that \( f(x_{1:K}) = 1 \) since \( f \) is not a constant function. Let us assume that \( q_1 = \ldots = q_l = r \). We have
\[
p_{1:l}(1) = \sum_{x_{1:K} \in [0,1]^K} \prod_{k=1}^{K} q_k^{x_k}(1 - q_k)^{1-x_k} f(x_{1:K})
\]
\[
= \sum_{x_{1:K} \in F} r^l \prod_{k=l+1}^{K} q_k^{x_k}(1 - q_k)^{1-x_k} f(x_{1:K})
\]
\[
= r^l q_{l+1:K}.
\](A.12)

where we write
\[
g(q_{l+1:K}) = \sum_{x_{1:K} \in F} \prod_{k=l+1}^{K} q_k^{x_k}(1 - q_k)^{1-x_k} f(x_{1:K}).
\](A.13)

Note that \( 0 < g(q_{l+1:K}) < 1 \) for all \( 0 < q_{l+1:K} < 1 \), since \( F \) includes \( x_{1:K} \) such that \( f(x_{1:K}) = 1 \) and \( f(x_{1:K}) = 0 \).

We have
\[
p_{1:l}(1|1) = \sum_{x_{1:K} \in [0,1]^K} \prod_{k=1}^{K} q_k^{x_k}(1 - q_k)^{1-x_k} f(x_{1:K})
\]
\[
= \sum_{x_{1:K} \in F} r^{l-1} \prod_{k=l+1}^{K} q_k^{x_k}(1 - q_k)^{1-x_k} f(x_{1:K})
\]
\[
= r^{l-1} g(q_{l+1:K}).
\](A.14)

for \( 1 \leq j \leq l \) and
\[
p_{1:j}(1|z_j) = \sum_{x_{1:K} \in [0,1]^K} \prod_{k=1}^{K} q_k^{x_k}(1 - q_k)^{1-x_k} f(x_{1:K})
\]
\[
= \sum_{x_{1:K} \in F} r^l \prod_{k=l+1}^{K} q_k^{x_k}(1 - q_k)^{1-x_k} f(x_{1:K})
\]
\[
= r h_{z_j}(q_{l+1:K} \ldots , q_k).
\](A.15)

for \( l + 1 \leq j \leq K \), where we write
\[
h_{z_j}(q_{l+1:K} \ldots , q_k) = \sum_{x_{1:K} \in F} \prod_{k=l+1}^{K} q_k^{x_k}(1 - q_k)^{1-x_k} f(x_{1:K}).
\](A.16)

We also have \( 0 < h_{z_j}(q_{l+1:K} \ldots , q_k) < 1 \) for all \( 0 < q_{l+1:K} < 1 \). Indeed, \( h_{z_j}(q_{l+1:K} \ldots , q_k) < 1 \) because \( z_{1:K} \in F \) and \( f(z_{1:K}) = 0 \). If there is no \( x_{1:K} \in F \) such that \( x_j = z_j \) and \( f(x_{1:K}) = 1 \), then \( F \) includes a strictly smaller face \( F' \) such that \( F' \) includes all \( x_{1:K} \) satisfying \( f(x_{1:K}) = 1 \). This contradicts the minimality of \( F \). Thus, we have \( 0 < h_{z_j}(q_{l+1:K} \ldots , q_k) \).

Let us conclude the proof by showing \( \text{LIT}(1; z_{1:K}) > \text{LIT}(0; z_{1:K}) \) as \( r \to 0 \) even though \( f(z_{1:K}) = 0 \). This implies that \( f \) does not satisfy \( (A.1) \) for a positive volume of the space of probabilities \( (q_0, \ldots, q_K) \) since the inequality is strict and \( \text{LIT}(1; z_{1:K}) \) and \( \text{LIT}(0; z_{1:K}) \) are continuous functions of \( q_0, \ldots, q_K \). Indeed, we have
This diverges as \( r \to 0 \) for every fixed \( 0 < q_{l+1} \ldots, q_{k} < 1 \) since \( \frac{\prod_{l=k+1}^{K} q_{l}^{h_{l}}}{g^{k-l}} \to 0 \). On the other hand,

\[
\prod_{k=1}^{K} \frac{p_{l|k}(1|x_{k})}{p_{l}(1)^{K}} = \left( \frac{r^{l-1}g^{l}}{r^{k}g^{k}} \right) \prod_{k=1}^{K} \frac{h_{k}}{r^{k-l}}.
\]

(A.17)

This diverges as \( r \to 0 \) for every fixed \( 0 < q_{l+1} \ldots, q_{k} < 1 \) since \( \frac{\prod_{l=k+1}^{K} q_{l}^{h_{l}}}{g^{k-l}} \to 0 \). On the other hand,

\[
\prod_{k=1}^{K} \frac{p_{l|k}(0|x_{k})}{p_{l}(0)^{K}} = \prod_{k=1}^{K} \frac{p_{l|k}(0|x_{k})}{(1 - p_{l}(1))^{k}} < \infty
\]

(A.18)

for \( r \to 1 \). Thus, the proof is completed. □

**Theorem 2.** Let \( f \) be a class \( l \geq 1 \) rule. Let \( \epsilon > 0 \). Let us flip the output of \( f \) with probability \( \epsilon < \frac{1}{2} \). Then, the LITM condition is still satisfied when \( f \) belongs to class \( l \). For \( l \geq 2 \), class \( l \) rules cannot satisfy the LITM condition. In particular, there is a positive volume in the space of probabilities \( q_{1}, \ldots, q_{k} \) such that (A.1) does not hold for some \( x_{l:k} \).

**Proof.** Let \( f \) be a class \( l \) rule. Without loss of generality, we can assume that \( f(x_{l:k}) = 1 \) if \( x_{l:j} = (1, \ldots, 1) \) and \( f(x_{l:k}) = 0 \) otherwise. Let \( p_{l,e}(y) \) be the probability that the output of \( f \) is \( y \) and \( p_{l|j,e}(y|z) \) the conditional probability that the output of \( f \) is \( y \) given that the \( j \)th input is \( z \) under the presence of noise of strength \( \epsilon \). It is clear that

\[
p_{l,e}(y) = (1 - \epsilon)p_{l}(y) + \epsilon p_{l}(1 - y)
\]

(A.19)

and

\[
p_{l|j,e}(y|z) = (1 - \epsilon)p_{l|j}(y|z) + \epsilon p_{l|j}(1 - y|z).
\]

(A.20)

First, let us consider the case \( l = 1 \). When \( l = 1 \), we have

\[
p_{l,e}(1) = (1 - \epsilon)q_{l} + \epsilon (1 - q_{l}) = (1 - 2\epsilon)q_{l} + \epsilon = q_{l,e},
\]

(A.21)

\[
p_{l|j,e}(1|1) = \begin{cases} (1 - \epsilon)q_{l} + \epsilon (1 - q_{l}) = q_{l,e} & \text{if } j = 1 \\ (1 - \epsilon)q_{l} + \epsilon (1 - q_{l}) = q_{l,e} & \text{otherwise} \end{cases}
\]

(A.22)

and

\[
p_{l|j,e}(1|0) = \begin{cases} (1 - \epsilon)q_{l} + \epsilon (1 - q_{l}) = q_{l,e} & \text{if } j = 1 \\ (1 - \epsilon)q_{l} + \epsilon (1 - q_{l}) = q_{l,e} & \text{otherwise}. \end{cases}
\]

(A.23)

Let us assume \( f(x_{l:k}) = 1 \). We have to show

\[
\prod_{k=1}^{K} \frac{p_{l|k,e}(1|x_{k})}{p_{l,e}(1)^{K}} > \prod_{k=1}^{K} \frac{p_{l|k,e}(0|x_{k})}{p_{l,e}(0)^{K}}.
\]

(A.24)

Since

\[
\text{lhs} = \frac{(1 - \epsilon)q_{l,e}^{K-1}}{q_{l,e}^{K}} = 1 - \epsilon
\]

(A.25)

and

\[
\text{rhs} = \frac{\epsilon (1 - q_{l,e})^{K-1}}{(1 - q_{l,e})^{K}} = \frac{\epsilon}{1 - q_{l,e}},
\]

(A.26)

we have

\[
\text{lhs} - \text{rhs} = \frac{1 - \epsilon}{q_{l,e}} - \frac{\epsilon}{1 - q_{l,e}} = \frac{(1 - 2\epsilon)(1 - q_{l})}{q_{l,e}(1 - q_{l,e})} > 0.
\]

(A.27)
Thus, (A.24) holds. Similarly, we can show that the following inequality holds when \( f(x_{1}; r) = 0 \):
\[
\prod_{k=1}^{K} \frac{p_{i_1k}^{(1)}(1|x_i)}{p_{i_1}(1)^K} \leq \prod_{k=1}^{K} \frac{p_{i_1k}^{(0)}(0|x_i)}{p_{i_1}(0)^K}.
\]
(A.28)
Thus, a class 1 rule \( f \) satisfies the LITM condition even when its output is flipped with probability \( \epsilon \).
Next, let us assume \( l \geq 2 \). Let \( q = \prod_{k=1}^{I} q_k \). We have
\[
p_{i_1}(1) = (1 - \epsilon)q + \epsilon (1 - q) = (1 - 2\epsilon)q + \epsilon = q_r,
\]
(A.29)
\[
p_{i_1}(1|1) = \begin{cases} (1 - \epsilon)\frac{q}{q_r} + \epsilon \left(1 - \frac{q}{q_r}\right) = \frac{(1 - 2\epsilon)q + \epsilon q}{q_r} & \text{if } 1 \leq j \leq I \\ (1 - \epsilon)q + \epsilon (1 - q) = q_r & \text{otherwise} \end{cases}
\]
(A.30)
and
\[
p_{i_1}(1|0) = \begin{cases} (1 - \epsilon) \times 0 + \epsilon \times 1 = \epsilon & \text{if } 1 \leq j \leq I \\ (1 - \epsilon)q + \epsilon (1 - q) = q_r & \text{otherwise} \end{cases}
\]
(A.31)
Let \( z_{1;K} = (1, 0, \ldots, 0) \). Since \( l \geq 2 \), we have \( f(z_{1;K}) = 0 \). However, the inequality
\[
\prod_{k=1}^{K} \frac{p_{i_1k}^{(1)}(1|z_k)}{p_{i_1}(1)^K} > \prod_{k=1}^{K} \frac{p_{i_1k}^{(0)}(0|z_k)}{p_{i_1}(0)^K}
\]
(A.32)
holds for a positive volume of the space of probabilities \( q_1, \ldots, q_K \). To see this, let us assume \( q_1 = \ldots = q_l = r \).
Then, we have
\[
\text{lhs} = \frac{((1 - 2\epsilon)q + \epsilon q_r)e^{l-1}}{q_r q_r} = \frac{((1 - 2\epsilon)r^{l-1} + \epsilon) e^{l-1}}{(1 - 2\epsilon)r^l + \epsilon}
\]
(A.33)
and
\[
\text{rhs} = \frac{(q_1 - (1 - 2\epsilon)q - q_r \epsilon)(1 - \epsilon)^{l-1}}{q_r (1 - q_r)}
\]
\[
= \frac{(1 - \epsilon - (1 - 2\epsilon)r^{l-1})(1 - \epsilon)^{l-1}}{1 - \epsilon - (1 - 2\epsilon)r^l}.
\]
(A.34)
The lowest degree term of the numerator of the difference \( \text{lhs} - \text{rhs} \) regarded as a polynomial of \( r \) is
\[
(1 - 2\epsilon)e^{l-1}(1 - \epsilon)^{l-1} - 1 > 0.
\]
(A.35)
Thus, if \( r > 0 \) is sufficiently close to 0, then we also have \( \text{lhs} - \text{rhs} > 0 \). Since the inequality is strict for such \( r \) and both sides are continuous functions of \( (q_1, \ldots, q_K) \), this proves that \( f \) does not satisfy (A.1) for a positive volume of the space of probabilities \( (q_1, \ldots, q_K) \) for \( z_{1;K} \).

ORCID iDs

K Nakajima @ https://orcid.org/0000-0001-5589-4054

References

[1] Cover T M and Thomas J A 1991 Elements of Information Theory (New York: Wiley)
[2] Schreiber T 2000 Phys. Rev. Lett. 85 461–4
[3] Bossomaier T, Barnett L, Harré M and Lizier J T 2016 An Introduction to Transfer Entropy: Information Flow in Complex Systems (Basel: Springer)
[4] Hidalgo J, Grilli J, Suweis S, Muñoz M A, Banavar J R and Maritan A 2014 Proc. Natl Acad. Sci. USA 111 10095–100
[5] Yamaguti Y and Tsuda I 2015 Neural Netw. 62 3–10
[6] Luque B and Faretra A 2000 Complex Syst. 12 241–52
[7] Ribeiro A S, Kauffman S A, Lloyd-Price J, Samuelsson B and Socolar J E 2008 Phys. Rev. E 77 011901
[8] Barnett L, Lizier J T, Harré M, Seth A K and Bossomaier T 2013 Phys. Rev. Lett. 111 177203
[9] Lizier J T, Prokopenko M and Zomaya A Y 2008 The information dynamics of phase transitions in random Boolean networks Proc. 11th Int. Conf. on the Simulation and Synthesis of Living Systems (Alife XI) ed S Bullock (Cambridge, MA: MIT Press) pp 574–81
[10] Lizier J T, Prokopenko M and Zomaya A Y 2008 Phys. Rev. E 77 026110
[11] Lizier J T, Prokopenko M and Zomaya A Y 2010 Chaos 20 037109
[12] Lizier J T, Prokopenko M and Zomaya A Y 2012 Inf. Sci. 208 39–54
[13] Kauffman S A 1969 J. Theor. Biol. 22 437–67
[14] Kauffman S A 1993 The Origins of Order: Self-Organization and Selection in Evolution (New York: Oxford University Press)
[15] Haruna T 2017 Artif. Life 23 105–18
[16] Lizier J T, Prokopenko M, Tanev I and Zomaya A Y 2008 Emergence of glider-like structures in a modular robotic system Proc. 11th Int. Conf. on the Simulation and Synthesis of Living Systems (Alife XI) ed S Bullock et al (Cambridge, MA: MIT Press) pp 366–73
[17] Obst O, Boedecker J and Asada M 2010 Improving recurrent neural network performance using transfer entropy Neural Information Processing. Models and Applications. ICONIP 2010 (Lecture Notes in Computer Science vol 6444) ed K W Wong et al (Berlin: Springer) pp 193–200
[18] Linsker R 1988 IEEE Comput. 21 105–17
[19] Drossel B 2008 Random Boolean networks Reviews of Nonlinear Dynamics and Complexity ed H G Schuster (Weinheim: Wiley)
[20] Arecchi F T 2011 Nonlinear Dyn. Psychol. Life Sci. 15 359–75
[21] Gunji Y P, Shinohara S, Haruna T and Basios V 2017 BioSystems 152 44–65
[22] Reichhardt C J O and Bassler K E 2007 J. Phys. A: Math. Theor. 40 4339
[23] Elowitz M B, Levine A J, Siggia E D and Swain P S 2002 Science 297 1183–6
[24] Waddington C H 1942 Nature 150 563–5
[25] Harris S E, Sawhill B K, Wensche A and Kauffman S 2002 Complexity 7 23–40
[26] Kauffman S, Peterson C, Samuelsson B and Troein C 2003 Proc. Natl Acad. Sci. USA 100 14796–9
[27] Bassler K E, Lee C and Lee Y 2004 Phys. Rev. Lett. 93 038101
[28] Balleza E, Alvarez-Buylla E R, Chaos A, Kauffman S, Shmulevich I and Aldana M 2008 PLoS One 3 e2456
[29] Aldana M, Balleza E, Kauffman S and Resendiz O 2007 J. Theor. Biol. 245 433–48
[30] Goudarzi A, Teuscher C, Gulbahce N and Rohlf T 2012 Phys. Rev. Lett. 108 128702
[31] Drossel B, Mihaljev T and Greil F 2005 Phys. Rev. Lett. 94 088701
[32] Derrida B and Pomeau Y 1986 Europhys. Lett. 1 45–9
[33] Bak P and Sneppen K 1993 Phys. Rev. Lett. 71 4083–6
[34] Gunji Y P 1995 BioSystems 35 33–62
[35] Matsuno K 1996 BioSystems 38 111–8
[36] Nakajima K and Haruna T 2018 submitted