Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Properties of the Statistical Complexity Functional and Partially Deterministic HMMs

(revised version: August 2009)

by

Wolfgang Löhr

Preprint no.: 24 2009

Properties of the Statistical Complexity Functional and Partially Deterministic HMMs*

Wolfgang Löhr†

August 05, 2009

Abstract

Statistical complexity is a measure of complexity of discrete-time stationary stochastic processes, which has many applications. We investigate its more abstract properties as a non-linear functional on the space of processes and show its close relation to Knight's prediction process. We prove lower semi-continuity, concavity, and a formula for the ergodic decomposition of statistical complexity. On the way, we show that the discrete version of the prediction process has a continuous Markov transition. We also prove that, given the past output of a partially deterministic hidden Markov model (HMM), the uncertainty of the internal state is constant over time and knowledge of the internal state gives no additional information on the future output. Using this fact, we show that the causal state distribution is the unique stationary representation on prediction space that may have finite entropy.

Keywords: prediction process, statistical complexity, lower semi-continuity, ergodic decomposition, concavity, partially deterministic hidden Markov model, HMM.

Contents

1 Introduction

An important task of complex systems sciences is to define "complexity". Measures that quantify complexity are of both theoretical (e.g. [16]) and practical interest. In applications, they are widely used to identify "interesting" parts of simulations and real-world data (e.g. [9]). There exist various measures of different kinds of complexity. In particular, *statistical complexity* constitutes a complexity measure for stationary stochastic processes in doubly infinite discrete time and discrete state space. It was introduced

*Accepted for publication in Entropy

†Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany
E-mail: Wolfgang.Loehr@mis.mpg.de, URL: http://personal-homepages.mis.mpg.de/loehr

by Jim Crutchfield and co-workers within a theory called *computational mechanics*[1] ([5, 18, 1]). Statistical complexity is applied to a variety of real-world data, e.g. in [4]. An important, closely related concept of computational mechanics is the so-called ε-machine. It is a particular partially deterministic HMM that encodes the mechanisms of prediction. Partially deterministic HMMs are often called *deterministic stochastic automata* to emphasise their close connection to a key concept of theoretical computer science, namely *deterministic finite state automata* ([8]).

In this paper, we look at more abstract features of statistical complexity as well as partially deterministic HMMs. We consider statistical complexity to be a non-linear functional from the space of Δ-valued stationary processes (Δ countable) to the set $\overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{\infty\}$ of non-negative extended real numbers. Here, we identify stationary processes with their law, i.e. with shift-invariant probability measures on the sequence space $\Delta^{\mathbb{Z}}$, and equip the space of measures with the usual weak-$*$ topology (often called "weak topology"). Because Δ is discrete, this topology is equal to the topology of finite-dimensional convergence. In ergodic theory, Kolmogorov-Sinai entropy is studied as a function of the (invariant) measure, and the questions of continuity properties, affinity, and behaviour under ergodic decomposition arise naturally (e.g. [10]). We believe that these questions are worthwhile considering also for complexity measures. A formula for the ergodic decomposition of excess entropy, another complexity measure for stochastic processes, was obtained in [6, 7]. Our results presented here include the corresponding formula for statistical complexity, and this formula directly implies concavity. The most important result is lower semi-continuity of statistical complexity. We consider this a desirable property for a complexity measure, as it means that a process cannot be complex if it can be approximated by non-complex ones.

In Section 2, we define statistical complexity and show its relations to a discrete version of Frank Knight's *prediction process* ([11, 13]). The prediction process is the measure-valued process of conditional probabilities of the future given the past. It takes values in the space $\mathcal{P}(\Delta^{\mathbb{N}})$ of probability measures on $\Delta^{\mathbb{N}}$, called prediction space. In our formulation, statistical complexity is the marginal entropy of the prediction process. This is equivalent to the classical definition as entropy of a certain partition of the past. We only replace equivalence classes with the respective induced probabilities on the future. In this section, we also show that the discrete (and thus technically vastly simplified) version of the prediction process has a continuous Markov transition kernel (Proposition 2.5).

In Section 3, we investigate properties of partially deterministic HMMs. Here, we use a general notion of HMM (sometimes called edge-emitting HMM), where new internal state and output symbol are jointly determined and may have dependencies conditioned on the last internal state. Partial determinism means that this dependence is extreme in the sense that last internal state and output together uniquely determine the following internal state. We show that, if one knows the past output trajectory, the remaining uncertainty (measured by entropy) of the internal state is constant over time, although it may depend on the ergodic component (Proposition 3.7). Furthermore, the distribution of future output is the same for any internal state that is compatible with the past output (Corollary 3.9). In Subsection 3.3, we construct a canonical Markov kernel, such that taking any measure ν on prediction space $\mathcal{P}(\Delta^{\mathbb{N}})$ (i.e. ν is a measure on measures) as initial distribution, we obtain a partially deterministic HMM of a process $P \in \mathcal{P}(\Delta^{\mathbb{N}})$. This process P coincides with the measure $r(\nu)$ represented by ν in the sense of integral representation theory, and if ν is appropriately chosen, we obtain the ε-machine of computational mechanics (or something isomorphic) as special case. Using the properties of partially deterministic HMMs, we obtain that there is no invariant representation on prediction space with finite entropy other than, possibly, the causal state distribution, which may have finite or infinite entropy (Proposition 3.12).

Section 4 contains our results about statistical complexity. We show that the complexity of a process is the average complexity of its ergodic components plus the entropy of the mixture (Proposition 4.1). As a direct consequence, statistical complexity is concave (Corollary 4.2) and non-continuous (even w.r.t. variational topology). But it does have a continuity property. Namely, using the results of the previous sections, we show in Theorem 4.7 that it is weak-$*$ lower semi-continuous.

[1] Here "computational mechanics" is unrelated to computer simulations of mechanical systems

2 Prediction dynamic & statistical complexity

For the whole article, fix a countable set Δ with at least two elements and discrete topology. We identify Δ-valued stochastic processes $X_{\mathbb{Z}} := (X_k)_{k\in\mathbb{Z}}$, defined on some probability space $(\Omega, \mathfrak{A}, \mathbb{P})$, with their respective laws $P := \mathbb{P} \circ X_{\mathbb{Z}}^{-1} \in \mathcal{P}(\Delta^{\mathbb{Z}})$. Here, $\mathcal{P}$ denotes the set of probability measures. If $X_{\mathbb{Z}}$ is stationary, P is in the set $\mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$ of shift-invariant probability measures. Let $X'_k \colon \Delta^{\mathbb{Z}} \to \Delta$ be the canonical projections. Then $X'_{\mathbb{Z}}$ is a process on $(\Delta^{\mathbb{Z}}, \mathfrak{B}(\Delta^{\mathbb{Z}}), P)$ with the same distribution as $X_{\mathbb{Z}}$. Here, $\mathfrak{B}$ denotes the Borel σ-algebra. We often decompose the time set $\mathbb{Z}$ into the "future" $\mathbb{N}$ and the "past" $\mathbb{Z} \setminus \mathbb{N} = -\mathbb{N}_0$, where $\mathbb{N}_0 = \mathbb{N} \cup \{\,0\,\}$. For simplicity of notation, we denote the canonical projections on $\Delta^{\mathbb{N}}$ with the same symbols, X'_k, as the projections on $\Delta^{\mathbb{Z}}$. If not stated otherwise, product spaces are equipped with product and spaces of probability measures are equipped with weak-$*$-topology. We use the arrow $\xrightarrow{*}$ to denote weak-$*$ convergence.

2.1 Discrete version of Knight's prediction process

Given a Lusin space valued, measurable stochastic process with time set $\mathbb{R}_+$, Frank Knight defines the corresponding *prediction process* as a process of conditional probabilities of the future given the past. This theory originated in [11] and was developed in [15, 12, 13]. The most important properties of the prediction process are that its paths are right continuous with left limits (cadlag), it has the strong Markov property and determines the original process. The continuous time set and the generality of the state space lead to a lot of technical difficulties. In our simpler, discrete setting, these difficulties mostly disappear, and useful properties of the prediction process, such as having cadlag paths, become meaningless. A new aspect, however, is added by considering infinite pasts of stationary processes via the time-set $\mathbb{Z}$. The marginal distribution (unique due to stationarity) of the prediction process is an important quantity, used to define statistical complexity. For this subsection, fix a *stationary* process $X_{\mathbb{Z}}$ with distribution $P \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$.

We use the following notation concerning Markov kernels and conditional probabilities. If K is a kernel from Ω to a measurable space M, we consider K as measurable function from Ω to $\mathcal{P}(M)$ and write $K(\omega;\, A) := K(\omega)(A)$.Given random variables X, Y on Ω, we write $K = \mathbb{P}(X \mid Y)$ if K is the conditional probability kernel of X given Y, i.e. $K(\omega;\, A) = \mathbb{P}\big(\{\, X \in A \,\} \,\big|\, Y\big)(\omega)$.

Definition 2.1. Let $Z_{\mathbb{Z}} = Z^P_{\mathbb{Z}}$ be the $\mathcal{P}(\Delta^{\mathbb{N}})$-valued stochastic process of conditional probabilities defined by $Z_k := P(X'_{[k+1,\infty[} \mid X'_{]-\infty,k]})$ for $k \in \mathbb{Z}$. Then $Z_{\mathbb{Z}}$ is called **prediction process** of $X_{\mathbb{Z}}$. $\mathcal{P}(\Delta^{\mathbb{N}})$ is called **prediction space**.

It is evident that the Markov property of the prediction process in continuous time also holds in discrete time. Nevertheless, we give a proof, because it is elementary in our discrete setting. The corresponding transition kernel works as follows. Assume the prediction process is in state $z \in \mathcal{P}(\Delta^{\mathbb{N}})$. The transition kernel maps z to a measure on measures, namely $P(Z_1 \mid Z_0 = z) \in \mathcal{P}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big)$. Note that z is a state of the prediction process but at the same time a probability measure. Thus it makes sense to consider the conditional probability given $X'_1 = d$ w.r.t. the measure z. It is intuitively plausible that the next state will be one of those conditional probabilities with d distributed according to the marginal of z. The resulting measure has to be shifted by one as time proceeds. With $\varsigma \colon \Delta^{\mathbb{N}} \to \Delta^{\mathbb{N}}$, we denote the left shift.

Proposition 2.2. *For $z \in \mathcal{P}(\Delta^{\mathbb{N}})$, let $\phi_z \colon \Delta^{\mathbb{N}} \to \mathcal{P}(\Delta^{\mathbb{N}})$, $\phi_z(\omega) := z(\varsigma^{-1}{}_{(\,\cdot\,)} \mid X'_1)(\omega)$. The prediction process $Z_{\mathbb{Z}}$ is a stationary Markov process. The kernel $S \colon \mathcal{P}(\Delta^{\mathbb{N}}) \to \mathcal{P}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big)$ with $S(z) = z \circ \phi_z^{-1}$, i.e.*

$$S(z)(B) \;:=\; S(z;\, B) \;:=\; z\big(\{\, \phi_z \in B \,\}\big), \qquad z \in \mathcal{P}(\Delta^{\mathbb{N}}),\; B \in \mathfrak{B}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big),$$

satisfies $P(Z_k \mid Z_{k-1}) = S \circ Z_{k-1}$ a.s. Thus, S is the transition kernel of the prediction process.

Proof. Stationarity is obvious from stationarity of $X_{\mathbb{Z}}$. We obtain a.s.

$$\begin{aligned} S(Z_0;\, B) \;&=\; Z_0\big(\{\, Z_0(\varsigma^{-1}{}_{(\,\cdot\,)} \mid X'_1) \in B \,\}\big) \;=\; P\Big(\Big\{\, P(X'_{[2,\infty[} \mid X'_{]-\infty,1]}) \in B \,\Big\} \,\Big|\, X'_{-\mathbb{N}_0}\Big) \\ &=\; P\big(\{\, Z_1 \in B \,\} \,\big|\, X'_{-\mathbb{N}_0}\big). \end{aligned}$$

In particular, $P\big(\{\, Z_1 \in B \,\} \;\big|\; X'_{-\mathbb{N}_0}\big)$ is $\sigma(Z_0)$-measurable (modulo P) and together with $\sigma(Z_0) \subseteq \sigma(X'_{-\mathbb{N}_0})$ we obtain

$$P\big(\{\, Z_1 \in B \,\} \;\big|\; Z_0\big) = P\big(\{\, Z_1 \in B \,\} \;\big|\; X'_{-\mathbb{N}_0}\big) = S(Z_0;\, B), \tag{1}$$

as claimed. We still have to verify the Markov property. But because the σ-algebra induced by $Z_{-\mathbb{N}_0}$ is nested between those induced by Z_0 and $X'_{-\mathbb{N}_0}$, i.e. $\sigma(Z_0) \subseteq \sigma(Z_{-\mathbb{N}_0}) \subseteq \sigma(X'_{-\mathbb{N}_0})$, we obtain the Markov property from the first equality in (1). □

Definition 2.3. We call the Markov transition S of the prediction process **prediction dynamic**.

Note that although the prediction process $Z_{\mathbb{Z}}$ obviously depends on P, prediction space $\mathcal{P}(\Delta^{\mathbb{N}})$ and prediction dynamic S do not. In the case of general Lusin state space, it is non-trivial to prove the existence of regular versions of conditional probability such that $\phi_z(\omega)$ is jointly measurable in (z, ω) (see [13]). For countable Δ, however, we even obtain essential continuity in an elementary way. This enables us to prove continuity of the prediction dynamic.

Lemma 2.4. *Let $z, z_n \in \mathcal{P}(\Delta^{\mathbb{N}})$ and $z_n \overset{*}{\rightharpoonup} z$. There is a clopen (i.e. closed and open) set $\Omega_z \subseteq \Delta^{\mathbb{N}}$ with $z(\Omega_z) = 1$ such that $\phi_{z_n} \overset{*}{\rightharpoonup} \phi_z$, uniformly on compact subsets of Ω_z.*

Proof. Let $A_\omega := {X'_1}^{-1}\big(X'_1(\omega)\big)$ and $\Omega_z := \big\{\, \omega \in \Delta^{\mathbb{N}} \;\big|\; z(A_\omega) > 0 \,\big\}$. Because Δ is discrete and countable, Ω_z is clopen with $z(\Omega_z) = 1$. Uniform convergence on compacta is equivalent to $\phi_{z_n}(\omega_n) \overset{*}{\rightharpoonup} \phi_z(\omega)$ whenever $\omega_n \to \omega$ in Ω_z. For sufficiently large n, $X'_1(\omega_n) = X'_1(\omega)$ and because ς^{-1} maps cylinder sets to cylinder sets, $\phi_{z_n}(\omega_n) = \frac{z_n(A_\omega \cap \varsigma^{-1}(\,\cdot\,))}{z_n(A_\omega)} \overset{*}{\rightharpoonup} \phi_z(\omega)$. □

Proposition 2.5. *The prediction dynamic S is continuous.*

Proof. Let $z_n, z \in \mathcal{P}(\Delta^{\mathbb{N}})$ with $z_n \overset{*}{\rightharpoonup} z$ and Ω_z as in Lemma 2.4. We have to show

$$\int g \,\mathrm{d}S(z_n) \;=\; \int g \circ \phi_{z_n} \,\mathrm{d}z_n \;\overset{n\to\infty}{\longrightarrow}\; \int g \circ \phi_z \,\mathrm{d}z \;=\; \int g \,\mathrm{d}S(z) \tag{2}$$

for any bounded continuous g. According to Prokhorov's theorem, $(z_n)_{n\in\mathbb{N}}$ is uniformly tight and we can restrict the integrations to compact subsets. Because $\lim_{n\to\infty} z_n(\Omega_z) = z(\Omega_z) = 1$, we can restrict to compact subsets of Ω_z. There, the convergence of ϕ_{z_n} is uniform, thus (2) holds. □

2.2 Statistical complexity

In integral representation theory, a measure $\nu \in \mathcal{P}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big)$ represents the measure $z \in \mathcal{P}(\Delta^{\mathbb{N}})$ if[2]

$$z \;=\; r(\nu) \;:=\; \int_{\mathcal{P}(\Delta^{\mathbb{N}})} \mathrm{id}_{\mathcal{P}(\Delta^{\mathbb{N}})} \,\mathrm{d}\nu, \tag{3}$$

where $r\colon \mathcal{P}(\mathcal{P}(\Delta^{\mathbb{N}})) \to \mathcal{P}(\Delta^{\mathbb{N}})$ is called **resolvent** or **barycentre map** (see [3]) and id is the identity map. $z = r(\nu)$ means that z is a mixture (convex combination) of other processes, and the mixture is described by ν. A trivial representation for z is given by δ_z, the Dirac measure in z. The measure ν is called S-**invariant** if $\nu S = \nu$, where $\nu S := \int S \,\mathrm{d}\nu$. In other words, it is S-invariant if iterating with the prediction dynamic S does not change it. We see in the following lemma that generally iterating with S shifts the represented measure, i.e. νS represents $z \circ \varsigma^{-1}$.

Lemma 2.6. $r(\nu S) \;=\; r(\nu) \circ \varsigma^{-1}$. *In particular, S-invariant ν represent stationary processes.*

Proof. Because $r(\nu S) = \int\int \mathrm{id}_{\mathcal{P}(\Delta^{\mathbb{N}})} \,\mathrm{d}S \,\mathrm{d}\nu$, it is sufficient to consider Dirac measures δ_z, $z \in \mathcal{P}(\Delta^{\mathbb{N}})$ (the general claim follows by integration over ν). For Dirac measures we have

$$r(\delta_z S) \;=\; \int \mathrm{id}_{\mathcal{P}(\Delta^{\mathbb{N}})} \,\mathrm{d}S(z) \;=\; \int \phi_z \,\mathrm{d}z \;=\; \int z\big(\varsigma^{-1}(\,\cdot\,) \;\big|\; X'_1\big) \,\mathrm{d}z \;=\; z \circ \varsigma^{-1}. \qquad \square$$

[2] Measure valued integrals are Gel'fand integrals. That is, $\mu = \int K \,\mathrm{d}\nu$ for some kernel K means $\int f \,\mathrm{d}\mu = \int\int f \,\mathrm{d}K(\,\cdot\,) \,\mathrm{d}\nu$ for all continuous, real-valued f or, equivalently, $\mu(B) = \int K(\,\cdot\,;\, B) \,\mathrm{d}\nu$ for all measurable sets B.

If ν is S-invariant, we also say that ν represents the stationary extension of $r(\nu)$ to $\Delta^{\mathbb{Z}}$. The marginal of the prediction process is an important such representation, which we call causal state distribution because of its close relation to the causal states of computational mechanics.

Definition 2.7. For $P \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$, the **causal state distribution** $\mu_{\mathfrak{C}}(P)$ is the marginal distribution of the prediction process, i.e. $\mu_{\mathfrak{C}}(P) := P \circ Z_0^{-1} \in \mathcal{P}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big)$.

The causal state distribution of P is an S-invariant representation of P.

Lemma 2.8. *Let $P \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$. Then $\mu_{\mathfrak{C}}(P)$ is S-invariant and represents P.*

Proof. From Proposition 2.2 we know that $P(Z_1 \mid Z_0) = S \circ Z_0$ and $Z_{\mathbb{Z}}$ is stationary. Thus

$$\int S \,\mathrm{d}\mu_{\mathfrak{C}}(P) \;=\; \int S \circ Z_0 \,\mathrm{d}P \;=\; \int P(Z_1 \mid Z_0) \,\mathrm{d}P \;=\; P \circ Z_1^{-1} \;=\; \mu_{\mathfrak{C}}(P).$$

Furthermore, $\mu_{\mathfrak{C}}(P)$ represents P because we have

$$r\big(\mu_{\mathfrak{C}}(P)\big) \;=\; \int Z_0 \,\mathrm{d}P \;=\; \int P(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0}) \,\mathrm{d}P \;=\; P \circ {X'_{\mathbb{N}}}^{-1}. \qquad \square$$

Remark. In computational mechanics, slightly different definitions are used. There, one works with equivalence classes of past trajectories (called *causal states*) instead of probability measures on the future. Because past trajectories $x, y \in \Delta^{-\mathbb{N}_0}$ are identified if $P(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0} = x) = P(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0} = y)$, the two approaches are equivalent. The advantage of working on prediction space $\mathcal{P}(\Delta^{\mathbb{N}})$ is that it has a natural topology and the prediction processes of all Δ-valued stochastic processes are described in a unified way on the same space with the same transition kernel.

Example 2.9. $\mu_{\mathfrak{C}}$ is *not* continuous. Let P be a non-deterministic i.i.d. (independent, identically distributed) process. Obviously, the causal state distribution of an i.i.d. process is the Dirac measure $\delta_{P_{\mathbb{N}}}$ in its restriction $P_{\mathbb{N}} := P \circ {X'_{\mathbb{N}}}^{-1}$ to positive time. According to [17], periodic measures are dense in the stationary measures and we find an approximating sequence $P_n \overset{*}{\rightharpoonup} P$ of periodic measures P_n. But the past of a periodic process determines its future. Thus its causal state distribution is supported by the set of Dirac measures on $\Delta^{\mathbb{N}}$. Because the set of Dirac measures is closed in $\mathcal{P}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big)$, the topological supports $\operatorname{supp} \mu_{\mathfrak{C}}(P_n)$ are disjoint from the support $\operatorname{supp} \mu_{\mathfrak{C}}(P) = \{\, P_{\mathbb{N}} \,\}$. Consequently, $\mu_{\mathfrak{C}}(P_n)$ cannot converge to $\mu_{\mathfrak{C}}(P)$. $\lozenge$

With statistical complexity, we measure complexity of a process P by the "diversity" of its expected futures, given observed pasts (i.e. of $\mu_{\mathfrak{C}}(P)$). As measure of "diversity" of a probability measure μ, Shannon entropy $H(\mu)$ is used. With $\varphi(x) := -x \log(x)$, it is defined as

$$H(\mu) \;:=\; \sup\Big\{ \sum_{i=1}^{n} \varphi\big(\mu(B_i)\big) \;\Big|\; n \in \mathbb{N},\ B_i \text{ disjoint, measurable} \Big\}. \tag{4}$$

Definition 2.10. For $P \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$, the quantity $C_{\mathfrak{C}}(P) := H\big(\mu_{\mathfrak{C}}(P)\big) \in \overline{\mathbb{R}}_+$ is called **statistical complexity** of P.

Note that if the probability space is sufficiently regular (e.g. separable, metrisable), $H(\mu)$ can only be finite if μ is supported by a countable set A. In this case

$$H(\mu) \;=\; \sum_{a \in A} \varphi\big(\mu(\{\, a \,\})\big).$$

Probably, lower semi-continuity of the entropy is well-known. We give a proof in the appendix.

Lemma 2.11. *Let M be a separable, metrisable space. Then the entropy $H\colon \mathcal{P}(M) \to \overline{\mathbb{R}}_+$ is weak-$*$ lower semi-continuous.*

3 Partially deterministic HMMs

Probability measures on prediction space induce hidden Markov models (HMMs) with an additional partial determinism property, and it turns out to be helpful to investigate such HMMs. In Section 3.1, we define HMMs and introduce the notation we need for the further discussion. In Section 3.2, we define the partial determinism property and obtain our results about the HMMs satisfying this property. In Section 3.3, we show how measures on prediction space induce partially deterministic HMMs and apply the results from Section 3.2 to prove that the causal state distribution is the only invariant representation on prediction space that can have finite entropy.

3.1 HMMs

We use the term HMM in a wide sense, meaning a pair (T, μ), where μ is an initial probability measure on some Polish space M of internal states and T is a Markov kernel from M to $\Delta \times M$. The HMM generates on $(\Omega, \mathfrak{A}, \mathbb{P})$ a Δ-valued output process $X_{\mathbb{N}}$ and a (coupled) M-valued internal process $W_{\mathbb{N}_0}$, such that W_0 is μ-distributed and the joint process is Markovian with

$$\mathbb{P}\big(\{\, X_k \in D,\ W_k \in B \,\} \,\big|\, X_{k-1}, W_{k-1}\big) \;=\; T(W_{k-1};\, D \times B) \qquad \text{a.s.}$$

We call (T, μ) an HMM **of** $z \in \mathcal{P}(\Delta^{\mathbb{N}})$ if $z = \mathbb{P} \circ X_{\mathbb{N}}^{-1}$. If $\mu(B) = \int T(\,\cdot\,;\, \Delta \times B)\, \mathrm{d}\mu$, we say that the HMM is **invariant** and extend the generated processes to stationary processes $X_{\mathbb{Z}}$ and $W_{\mathbb{Z}}$. We need some further notation.

Definition 3.1. Let (T, μ) be an HMM, $m \in M$, $d \in \Delta$, and $\nu \in \mathcal{P}(M)$.

a) The **output kernel** $K\colon M \to \mathcal{P}(\Delta)$ is defined by $K(m) := K_m := T(m;\, \cdot \times M) \in \mathcal{P}(\Delta)$. We also use the notations $\widehat{K}_d(m) := K_m(d) := K_m\big(\{\, d \,\}\big)$ and $K_\nu := \int K \,\mathrm{d}\nu$.

b) The **internal operators** $L_d\colon \mathcal{P}(M) \to \mathcal{P}(M) \cup \{\, 0 \,\}$ are defined as follows. $L_d(\nu) = 0$ if $K_\nu(d) = 0$ and

$$L_d(\nu)(B) \;:=\; \frac{\int T(\,\cdot\,;\, \{\, d \,\} \times B)\, \mathrm{d}\nu}{K_\nu(d)} \qquad \text{otherwise.}$$

Remark. a) K_m is the distribution of the next output symbol when the internal state is m, i.e. $K_m = \mathbb{P}(X_1 \mid W_0 = m)$ a.s. Further, K_μ is the law of X_1.

b) The internal operator L_d describes the update of knowledge of the internal state when the symbol $d \in \Delta$ is observed. For Dirac measures, we obtain

$$L_d(\delta_m) \;=\; \mathbb{P}(W_1 \mid W_0 = m,\ X_1 = d) \qquad \text{a.s.}$$

Be warned that L_d is *not* induced by a kernel in the following sense. There is no kernel $l_d\colon M \to \mathcal{P}(M)$ such that $L_d(\nu) = \int l_d \,\mathrm{d}\nu$. To see this, note that $L_d(\nu) \neq \int L_d \circ \iota \,\mathrm{d}\nu$ for $\iota(m) = \delta_m$, because $L_d(\nu)$ is normalised outside the integral as opposed to an individual normalisation of the $L_d(\delta_m)$ inside the integral on the right-hand side.

It directly follows from the definition of $(X_{\mathbb{N}}, W_{\mathbb{N}_0})$ by a Markov kernel that the conditional probability, given that the internal state is m, is obtained by starting the HMM in m. In other words, it is generated by the HMM (T, δ_m). Similarly, the conditional probability given an observed symbol $X_1 = d$ is obtained by starting the HMM in the updated initial distribution $L_d(\mu)$. We formulate these observations in the following lemma and give a formal proof in the appendix.

Lemma 3.2. *Let (T, μ) be an HMM with internal and output processes $W_{\mathbb{N}_0}$, $X_{\mathbb{N}}$. Then a.s. $(T, \delta_{W_0(\omega)})$ is an HMM of $\mathbb{P}(X_{\mathbb{N}} \mid W_0)(\omega)$, and $\big(T, L_{X_1(\omega)}(\mu)\big)$ is an HMM of $\mathbb{P}(X_{[2,\infty[} \mid X_1)(\omega)$.*

Definition 3.3 ($Y_{\mathbb{Z}}$ and $H_{\mathbb{Z}}$). Given an invariant HMM, let $Y_{\mathbb{Z}}$ be the $\mathcal{P}(M)$-valued process of expectations over internal states given by $Y_k := \mathbb{P}(W_k \mid X_{]-\infty,k]})$. Let $H_{\mathbb{Z}}$ be the process of entropies of the random measures Y_k, i.e. $H_k(\omega) := H\big(Y_k(\omega)\big)$, where entropy H is defined by (4).

Remark. Y_k describes the current knowledge of the internal state, given the past. H_k is the entropy of the *value* of Y_k and measures "how uncertain" the knowledge of the internal state is. It is important to bear in mind that this is different from the entropy of the *random variable* Y_k. To avoid confusion, we always write $H^{\mathbb{P}}(X)$ when referring to the entropy of a random variable X defined on a probability space with measure $\mathbb{P}$.

The following lemma justifies the idea of the internal operator L_d being an update of knowledge of the internal state. Furthermore, it enables us to condition on Y_0 instead of $X_{-\mathbb{N}_0}$. The conditional probability of the internal state given the past, Y_0, contains as much information about X_1 (and in fact $X_{\mathbb{N}}$, but we do not need that here) as the past $X_{-\mathbb{N}_0}$ does.

Lemma 3.4. *a)* $Y_1(\omega) \;=\; L_{X_1(\omega)}\big(Y_0(\omega)\big)$ *a.s.*

b) $\mathbb{P}\big(\{\, X_1 = d \,\} \;\big|\; Y_0\big)(\omega) \;=\; \mathbb{P}\big(\{\, X_1 = d \,\} \;\big|\; X_{-\mathbb{N}_0}\big)(\omega) \;=\; K_{Y_0(\omega)}(d)$ *a.s.*

Proof. Conditional independence of (X_1, W_1) and $X_{-\mathbb{N}_0}$ given W_0 implies that a.s. $\mathbb{P}(X_1,\, W_1 \mid W_0) = \mathbb{P}(X_1,\, W_1 \mid W_0,\, X_{-\mathbb{N}_0})$ and thus

$$\int T \,\mathrm{d}Y_0 \;=\; \int \mathbb{P}(X_1,\, W_1 \mid W_0) \,\mathrm{d}\mathbb{P}(\,\cdot \mid X_{-\mathbb{N}_0}) \;=\; \mathbb{P}(X_1,\, W_1 \mid X_{-\mathbb{N}_0}). \tag{5}$$

a) Let $d = X_1(\omega)$ and for $B \in \mathfrak{B}(M)$ set $F_B := \{\, X_1 = d,\; W_1 \in B \,\}$. We obtain a.s.

$$L_d(Y_0)(B) \overset{(5)}{=} \frac{\mathbb{P}(F_B \mid X_{-\mathbb{N}_0})}{\mathbb{P}(F_M \mid X_{-\mathbb{N}_0})} \overset{(d\,=\,X_1(\omega))}{=} \mathbb{P}\big(\{\, W_1 \in B \,\} \;\big|\; X_{-\mathbb{N}_0},\, X_1\big) \;=\; Y_1(\,\cdot\,)(B).$$

b) The second equality follows directly from (5). The first follows because, due to the second equality, $\mathbb{P}\big(\{\, X_1 = d \,\} \;\big|\; X_{-\mathbb{N}_0}\big)$ is $\sigma(Y_0)$-measurable modulo $\mathbb{P}$. □

The previous lemma enables us to prove that $Y_{\mathbb{Z}}$ is Markovian and compute its transition kernel. We already know that $L_d(\nu)$ is the updated expectation of the internal state when it previously was ν and d is observed. Thus it is not surprising that the conditional probability of Y_k given $Y_{k-1} = \nu$ is a convex combination of Dirac measures in $L_d(\nu)$ for different d (note that Y_k is a measure-valued random variable, thus its conditional probability distribution is indeed a measure on measures). The mixture is given by the output kernel K, more precisely by K_ν.

Lemma 3.5. *For an invariant HMM, $Y_{\mathbb{Z}}$ and $H_{\mathbb{Z}}$ are stationary. $Y_{\mathbb{Z}}$ is a Markov process with transition kernel*

$$\mathbb{P}(Y_{k+1} \mid Y_k = \nu) \;=\; \sum_{d\in\Delta} K_\nu(d) \cdot \delta_{L_d(\nu)} \;\in\; \mathcal{P}\big(\mathcal{P}(M)\big) \qquad \forall \nu \in \mathcal{P}(M).$$

Proof. Stationarity is obvious. For $\nu_0, \ldots, \nu_k \in \mathcal{P}(M)$ and $\nu := \nu_k$ we obtain

$$\begin{aligned} \mathbb{P}(Y_{k+1} \mid Y_{[0,k]} = \nu_{[0,k]}) &\overset{\text{(lem. 3.4a)}}{=} \mathbb{P}\big(L_{X_{k+1}(\,\cdot\,)}(\nu) \;\big|\; Y_{[0,k]} = \nu_{[0,k]}\big) \\ &= \sum_{d\in\Delta} \mathbb{P}\big(\{\, X_{k+1} = d \,\} \;\big|\; Y_{[0,k]} = \nu_{[0,k]}\big) \cdot \delta_{L_d(\nu)}. \end{aligned}$$

$\sigma(Y_{[0,k]})$ is nested between $\sigma(Y_k)$ and $\sigma(X_{]-\infty,k]})$. Therefore, Lemma 3.4 b) implies that we have $\mathbb{P}\big(\{\, X_{k+1} = d \,\} \;\big|\; Y_{[0,k]} = \nu_{[0,k]}\big) = K_{\nu_k} = K_\nu$ and hence the claim follows. □

3.2 Partial determinism

If the transition T of an HMM is deterministic, i.e. if the internal state determines the next state and output (and thus the whole future) uniquely, the HMM is called *(completely) deterministic*. In a deterministic HMM, all randomness is due to the initial distribution. This is a very strong property, and a weaker partial determinism property is useful. In a partially deterministic HMM, the output symbol is determined randomly, but the new internal state is a function $f(m, d)$ of the last internal state m and the new output

symbol d. If the internal space M is finite, such HMMs are stochastic versions of *deterministic finite state automata (DFAs)*, an important concept of theoretical computer science (see [8, Chap. 2]). The function f directly corresponds to the transition function of the DFA, but the start state is replaced by the initial distribution and the HMM assigns probabilities to the outputs via the output kernel K. A difference in interpretation is that the symbols from Δ are considered *input* of the DFA and *output* of HMMs. To emphasise their close connection to DFAs, partially deterministic HMMs are often called *deterministic stochastic automata*, although they are not completely deterministic.

Definition 3.6. An HMM (T, μ) is called **partially deterministic** if there is a measurable function $f\colon M \times \Delta \to M$, called **transition function**, such that $T(m) = K_m \otimes \delta_{f_m(\,\cdot\,)}$ for all $m \in M$, i.e.

$$T(m;\, D \times B) \;=\; K_m\big(D \cap f_m^{-1}(B)\big) \qquad \forall m \in M,\; D \subseteq \Delta,\; B \in \mathfrak{B}(M),$$

where $f_m(d) := \hat{f}_d(m) := f(m,d)$ and $\mathfrak{B}(M)$ is the Borel σ-algebra on M.

Remark. For partially deterministic HMMs we obtain

$$L_d(\nu)(B) \;=\; \frac{1}{K_\nu(d)} \int_{\hat{f}_d^{-1}(B)} \widehat{K}_d \,\mathrm{d}\nu \qquad \text{and} \qquad L_d(\delta_m) \;=\; \delta_{f_m(d)}. \tag{6}$$

The second equation implies $W_k = f_{W_{k-1}}(X_k)$ a.s., justifying the name transition function for f.

The following proposition is crucial for understanding partially deterministic representations. It states that, given the past output, the uncertainty $H_k = H(Y_k)$ about the internal state is constant over time and the next output symbol is independent of the internal state. The proof is along the following lines. If we know the internal state at one point in time, we can maintain knowledge of the internal state due to partial determinism. More generally, the uncertainty H_k of the internal state cannot decrease on average and thus is a supermartingale. But because it is also stationary, the trajectories have to be constant. If two possible internal states would lead to different probabilities for the next output symbol, we could increase our knowledge of the internal state by observing the next output. But because of partial determinism, this would also decrease the uncertainty of the following internal state, in contradiction to the constant trajectories of $H_{\mathbb{Z}}$.

Proposition 3.7. *Let (T, μ) be a partially deterministic, invariant HMM with $H(\mu) < \infty$. Then $H_{\mathbb{Z}}$ has a.s. constant trajectories, i.e. $H_k = H_0$ a.s., and the restriction $K\restriction_{\mathrm{supp}(Y_0)}$ of the output kernel K to the support $\mathrm{supp}(Y_0) \subseteq M$ of the random measure Y_0 is a.s. a constant kernel, i.e.*

$$K_m \;=\; K_{\hat{m}} \qquad \forall m, \hat{m} \in \mathrm{supp}\big(Y_0(\omega)\big) \qquad \textit{a.s.} \tag{7}$$

Proof. We show that $H_{\mathbb{Z}}$ is a supermartingale to use the following well-known property.

Lemma. Every stationary supermartingale has a.s. constant trajectories.

Because $H(\mu) < \infty$, we may assume w.l.o.g. that M is countable. Note that $\varphi(x) = -x\log(x)$ satisfies $\varphi(\sum x_i) \le \sum \varphi(x_i)$. We obtain

$$H\big(L_d(\nu)\big) \overset{(6)}{=} \sum_{\hat{m} \in M} \varphi\left(\sum_{m \in \hat{f}_d^{-1}(\hat{m})} \nu(m) \frac{K_m(d)}{K_\nu(d)}\right) \;\le\; \sum_{m \in \hat{f}_d^{-1}(M) = M} \varphi\left(\nu(m)\frac{K_m(d)}{K_\nu(d)}\right).$$

We use the filtration $\mathcal{F}_k := \sigma(Y_{]-\infty,k]})$. Markovianity of $Y_{\mathbb{Z}}$ yields $\mathrm{E}(H_{k+1} \mid \mathcal{F}_k) = \mathrm{E}(H_{k+1} \mid Y_k)$.

$$\begin{aligned} \mathrm{E}\big(H_{k+1} \,\big|\, Y_k = \nu\big) &\overset{\text{(lem. 3.5)}}{=} \sum_{d \in \Delta} K_\nu(d) \cdot H\big(L_d(\nu)\big) \;\le\; -\sum_{d,m} \nu(m) K_m(d) \cdot \log\left(\nu(m)\frac{K_m(d)}{K_\nu(d)}\right) \\ &= H^{\mathbb{P}}(W_k \mid X_{k+1},\, Y_k = \nu) \;\le\; H^{\mathbb{P}}(W_k \mid Y_k = \nu) \;=\; H(\nu), \end{aligned} \tag{8}$$

where the second equality holds because $\mathbb{P}\big(\{\, W_k = m,\; X_{k+1} = d \,\} \;\big|\; Y_k = \nu\big) = \nu(m)K_m(d)$ and $\mathbb{P}\big(\{\, X_{k+1} = d \,\} \;\big|\; Y_k = \nu\big) = K_\nu(d)$. Thus $H_{\mathbb{Z}}$ is a supermartingale w.r.t. $(\mathcal{F}_k)_{k\in\mathbb{Z}}$ and has a.s. constant trajectories. In particular, inequality (8) is actually an equality. Because $H(\mu) < \infty$ and $\mu = \int Y_k \,\mathrm{d}\mathbb{P}$, the entropy of $Y_k(\omega)$ is a.s. finite. Thus, $H^{\mathbb{P}}(W_k \mid X_{k+1}, Y_k = \nu) = H^{\mathbb{P}}(W_k \mid Y_k = \nu)$ implies that W_k and X_{k+1} are independent given $Y_k = \nu$, i.e. $K\restriction_{\mathrm{supp}(\nu)}$ is constant. ∎

The finite-entropy assumption is indeed necessary for the second statement of Proposition 3.7. The shift, for example, defines a deterministic HMM that does not (in general) satisfy (7).

Example 3.8 (shift HMM). The shift HMM is defined as follows. The internal state consists of the whole trajectory, $M := \Delta^{\mathbb{Z}}$. $T = T^{\varsigma}$ outputs the symbol at position one and shifts the sequence to the left. More formally with $m = (m_k)_{k\in\mathbb{Z}} \in M$ and $\varsigma(m) = (m_{k+1})_{k\in\mathbb{Z}}$ we have

$$T^{\varsigma}(m) \;=\; \delta_{m_1} \otimes \delta_{\varsigma(m)} \;=\; \delta_{(m_1,\varsigma(m))}.$$

If $P \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$, it is obvious that (T^{ς}, P) is an invariant, deterministic (in particular partially deterministic) HMM of P. Here, P is the law of both $X_{\mathbb{Z}}$ and W_0; in fact even $X_{\mathbb{Z}} = W_0$. We claim that, generically, (T^{ς}, P) does not satisfy (7) (and of course the internal state entropy $H(P)$ is infinite). Indeed, $K_m = \delta_{m_1}$ and thus $K_m = K_{\hat{m}}$ implies $m_1 = \hat{m}_1$. Because $Y_0(\omega) = \mathbb{P}(X_{\mathbb{Z}} \mid X_{-\mathbb{N}_0})(\omega)$, equation (7) implies that $X_{-\mathbb{N}_0}$ determines X_1 uniquely, which is generically not true. The analogously defined one-sided shift on $M = \Delta^{\mathbb{N}}$ also does not satisfy (7). Note that, because future trajectories are equivalent to internal states, the associated process $Y_{\mathbb{Z}}$ is essentially the prediction process in the sense that $Y_k = Z_k \circ X_{\mathbb{Z}}$. $\Diamond$

Proposition 3.7 tells us that the next output symbol of a partially deterministic HMM is conditionally independent of the internal state, given the past output. But even more is true. The whole future output is conditionally independent of the internal state. Thus, if we know the past, the internal state provides no additional information useful for predicting the future output.

Corollary 3.9. *Let (T, μ) be partially deterministic, invariant, and $H(\mu) < \infty$. Then*

$$\mathbb{P}(X_{\mathbb{N}} \mid W_0 = m) \;=\; \mathbb{P}(X_{\mathbb{N}} \mid W_0 = \hat{m}) \qquad \forall m, \hat{m} \in \operatorname{supp}(Y_0) \qquad \textit{a.s.}$$

Proof. According to Proposition 3.7, $\mathbb{P}(X_1 \mid W_0 = \cdot) = K$ is constant on $\operatorname{supp}(Y_0)$. To obtain the statement for $X_{[1,n]}$, consider the n-tuple HMM defined as follows. The output space is Δ^n, the internal space is M, whereas the output and internal processes $\widehat{X}_{\mathbb{Z}}$ and $\widehat{W}_{\mathbb{Z}}$ are given by $\widehat{X}_k = X_{[(k-1)n+1,kn]}$ and $\widehat{W}_k = W_{nk}$. This is achieved by the HMM $(\widehat{T}, \mu)$ with $\widehat{T}\colon M \to \mathcal{P}(\Delta^n \times M)$, $\widehat{T}(m) = \mathbb{P}(X_{[1,n]}, W_n \mid W_0 = m)$. The HMM is obviously partially deterministic with transition function $f_{d_n} \circ \cdots \circ f_{d_1}$ and invariant. Thus Proposition 3.7 implies that $\mathbb{P}(X_{[1,n]} \mid W_0 = \cdot) = \mathbb{P}(\widehat{X}_1 \mid \widehat{W}_0 = \cdot)$ is constant on $\operatorname{supp}(\widehat{Y}_0)$. Because we can couple the processes such that $\widehat{Y}_0 = Y_0$, the claim follows. $\square$

3.3 Representations on prediction space

We can interpret any probability measure μ on prediction space $\mathcal{P}(\Delta^{\mathbb{N}})$ as initial distribution of an HMM. The "internal state update" of the corresponding transition $T^{\mathfrak{C}}$ follows the same rule as the prediction dynamic S, described by the conditional probability given the last observation. The difference is that now we include output symbols from Δ. We want to construct the HMM in such a way that if it is started in the internal state $z \in \mathcal{P}(\Delta^{\mathbb{N}})$, its output process is distributed according to z (which is also a measure on the future). Thus, the distribution of the next output d has to be equal to the marginal of z. The next internal state has to be the conditional z-probability of the future given $X'_1 = d$. Recall that $\phi_z(\omega) = z(\varsigma^{-1}{}_{(\,\cdot\,)} \mid X'_1)(\omega)$.

Definition 3.10. We define the Markov kernel $T^{\mathfrak{C}}$ from $\mathcal{P}(\Delta^{\mathbb{N}})$ to $\Delta \times \mathcal{P}(\Delta^{\mathbb{N}})$ by

$$T^{\mathfrak{C}}(z;\, D \times B) \;:=\; z\big(\{\, X'_1 \in D,\; \phi_z \in B \,\}\big), \qquad z \in \mathcal{P}(\Delta^{\mathbb{N}}),\; D \subseteq \Delta,\; B \in \mathfrak{B}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big).$$

Note that $T^{\mathfrak{C}}(z;\, \Delta \times B) = S(z;\, B)$, i.e. marginalising $T^{\mathfrak{C}}(z)$ to the internal component yields the prediction dynamic. Thus, if $\mu = \mu_{\mathfrak{C}}(P)$ is the causal state distribution (Definition 2.7) of some process $P \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$, then the internal state process of the induced HMM $(T^{\mathfrak{C}}, \mu)$ coincides with the prediction process $Z_{\mathbb{Z}}$ of P. From the following lemma we conclude that the output process $X_{\mathbb{Z}}$ is, as expected, distributed according to P. More generally, if $\mu \in \mathcal{P}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big)$ represents a process $z \in \mathcal{P}(\Delta^{\mathbb{N}})$ in the sense of integral representation theory as a mixture of other processes, it also induces an HMM of z, namely $(T^{\mathfrak{C}}, \mu)$. Recall that r is the resolvent, defined in (3), and associates the represented process to μ.

Lemma 3.11. *Let $\mu \in \mathcal{P}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big)$. Then $(T^{\mathfrak{C}}, \mu)$ is a partially deterministic HMM of $r(\mu)$. In particular, $\big(T^{\mathfrak{C}}, \mu_{\mathfrak{C}}(P)\big)$ is an invariant HMM of $P \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$.*

Proof. Partial determinism follows directly from the definition of $T^{\mathfrak{C}}$, with $K_z = z \circ {X'_1}^{-1}$ and transition function f given by $f_z \circ X'_1 := \phi_z$, which is well defined due to the $\sigma(X'_1)$-measurability of ϕ_z. Obviously $T^{\mathfrak{C}}(z;\, D \times B) = K_z\big(D \cap f_z^{-1}(B)\big)$. We assume w.l.o.g. that μ is a Dirac measure (the general claim follows by integration over μ). Thus let $\mu = \delta_z$ with $z = r(\mu)$. Recall that, according to Lemma 3.2, $\big(T^{\mathfrak{C}}, T^{\mathfrak{C}}_d(\delta_z)\big)$ is an HMM of the conditional probability of $X'_{[2,\infty[}$ given $X'_1 = d$ (w.r.t. the output process of $(T^{\mathfrak{C}}, \delta_z)$). With $T^{\mathfrak{C}}\big(z;\, \{\, d \,\} \times \mathcal{P}(\Delta^{\mathbb{N}})\big) = z\big(\{\, X'_1 = d \,\}\big)$ and

$$r\big(T^{\mathfrak{C}}_d(\delta_z)\big) \overset{(6)}{=} r\big(\delta_{f_z(d)}\big) \;=\; f_z(d) \;=\; z\big(\varsigma^{-1}(\,\cdot\,) \mid X'_1 = d\big),$$

the claim follows by induction. □

Remark (ε-machine). $(T^{\mathfrak{C}}, \mu_{\mathfrak{C}}(P))$ corresponds to the ε-machine of computational mechanics. It is in some sense a minimal predictive model but not always the minimal HMM of P (see [14]).

Given a process $P \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$, there are (usually) many invariant representations on prediction space (i.e. S-invariant $\nu \in \mathcal{P}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big)$ with $r(\nu) = P_{\mathbb{N}}$). The next proposition shows that the causal state distribution of P is distinguished among them as the *only* one that can have finite entropy.

Proposition 3.12. *Let $\nu \in \mathcal{P}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big)$ be S-invariant, and $P \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$ the measure it represents. If $\nu \neq \mu_{\mathfrak{C}}(P)$, then $H(\nu) = \infty$.*

Proof. Let $H(\nu) < \infty$. According to Lemma 3.11, $(T^{\mathfrak{C}}, \nu)$ is an invariant HMM of P and satisfies the conditions of Corollary 3.9. Let $W_{\mathbb{Z}}$ be the corresponding $M = \mathcal{P}(\Delta^{\mathbb{N}})$-valued internal process. For a.e. fixed ω, Lemma 3.2 tells us that $(T^{\mathfrak{C}}, \delta_{W_0(\omega)})$ is an HMM of $\mathbb{P}(X_{\mathbb{N}} \mid W_0)(\omega)$, but it is also an HMM of $r(\delta_{W_0(\omega)}) = W_0(\omega)$ due to Lemma 3.11. Thus, $\mathbb{P}(X_{\mathbb{N}} \mid W_0) = W_0$ and

$$z \;=\; \mathbb{P}(X_{\mathbb{N}} \mid W_0 = z) \overset{\text{(cor. 3.9)}}{=} \mathbb{P}(X_{\mathbb{N}} \mid W_0 = \hat{z}) \;=\; \hat{z} \qquad \forall z, \hat{z} \in \operatorname{supp}\big(Y_0(\omega)\big).$$

This means $\big|\operatorname{supp}(Y_0)\big| = 1$, i.e. $Y_0(\omega)$ is a Dirac measure. Thus $Y_0 = \mathbb{P}(W_0 \mid X_{-\mathbb{N}_0}) = \delta_{W_0}$ a.s. and

$$Z_0 \circ X_{\mathbb{Z}} \;=\; \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \;=\; \int \mathbb{P}(X_{\mathbb{N}} \mid W_0 = \,\cdot\,)\, \mathrm{d}Y_0 \;=\; \mathbb{P}(X_{\mathbb{N}} \mid W_0) \;=\; W_0 \qquad \text{a.s.}$$

Because W_0 is ν-distributed and $\mu_{\mathfrak{C}}(P)$ is the law of Z_0, we obtain $\nu = \mu_{\mathfrak{C}}(P)$. □

We conclude this section with two examples of representations on prediction space. They are extreme cases. The first one, ν_1, is maximally concentrated, namely ν_1 is the Dirac measure in (the future of) the process we want to represent. Thus it has no uncertainty in itself, but the (unique) process in its support can be arbitrary. The second example, ν_2, is supported by maximally concentrated processes, i.e. by Dirac measures on $\Delta^{\mathbb{N}}$, but the mixture ν_2 is as diverse as the original process. The HMM corresponding to ν_2 is equivalent to the one-sided shift (Example 3.8).

Example 3.13. Let $P \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$, $P_{\mathbb{N}} = P \circ X_{\mathbb{N}}^{-1}$ and $\nu = \delta_{P_{\mathbb{N}}}$. Then ν is a representation of $P_{\mathbb{N}}$ with $H(\nu) = 0$. This is no contradiction to Proposition 3.12 because ν is not S-invariant (if P is not i.i.d.) ♢

Example 3.14 (lifted shift). Let $P \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$ and $\nu = P_{\mathbb{N}} \circ \iota^{-1}$, where $\iota\colon \Delta^{\mathbb{N}} \to \mathcal{P}(\Delta^{\mathbb{N}})$, $\iota(x) = \delta_x$ is the embedding as Dirac measures. ν is an S-invariant representations of P and $(T^{\mathfrak{C}}, \nu)$ is equivalent to the one-sided shift. The only difference is that trajectories $x \in \Delta^{\mathbb{N}}$ are replaced by corresponding Dirac measures $\delta_x \in \mathcal{P}(\Delta^{\mathbb{N}})$. In other words, ι is an isomorphism. This is no contradiction to Proposition 3.12 because $H(\nu) = \infty$ (if P is not concentrated on countably many trajectories). ♢

4 Properties of the statistical complexity functional

Recall that the statistical complexity $C_{\mathfrak{C}}(P)$ (Definition 2.10) of a process $P \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$ is defined as the entropy $H\big(\mu_{\mathfrak{C}}(P)\big)$ of its causal state distribution. In this section, we investigate $C_{\mathfrak{C}}$ as a functional on the space of processes. First, we consider the problem of ergodic decomposition. With ergodic decomposition of P, we denote a probability measure ν on the ergodic measures $\mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}}) \subseteq \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$ that satisfies

$$P \;=\; r(\nu) \;=\; \int_{\mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}})} \mathrm{id}_{\mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}})}\, \mathrm{d}\nu.$$

Such a measure ν always exists and is uniquely determined by P. In [6, 7], Łukasz Dębowski investigated another complexity measure, *excess entropy*, and gave a formula for its ergodic decomposition. Here, we obtain the corresponding result for statistical complexity. It is the average complexity of the ergodic components plus the entropy of the mixture.

Proposition 4.1 (ergodic decomposition). *Let $\nu \in \mathcal{P}\big(\mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}})\big)$ be the ergodic decomposition of $P \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$. Then*

$$C_{\mathfrak{C}}(P) \;=\; \int C_{\mathfrak{C}}\, \mathrm{d}\nu + H(\nu).$$

Proof. First note that $\mu_{\mathfrak{C}}(P_1)$ and $\mu_{\mathfrak{C}}(P_2)$ are singular for distinct ergodic $P_1, P_2 \in \mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}})$. Indeed, there exist disjoint $A_1, A_2 \in \sigma(X'_{-\mathbb{N}_0})$ and disjoint $B_1, B_2 \in \sigma(X'_{\mathbb{N}})$ s.t. $P_k(A_k) = 1$ and $P_k(B_k \mid X'_{-\mathbb{N}_0})\restriction_{A_k} \equiv 1$. Consequently, if ν is not supported by a countable set, $\mu_{\mathfrak{C}}(P)$ cannot be supported by a countable set and $C_{\mathfrak{C}}(P) = H(\nu) = \infty$. Thus assume $\nu = \sum_{k\in\mathbb{N}} \nu_k \delta_{P_k}$ for some $\nu_k \geq 0$ and distinct $P_k \in \mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}})$. Then there are disjoint $A_k \in \sigma(X'_{-\mathbb{N}_0})$ s.t. $P_k(A_k) = 1$. We claim

$$P(\,\cdot \mid X'_{-\mathbb{N}_0}) \;=\; \sum_{k\in\mathbb{N}} 1_{A_k} P_k(\,\cdot \mid X'_{-\mathbb{N}_0}) \qquad P\text{-a.s.}$$

Indeed, the $\sigma(X'_{-\mathbb{N}_0})$-measurability is clear, and for $A \in \sigma(X'_{-\mathbb{N}_0})$, $F \in \mathfrak{B}(\Delta^{\mathbb{Z}})$ we have

$$\begin{aligned}
\int_A \sum_{k\in\mathbb{N}} 1_{A_k} P_k(F \mid X'_{-\mathbb{N}_0})\, \mathrm{d}P \qquad &= \qquad \sum_{j\in\mathbb{N}} \nu_j \int_A \sum_{k\in\mathbb{N}} 1_{A_k} P_k(F \mid X'_{-\mathbb{N}_0})\, \mathrm{d}P_j \\
&\overset{(P_j(A_j)\,=\,1)}{=} \quad \sum_j \nu_j \int_{A\cap A_j} P_j(F \mid X'_{-\mathbb{N}_0})\, \mathrm{d}P_j \\
&= \qquad \sum_j \nu_j P_j(F \cap A \cap A_j) \;=\; P(F \cap A).
\end{aligned}$$

As $P(A_k) = \nu_k$, it follows that $\mu_{\mathfrak{C}}(P) = \sum_k \nu_k \mu_{\mathfrak{C}}(P_k)$. Mutual singularity of the $\mu_{\mathfrak{C}}(P_k)$ implies

$$C_{\mathfrak{C}}(P) \;=\; H\left(\sum_k \nu_k \mu_{\mathfrak{C}}(P_k)\right) \;=\; \sum_k \nu_k H\big(\mu_{\mathfrak{C}}(P_k)\big) + H(\nu). \qquad \square$$

Several corollaries follow directly from this proposition. The set $\mathcal{P}_{\mathfrak{C}} := C_{\mathfrak{C}}^{-1}(\mathbb{R})$ of stationary processes with finite statistical complexity is convex, $C_{\mathfrak{C}}$ is concave but not continuous, and the set $\mathcal{P}_\infty := \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}}) \setminus \mathcal{P}_{\mathfrak{C}}$ of processes with infinite statistical complexity is dense.

Corollary 4.2 (concavity). *$\mathcal{P}_{\mathfrak{C}}$ is a convex set and $C_{\mathfrak{C}}$ is concave. Moreover, for all $\nu \in \mathcal{P}(\mathbb{N})$, $\nu_k := \nu(k)$ and $P_k \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$*

$$\sum_{k\in\mathbb{N}} \nu_k C_{\mathfrak{C}}(P_k) \;\leq\; C_{\mathfrak{C}}\Big(\sum_{k\in\mathbb{N}} \nu_k P_k\Big) \;\leq\; \sum_{k\in\mathbb{N}} \nu_k C_{\mathfrak{C}}(P_k) + H(\nu).$$

Proof. Use ergodic decomposition of the P_k and Proposition 4.1. $\square$

Corollary 4.3 (non-continuity). *$C_{\mathfrak{C}}\restriction_{\mathcal{P}_{\mathfrak{C}}}$ is not continuous in any $P \in \mathcal{P}_{\mathfrak{C}}$ w.r.t. variational topology, let alone w.r.t. weak-$*$ topology.*

Proof. Let $Q_n \in \mathcal{P}_{\mathfrak{C}}$ with $\lim_{n\to\infty} \frac{1}{n} C_{\mathfrak{C}}(Q_n) \to \infty$ and $P_n := \frac{n-1}{n} P + \frac{1}{n} Q_n$. Then $P_n \to P$ in variational topology, but $C_{\mathfrak{C}}(P_n) \geq \frac{1}{n} C_{\mathfrak{C}}(Q_n) \to \infty$ by Corollary 4.2. □

Corollary 4.4. *$\mathcal{P}_\infty$ is dense in $\mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$ w.r.t. variational- and a fortiori w.r.t. weak-$*$-topology.*

Proof. Let $P, Q \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$ with $C_{\mathfrak{C}}(Q) = \infty$. Then $\mathcal{P}_\infty \ni \frac{n-1}{n} P + \frac{1}{n} Q \to P$. □

We give a simple example of a situation where statistical complexity is not continuous.

Example 4.5 (non-continuity). Let Q_p be the Bernoulli process on $\Delta = \{0, 1\}$ with parameter $0 < p < 1$, i.e. $Q_p(X'_1 = 1) = p$. Consider the process of throwing a coin which is either slightly biased to 0 or 1, each with probability $\frac{1}{2}$, i.e. $P_\varepsilon = \frac{1}{2} Q_{\frac{1}{2}+\varepsilon} + \frac{1}{2} Q_{\frac{1}{2}-\varepsilon}$ with $0 < \varepsilon < \frac{1}{2}$. Then $P_\varepsilon \overset{*}{\to} P_0 = Q_{\frac{1}{2}}$ for $\varepsilon \to 0$, but $C_{\mathfrak{C}}(P_\varepsilon) = \log(2)$ for $\varepsilon > 0$ and $C_{\mathfrak{C}}(P_0) = 0$. ◊

The proof of our most important result about statistical complexity, namely its lower semi-continuity, makes use of the propositions given in Sections 2.1 and 3 It also uses a compactness argument. To this end we need, in the case of infinite Δ, a lemma guaranteeing that $\mu_{\mathfrak{C}}$ preserves relative compactness.

Lemma 4.6. *Let $\mathcal{M} \subseteq \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$ be relatively compact. Then $\mu_{\mathfrak{C}}(\mathcal{M}) := \{ \mu_{\mathfrak{C}}(P) \mid P \in \mathcal{M} \}$ is relatively compact in $\mathcal{P}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big)$.*

Proof. Using Prokhorov's theorem, we have to show that $\mu_{\mathfrak{C}}(\mathcal{M})$ is tight provided that $\mathcal{M}$ is tight. Let $\varepsilon > 0$ and $K_n \subseteq \Delta^{\mathbb{Z}}$ compact with $P(K_n) \geq 1 - \varepsilon \frac{2^{-n}}{n}$ for all $P \in \mathcal{M}$. We define $K'_n := X'_{\mathbb{N}}(K_n)$, $\widetilde{K} := \{ z \in \mathcal{P}(\Delta^{\mathbb{N}}) \mid z(K'_n) \geq 1 - \frac{1}{n} \; \forall n \in \mathbb{N} \}$ and $f_n := P\big(\{ X'_{\mathbb{N}} \in K'_n \} \mid X'_{-\mathbb{N}_0}\big)$. For $P \in \mathcal{M}$

$$\int f_n \,\mathrm{d}P \;\geq\; \int P(K_n \mid X'_{-\mathbb{N}_0}) \,\mathrm{d}P \;=\; P(K_n) \;\geq\; 1 - \varepsilon \tfrac{2^{-n}}{n}.$$

We obtain $P\big(\bigcup_n \{ f_n < 1 - \frac{1}{n} \}\big) \leq \sum_n n(1 - \int f_n \,\mathrm{d}P) \leq \sum \varepsilon\, 2^{-n} = \varepsilon$ and, as a consequence, $\mu_{\mathfrak{C}}(P)(\widetilde{K}) = P\big(\{ Z_0 \in \widetilde{K} \}\big) = P\big(\bigcap_n \{ f_n \geq 1 - \frac{1}{n} \}\big) \geq 1 - \varepsilon$ for all $P \in \mathcal{M}$. We still have to show compactness of $\widetilde{K}$. It is closed because $z_k \overset{*}{\to} z$ implies $z(K'_n) \geq \limsup_k z_k(K'_n)$ due to closedness of K'_n. It is tight by definition because the K'_n are compact. Therefore, $\widetilde{K}$ is compact. □

Theorem 4.7 (lower semi-continuity). *The statistical complexity functional, $C_{\mathfrak{C}} \colon \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}}) \to \overline{\mathbb{R}}_+$, is weak-$*$ lower semi-continuous.*

Proof. Let $P_n, P \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$ with $P_n \overset{*}{\to} P$. Every subsequence of $\big(\mu_{\mathfrak{C}}(P_n)\big)_{n\in\mathbb{N}}$ has an accumulation point (a.p.), according to Lemma 4.6. Consequently,

$$\liminf_{n\to\infty} C_{\mathfrak{C}}(P_n) \;=\; \liminf_{n\to\infty} H\big(\mu_{\mathfrak{C}}(P_n)\big) \overset{(H \text{ lsc})}{\geq} \inf\big\{ H(\nu) \mid \nu \text{ a.p. of } (\mu_{\mathfrak{C}}(P_n))_{n\in\mathbb{N}} \big\}.$$

Every $\mu_{\mathfrak{C}}(P_n)$ is S-invariant. According to Proposition 2.5, S is continuous and thus every a.p. ν of $\big(\mu_{\mathfrak{C}}(P_n)\big)_{n\in\mathbb{N}}$ is also S-invariant. The resolvent $r \colon \mathcal{P}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big) \to \mathcal{P}(\Delta^{\mathbb{N}})$ is continuous (see [3]), and thus ν represents P. Therefore, according to Proposition 3.12, $H(\nu) \geq C_{\mathfrak{C}}(P)$. In total we obtain

$$\liminf_{n\to\infty} C_{\mathfrak{C}}(P_n) \;\geq\; C_{\mathfrak{C}}(P).$$ □

We argue that, from a theoretical point of view, every complexity measure should be lower semi-continuous. While it is not counter intuitive that it is possible to approximate a simple system by unnecessarily complex ones (and hence the complexity is not continuous), it would be strange to consider a process complex if there is an approximating sequence with (uniformly) simple processes. Therefore, an axiomatic characterisation of complexity measures (although, of course, we are far from having such a characterisation) should include lower semi-continuity. There are also slightly more practical reasons why semi-continuity is a nice property.

In a model selection task, for instance, it might be desirable to impose some upper bound $a \in \mathbb{R}_+$ on the complexity of considered processes (e.g. to avoid overfitting). An important consequence of lower semi-continuity is that the set $C_{\mathfrak{C}}^{-1}\big([0,a]\big) = \big\{\, P \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}}) \;\big|\; C_{\mathfrak{C}}(P) \leq a \,\big\}$ of processes with complexity bounded by a is closed. This makes the complexity constraint technically easier. Consider any complete metric on $\mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$ compatible with weak-$*$ (or any stronger) topology (e.g. Prokhorov, Kantorovich-Rubinshtein or variational metric). Then due to closedness, for every $P \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$ with arbitrary complexity, there is a (not necessarily unique) closest "sufficiently simple" process P_a with complexity not exceeding a. Another consequence is that the set of processes with infinite complexity is generic in the following sense.

Corollary 4.8. *$\mathcal{P}_\infty$ contains a dense $\mathcal{G}_\delta$-set.*

Proof. Because all $C_{\mathfrak{C}}^{-1}\big([0,n]\big)$ are closed, $\mathcal{P}_\infty$ is a $\mathcal{G}_\delta$-set. It is dense according to Corollary 4.4. □

Example 4.9. Consider the experiment of first choosing a random coin with success probability p uniformly in $[0,1]$ and then generating an i.i.d. sequence with this coin. More precisely, let Q_p be the Bernoulli process with parameter p on $\Delta = \{\,0,1\,\}$ and $P = \int Q_p \,\mathrm{d}p$. Then P has infinite statistical complexity according to Proposition 4.1. We might approximate P by $P_n \overset{*}{\rightharpoonup} P$ (e.g. with ergodic P_n). Then Theorem 4.7 implies that the complexity of P_n necessarily tends to infinity. ◊

Example 4.10. Let Δ be finite, then $\mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$ is compact. Assume we made observations of a Δ-valued process and want to fit some $P \in \mathcal{P}_{\mathsf{inv}}(\Delta^{\mathbb{Z}})$. From the observations, we might derive a set of closed constraints, e.g. $P\big(\{\, X_1' = X_2' \,\}\big) \in [a,b]$, $P\big(\{\, X_1' = d \,\}\big) \geq \varepsilon$, and $P\big(\{\, X_2' = d \,\} \;\big|\; X_1' = d\big) \in [a,b]$ (the third is closed only in presence of the second). Further closed constraints may be given by modelling assumptions. Because the resulting set of admissible processes is compact, lower semi-continuity implies that there is at least one process of minimal complexity satisfying all constraints. ◊

A Appendix

Proof of Lemma 2.11 (lower semi-continuity of the entropy). Recall that $\varphi(x) := -x\log(x)$ and denote the boundary of a set B by ∂B. Define

$$\widehat{H}(\mu) \;:=\; \sup\Big\{\, \sum_{i=1}^{n} \varphi\big(\mu(B_i)\big) \;\Big|\; n \in \mathbb{N},\; B_i \text{ disjoint},\; \mu(\partial B_i) = 0 \,\Big\}.$$

Obviously, $\widehat{H} \leq H$. Recall that $\mu_n \overset{*}{\rightharpoonup} \mu$ implies $\mu_n(A) \to \mu(A)$ for all A with $\mu(\partial A) = 0$ (e.g. [2]). Thus $\widehat{H}$ is clearly lower semi-continuous and it is sufficient to show

$$H(\mu) \;\leq\; \widehat{H}(\mu).$$

If μ is not supported by any countable set, $\widehat{H}(\mu) = \infty$ due to separability of M. Let $\mu = \sum_{i=1}^{\infty} a_i \delta_{x_i}$ ($a_i \in [0,1]$, $x_i \in M$), and d a compatible metric on M. For fixed $n \in \mathbb{N}$, we can choose a radius $r_n > 0$, such that $B_i^n := \{\, x \in M \mid d(x_i,x) < r_n \,\}$, $i = 1,\dots,n$, are disjoint and $\mu(\partial B_i^n) = 0$. We get

$$\sum_{i=1}^{n} \varphi(a_i) \overset{(\varphi' \geq -1)}{\leq} \sum_{i=1}^{n} \varphi\big(\mu(B_i^n)\big) + \sum_{i=1}^{n} \big(\mu(B_i^n) - a_i\big) \;\leq\; \widehat{H}(\mu) + \sum_{i=n+1}^{\infty} a_i.$$

Therefore, $H(\mu) = \lim_{n\to\infty} \sum_{i=1}^{n} \varphi(a_i) \leq \widehat{H}(\mu)$. □

Proof of Lemma 3.2. We first prove that (T, δ_{W_0}) is an HMM of $\mathbb{P}(X_{\mathbb{N}} \mid W_0)$. Let $G_T(m) \in \mathcal{P}(\Delta^{\mathbb{N}})$ be the distribution of the output process of (T,δ_m). Because G_T is measurable, $G_T \circ W_0$ is $\sigma(W_0)$-measurable. From the definition of $(W_{\mathbb{N}_0}, X_{\mathbb{N}})$ it follows for measurable $B \subseteq M$, $A \subseteq \Delta^{\mathbb{N}}$ that

$$\mathbb{P}\big(\{\, W_0 \in B \,\} \cap \{\, X_{\mathbb{N}} \in A \,\}\big) \;=\; \int_B G_T(\,\cdot\,;A)\,\mathrm{d}\mu \;=\; \int_{W_0^{-1}(B)} G_T\big(W_0(\,\cdot\,);A\big)\,\mathrm{d}\mathbb{P},$$

where the second equality holds because W_0 is distributed according to μ. Thus $G_T \circ W_0$ is the claimed conditional probability. To see that $\big(T, L_{X_1}(\mu)\big)$ is an HMM of $\mathbb{P}(X_{[2,\infty[} \mid X_1)$, let $d \in \Delta$ and observe

$$\int G_T(\,\cdot\,;A)\,\mathrm{d}L_d(\mu) \;=\; \frac{1}{K_\mu(d)} \int\int_{\{d\}\times M} G_T(\,\cdot\,;A)\,\mathrm{d}T\,\mathrm{d}\mu \;=\; \frac{\mathbb{P}\big(\{\, X_1 = d,\; X_{[2,\infty[} \in A \,\}\big)}{\mathbb{P}\big(\{\, X_1 = d \,\}\big)}.$$ □

Acknowledgements

I am thankful to Nihat Ay for introducing me to computational mechanics, discussions, and all kinds of scientific support. I also thank the anonymous reviewers for their many helpful comments.

References

[1] Nihat Ay and James P. Crutchfield. Reductions of hidden information sources. *Journal of Statistical Physics*, 120:659–684, 2005.

[2] Patrick Billingsley. *Convergence of Probability Measures*. Wiley, New York, second edition, 1968.

[3] Gustave Choquet. *Lectures on Analysis, Volume II (Representation Theory)*. W. A. Benjamin, Inc., New York, 1969.

[4] Richard W. Clarke, Mervyn P. Freeman, and Nicholas W. Watkins. Application of computational mechanics to the analysis of natural data: An example in geomagnetism. *Phys. Rev. E*, 67(1):016203, 2003.

[5] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105–108, 1989.

[6] Łukasz Dębowski. Ergodic decomposition of excess entropy and conditional mutual information. IPI PAN Reports, nr 993, 2006.

[7] Łukasz Dębowski. A general definition of conditional information and its application to ergodic decomposition. *Statistics & Probability Letters*, 79:1260–1268, 2009.

[8] John Hopcroft and Jeffrey Ullman. *Introduction to Automata Theory, Language, and Computation*. Addison-Wesely, Reading, Massachusetts, 1979.

[9] Heike Jänicke, Alexander Wiebel, Gerik Scheuermann, and Wolfgang Kollmann. Multifield visualization using local statistical complexity. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1384–1391, 2007.

[10] Gerhard Keller. *Equilibrium States in Ergodic Theory*. London Mathematical Society, New York, 1998.

[11] Frank Knight. A predictive view of continuous time processes. *The Annals of Probability*, pages 573–96, 1975.

[12] Frank Knight. *Essays on the Prediction Process*, volume 1 of *Lecture Notes Series*. Institute of Mathematical Statistics, Hayward, CA, 1981.

[13] Frank Knight. *Foundations of the Prediction Process*. Oxford Science Publications, New York, 1992.

[14] Wolfgang Löhr and Nihat Ay. On the generative nature of prediction. *Advances in Complex Systems*, 12(2), 2009.

[15] P. Meyer. La théorie de la prédiction de F. Knight. *Seminaire de Probabilités*, X:86–103, 1976.

[16] Eckehard Olbrich, Nils Bertschinger, Nihat Ay, and Jürgen Jost. How should complexity scale with system size? *European Physical Journal B*, 63:407–415, 2008.

[17] Parthasarathy. On the category of ergodic measures. *Illinois J. Math.*, 5:648–656, 1961.

[18] Cosma R. Shalizi and James P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104:817–879, 2001.