Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Information geometry on complexity and stochastic interaction

by

Nihat Ay

Preprint no.: 95 2001

Information Geometry on Complexity and Stochastic Interaction

Nihat Ay

Max-Planck-Institute for Mathematics in the Sciences

Inselstr. 22-26

04103 Leipzig, Germany

E-mail: nay@mis.mpg.de

30. November 2001

Abstract

Interdependencies of stochastically interacting units are usually quantified by the Kullback-Leibler divergence of a stationary joint probability distribution on the set of all configurations from the corresponding factorized distribution. This is a spatial approach which does not describe the intrinsically temporal aspects of interaction. In the present paper the setting is extended to a dynamical version where temporal interdependencies are also captured by using information geometry of Markov-chain manifolds.

Key words and phrases. stochastic interaction, complexity, information geometry, Kullback-Leibler divergence, separability, Markov chains, random fields.

Contents

1 Introduction

"The whole is more than the sum of its elementary parts." This statement characterizes the present approach to *complexity*. Let's put it in a more formal setting. Assume that we have a system consisting of elementary units $\nu \in V$. With each non-empty subsystem $S \subset V$ we associate a set $\mathcal{O}_S$ of objects that can be generated by S. Examples for such objects are (deterministic) dynamical systems, stochastic processes, and probability distributions. Furthermore we assume that there is a "composition" map $\otimes : \prod_{\nu \in V} \mathcal{O}_{\{\nu\}} \hookrightarrow \mathcal{O}_V$ that defines how to put objects of the individual units together in order to describe a global object without any interrelations. The image of $\otimes$ consists of the *split* global objects which are completely characterized by the individual ones, and therefore represent the absence of complexity. In order to quantify complexity, assume that there is given a function $D : (x, y) \mapsto D(x \,\|\, y)$, that measures the divergence of global objects $x, y \in \mathcal{O}_V$. We define the complexity of $x \in \mathcal{O}_V$ to be the divergence from being split:

$$\mathit{Complexity}(x) \;:=\; \inf_{y \;\mathit{split}} D(x \,\|\, y). \tag{1}$$

Of course, this approach is very general, and there are many ways to define complexity following this concept. Is there a canonical way? At least, within the probabilistic setting, *information geometry* [1], [3] provides a very convincing framework for this. In the context of random fields, it leads to a measure

for "spatial" interdependencies: Given state sets Ω_ν, $\nu \in V$, we define the set $\mathcal{O}_S$ of objects that are generated by a subsystem $S \subset V$ to be the probability distributions on the product set $\prod_{\nu\in S} \Omega_\nu$. A family of individual probability distributions $p^{(\nu)}$ on Ω_ν can be considered as a distribution on the whole configuration set $\prod_{\nu\in V} \Omega_\nu$ by identifying it with the product $\otimes_{\nu\in V} p^{(\nu)} \in \mathcal{O}_V$. In order to define the complexity of a distribution $p \in \mathcal{O}_V$ on the whole system, according to (1) we have to choose a divergence function. A canonical choice for D is given by the *Kullback-Leibler divergence* [15], [9]:

$$\mathit{Complexity}(p) \;:=\; I(p) \;:=\; \inf_{p^{(\nu)}\in\mathcal{O}_\nu,\, \nu\in V} D\big(p \,\|\, \otimes_{\nu\in V}\; p^{(\nu)}\big) \tag{2}$$

It is well known that $I(p)$ quantifies spatial interdependencies [2]. It vanishes exactly when the units are stochastically independent with respect to p. Such split distributions are called *factorizable* in this context. In Figure 1, the example of two binary units with the state sets $\{0,1\}$ is illustrated.

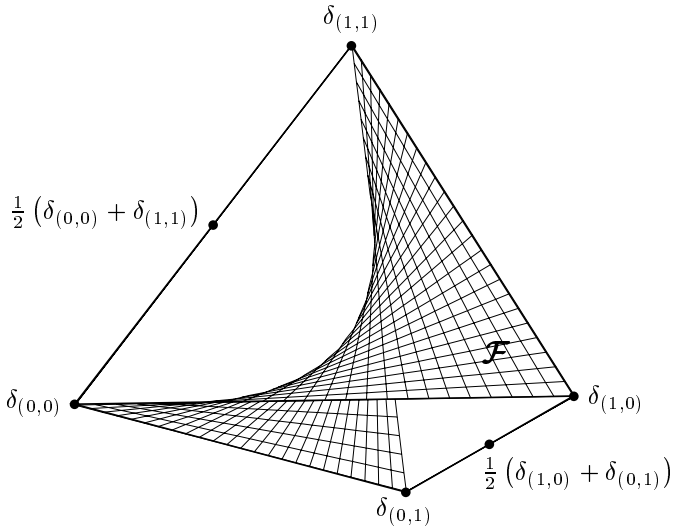

FIG. 1. *$\mathcal{F}$ denotes the set of factorizable distributions on $\{0,1\} \times \{0,1\}$.*

The distributions with maximal interdependence (complexity) are given by

$$\frac{1}{2}\left(\delta_{(0,0)}+\delta_{(1,1)}\right) \qquad \text{and} \qquad \frac{1}{2}\left(\delta_{(1,0)}+\delta_{(0,1)}\right).$$

Spatial interdependence has been studied by Amari [2] and Ay [4], [5] from the information geometric point of view, where it is referred to as (*stochastic*) *interaction* and discussed in view of neural networks. The aim of the present paper is to use the concept of complexity that is formalized by (1) in order to extend spatial interdependence to a dynamical notion of interaction, where the evolution in time is taken into account. Therefore, the term "stochastic interaction" is mainly used in the context of spatio-temporal interdependence.

The present paper is organized as follows. After a brief introduction into the information geometric description of finite probability spaces in Section 2, the general notion of separability is introduced for Markovian transition kernels, and information geometry is used for quantifying non-separability as divergence from separability (Section 3). In Section 4, the presented theoretical framework is used to derive a dynamical version of (2), where spatio-temporal interdependencies are quantified and referred to as *stochastic interaction*. This is illustrated by some simple but instructive examples.

2 Preliminaries on Finite Information Geometry

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

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

Following the information geometric description of finite probability spaces, its interior $\mathcal{P}(\Omega)$ can be considered as a differentiable submanifold of $\mathbb{R}^\Omega$ with dimension $|\Omega| - 1$ and the basis-point independent tangent space[1]

$$\mathrm{T}(\Omega) := \left\{ x \in \mathbb{R}^\Omega \ : \ \sum_{\omega\in\Omega} x(\omega) = 0 \right\}.$$

With the *Fisher metric* $\langle \cdot, \cdot \rangle_p : \mathrm{T}(\Omega) \times \mathrm{T}(\Omega) \to \mathbb{R}$ in $p \in \mathcal{P}(\Omega)$ defined by

$$(x, y) \ \mapsto \ \langle x, y \rangle_p := \sum_{\omega\in\Omega} \frac{1}{p(\omega)}\, x(\omega) y(\omega),$$

$\mathcal{P}(\Omega)$ becomes a Riemannian manifold [17].[2] The most important additional structure studied in information geometry is given by a pair of dual affine connections on the manifold. Application of such a dual structure to the present

[1] If one considers $\mathcal{P}(\Omega)$ as an "abstract" differentiable manifold, there are many ways to represent it as a submanifold of $\mathbb{R}^\Omega$. In information geometry, the natural embedding presented here is called (-1)- respectively (m)-*representation*.

[2] In mathematical biology this metric is also known as *Shahshahani metric* [11], [14].

situation leads to the notion of (-1)- and $(+1)$-geodesics: Each two points $p, q \in \mathcal{P}(\Omega)$ can be connected by the geodesics $\gamma^{(\alpha)} = \left(\gamma_\omega^{(\alpha)}\right)_{\omega\in\Omega} : [0,1] \to \mathcal{P}(\Omega)$, $\alpha \in \{-1, +1\}$, with

$$\gamma_\omega^{(-1)}(t) \;:=\; (1-t)\,p(\omega) + t\,q(\omega) \quad \text{and} \quad \gamma_\omega^{(+1)}(t) \;:=\; r(t)\;p(\omega)^{1-t}\,q(\omega)^t.$$

Here, $r(t)$ denotes the normalization factor.

A submanifold $\mathcal{E}$ of $\mathcal{P}(\Omega)$ is called *exponential family* if there exist a point $p_0 \in \mathcal{P}(\Omega)$ and vectors $v_1, \ldots, v_d \in \mathbb{R}^\Omega$, such that it can be expressed as the image of the map $\mathbb{R}^d \to \mathcal{P}(\Omega)$, $\theta = (\theta_1, \ldots, \theta_d) \mapsto p_\theta$, with

$$p_\theta(\omega) \;:=\; \frac{p_0(\omega)\,\exp\left(\sum_{i=1}^d \theta_i\,v_i(\omega)\right)}{\sum_{\omega'\in\Omega} p_0(\omega')\,\exp\left(\sum_{i=1}^d \theta_i\,v_i(\omega')\right)}. \tag{3}$$

Let p be a probability distribution in $\mathcal{P}(\Omega)$. An element $p' \in \mathcal{E}$ is called (-1)-*projection* of p onto $\mathcal{E}$ iff the (-1)-geodesic connecting p and p' intersects $\mathcal{E}$ orthogonally with respect to the Fisher metric. Such a point p' is unique ([1], Theorem 3.9, p. 91) and can be characterized by the *Kullback-Leibler divergence* [15], [9] (This is a special case of Csiszár's f-*divergence* [8])

$$D : \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \to \mathbb{R}_+, \qquad (p,q) \;\mapsto\; D(p\,\|\,q) \;:=\; \sum_{\omega\in\Omega} p(\omega)\,\ln\frac{p(\omega)}{q(\omega)}.$$

We define the distance $D(\cdot\,\|\,\mathcal{E}) : \mathcal{P}(\Omega) \to \mathbb{R}_+$ from $\mathcal{E}$ by

$$p \;\mapsto\; D(p\,\|\,\mathcal{E}) \;:=\; \inf_{q\in\mathcal{E}} D(p\,\|\,q).$$

It is well known that a point $p' \in \mathcal{E}$ is the (-1)-projection of p onto $\mathcal{E}$ if and only if it satisfies the minimizing property $D(p\,\|\,\mathcal{E}) = D(p\,\|\,p')$ ([1], Theorem 3.8, p. 90; [3], Corollary 3.9, p. 63).

In the present paper, the set of states is given by the Cartesian product of individual state sets Ω_ν, $\nu \in V$, where V denotes the set of *units*. In the following, the unit set and the corresponding state sets are assumed to be non-empty and finite. For a subsystem $S \subset V$, $\Omega_S := \prod_{\nu \in S} \Omega_\nu$ denotes the set of all configurations on S. The elements of $\bar{\mathcal{P}}(\Omega_S)$ are the *random fields* on S. One has the natural restriction $X_S : \Omega_V \to \Omega_S$, $\omega = (\omega_\nu)_{\nu \in V} \mapsto \omega_S := (\omega_\nu)_{\nu \in S}$, which induces the projection $\bar{\mathcal{P}}(\Omega_V) \to \bar{\mathcal{P}}(\Omega_S)$, $p \mapsto p_S$, where p_S denotes the image measure of p under the variable X_S. If the subsystem S consists of exactly one unit ν, we write p_ν instead of $p_{\{\nu\}}$.

The following example, which allows to put the definition (2) into the information geometric setting, represents the main motivation for the present approach to stochastic interaction. It will be generalized in Section 4.

Example 2.1. (Factorizable Distributions and Spatial Interdependence) Let V be a finite set of units and Ω_ν, $\nu \in V$, corresponding state sets. Consider the *tensorial map*

$$\prod_{\nu \in V} \mathcal{P}(\Omega_\nu) \;\hookrightarrow\; \mathcal{P}(\Omega_V), \qquad (p^{(\nu)})_{\nu \in V} \;\mapsto\; \otimes_{\nu \in V} p^{(\nu)},$$

with

$$\left(\otimes_{\nu \in V} p^{(\nu)}\right)(\omega) \;:=\; \prod_{\nu \in V} p^{(\nu)}(\omega_\nu)$$

The image $\mathcal{F} := \mathcal{F}(\Omega_V) := \left\{\otimes_{\nu \in V} p^{(\nu)} \;:\; p^{(\nu)} \in \mathcal{P}(\Omega_\nu),\;\; \nu \in V\right\}$ of this map, which consists of all factorizable and strictly positive probability distributions, is

an exponential family in $\mathcal{P}(\Omega_V)$ with $\dim \mathcal{F} = \sum_{\nu \in V}(|\Omega_\nu| - 1)$. For the particular case of binary units, that is $|\Omega_\nu| = 2$ for all ν, the dimension of $\mathcal{F}$ is equal to the number $|V|$ of units. The following statement is well known [2]:

The (-1)-projection of a distribution $p \in \mathcal{P}(\Omega_V)$ on $\mathcal{F}$ is given by $\otimes_{\nu \in V} p_\nu$ (the p_ν, $\nu \in V$, are the marginal distributions), and one has the representation

$$I(p) \;=\; D(p \,\|\, \mathcal{F}) \;=\; \sum_{\nu \in V} H(p_\nu) - H(p),$$

where H denotes the *Shannon entropy* [19]. As stated in the introduction, $I(p)$ is a measure for the spatial interdependencies of the units. It vanishes exactly when the units are stochastically independent.

Bofore extending the spatial notion of interaction to a dynamical one, in Section 3 we consider the more general concept of separability of transition kernels.

3 Quantifying Non-Separability

3.1 Manifolds of Separable Transition Kernels

Consider a finite set V of units, corresponding state sets Ω_ν, $\nu \in V$, and two subsets $A, B \subset V$ with $B \neq \emptyset$. A function

$$K : \Omega_A \times \Omega_B \to [0,1], \qquad (\omega, \omega') \mapsto K(\omega' \,|\, \omega),$$

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

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

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

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

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

For two random variables X, Y with $\mathrm{Prob}\{X = \omega\} = p(\omega)$ for all $\omega \in \Omega_A$, and $\mathrm{Prob}\{Y = \omega' \,|\, X = \omega\} = K(\omega' \,|\, \omega)$ for all $\omega \in \Omega_A$ with $p(\omega) > 0$ and all $\omega' \in \Omega_B$, we set $H(Y \,|\, X) := H(p, K)$.

In the present paper, the set $\bar{\mathcal{K}}(\Omega_V)$ is interpreted as a model for the dynamics of interacting units, and the information flow associated with this dynamics is studied in Section 4. In the present section, we introduce a general notion of separability of transition kernels in order to capture all examples that are discussed in the paper in a unified way.

Consider a family $\mathscr{S} := \{(A_1, B_1), (A_2, B_2), \ldots, (A_n, B_n)\}$ where the A_i and B_i are subsets of V. We assume that $\{B_1, \ldots, B_n\}$ is a partition of V, that is $B_i \neq \emptyset$ for all i, $B_i \cap B_j = \emptyset$ for all $i \neq j$, and $V = B_1 \uplus \cdots \uplus B_n$. Now consider the corresponding *tensorial map*

$$\otimes_{\mathscr{S}} : \prod_{(A,B)\in\mathscr{S}} \mathcal{K}(\Omega_B \,|\, \Omega_A) \;\hookrightarrow\; \mathcal{K}(\Omega_V), \quad \big(K_B^A\big)_{(A,B)\in\mathscr{S}} \;\mapsto\; \otimes_{(A,B)\in\mathscr{S}} K_B^A$$

with

$$\big(\otimes_{(A,B)\in\mathscr{S}} K_B^A\big)\,(\omega' \,|\, \omega) \;:=\; \prod_{(A,B)\in\mathscr{S}} K_B^A(\omega'_B \,|\, \omega_A), \qquad \text{for all} \quad \omega, \omega' \in \Omega_V.$$

The image $\mathcal{K}_{\mathscr{S}}(\Omega_V)$ of $\otimes_{\mathscr{S}}$ is a submanifold of $\mathcal{K}(\Omega_V)$ with

$$\dim \mathcal{K}_{\mathscr{S}}(\Omega_V) \;=\; \sum_{(A,B)\in\mathscr{S}} |\Omega_A|\big(|\Omega_B| - 1\big).$$

Its elements are the *separable* transition kernels with respect to $\mathscr{S}$.

Here are the most important examples:

Examples and Definitions 3.1.

(1) If we set $\mathscr{S} := \{(V, V)\}$, the tensorial map is nothing but the identity $\mathcal{K}(\Omega_V) \to \mathcal{K}(\Omega_V)$, and therefore one has $\mathcal{K}_{\mathscr{S}}(\Omega_V) = \mathcal{K}(\Omega_V)$.

(2) Consider the case where no temporal information is transmitted but all spatial information: $\mathscr{S} := ind := \{(\emptyset, V)\}$. In that case the tensorial map $\otimes_{\mathscr{S}}$ reduces to the natural embedding

$$\mathcal{K}(\Omega_V \,|\, \Omega_\emptyset) = \mathcal{P}(\Omega_V) \hookrightarrow \mathcal{K}(\Omega_V)$$

which assigns to each probability distribution p the kernel

$$K(\omega' \,|\, \omega) := p(\omega'), \qquad \omega, \omega' \in \Omega_V.$$

Therefore, we write $\mathcal{K}_{ind}(\Omega_V) = \mathcal{P}(\Omega_V)$.

(3) In addition to the splitting in time which is described in example (2), consider also a complete splitting in space: $\mathscr{S} := fac := \{(\emptyset, \{\nu\}) \ : \ \nu \in V\}$. Then we recover the tensorial map of Example 2.1. Thus, $\mathcal{K}_{fac}(\Omega_V)$ can be identified with $\mathcal{F}(\Omega_V)$.

(4) To model the important class of *parallel information processing*, we set $\mathscr{S} := par := \{(V, \{\nu\}) \ : \ \nu \in V\}$. Here, each unit "computes" its new state on the basis of all current states according to a kernel $K^{(\nu)} \in \mathcal{K}(\Omega_\nu \,|\, \Omega_V)$. The transition from a configuration $\omega = (\omega_\nu)_{\nu \in V}$ of the whole system to a new configuration $\omega' = (\omega'_\nu)_{\nu \in V}$ is done according to the following composed kernel in $\mathcal{K}(\Omega_V)$:

$$K(\omega' \,|\, \omega) \ = \ \prod_{\nu \in V} K^{(\nu)}({\omega'}_\nu \,|\, \omega), \qquad \omega, \omega' \in \Omega_V.$$

(5) In applications, parallel processing is adapted to a graph $G = (V, E)$ – here, $E \subset V \times V$ denotes the set of edges – in order to model constraints for the information flow in the system. This is represented by $\mathscr{S} := \mathscr{S}(G) := \{(\mathrm{pa}(\nu), \{\nu\}) \ :$

$\nu \in V\}$. Each unit ν is supposed to process only information from its parents $\mathrm{pa}(\nu) = \{\mu \in V \ : \ (\mu,\nu) \in E\}$, which is modelled by a transition kernel $K^{(\nu)} \in \mathcal{K}(\Omega_\nu \,|\, \Omega_{\mathrm{pa}(\nu)})$. The parallel transition of the whole system is then described by

$$K(\omega' \,|\, \omega) \ = \ \prod_{\nu \in V} K^{(\nu)}({\omega'}_\nu \,|\, \omega_{\mathrm{pa}(\nu)}), \qquad \omega, \omega' \in \Omega_V .$$

(6) Now, we introduce the example of parallel processing that plays the most important role in the present paper: Consider non-empty and pairwise distinct subsystems $S_1, \ldots, S_n$ of V with $V = S_1 \uplus \cdots \uplus S_n$ and define $\mathscr{S} := \mathscr{S}(S_1, \ldots, S_n) := \{(S_i, S_i) \ : \ i = 1, \ldots, n\}$. It describes $\{S_1, \ldots, S_n\}$-*split information processing*, where the subsystems do not interact with each other. Each subsystem S_i only processes information from its own current state according to a kernel $K^{(i)} \in \mathcal{K}(\Omega_{S_i})$. The composed transition of the whole system is then given by

$$K(\omega' \,|\, \omega) \ = \ \prod_{i=1}^{n} K^{(i)}({\omega'}_{S_i} \,|\, \omega_{S_i}), \qquad \omega, \omega' \in \Omega_V .$$

For the completely split case, where the subsystems are the elementary units, we define $\mathit{spl} := \mathscr{S}(\{\nu\},\ \nu \in V) = \{(\{\nu\}, \{\nu\}) \ : \ \nu \in V\}$.

3.2 Non-Separability as Divergence from Separability

Consider a Markov chain $X_n = (X_{\nu,n})_{\nu \in V}$, $n = 0, 1, 2, \ldots$, that is given by an initial distribution $p \in \bar{\mathcal{P}}(\Omega_V)$ and a kernel $K \in \bar{\mathcal{K}}(\Omega_V)$. The probabilistic

properties of this stochastic process are determined by the following set of finite marginals:

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

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

$$\overline{\mathrm{MC}}(\Omega_V) := \bar{\mathcal{P}}(\Omega_V) \times \bar{\mathcal{K}}(\Omega_V),$$

and we also use the notation $\{X_n\} = \{X_0, X_1, X_2, \ldots\}$ instead of (p, K). The interior $\mathrm{MC}(\Omega_V)$ of the set of Markov chains carries the natural dualistic structure from $\mathcal{P}(\Omega_V \times \Omega_V)$, which is induced by the diffeomorphic composition map $\otimes : \mathrm{MC}(\Omega_V) \to \mathcal{P}(\Omega_V \times \Omega_V)$,

$$(p, K) \mapsto p \otimes K, \quad \text{with} \quad (p \otimes K)(\omega, \omega') \;:=\; p(\omega)\, K(\omega' \,|\, \omega).$$

($\otimes$ can be extended to a continuous surjective map $\overline{\mathrm{MC}}(\Omega_V) \to \bar{\mathcal{P}}(\Omega_V \times \Omega_V)$). Thus, we can talk about exponential families and (-1)-projections in $\mathrm{MC}(\Omega_V)$. The "distance" $D((p, K) \,\|\, (p', K'))$ from a Markov chain (p, K) to another one (p', K') is given by

$$D(p \otimes K \,\|\, p' \otimes K') \;=\; D(p \,\|\, p') + D_p(K \,\|\, K'),$$

with

$$D_p(K \,\|\, K') \;:=\; \sum_{\omega \in \Omega} p(\omega)\, D\big(K(\cdot \,|\, \omega) \,\|\, K'(\cdot \,|\, \omega)\big).$$

For a set $\mathscr{S} = \{(A_1, B_1), (A_2, B_2), \ldots, (A_n, B_n)\}$, we introduce the exponential family (see Proposition 6.1)

$$\mathrm{MC}_{\mathscr{S}}(\Omega_V) \;:=\; \mathcal{P}(\Omega_V) \times \mathcal{K}_{\mathscr{S}}(\Omega_V) \;\subset\; \mathrm{MC}(\Omega_V),$$

which has dimension $\big(|\Omega_V| - 1\big) + \sum_{(A,B)\in\mathscr{S}} |\Omega_A|\big(|\Omega_B| - 1\big)$.

The set of all these exponential families is partially ordered by inclusion with $\mathrm{MC}(\Omega_V)$ as the greatest element and $\mathrm{MC}_{fac}(\Omega_V)$ as the least one. This ordering is connected with the following partial ordering $\preceq$ of the sets $\mathscr{S}$:
Given $\mathscr{S} = \{(A_1, B_1), \ldots, (A_m, B_m)\}$ and $\mathscr{S}' = \{(A'_1, B'_1), \ldots, (A'_n, B'_n)\}$, we write $\mathscr{S} \preceq \mathscr{S}'$ ($\mathscr{S}'$ *coarser* than $\mathscr{S}$) iff for all $(A, B) \in \mathscr{S}$ there exists a pair $(A', B') \in \mathscr{S}'$ with $A \subset A'$ and $B \subset B'$. One has

$$\mathscr{S} \;\preceq\; \mathscr{S}' \quad\Rightarrow\quad \mathcal{K}_{\mathscr{S}}(\Omega_V) \;\subseteq\; \mathcal{K}_{\mathscr{S}'}(\Omega_V). \tag{4}$$

Thus, coarsening enlarges the corresponding manifold (The proof is given in the appendix).

Now, we describe the (-1)-projections on the exponential families $\mathrm{MC}_{\mathscr{S}}(\Omega_V)$:

Proposition 3.2. *Let (p, K) be a Markov chain in $\mathrm{MC}(\Omega_V)$ and $\mathscr{S} \preceq \mathscr{S}'$. Then:*

(i) (Projection) *The (-1)-projection of (p, K) on $\mathrm{MC}_{\mathscr{S}}(\Omega_V)$ is given by $(p, K_{\mathscr{S}})$*

with $K_{\mathscr{S}} := \otimes_{(A,B)\in\mathscr{S}} K_B^A$. *Here, the kernels* $K_B^A \in \mathcal{K}(\Omega_B \,|\, \Omega_A)$ *denote the corresponding* marginals *of* K*:*

$$K_B^A(\omega' \,|\, \omega) \;:=\; \frac{\sum_{\substack{\sigma,\sigma'\in\Omega_V \\ \sigma_A=\omega,\ \sigma'_B=\omega'}} p(\sigma)\, K(\sigma' \,|\, \sigma)}{\sum_{\substack{\sigma\in\Omega_V \\ \sigma_A=\omega}} p(\sigma)}, \qquad \omega \in \Omega_A,\ \omega' \in \Omega_B.$$

$K_{\mathscr{S}}$ *is the* projection of K on $\mathcal{K}_{\mathscr{S}}(\Omega_V)$ with respect to p.

(ii) (Entropic Representation) *The corresponding divergence is given by*

$$\begin{aligned} D\big((p,K) \,\|\, \mathrm{MC}_{\mathscr{S}}(\Omega_V)\big) \;&=\; D_p(K \,\|\, K_{\mathscr{S}}) \\ &=\; \sum_{(A,B)\in\mathscr{S}} H\big(p_A, K_B^A\big) \;-\; H(p,K). \end{aligned}$$

(iii) (Pythagorian Relation) *One has*

$$D_p(K \,\|\, K_{\mathscr{S}}) \;=\; D_p(K \,\|\, K_{\mathscr{S}'}) + D_p(K_{\mathscr{S}'} \,\|\, K_{\mathscr{S}}).$$

If $K \in \mathcal{P}(\Omega_V)$, that is $K(\omega' \,|\, \omega) = p(\omega)$, $\omega, \omega' \in \Omega_V$, with a probability distribution $p \in \mathcal{P}(\Omega_V)$, then the divergence $D_p(K \,\|\, K_{fac})$ is nothing but the measure $I(p)$ for spatial interdependencies that has been discussed in the introduction and in Example 2.1. More generally, we interpret the divergence $D_p(K \,\|\, K_{\mathscr{S}})$ as a natural measure for the non-separability of (p,K) with respect to $\mathscr{S}$. The corresponding function $I_{\mathscr{S}} : (p,K) \mapsto I_{\mathscr{S}}(p,K) := D_p(K \,\|\, K_{\mathscr{S}})$ has a unique continuous extension to the set $\overline{\mathrm{MC}}(\Omega_V)$ of all Markov chains which is also denoted by $I_{\mathscr{S}}$ (see Lemma 4.2 in [4]). Thus, non-separability is defined for not necessarily strictly positive Markov chains.

4 Application to Stochastic Interaction

4.1 The Definition of Stochastic Interaction

As stated in the introduction we use the concept of complexity that is described by the formal definition (1) in order to define stochastic interaction.

Let V be a set of units and Ω_ν, $\nu \in V$, corresponding state sets. Furthermore, consider non-empty and pairwise distinct subsystems $S_1, \ldots, S_n \subset V$ with $V = S_1 \uplus \cdots \uplus S_n$. The stochastic interaction of $S_1, \ldots, S_n$ with respect to $(p, K) \in \overline{\mathrm{MC}}(\Omega_V)$ is quantified by the divergence of (p, K) from the set of Markov chains that represent $\{S_1, \ldots, S_n\}$-split information processing, where the subsystems do not interact with each other (see Examples and Definitions 3.1 (6)). More precisely, we define the *stochastic interaction* (*of the subsystems* $S_1, \ldots, S_n$) to be the function $I_{S_1,\ldots,S_n} : \overline{\mathrm{MC}}(\Omega_V) \to \mathbb{R}_+$ with

$$I_{S_1,\ldots,S_n}(p, K) \;:=\; I_{\mathscr{S}(S_1,\ldots,S_n)}(p, K) \;=\; \inf_{K' \ \{S_1,\ldots,S_n\}\text{-split}} D_p(K \,\|\, K'). \qquad (5)$$

In the case of complete splitting of $V = \{\nu_1, \ldots, \nu_n\}$ into the elementary units, that is $S_i := \{\nu_i\}$, $i = 1, \ldots, n$, we simply write I instead of $I_{\{\nu_1\},\ldots,\{\nu_n\}}$.

The definition of stochastic interaction given by (5) is consistent with the complexity concept that is discussed in the introduction.

Here are some basic properties of I, which are well known in the spatial setting of Example 2.1:

PROPOSITION 4.1. *Let V be a set of units, Ω_ν, $\nu \in V$, corresponding state sets, and $X_n = (X_{\nu,n})_{\nu\in V}$, $n = 0, 1, 2, \ldots$, a Markov chain on Ω_V. For a subsystem $S \subset V$, we write $X_{S,n} := (X_{\nu,n})_{\nu\in S}$. Assume that the chain is given by $(p, K) \in \overline{\mathrm{MC}}(\Omega_V)$, where p is a stationary distribution with respect to K. Then the following holds:*

(i)

$$I\{X_n\} \;=\; \sum_{\nu\in V} H(X_{\nu,n+1} \,|\, X_{\nu,n}) \;-\; H(X_{n+1} \,|\, X_n). \tag{6}$$

(ii) $A, B \subset V,\;\; A, B \neq \emptyset,\;\; A \cap B = \emptyset,\;\; A \uplus B = V \quad \Rightarrow$

$$I\{X_n\} \;=\; I\{X_{A,n}\} + I\{X_{B,n}\} + I_{A,B}\{X_n\}$$

(iii) If the process is parallel, then

$$\begin{aligned} I\{X_n\} \;&=\; \sum_{\nu\in V} \Big(H(X_{\nu,n+1} \,|\, X_{\nu,n}) - H(X_{\nu,n+1} \,|\, X_n) \Big) \\ &=\; \sum_{\nu\in V} MI(X_{\nu,n+1}; X_{V\setminus\nu,n} \,|\, X_{\nu,n}). \end{aligned} \tag{7}$$

(iv) If the process is adapted to a graph (V, E) then

$$\begin{aligned} I\{X_n\} \;&=\; \sum_{\nu\in V} \Big(H(X_{\nu,n+1} \,|\, X_{\nu,n}) - H(X_{\nu,n+1} \,|\, X_{\mathrm{pa}(\nu),n}) \Big) \\ &=\; \sum_{\nu\in V} MI(X_{\nu,n+1}; X_{\mathrm{pa}(\nu)\setminus\nu,n} \,|\, X_{\nu,n}). \end{aligned} \tag{8}$$

In the statements (iii) and (iv), the *conditional mutual information* $MI(X; Y \,|\, Z)$ of two random variables X, Y with respect to a third one Z is defined to be the difference $H(X \,|\, Z) - H(X \,|\, Y, Z)$ (see [10], p. 22).

If X_{n+1} is independent from X_n for all n, the stochastic interaction $I\{X_n\}$ reduces to the measure $I(p)$ for spatial interdependencies with respect to the stationary distribution p of $\{X_n\}$ (see Example 2.1). Thus, the dynamical notion of stochastic interaction is a generalization of the spatial one. Geometrically, this can be illustrated as follows.

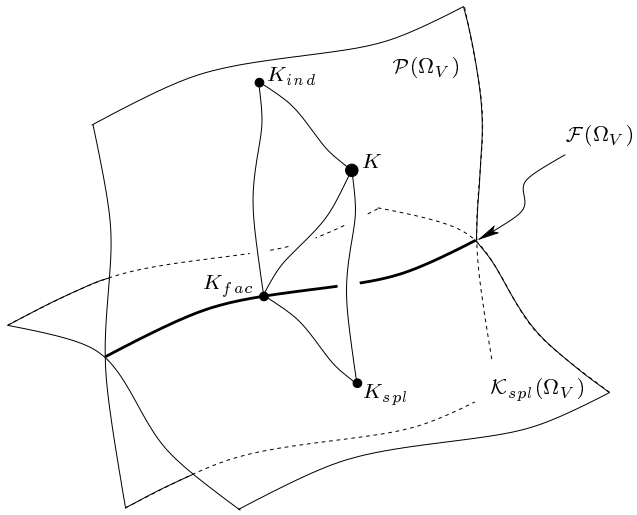

FIG. 2.

In addition to the projection K_{spl} of the kernel $K \in \mathrm{MC}(\Omega_V)$ with respect to a distribution $p \in \mathcal{P}(\Omega_V)$ on the set of split kernels, we consider its projections K_{ind} and K_{fac} on the set $\mathcal{P}(\Omega_V)$ of independent kernels and on the subset $\mathcal{F}(\Omega_V)$, respectively. From Proposition 3.2 we know

$$D_p(K \,\|\, K_{ind}) \;=\; H(X_{n+1}) - H(X_{n+1} \,|\, X_n)$$

$$((global)\ transinformation)$$

$$I(p) \;=\; D_p(K_{ind} \,\|\, K_{fac}) \;=\; \sum_{\nu \in V} H(X_{\nu,\,n+1}) \;-\; H(X_{n+1})$$
$$(\textit{spatial interdependence})$$

$$D_p(K_{spl} \,\|\, K_{fac}) \;=\; \sum_{\nu \in V} \big(H(X_{\nu,\,n+1}) - H(X_{\nu,\,n+1} \,|\, X_{\nu,\,n}) \big)$$
$$(\textit{sum of individual transinformations}).$$

According to the Pythagorian relation (Proposition 3.2 (iii)), we get the following representation of stochastic interaction:

$$\begin{aligned} I\{X_n\} \;&=\; D_p(K \,\|\, K_{spl}) \\ &=\; I(p) \;+\; D_p(K \,\|\, K_{ind}) - D_p(K_{spl} \,\|\, K_{fac}). \end{aligned} \tag{9}$$

In the particular case of an independent process, the divergences $D_p(K \,\|\, K_{ind})$ and $D_p(K_{spl} \,\|\, K_{fac})$ in (9) vanish, and the stochastic interaction coincides with spatial interdependence.

4.2 Some Examples

Example 4.2. (Source and Receiver) Consider two units $1 = \textit{source}$ and $2 = \textit{receiver}$ with the state sets Ω_1 and Ω_2. Assume that the information flow is adapted to the graph $G = \{\{1,2\},\{(1,2)\}\}$, which only allows a transmission from the first unit to the second. In each transition from time n to $n+1$, a state $X_{1,\,n+1}$ of the first unit is chosen independently from $X_{1,\,n}$ according to a probability distribution $p \in \mathcal{P}(\Omega_1)$. The state $X_{2,\,n+1}$ of the second unit at time

$n+1$ is "computed" from $X_{1,n}$ according to a kernel $K \in \mathcal{K}(\Omega_2 \,|\, \Omega_1)$. Using formula (8), we have

$$I\{X_n\} \;=\; H(X_{2,n+1}) - H(X_{2,n+1} \,|\, X_{1,n}).$$

This is the well-known *mutual information* of the variables $X_{2,n+1}$ and $X_{1,n}$, which has a temporal interpretation within the present approach. It plays an important role in *coding and information theory* [10], [18].

EXAMPLE 4.3. (TWO BINARY UNITS I) Consider two units with the state sets $\{0,1\}$. Each unit copies the state of the other unit with probability $1-\varepsilon$. The transition probabilities for the units are given by the following tables:

$K^{(1)}(x' \,\|\, (x,y))$	0	1
$(0,0)$	$1-\varepsilon$	ε
$(0,1)$	ε	$1-\varepsilon$
$(1,0)$	$1-\varepsilon$	ε
$(1,1)$	ε	$1-\varepsilon$

$K^{(2)}(y' \,\|\, (x,y))$	0	1
$(0,0)$	$1-\varepsilon$	ε
$(0,1)$	$1-\varepsilon$	ε
$(1,0)$	ε	$1-\varepsilon$
$(1,1)$	ε	$1-\varepsilon$

The transition kernel $K \in \bar{\mathcal{K}}_{par}(\{0,1\} \times \{0,1\})$ for the corresponding parallel dynamics of the whole system is then given by

$K((x',y') \mid (x,y))$	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$(0,0)$	$(1-\varepsilon)^2$	$(1-\varepsilon)\varepsilon$	$\varepsilon(1-\varepsilon)$	ε^2
$(0,1)$	$\varepsilon(1-\varepsilon)$	ε^2	$(1-\varepsilon)^2$	$(1-\varepsilon)\varepsilon$
$(1,0)$	$(1-\varepsilon)\varepsilon$	$(1-\varepsilon)^2$	ε^2	$\varepsilon(1-\varepsilon)$
$(1,1)$	ε^2	$\varepsilon(1-\varepsilon)$	$(1-\varepsilon)\varepsilon$	$(1-\varepsilon)^2$

Note that for $\varepsilon \in \{0,1\}$, K corresponds to the deterministic transformations

$$\varepsilon = 0: \quad (x,y) \mapsto (y,x), \qquad \text{and} \qquad \varepsilon = 1: \quad (x,y) \mapsto (1-y, 1-x),$$

which in an intuitive sense describe complete information exchange of the units. With the unique stationary probability distribution $p = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ one can easily compute the marginal kernels

$K_1(x' \mid x)$	0	1
0	$\frac{1}{2}$	$\frac{1}{2}$
1	$\frac{1}{2}$	$\frac{1}{2}$

$K_2(y' \mid y)$	0	1
0	$\frac{1}{2}$	$\frac{1}{2}$
1	$\frac{1}{2}$	$\frac{1}{2}$

which describe the split dynamics according to $K_{spl} = K_1 \otimes K_2$. With (7) we finally get

$$I\{X_n\} \;=\; 2\Big(\ln 2 + (1-\varepsilon)\ln(1-\varepsilon) + \varepsilon \ln \varepsilon\Big).$$

The shape of this function is shown in the following picture:

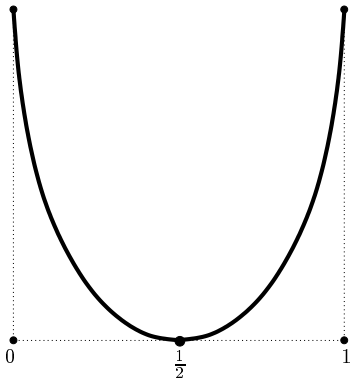

FIG. 3.

This function is symmetric around $\varepsilon = \frac{1}{2}$ where it vanishes. In $\varepsilon = 0$ and $\varepsilon = 1$ it attains its maximal value $2 \ln 2$. As stated above, this corresponds to the deterministic transformations with complete information exchange.

EXAMPLE 4.4. (TWO BINARY UNITS II) Consider again two binary units with the state sets $\{0, 1\}$ and the transition probabilities

$K^{(1)}(x' \mid (x, y))$	0	1
$(0, 0)$	1	0
$(0, 1)$	$1 - \varepsilon$	ε
$(1, 0)$	ε	$1 - \varepsilon$
$(1, 1)$	0	1

$K^{(2)}(y' \mid (x, y))$	0	1
$(0, 0)$	0	1
$(0, 1)$	$1 - \varepsilon$	ε
$(1, 0)$	ε	$1 - \varepsilon$
$(1, 1)$	1	0

The transition kernel $K \in \bar{\mathcal{K}}(\{0, 1\} \times \{0, 1\})$ of the corresponding parallel dynamics is given by

$K((x',y') \mid (x,y))$	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$(0,0)$	0	1	0	0
$(0,1)$	$(1-\varepsilon)^2$	$(1-\varepsilon)\varepsilon$	$\varepsilon(1-\varepsilon)$	ε^2
$(1,0)$	ε^2	$\varepsilon(1-\varepsilon)$	$(1-\varepsilon)\varepsilon$	$(1-\varepsilon)^2$
$(1,1)$	0	0	1	0

Note that for $\varepsilon \in \{0,1\}$, K corresponds to the deterministic transformations

$$\varepsilon = 0: \ \ (x,y) \mapsto (x, 1-y), \qquad \text{and} \qquad \varepsilon = 1: \ \ (x,y) \mapsto (y, 1-x).$$

Thus in an intuitive sense, for $\varepsilon = 1$ the units completely interact with each other, and for $\varepsilon = 0$ there is no interaction. For $\varepsilon \in]0,1[$ we compute the interaction with respect to the unique stationary probability distribution

$$p \ = \ \tfrac{1}{4(\varepsilon^2-\varepsilon+1)}\,(2\varepsilon^2 - 2\varepsilon + 1 \ , \ 1 \ , \ 1 \ , \ 2\varepsilon^2 - 2\varepsilon + 1).$$

With the corresponding marginal kernels

$K_1(x' \mid x)$	0	1
0	$1 - \frac{\varepsilon}{2(\varepsilon^2-\varepsilon+1)}$	$\frac{\varepsilon}{2(\varepsilon^2-\varepsilon+1)}$
1	$\frac{\varepsilon}{2(\varepsilon^2-\varepsilon+1)}$	$1 - \frac{\varepsilon}{2(\varepsilon^2-\varepsilon+1)}$

$K_2(y' \mid y)$	0	1
0	$\frac{\varepsilon}{2(\varepsilon^2-\varepsilon+1)}$	$1 - \frac{\varepsilon}{2(\varepsilon^2-\varepsilon+1)}$
1	$1 - \frac{\varepsilon}{2(\varepsilon^2-\varepsilon+1)}$	$\frac{\varepsilon}{2(\varepsilon^2-\varepsilon+1)}$

and formula (7), we get

$$\begin{aligned} I\{X_n\} \ &= \ \frac{\varepsilon}{\varepsilon^2-\varepsilon+1}\Big(-(2\varepsilon^2-3\varepsilon+2)\ln(2\varepsilon^2-3\varepsilon+2) \\ &\qquad + 2(\varepsilon^2-\varepsilon+1)\ln 2(\varepsilon^2-\varepsilon+1) \ + \ (1-\varepsilon)\ln(1-\varepsilon)\Big) \end{aligned}$$

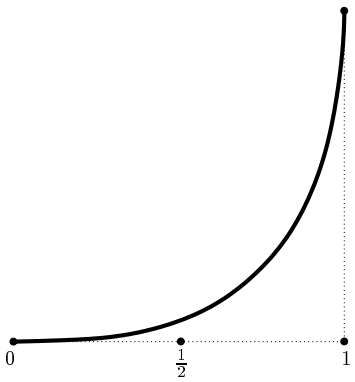

FIG. 4.

This function is monotonically increasing from the minimal value 0 (no interaction) in $\varepsilon = 0$ to its maximal value $2\ln 2$ (complete interaction) in $\varepsilon = 1$.

5 Conclusion

Following the general concept that complexity is characterized by the divergence of a composed system from the superposition of its elementary parts, information geometry has been used to derive a measure for spatio-temporal interdependencies among a finite set of units, which is referred to as *stochastic interaction*. This generalizes the well-known measure for spatial interdependence that is quantified by the Kullback-Leibler divergence of a probability distribution from its factorization [2], [4]. Thereby, previous work by Ay [5] is continued, where the optimization of dependencies among stochastic units has been proposed as a principle for neural organization in feed-forward networks. Of course, the present setting is much more general and provides a way to consider also recurrent networks (This work is in progress). The dynamical properties of strongly interacting units in the sense of the present paper are studied by Ay and Wennekers in [6], where the emergence of determinism and structure in such systems is demonstrated.

6 Appendix: Proofs

PROPOSITION 6.1. *The manifold* $\mathrm{MC}_{\mathscr{S}}(\Omega_V)$ *is an exponential family in* $\mathrm{MC}(\Omega_V)$.

PROOF. To see this, consider the functions $\Omega_V \times \Omega_V \to \mathbb{R}$

$$v_\sigma(\omega, \omega') := \begin{cases} 1, & \text{if } \omega = \sigma \\ 0, & \text{otherwise} \end{cases}, \qquad \sigma \in \Omega_V,$$

and

$$v_{\sigma,\sigma'}(\omega, \omega') := \begin{cases} 1, & \text{if } \omega_A = \sigma,\ \omega'_B = \sigma' \\ 0, & \text{otherwise} \end{cases}, \qquad (A,B) \in \mathscr{S},\ \sigma \in \Omega_A,\ \sigma' \in \Omega_B.$$

It is easy to verify that the image of $\mathrm{MC}_{\mathscr{S}}(\Omega_V)$ under the map $\otimes$ is the following exponential family in $\mathcal{P}(\Omega_V \times \Omega_V)$:

$$\exp\left\{ \sum_{\sigma \in \Omega_V} \lambda_\sigma\ v_\sigma + \sum_{(A,B)\in\mathscr{S}} \sum_{\sigma\in\Omega_A,\ \sigma'\in\Omega_B} \lambda_{\sigma,\sigma'}\ v_{\sigma,\sigma'}\ -\ \Theta \right\}, \qquad \lambda_\sigma, \lambda_{\sigma,\sigma'} \in \mathbb{R}.$$

Here, Θ denotes the normalization factor, which depends on the λ-parameters. In particular, each element in $\mathrm{MC}_{\mathscr{S}}(\Omega_V)$ can be expressed in the following way

$$\begin{aligned} & p(\omega) \prod_{(A,B)\in\mathscr{S}} K_B^A(\omega'_B \,|\, \omega_A) \\ = \ & \exp\left\{ \ln p(\omega) + \sum_{(A,B)\in\mathscr{S}} \ln K_B^A(\omega'_B \,|\, \omega_A) \right\} \\ = \ & \exp\left\{ \sum_{\sigma\in\Omega_V} \ln p(\sigma)\ v_\sigma(\omega,\omega') + \sum_{(A,B)\in\mathscr{S}} \sum_{\sigma\in\Omega_A,\ \sigma'\in\Omega_B} \ln K_B^A(\sigma' \,|\, \sigma)\ v_{\sigma,\sigma'}(\omega,\omega') \right\}. \end{aligned}$$

□

PROOF OF IMPLICATION (4). If

$$\mathscr{S} = \{(A_1, B_1), \ldots, (A_m, B_m)\} \preceq \mathscr{S}'\{(A'_1, B'_1), \ldots, (A'_n, B'_n)\},$$

then there exists a partition M_i, $i = 1, \ldots, n$, of the index set $\{1, \ldots, m\}$ such that

$$B'_i = \biguplus_{j \in M_i} B_j, \qquad i = 1, \ldots, n.$$

Let (p, K) be a Markov chain in $\mathrm{MC}_{\mathscr{S}}(\Omega_V)$. Then there exist $K_B^A \in \mathcal{K}(\Omega_B \,|\, \Omega_A)$ with

$$\begin{aligned} K(\omega' \,|\, \omega) &= \prod_{(A,B)\in\mathscr{S}} K_B^A(\omega'_B \,|\, \omega_A) \\ &= \prod_{i=1}^{n} \underbrace{\prod_{\substack{(A_j,B_j)\in\mathscr{S} \\ j\in M_i}} K_{B_j}^{A_j}(\omega'_{B_j} \,|\, \omega_{A_j})}_{=:\, K_{B'_i}^{A'_i}(\omega'_{B'_i} \,|\, \omega_{A'_i})}, \qquad \omega, \omega' \in \Omega_V. \end{aligned}$$

The kernels $K_{B'_i}^{A'_i}$ are contained in $\mathcal{K}_{\mathscr{S}'}$, and therefore we get $(p, K) \in \mathrm{MC}_{\mathscr{S}'}(\Omega_V)$.

□

PROOF OF PROPOSITION 3.2.

(i) Consider the following strictly convex function ($\mathbb{R}_+^*$ denotes the set of positive real numbers)

$$F : \left(\mathbb{R}_+^*\right)^{\Omega_V} \times \left(\prod_{(A,B)\in\mathscr{S}} \left(\mathbb{R}_+^*\right)^{\Omega_A \times \Omega_B} \right) \to \mathbb{R},$$

$$(x, y) = (x_\omega,\, \omega \in \Omega_V; y_{\omega_A,\omega_B},\, \omega_A \in \Omega_A, \omega_B \in \Omega_B) \mapsto$$

$$F(x,y) \;:=\; \sum_{\omega\in\Omega_V} p(\omega)\ln\frac{p(\omega)}{x_\omega} \;+\; \sum_{\omega,\,\omega'\in\Omega_V} p(\omega)\,K(\omega'\,|\,\omega)\ln\frac{K(\omega'\,|\,\omega)}{\prod_{(A,B)\in\mathscr{S}} y_{\omega_A,\omega'_B}}$$
$$+\,\lambda\left(\sum_{\omega\in\Omega_V} x_\omega \;-\; 1\right) \;+\; \sum_{(A,B)\in\mathscr{S}}\sum_{\omega_A\in\Omega_A}\lambda^B_{\omega_A}\left(\sum_{\omega'_B\in\Omega_B} y_{\omega_A,\omega'_B} - 1\right).$$

Here, λ and the $\lambda^B_{\omega_A}$ are Lagrangian parameters. Note that in the case $x\in\mathcal{P}(\Omega_V)$ and $y\in\prod_{(A,B)\in\mathscr{S}}\mathcal{K}(\Omega_B\,|\,\Omega_A)$, the value $F(x,y)$ is nothing but the divergence of (p,K) from $(x,\otimes_{\mathscr{S}}(y))$. In order to get the Markov chain that minimizes the divergence we have to compute the partial derivatives of F:

$$\begin{aligned}\frac{\partial F}{\partial x_\sigma}(x,y) &= -\sum_{\omega\in\Omega_V} p(\omega)\,\frac{1}{x_\omega}\,\delta_{\sigma,\,\omega} \;+\; \lambda\\ &= -\frac{p(\sigma)}{x_\sigma} \;+\; \lambda,\end{aligned}$$

and

$$\begin{aligned}\frac{\partial F}{\partial y_{\sigma_C,\sigma'_D}}(x,y) &= -\sum_{\omega,\omega'\in\Omega_V} p(\omega)\,K(\omega'\,|\,\omega)\sum_{(A,B)\in\mathscr{S}}\frac{1}{y_{\omega_A,\omega'_B}}\delta_{(\omega_A,\omega'_B),(\sigma_C,\sigma'_D)}\\ &\quad+\sum_{(A,B)\in\mathscr{S}}\sum_{\omega_A\in\Omega_A}\lambda^B_{\omega_A}\sum_{\omega'_B\in\Omega_B}\delta_{(\omega_A,\omega'_B),(\sigma_C,\sigma'_D)}\\ &= -\sum_{\omega,\omega'\in\Omega_V} p(\omega)\,K(\omega'\,|\,\omega)\frac{1}{y_{\omega_C,\omega'_D}}\delta_{(\omega_C,\omega'_D),(\sigma_C,\sigma'_D)} \;+\; \lambda^D_{\sigma_C}\\ &= -\frac{1}{y_{\sigma_C,\sigma'_D}}\sum_{\substack{\omega,\omega'\in\Omega_V\\ \omega_C=\sigma_C,\,\omega'_D=\sigma'_D}} p(\omega)\,K(\omega'\,|\,\omega) \;+\; \lambda^D_{\sigma_C}.\end{aligned}$$

For a critical point (x,y), the partial derivatives vanish. We get the following solution:

$$x_\sigma \;=\; p(\sigma), \qquad \sigma\in\Omega_V,$$

and

$$y_{\sigma_C,\sigma'_D} \;=\; \frac{1}{\sum_{\omega_C\in\Omega_C} p(\omega)} \sum_{\substack{\omega,\omega'\in\Omega_V\\ \omega_C=\sigma_C\,,\,\omega'_D=\sigma'_D}} p(\omega)\,K(\omega'\,|\,\omega) \qquad \sigma_C\in\Omega_C,\ \sigma'_D\in\Omega_D.$$

From Theorem 3.10 in [3] we know that this solution is the (-1)-projection of (p,K) onto $\mathrm{MC}_{\mathscr{S}}(\Omega_V)$. It is given by the initial distribution p and the corresponding marginals K^A_B, $(A,B)\in\mathscr{S}$, of K.

(ii) With (i) we get

$$\begin{aligned}
& D\big((p,K)\,\|\,\mathrm{MC}_{\mathscr{S}}(\Omega_V)\big)\\
=\;& D_p(K\,\|\,K_{\mathscr{S}})\\
=\;& \sum_{\omega,\omega'\in\Omega_V} p(\omega)\,K(\omega'\,|\,\omega)\ln\frac{K(\omega'\,|\,\omega)}{\prod_{(A,B)\in\mathscr{S}} K^A_B(\omega'_B\,|\,\omega_A)}\\
=\;& -H(p,K)\\
& -\sum_{(A,B)\in\mathscr{S}}\sum_{\omega,\omega'\in\Omega_V} p(\omega)\,K(\omega'\,|\,\omega)\ln K^A_B(\omega'_B\,|\,\omega_A)\\
=\;& -H(p,K)\\
& -\sum_{(A,B)\in\mathscr{S}}\sum_{\omega\in\Omega_A,\omega'\in\Omega_B} \ln K^A_B(\omega'\,|\,\omega)\underbrace{\sum_{\substack{\sigma,\sigma'\in\Omega_V\\ \sigma_A=\omega,\ \sigma'_B=\omega'}} p(\sigma)\,K(\sigma'\,|\,\sigma)}_{p_A(\omega)\,K^A_B(\omega'\,|\,\omega)}\\
=\;& \sum_{(A,B)\in\mathscr{S}} H(p_A,K^A_B)\;-\;H(p,K).
\end{aligned}$$

(iii) According to (4) we have $\mathrm{MC}_{\mathscr{S}}(\Omega_V)\subseteq\mathrm{MC}_{\mathscr{S}'}(\Omega_V)$, and the statement follows from the Pythagorian relation ([3], p. 62, Theorem 3.8). □

Proof of Proposition 4.1.

(i) This follows from Proposition 3.2 (ii).

(ii) We apply (i):

$$\begin{aligned} I\{X_n\} &\overset{(i)}{=} \sum_{\nu\in V} H(X_{\nu,\,n+1}\,|\,X_{\nu,\,n}) \;-\; H(X_{n+1}\,|\,X_n) \\ &= \left(\sum_{\nu\in A} H(X_{\nu,\,n+1}\,|\,X_{\nu,\,n}) \;-\; H(X_{A,\,n+1}\,|\,X_{A,\,n})\right) \\ &\quad + \left(\sum_{\nu\in B} H(X_{\nu,\,n+1}\,|\,X_{\nu,\,n}) \;-\; H(X_{B,\,n+1}\,|\,X_{B,\,n})\right) \\ &\quad + \Big(H(X_{A,\,n+1}\,|\,X_{A,\,n}) + H(X_{B,\,n+1}\,|\,X_{B,\,n}) - H(X_{n+1}\,|\,X_n)\Big) \\ &\overset{(i)}{=} I\{X_{A,\,n}\} + I\{X_{B,\,n}\} + I_{A,B}\{X_n\}. \end{aligned}$$

(iii) For parallel processing, one has

$$H(X_{n+1}\,|\,X_n) \;=\; \sum_{\nu\in V} H(X_{\nu,\,n+1}\,|\,X_n).$$

The statement is then implied by (i).

(iv) This follows from (iii) and the Markov property for (V,E)-adapted Markov chains.

References

[1] S. Amari, *Differential-Geometric Methods in Statistics (Lecture Notes in Statistics)*, Berlin: Springer, 1985, vol. 28.

[2] ——, "Information Geometry on Hierarchy of Probability Distributions," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1701-1711, July 2001.

[3] S. Amari and H. Nagaoka, *Methods of Information Geometry (Translations of Mathematical Monographs)*, New York: AMS and Oxford University Press, 2000, vol. 191.

[4] N. Ay, "An Information Geometric Approach to a Theory of Pragmatic Structuring," *The Annals of Probability*, 2001 (to appear).

[5] ——, "Locality of global stochastic interaction in directed acyclic networks," submitted to Neural Computation, 2001.

[6] N. Ay and T. Wennekers, "Dynamical Properties of Strongly Interacting Markov Chains," in preparation.

[7] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter, *Probabilistic Networks and Expert Systems (Statistics for Engineering and Information Science)*, New York, Berlin, Heidelberg: Springer, 1999.

[8] I. Csiszár, "On topological properties of f-divergence," *Studia Sci. Math. Hungar.*, vol. 2, pp. 329-339, 1967.

[9] ——, "I-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, vol. 3, pp. 146-158, 1975.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications)*, New York: Wiley-Interscience, 1991.

[11] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, 1998.

[12] J. Jost, "On the Notion of Complexity," *Theory in Biosciences*, vol. 117, pp. 161-171, 1998.

[13] ——, *Komplexe Systeme*, Lecture given at the University of Leipzig, 2000/2001.

[14] ——, *Biologie und Mathematik*, Lecture given at the University of Leipzig, 2001.

[15] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79-86, 1951.

[16] S. L. Lauritzen, *Graphical Models (Oxford Statistical Science Series)*, Clarendon Press: 1996, vol. 17.

[17] C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters," *Bull. Calcutta Math. Soc.*, vol. 37, pp. 81-91, 1945.

[18] S. Roman, *Coding and Information Theory (Graduate Texts in Mathematics)*, New York, Berlin, Heidelberg: Springer-Verlag, 1992, vol. 134.

[19] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. Journal*, vol. 27, pp. 379-423 and 623-656, 1948.

[20] G. Tononi, O. Sporns, G. M. Edelmann, "A measure for brain complexity: Relating functional segregation and integration in the nervous system," *Proc. Natl. Acad. Sci. USA*, vol. 91, pp. 5033-5037, 1994.

Nihat Ay

MPI for Mathematics in the Sciences

Inselstr. 22-26

04103 Leipzig, Germany

E-mail: nay@mis.mpg.de