Models of Discrete-Time Stochastic Processes and Associated Complexity Measures

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

im Fachgebiet

Mathematik

vorgelegt

von Diplommathematiker Wolfgang Löhr
geboren am 30. 12. 1980 in Nürnberg

Die Annahme der Dissertation wurde empfohlen von:

1. Prof. Dr. Jürgen Jost, Universität Leipzig

2. Prof. Dr. Gerhard Keller, Universität Erlangen-Nürnberg

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 12. 05. 2010 mit dem Gesamtprädikat magna cum laude

Abstract

Many complexity measures are defined as the size of a minimal representation in a specific model class. One such complexity measure, which is important because it is widely applied, is statistical complexity. It is defined for discrete-time, stationary stochastic processes within a theory called computational mechanics. Here, a mathematically rigorous, more general version of this theory is presented, and abstract properties of statistical complexity as a function on the space of processes are investigated. In particular, weak-$*$ lower semi-continuity and concavity are shown, and it is argued that these properties should be shared by all sensible complexity measures. Furthermore, a formula for the ergodic decomposition is obtained.

The same results are also proven for two other complexity measures that are defined by different model classes, namely process dimension and generative complexity. These two quantities, and also the information theoretic complexity measure called excess entropy, are related to statistical complexity, and this relation is discussed here.

It is also shown that computational mechanics can be reformulated in terms of Frank Knight's prediction process, which is of both conceptual and technical interest. In particular, it allows for a unified treatment of different processes and facilitates topological considerations. Continuity of the Markov transition kernel of a discrete version of the prediction process is obtained as a new result.

Acknowledgements

First of all, I want to thank my advisor Nihat Ay, without whom this thesis would obviously not have been possible. He always gave me advise when I needed it and let me go my way when I wanted to. I am also grateful to my official thesis advisor Jürgen Jost, and to Arleta Szkoła for fruitful discussions and advice. Further thanks go to William Kirwin for pointing out some language mistakes and to Arleta Szkoła for proofreading parts of the thesis. Last but not least, I thank the secretary of our group, Antje Vandenberg, who does a great job shielding everyone from any administrative problems.

Contents

Chapter 1

Introduction

An important task of complex systems sciences is to define complexity. Measures that quantify complexity are of both theoretical ([OBAJ08]) and practical interest. In applications, they are widely used to identify "interesting" parts of simulations and real-world data ([JWSK07]). There exist various measures of different kinds of complexity for different kinds of objects.

The main idea behind many complexity measures, such as statistical complexity discussed below, is the same that gave rise to the famous Kolmogorov complexity. Namely, the complexity is the "size" of some minimal "representation" of the object of interest. Different complexity measures are based on different precise definitions of these terms. For Kolmogorov complexity, for instance, representations are Turing machine programs computing individual binary strings, and the size is their length. For statistical complexity, on the contrary, the objects of interest are distributions of stochastic processes instead of individual strings, and the representations are particular kinds of prediction models. Their size is measured by the Shannon entropy of the internal states of the model.

In this thesis, the objects we are interested in are discrete-time, stationary stochastic processes with values in a state space Δ. In some parts, we have to restrict Δ to be countable (with discrete topology), but in most parts, we allow it to be a much more general space, namely a Souslin space. We aim to improve our understanding of some complexity measures and the classes of models used for their definitions. Our particular focus is on *statistical complexity* and the theory of prediction models called *computational mechanics* it is based on. Here, computational mechanics is a theory introduced by Jim Crutchfield and co-workers ([CY89, SC01, AC05]) that is unrelated to computer simulations of mechanical systems. It is applied to a variety of real-world data, e.g. in [CFW03]. In the present work, however, we are not considering applications, but are rather interested in a general, mathematically rigorous formulation of the theory.

Computational mechanics considers the following situation. Given a stationary stochastic process with time set $\mathbb{Z}$, the semi-infinite "past" (or "history") of the process (at all times up to and including zero) has been observed. Now the "future" (all positive times) has to be predicted as accurately as possible. The central objects of the theory are the *causal states*. They are defined as the elements of the minimal partition of the past that is sufficient for predicting the future of the process. An important, closely related concept is the so-called *ε-machine*, which is a particular hidden Markov model (HMM) on the causal states that encodes the mechanisms of prediction. Here, we show that causal states and ε-machine can be represented on the *prediction space* $\mathcal{P}(\Delta^{\mathbb{N}})$ of probability measures on the "future" $\Delta^{\mathbb{N}}$, making their close relation to a discrete-time version of Frank Knight's prediction process

([Kni92]) obvious. This representation underlines their importance, but it is also technically convenient and allows for a unified description of the ε-machines of different processes.

Statistical complexity is the entropy of the causal states or, equivalently, the internal state entropy of the ε-machine. While the causal states are the minimal sufficient partition of the past, it is an important fact that the ε-machine is not the minimal HMM of a given process. Namely, there can be HMMs with fewer internal states and lower internal state entropy. We take this observation as starting point to find on one hand a sub-class of HMMs in which the ε-machine is minimal, which turns out to be the case for *partially deterministic HMMs* (also known as deterministic stochastic automata). On the other hand, we provide a predictive interpretation of the potentially smaller HMMs.

Besides statistical complexity, we also discuss related quantities and their relation to statistical complexity, namely *excess entropy*, *generative complexity*, and *process dimension*. Excess entropy is a well-established, information theoretic complexity measure that can either be interpreted as the asymptotic amount of entropy exceeding the part determined by the entropy rate, or as the mutual information between the past and the future of the process. Generative complexity is a complexity measure based on minimal HMMs. Namely, it is the minimal internal state entropy of an HMM generating the given process. It was introduced recently by the author together with Nihat Ay in [LA09a]. Process dimension is a characteristic of the process ([Jae00]) that arises in the study of algebraic models called *observable operator models* (*OOMs*). OOMs are generalisations of HMMs, where the stochastic process of internal states is replaced by a linear evolution on an internal vector space. Process dimension is called *minimum effective degree of freedom* in [IAK92], but, to the best of our knowledge, it has not previously been interpreted as a complexity measure.

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. [Kel98]). We believe that these questions are worthwhile considering also for complexity measures. A formula for the ergodic decomposition of excess entropy was obtained in [Dęb06]. Our results presented here include the corresponding formula for statistical complexity and generative complexity. This formula directly implies concavity. The most important results in this direction are that all four quantities under consideration, excess entropy, statistical complexity, generative complexity and process dimension, are lower semi-continuous. Here, we equip the space of stochastic processes with the usual weak-$*$ topology (often called weak topology) and note that it is the most natural topology in our situation. Semi-continuity is a much stronger property in this topology than in the finer variational or information topology. While semi-continuity of excess entropy is more or less obvious, our proof in the case of statistical complexity uses results about partially deterministic HMMs and the prediction process obtained in earlier chapters. Our semi-continuity results for statistical complexity and process dimension cover only the case of a countable state space, but the corresponding result about generative complexity is more general. We consider lower semi-continuity to be an essential property for complexity measures, because it means that a process cannot be complex if it can be approximated by non-complex ones.

1.1 Structure and main results

Many of the important results presented in this thesis have been published recently by the author in [Löh09b], the author together with his advisor Nihat Ay in [LA09b, LA09a], or are

submitted for publication in [Löh09a].

Chapter 2 contains a review of some generative models of stochastic processes, namely HMMs and OOMs. These model classes are important for the definition of complexity measures and for a better understanding of predictive models. We introduce some notation and our technical framework. In particular, there are several slightly different but essentially equivalent definitions of HMMs in the literature and, after highlighting the differences, we define the version of HMMs that is used for the rest of the thesis. More specifically, our type of HMM is called transition-emitting Souslin HMM. In Sections 2.1.6 and 2.1.7, we consider the process of expectations of the internal state given the observed past, in particular in the special case of partially deterministic HMMs. The sub-class of partially deterministic HMMs is known better in the context of finite state stochastic automata. We extend the definition to Souslin spaces and obtain a new result in the countable case (Theorem 2.27, Corollary 2.29). Namely, the uncertainty of the internal state given the past output remains constant over time, and all internal states that are compatible with the observed past output induce the same expectation on the future output. This result has also been presented in [Löh09b].

Chapter 3 contains our discussion of predictive models. First, we review some information theoretic quantities that are necessary for the following sections, among them the excess entropy. In Section 3.2, we introduce and generalise the theory called computational mechanics, which is in particular used to define statistical complexity. This theory was until now only formulated for those processes with values in a countable space Δ that have countably many so-called causal states. The focus was primarily on applications and justifications from a physical point of view, rather than on a rigorous mathematical foundation. Therefore, the precise meaning of statements claiming minimality of the ε-machine remained unclear and lead to the misperception of the ε-machine as minimal generative HMM, although counterexamples have been known for a long time. Here, our contribution is the following. First, we extend the theory to arbitrary processes with values in a Souslin space and the considered model class from deterministic memory maps to stochastic memory functions, i.e. to Markov kernels (Propositions 3.10, 3.20 and 3.25). Second, we make the relation to generative HMMs more explicit (Propositions 3.12 and 3.14). Third, we compare the traditional approach of considering measurable partitions of the past with considering sub-σ-algebras, which might seem more appropriate from a measure theoretic perspective. The result is that both approaches are equivalent, provided we make a restriction to countably generated σ-algebras, which directly corresponds to the Souslin property of the space of memory states (Proposition 3.16 and the surrounding discussion). Fourth, we briefly show how to modify the definition of causal states using random times in order to deal with finite but varying observation lengths (Section 3.2.5). Most of these results, with the exception of Propositions 3.14 and 3.16, have been presented in the appendix of [LA09b].

In Section 3.3, we present a new predictive interpretation of HMMs, which was introduced in [LA09b]. We use these concepts and the results about partially deterministic HMMs developed in Chapter 2 to prove a minimality property of the ε-machine. Namely, it is the minimal partially deterministic HMM (Theorem 3.41, Corollary 3.42). The idea that partial determinism plays a crucial role is not new, but we do not know of any former mathematical proof that it ensures minimality of the ε-machine. This result is submitted for publication in [Löh09a]. Following [LA09a], we also suggest to consider, in analogy to statistical complexity, the minimal internal state entropy of a generative HMM as complexity measure called *generative complexity* (Section 3.3.2).

In Section 3.4, we show that the concepts of computational mechanics are closely related to a discrete-time version of Knight's prediction process and that there are representations of causal states and ε-machine on prediction space $\mathcal{P}(\Delta^{\mathbb{N}})$. This viewpoint was introduced in [Löh09b]. We call the prediction space versions of causal states and ε-machine *effect space* and *prediction HMM* respectively and show in Proposition 3.52 that the prediction space versions are indeed isomorphic to the classical ones. The terminology does not follow the one used in [Löh09b], because the representations on prediction space do not admit the intuition of "causal" anymore, and thus new names seem appropriate. We call the prediction space version of the distribution of the causal states *effect distribution* and prove the following remarkable property (Proposition 3.53). There may be many measures on prediction space that are invariant w.r.t. the prediction dynamic and represent a given process in the sense of integral representation theory. But all of them have infinite entropy, except, possibly, the effect distribution. The formulations on prediction space have several advantages from a theoretical point of view, such as providing a natural topology and describing different processes on a common space. They are also very convenient for comparison with the canonical OOM and the excess entropy. In this regard, we obtain a close relationship between the effect space and the canonical OOM that is not at all obvious when one thinks of causal states as equivalence classes of past trajectories. Namely, the weak-$*$ closure of the linear hull of the effect space coincides with the canonical OOM vector space (Theorem 3.56). In Section 3.4.4, we express the excess entropy as function of the effect distribution. The results of Section 3.4, with the exception of the unpublished Theorem 3.56 and the discussion in Section 3.4.4, are published in [Löh09b].

Chapter 4 contains results about the complexity measures excess entropy, statistical complexity, and generative complexity considered as functions on the space of stochastic processes. All three of them are lower semi-continuous in the weak-$*$ topology, concave, and satisfy the following ergodic decomposition formula. The complexity of a process is the average complexity of its ergodic components plus the entropy of the mixture. We call complexity measures with this ergodic decomposition behaviour entropy-based and show in Proposition 4.6 that all of them are concave, non-continuous and generically infinite. That excess entropy has the above mentioned properties was already known and is considered briefly for completeness. The lower semi-continuity and ergodic decomposition results for statistical complexity (Theorems 4.10 and 4.12) are published in [Löh09b]. Our semi-continuity result covers only the case of a countable state space Δ. The corresponding results for generative complexity (Theorems 4.13 and 4.15) are submitted for publication in [Löh09a] under the assumption of finite Δ. In this thesis, we treat the more general case of a Souslin space Δ.

In Section 4.5, we suggest to consider process dimension as a complexity measure. In the case of countable Δ, we show that it is lower semi-continuous (Theorem 4.16) and, although it is not entropy-based, satisfies a simple ergodic decomposition formula (Theorem 4.17). The dimension of a process only depends on the ergodic components and not on their weights. More precisely, it is the sum of the dimensions of the ergodic components. These properties of process dimension are not yet published.

The **appendix** provides proofs of some technical results that are needed in the main part and (presumably) well-known but not so easy to locate explicitly in the literature. We also recall some properties of Souslin spaces and the extension results of Kolmogorov and Ionescu-Tulcea in the appendix.

1.2 Notation

In this section, we introduce some notation that is used throughout the thesis.

Measures and topology: In this thesis, $(\Delta, \mathcal{D})$ and $(\Gamma, \mathcal{G})$ are always measurable spaces and usually assumed to be separable, metrisable topological spaces. In this case, we implicitly assume that $\mathcal{D} = \mathfrak{B}(\Delta)$ and $\mathcal{G} = \mathfrak{B}(\Gamma)$ are the respective Borel σ-algebras. With $\mathcal{P}(\Delta)$, we denote the space of probability measures on $(\Delta, \mathcal{D})$ and equip $\mathcal{P}(\Delta)$ with the σ-algebra $\sigma\big(\mu \mapsto \mu(D),\ D \in \mathcal{D}\big)$, generated by the evaluations in measurable sets. Here, σ denotes the generated σ-algebra. If Δ is a topological space, we always impose the weak-$*$ topology (often simply called weak topology) on $\mathcal{P}(\Delta)$. Note that if Δ is separable and metrisable, the Borel σ-algebra $\mathfrak{B}\big(\mathcal{P}(\Delta)\big)$ coincides with the σ-algebra of evaluations. We use the arrow $\xrightarrow{*}$ to denote weak-$*$ convergence. If Δ is countable, we implicitly assume discrete topology and for $d \in \Delta$, $\mu \in \mathcal{P}(\Delta)$ we sometimes write $\mu(d)$ instead of $\mu\big(\{\, d \,\}\big)$.

Integrals: If $f\colon \Gamma \to \mathbb{R}$ is integrable and $\mu \in \mathcal{P}(\Gamma)$ we use the notation

$$\int f \,\mathrm{d}\mu \;=\; \int_{x\in\Gamma} f(x) \,\mathrm{d}\mu.$$

Note that in our notation $\int f \,\mathrm{d}\mu(x)$ never means that x is the integration variable, but that the measure $\mu(x) \in \mathcal{P}(\Gamma)$ depends on x. If K is a measure-valued measurable function, i.e. $K\colon \Gamma \to \mathcal{P}(\Delta)$, then

$$\nu = \int K \,\mathrm{d}\mu \qquad \text{means} \qquad \nu(D) \;=\; \int_{g\in\Gamma} K(g)(D) \,\mathrm{d}\mu \quad \forall D \in \mathcal{D}.$$

Note that due to the dominated convergence theorem, ν is a well-defined probability measure $\nu \in \mathcal{P}(\Delta)$. The integral can be seen as Gel'fand integral, that is we have

$$\int f \,\mathrm{d}\big(\textstyle\int K \,\mathrm{d}\mu\big) \;=\; \displaystyle\int\int f \,\mathrm{d}K \,\mathrm{d}\mu$$

for bounded measurable f. Recall that if Γ and Δ are separable, metrisable spaces and K is continuous, then the function $\mu \mapsto \int K \,\mathrm{d}\mu$ is continuous as well.

Markov kernels: We consider a Markov kernel (transition probability) K from Γ to Δ to be a measurable function $K\colon \Gamma \to \mathcal{P}(\Delta)$. This definition is obviously equivalent to the perhaps more common definition as function $\Gamma \times \mathcal{D} \to \mathbb{R}$. We use the notation $K(g;\, D) := K(g)(D)$ for the probability of $D \in \mathcal{D}$ w.r.t. the measure $K(g)$, where $g \in \Gamma$. If $\mu \in \mathcal{P}(\Gamma)$, we define the product $\mu \otimes K \in \mathcal{P}(\Gamma \times \Delta)$ by

$$\mu \otimes K(G \times D) \;:=\; \int_G K(\,\cdot\,;\, D) \,\mathrm{d}\mu \qquad \forall G \in \mathcal{G},\ D \in \mathcal{D}.$$

The product between kernels $K_1\colon \Gamma \to \mathcal{P}(\Delta_1)$ and $K_2\colon \Delta_1 \to \mathcal{P}(\Delta_2)$ is defined as the kernel $K_1 \otimes K_2\colon \Gamma \to \mathcal{P}(\Delta_1 \times \Delta_2)$ with

$$(K_1 \otimes K_2)(g;\, D_1 \times D_2) \;:=\; \big(K_1(g) \otimes K_2\big)(D_1 \times D_2) \;=\; \int_{D_1} K_2(\,\cdot\,;\, D_2) \,\mathrm{d}K_1(g)$$

for $g \in \Gamma$, $D_1 \in \mathcal{D}_1$ and $D_2 \in \mathcal{D}_2$.

Conditional probability kernels: If not explicitly stated otherwise, random variables are defined on a common probability space $(\Omega, \mathfrak{A}, \mathbb{P})$ and ω is always an element of Ω. The distribution of a random variable X is denoted by

$$\mathbb{P}_X \;:=\; \mathbb{P} \circ X^{-1}.$$

Let X be a Γ-valued and Y a Δ-valued random variable. We usually impose restrictions on Δ that guarantee the existence of regular versions of conditional probability of Y. In such a case, we implicitly assume that a regular version is chosen and denote the conditional probability kernel from Ω to Δ by $\mathbb{P}(Y \mid X)$. Thus, $K = \mathbb{P}(Y \mid X)$ means $K(\omega;\, A) = \mathbb{P}\big(\{\, Y \in A \,\} \;\big|\; X\big)(\omega)$. Similarly, the corresponding kernel from Γ to Δ is denoted by $\mathbb{P}(Y \mid X = \,\cdot\,)$.

Stochastic processes: We consider Δ-valued stochastic processes in discrete time, $X_{\mathbb{Z}} := (X_k)_{k\in\mathbb{Z}}$ or $X_{\mathbb{N}} := (X_k)_{k\in\mathbb{N}}$. Sometimes, we also call the distribution $P = \mathbb{P}_{X_{\mathbb{Z}}} \in \mathcal{P}(\Delta^{\mathbb{Z}})$ of $X_{\mathbb{Z}}$ stochastic process. If $X_{\mathbb{Z}}$ is stationary, P is in the subset $\mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}}) \subseteq \mathcal{P}(\Delta^{\mathbb{Z}})$ of shift-invariant probability measures. Let $X'_k \colon \Delta^{\mathbb{Z}} \to \Delta$, $k \in \mathbb{Z}$, 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 $X_{\mathbb{Z}}$. For simplicity of notation, we denote the canonical projections on $\Delta^{\mathbb{N}}$ with the same symbols, X'_k, as the projections on $\Delta^{\mathbb{Z}}$. The distribution of the restriction to positive times is denoted by $P_{\mathbb{N}} := P_{X'_{\mathbb{N}}} = \mathbb{P}_{X_{\mathbb{N}}}$. We use interval notation also for discrete intervals, e.g. $[1,n] = \{\, 1, \ldots, n \,\}$ and for $D_1, \ldots, D_n \subseteq \Delta$ we define $D_{[1,n]} := D_1 \times \cdots \times D_n$. We denote the corresponding cylinder set by

$$[D_1 \times \cdots \times D_n] \;:=\; \{\, X'_{[1,n]} \in D_{[1,n]} \,\} \;=\; \{\, X'_k \in D_k,\; k = 1, \ldots, n \,\}$$

or, in the case of countable Δ and $d_1, \ldots, d_n \in \Delta$, by

$$[d_1, \ldots, d_n] \;:=\; \big[\{\, d_1 \,\} \times \cdots \times \{\, d_n \,\}\big].$$

Given a process $X_{\mathbb{Z}}$, we interpret $X_{\mathbb{N}}$ as future and $X_{-\mathbb{N}_0}$ as past of the process. We often need the conditional probability kernel of the future given the past. Sometimes, we abbreviate

$$\mathbb{P}_{X_{\mathbb{N}}}^{X_{-\mathbb{N}_0}} \;=\; \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \qquad \text{and} \qquad P_{\mathbb{N}}^{-\mathbb{N}_0} \;=\; P(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0}).$$

Chapter 2

Generative models

In this chapter, we consider the task of generating a discrete-time stochastic process. More precisely, we compare different models that are able and commonly used to represent stochastic processes. Given such a model, it is possible to simulate the process by producing sample trajectories efficiently. We consider two different model classes. The first one consists of different flavours of hidden Markov models (HMMs) and introduces a hidden *Markovian* dynamics of unobservable internal states. All components, including the constructed internal one, have a probabilistic interpretation as stochastic processes. The second class of more algebraic representations, called observable operator models (OOMs), is closely related to HMMs. It admits potentially more concise representations but the internal evolution is no longer described by a stochastic process.

2.1 Hidden Markov models

There is an extensive literature about Markov processes and the Markov property allows to solve a lot of problems that are intractable for general processes. But of course, not all processes of interest are Markovian. The idea of hidden Markov models (HMMs) is to model more general processes as "observable" parts of larger Markovian systems with an internal and an observable component. The internal component is often assumed to be finite, but we will not generally make this restriction. In the literature, many definitions of HMMs are in use and they differ in several details. In Section 2.1.2, we compare the main differences. From Section 2.1.4 on, we consider only one type of HMM, namely transition-emitting HMMs, although this is not the most common type. It is, however, the most convenient one for our purposes and the one used in computational mechanics.

In many applications of HMMs, the internal states have a concrete physical or conceptual meaning. Even more, they often are the objects of interest that are to be inferred. Thus, the internal states cannot be chosen freely but are an essential part of the modelling. This is commonly the case for HMMs in computational biology ([Kos01, HSF97]). The HMM is used to compute, for each observed sequence, an estimate of the sequence of internal states. This calculation is known as *smoothing* and can be solved by the famous *forward-backward algorithm*. A related task is parameter estimation. Usually, only the architecture of an HMM is fixed by the design process and the actual values for the (or some) probabilities have to be learnt from training data. This can be achieved by the *EM algorithm*. For an extensive treatment of these and other algorithms, see [CMR05].

Here, we take a different point of view and are more interested in HMMs as generative models. We consider the internal component to be hypothetical. For us, its main purpose is to allow for a compact description of the observable process and an efficient computation of its finite-dimensional marginal probabilities. This interpretation is sometimes used for HMMs in speech recognition ([Jel99]).

2.1.1 Markov processes

We assume the reader to be familiar with the concept of Markov processes and the main purpose of this section is to fix notation. Let $(\Delta, \mathcal{D})$, for the moment, be an arbitrary measurable space. A Δ-valued stochastic process $(X_n)_{n\in\mathbb{N}}$, denoted for brevity by $X_{\mathbb{N}}$, satisfies the **Markov property** if

$$\mathbb{P}\big(\{\, X_{k+1} \in D \,\} \;\big|\; X_{[1,k]}\big) \;=\; \mathbb{P}\big(\{\, X_{k+1} \in D \,\} \;\big|\; X_k\big) \qquad \text{a.s. } \forall k \in \mathbb{N},\ D \in \mathcal{D},$$

where $[1,k] = 1,\ldots,k$ denotes the discrete interval and $X_I = (X_n)_{n\in I}$ for every index set I. The standard way to specify a process with the Markov property is in terms of an initial distribution $\mu \in \mathcal{P}(\Delta)$ and Markov kernels (transition probabilities) T_k from Δ to Δ. μ determines the distribution of X_1, and T_k specifies the conditional distributions of X_{k+1} given X_k. The initial distribution μ together with the Markov kernels T_k define, according to the Ionescu-Tulcea extension theorem (e.g. [Nev65, Prop. V.1.1]), a unique probability measure $P \in \mathcal{P}(\Delta^{\mathbb{N}})$, satisfying

$$P_{[1,n]} \;:=\; P \circ {X'_{[1,n]}}^{-1} \;=\; \mu \otimes \bigotimes_{k=1}^{n} T_k, \tag{2.1}$$

where $\Delta^{\mathbb{N}}$ is equipped with the product σ-algebra. Note that we cannot use the Kolmogorov extension theorem instead of Ionescu-Tulcea's, unless we impose restrictions on Δ. See Appendix A.2 for a short discussion of the extension theorems. Any process with distribution P as defined by (2.1) satisfies the Markov property and

$$\mathbb{P}\big(\{\, X_{k+1} \in D \,\} \;\big|\; X_k\big)(\omega) \;=\; T_k\big(X_k(\omega);\, D\big) \qquad \text{a.s.} \qquad \forall D \in \mathcal{D}.$$

Even more, the right-hand side is a regular version of conditional probability and thus we may assume that both sides agree:

$$\mathbb{P}(X_{k+1} \mid X_k) \;=\; T_k \circ X_k.$$

In full generality of Δ, not every process with the Markov property arises from kernels T_k and initial distribution μ as above. The reason is that regular versions of conditional probability need not exist. In nearly all parts of this thesis, however, we impose restrictions on the measurable spaces that ensure the existence of regular versions. In this case, all distributions of processes with the Markov property satisfy (2.1) with

$$T_k \;:=\; \mathbb{P}(X_{k+1} \mid X_k) \qquad \text{and} \qquad \mu \;:=\; \mathbb{P}_{X_1} \;=\; \mathbb{P} \circ X_1^{-1}.$$

Since we are interested in *generative* models, we take the existence of the Markov kernels T_k as part of our definition of Markov process (T_k is interpreted as generative mechanism). We restrict ourselves to (time) homogeneous Markov processes, i.e. the case where $T_k = T$ for all k, and the term Markov process shall always mean homogeneous Markov process.

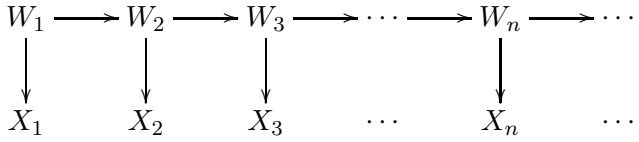
Definition 2.1. A **Markov model** of $P \in \mathcal{P}(\Delta^{\mathbb{N}})$ is a pair (T, μ), where T is a Markov kernel from Δ to Δ and $\mu \in \mathcal{P}(\Delta)$ is the initial distribution such that (2.1) is satisfied for $T_k = T$. The measure P, as well as any process $X_{\mathbb{N}}$ with distribution P, is called generated by (T, μ). A process $X_{\mathbb{N}}$ is called (homogeneous) **Markov process** if there is a Markov model generating it.

2.1.2 Different types of HMMs

The oldest type of HMM (the term "HMM" was introduced much later) is also the most restrictive one. It is called function of a Markov chain (sometimes functional of a Markov chain). The intuition is that we cannot observe the Markov process directly but only a function (coarse graining) of it. A Δ-valued process $X_{\mathbb{N}}$ is a function of a Markov chain if there is a Markov process $W_{\mathbb{N}}$ with values in some measurable space $(\Gamma, \mathcal{G})$, and a measurable function $f\colon \Gamma \to \Delta$ such that $X_k = f(W_k)$. Of course, if we do not impose restrictions on Γ, *every* process $X_{\mathbb{N}}$ is a function of a Markov chain (a possible representation is the shift, see Example 2.7). Usually, in the literature, Γ is assumed to be finite and in this case, not all processes are functions of finite Markov chains. We do, however, not always restrict to the finite case. The Markov process $W_{\mathbb{N}}$, also called internal process, is specified by a Markov model (T, μ), i.e. by initial distribution and transition kernel.

Definition 2.2. A **functional HMM** with internal space Γ and output space Δ is a triple (T, μ, f), where $\mu \in \mathcal{P}(\Gamma)$ and both $T\colon \Gamma \to \mathcal{P}(\Gamma)$ and $f\colon \Gamma \to \Delta$ are measurable. The process $W_{\mathbb{N}}$ generated by the Markov model (T, μ) is called **internal process** and the process $X_{\mathbb{N}}$ defined by $X_k := f(W_k)$ is called **function of a Markov chain** or **output process** of the functional HMM.

A natural generalisation of functions of Markov chains is to consider stochastic instead of deterministic functions. This corresponds to a noisy observation channel ([BP66]). The resulting model is the most common type of HMM. More specifically, we call this type state-emitting HMM, because we interpret the observed symbols from Δ as emitted by a machine that is described by the HMM. And the probability distribution for the emitted symbol only depends on the current internal *state* (as opposed to the whole *transition* from one internal state to the next one, which we consider below). In terms of graphical models,[1] the dependence structure is visualised as



Definition 2.3. (T, μ, K) is called **state-emitting HMM** if $T\colon \Gamma \to \mathcal{P}(\Gamma)$, $K\colon \Gamma \to \mathcal{P}(\Delta)$ are measurable and $\mu \in \mathcal{P}(\Gamma)$.

For our purposes, a less restrictive version of HMM is more convenient. Given a state-emitting HMM (T', μ, K), we can combine the kernels T' and K into one joint kernel T from Γ to $\Gamma \times \Delta$, namely $T := T' \otimes K$. Now it seems natural to consider arbitrary, not necessarily factorising such kernels, describing the joint production of the output symbol and the next

[1] We do not require knowledge of graphical models, but use it only for visualisations that should be intuitive enough. See, [Lau96] for a treatment of graphical models.

internal state. Because new internal state and output symbol are jointly determined, the distribution of the output symbol depends on the whole transition instead of just one of the internal states. Therefore, in contrast to state-emitting HMMs, we call such HMMs transition-emitting. They are, for instance, used in [Jel99]. Transition-emitting HMMs are also known under the name of *stochastic output automata* (see [Buk95]), a name most directly linked to the intuition of a "machine" that has internal (unobservable) states Γ and, at each time step, emits a symbol from the space Δ while updating its internal state.

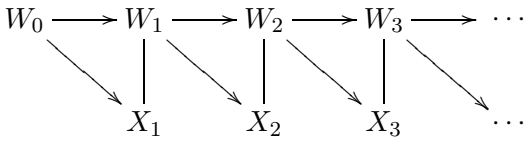
In the definitions of partially deterministic HMMs in Section 2.1.7, it is notationally more convenient to change the order of output symbol and new internal state. Thus we interpret T as kernel from Γ to $\Delta \times \Gamma$. The pair (T, μ) generates an internal process $W_{\mathbb{N}_0}$ ($\mathbb{N}_0 := \mathbb{N} \cup \{0\}$) on Γ and a (coupled) output process $X_{\mathbb{N}}$ on Δ, such that W_0 is μ-distributed and the joint process is Markovian with

$$\mathbb{P}\big(\{ X_{k+1} \in G,\ W_{k+1} \in D, \} \,\big|\, W_k, X_k\big) \;=\; T(W_k;\, D \times G), \qquad \forall D \in \mathcal{D},\ G \in \mathcal{G},$$

where we can assume, as in Section 2.1.1, that the equality always (not only a.s.) holds and write

$$\mathbb{P}(X_{k+1}, W_{k+1} \mid W_k, X_k) \;=\; T \circ W_k.$$

The dependence structure can be visualised as



Remark. In our definition of transition-emitting HMMs, the internal process starts one time step earlier than the output process. Thus, if we want to interpret a state-emitting HMM (T', μ', K) of a non-stationary process as transition-emitting HMM by defining $T = T' \otimes K$, there is a minor issue concerning the first output symbol. (T, μ') generates the shifted process, where the first symbol is dropped and there may not exist a $\mu \in \mathcal{P}(\Gamma)$ such that the one-step iterate $\int T \,\mathrm{d}\mu$ has the correct marginals. This problem can be solved by adding an additional start state to Γ.

If we define $T'(d, g) := T(g)$, the joint distribution of $W_{\mathbb{N}_0}$ and $X_{\mathbb{N}}$ generated by a transition-emitting HMM is given by $\mu \otimes \bigotimes_{k \in \mathbb{N}} T'$. More explicitly, we obtain for finite-dimensional sets

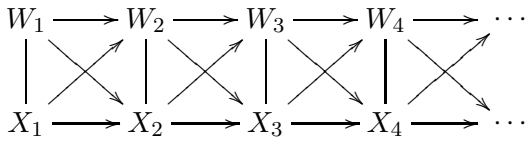
$$\begin{aligned}&\mathbb{P}\big(\{ W_{[0,n]} \in G_{[0,n]},\ X_{[1,n]} \in D_{[1,n]} \}\big)\\ &\qquad = \int_{g_0 \in G_0} \int_{(d_1, g_1) \in D_1 \times G_1} \cdots \int_{(d_n, g_n) \in D_n \times G_n} 1 \,\mathrm{d}T(g_{n-1}) \,\cdots\, \mathrm{d}T(g_0) \,\mathrm{d}\mu.\end{aligned}$$

Definition 2.4. A **transition-emitting HMM** is a pair (T, μ) with $\mu \in \mathcal{P}(\Gamma)$ and measurable $T\colon \Gamma \to \mathcal{P}(\Gamma \times \Delta)$. μ is called **initial distribution**, and T is called **generator**. We say that (T, μ) is an HMM of $P \in \mathcal{P}(\Delta^{\mathbb{N}})$ if $P = \mathbb{P}_{X_{\mathbb{N}}}$ and that (T, μ) generates the output process $X_{\mathbb{N}}$ or its distribution P.

The even more general notion of partially observed Markov process[2] allows the next internal state and output symbol to depend on the last output symbol as well as the internal state.

[2]The term "partially observed Markov process" sometimes also refers to transition- or state-emitting HMMs.

Thus it is just a (homogeneous) Markov process on a product space, where only one component (Δ) is considered to be observable, whereas the other component (Γ) consists of hidden states. Note that here *both* marginal processes need not be Markovian. The dependence structure can be visualised as



Definition 2.5. A **partially observed Markov model** is a Markov model (T, μ) on a product space $\Gamma \times \Delta$, where only the Δ-component is considered observable.

It is a trivial but important observation that the four discussed flavours of HMMs, partially observed Markov models, transition-emitting HMMs (stochastic automata), state-emitting HMMs, and functional HMMs (functions of Markov chains) are essentially equivalent in the following sense. To every partially observed Markov process (the most general notion), one can canonically associate a functional HMM (the most restrictive notion) such that the cardinality of the internal state space increases only by the constant factor of the cardinality of the output space. In fact, the new set of internal states is the product space $\Gamma' = \Gamma \times \Delta$. In particular, if Δ is finite, the classes of processes generated by finite functional HMMs, finite state-emitting HMMs, finite transition-emitting HMMs and finite partially observed Markov models coincide.

Proposition 2.6. *Let (T, μ) be a partially observed Markov model with space Γ of internal states and output space Δ. Then there is a functional HMM with internal state space $\Gamma' := \Gamma \times \Delta$ that generates the same output process.*

Proof. Let $f\colon \Gamma' \to \Delta$ be the canonical projection. Then (T, μ, f) is obviously a functional HMM of the same process. □

Transition-emitting HMMs are best suited for the following discussion, in particular the partial determinism property discussed in Section 2.1.7 is most natural for this class of HMMs. Furthermore, transition-emitting HMMs are more closely related to the algebraic models discussed in Section 2.2 below than the other types of HMMs are. Due to the essential equivalence of the different types, we feel free to restrict ourselves to transition-emitting HMMs and, from now on, all considered HMMs are transition-emitting. For every stochastic process $X_{\mathbb{N}}$, there exists an HMM that generates it. The most basic one is the (one-sided) shift.

Example 2.7 (shift). Let $P \in \mathcal{P}(\Delta^{\mathbb{N}})$ be the distribution of $X_{\mathbb{N}}$. Define $\Gamma := \Delta^{\mathbb{N}}$ and let σ denote the left-shift on $\Delta^{\mathbb{N}}$. Then there is an HMM of $X_{\mathbb{N}}$ with internal space Γ, such that the internal state coincides with the output trajectory. If the HMM is in the internal state $g = (g_1, g_2, \ldots) \in \Gamma$, the output symbol is g_1 and the next internal state is $\sigma(g) = (g_2, g_3, \ldots)$. More precisely, let $T^\sigma\colon \Gamma \to \mathcal{P}(\Delta \times \Gamma)$ with

$$T^\sigma(g) \;:=\; \delta_{X_1'(g)} \otimes \delta_{\sigma(g)} \;=\; \delta_{(g_1, \sigma(g))},$$

where δ_x is the Dirac measure in x, i.e. $\delta_x = 1_A(x)$ is one if and only if $x \in A$ and zero otherwise. We call the transition-emitting HMM (T^σ, P) one-sided shift HMM of P. It is obvious that it indeed generates P. ◊

2.1.3 Countable HMMs

In this section, we briefly look at the simplified, important special case where both the internal space Γ and the output space Δ are countable. In this case, probability measures can be regarded as vectors and transition probabilities as matrices. In this section, we use the most frequently used convention, which is to write measures as row vectors and multiply them to transition matrices from the left.

Definition 2.8. An HMM is called **countable HMM** if the internal space Γ and the output space Δ are countable. It is called **finite** if both spaces are finite.

Let (T, μ) be a countable, transition-emitting HMM. For $w = (w_0, \dots, w_n) \in \Gamma^{n+1}$ and $x = (x_1, \dots, x_n) \in \Delta^n$, the joint probability that internal and output process start with the values w and x respectively is given by

$$\mathbb{P}\big(\{\, W_{[0,n]} = w,\ X_{[1,n]} = x \,\}\big) \;=\; \mu(w_0) \prod_{k=1}^{n} T(w_{k-1};\, x_k, w_k),$$

and the formula for the distribution P of the observable process is

$$P\big([x_1, \dots, x_n]\big) \;=\; \sum_{g_0,\dots,g_n \in \Gamma} \mu(g_0) \prod_{k=1}^{n} T(g_{k-1};\, x_k, g_k).$$

Let l and m denote the number of elements of Δ and Γ respectively. Even if $l, m < \infty$, the number of terms in the above sum increases exponentially with n. Therefore, to actually compute the probability of an output sequence x, a different method is required. In order to use the properties of matrix multiplication, we split the $m \times l \cdot m$ matrix $\big(T(g;\, d, \hat{g})\big)_{g \in \Gamma, (d,\hat{g}) \in \Delta \times \Gamma}$ describing T into l different $m \times m$ matrices T_d, one for each output symbol $d \in \Delta$ (m and l may be infinite). We define

$$T_d \;:=\; \big(T_d(g, \hat{g})\big)_{g,\hat{g} \in \Gamma} \qquad \text{with} \qquad T_d(g, \hat{g}) \;:=\; T(g;\, d, \hat{g}).$$

T_d is a sub-stochastic matrix and $T_d(g, \hat{g}) = \mathbb{P}\big(\{\, X_k = d,\ W_k = \hat{g} \,\} \;\big|\; W_{k-1} = g\big)$. Multiplication with the initial distribution μ (interpreted as row-vector) yields

$$\mu \cdot T_d \;=\; \mathbb{P}\big(\{\, X_1 = d \,\}\big) \cdot \mathbb{P}(W_1 \mid X_1 = d). \tag{2.2}$$

Note that here we interpret the probability measure $\mathbb{P}(W_1 \mid X_1 = d)$ on Γ again as m-dimensional row vector. It should be plausible, and we prove it in Lemma 2.13 below in a more general context, that $\big(T, \mathbb{P}(W_1 \mid X_1 = d)\big)$ is an HMM of the conditional process $\mathbb{P}(X_{[2,\infty[} \mid X_1 = d)$. Consequently, applying (2.2) several times, we obtain for $x = (x_1, \dots, x_n) \in \Delta^n$

$$\begin{aligned} \mu T_{x_1} \cdots T_{x_n} &= \mathbb{P}\big(\{\, X_1 = x_1 \,\}\big) \cdots \mathbb{P}\big(\{\, X_n = x_n \,\} \;\big|\; X_{[1,n-1]} = x_{[1,n-1]}\big) \cdot \mathbb{P}(W_n \mid X_{[1,n]} = x) \\ &= \mathbb{P}\big(\{\, X_{[1,n]} = x \,\}\big) \cdot \mathbb{P}(W_n \mid X_{[1,n]} = x). \end{aligned}$$

With $\mathbb{1}$ denoting the m-dimensional column vector with all entries one, this yields

$$P\big([x_1, \dots, x_n]\big) \;=\; \mu T_{x_1} \cdots T_{x_n} \mathbb{1}$$

and the number of steps necessary to compute the probability of the output grows only linear in n. The family $(T_d)_{d \in \Delta}$ of matrices obviously determines the matrix T uniquely. On the

other hand, a family (T_d) of sub-stochastic matrices comes from an HMM if and only if their sum $\sum_{d\in\Delta} T_d$ is a stochastic matrix. Below, in Section 2.2.1, we see that the reformulation of the HMM in terms of T_d instead of T corresponds to an interpretation as observable operator model.

If Δ is countable, every Δ-valued process $X_{\mathbb{N}}$ is generated by a countable HMM. This is true, because the time set is only semi-infinite and thus a countable Γ can store the complete history of output symbols in the internal state (see the following example). This situation changes when we construct HMMs of processes $X_{\mathbb{Z}}$ in doubly infinite time. Then, an uncountable internal space may be necessary, even if the output space is countable (see Section 2.1.5).

Example 2.9. Let Δ be countable and $X_{\mathbb{N}}$ an arbitrary Δ-valued process with distribution $P \in \mathcal{P}(\Delta^{\mathbb{N}})$. We construct a countable transition-emitting HMM (T, μ) as follows. Let $\Gamma := \Delta^* := \biguplus_{n\in\mathbb{N}_0} \Delta^n = \biguplus_{n\in\mathbb{N}} \Delta^n \uplus \{\, e \,\}$, where $\uplus$ denotes the disjoint union and e is the "empty word" (the single element in Δ^0). The internal state of the HMM stores the past of the process, and the following output symbol is determined by the corresponding conditional probability. To achieve this, the initial distribution is $\mu := \delta_e$, the Dirac measure in the empty word. The generator $T\colon \Gamma \to \mathcal{P}(\Delta \times \Gamma)$ is for $g = (g_1, \ldots, g_n) \in \Delta^n \subseteq \Gamma$, $d \in \Delta$ and $\hat{g} \in \Gamma$ defined by

$$T(g;\, d, \hat{g}) \;:=\; P\big(\{\, X_{n+1} = d \,\} \;\big|\; X_{[1,n]} = g\big) \cdot \delta_{gd}(\hat{g}),$$

where $gd := (g_1, \ldots, g_n, d) \in \Delta^{n+1} \subseteq \Gamma$. It is easy to see that (T, μ) is indeed an HMM of $X_{\mathbb{N}}$. Note that even if $X_{\mathbb{N}}$ is stationary, the internal process of the above HMM is non-stationary and transient. ◊

2.1.4 Souslin HMMs

We do not always want to restrict to countable HMMs, even in the parts of the thesis, where Δ is assumed to be countable. The reason is that spaces like $\Delta^{\mathbb{Z}}$ or $\mathcal{P}(\Delta^{\mathbb{N}})$ are naturally occurring as internal spaces of HMMs and not all processes in doubly infinite time admit countable HMMs. Furthermore, many of the concepts we introduce do not require countability of Δ and Γ. We need, however, some technical restrictions on the measurable spaces in order to guarantee the existence of regular versions of conditional probability. In addition, countably generated σ-algebras are necessary to interchange "a.s." and quantifications over measurable sets. A standard assumption in probability theory guaranteeing these properties is that the occurring spaces are Polish[3]. While this assumption would be satisfactory for Δ, we need a slightly less restrictive one for Γ due to the following reasons. First, if X and Y are Polish spaces and $f\colon X \to Y$ is measurable, the image $f(X)$ of f needs neither be a Polish space nor a measurable subset of Y (thus it is also not Borel isomorphic to a Polish space in general). But the space of so-called causal states, discussed in Section 3.2.3, is isomorphic to a measurable image and we do not know if it is Polish in general. Second, the more general class of Souslin spaces arises naturally when we consider countably generated sub-σ-algebras in Section 3.2.2.

Definition 2.10. A metrisable topological space Γ is called **Souslin space** if it is the continuous image of a Polish space, i.e. there is a Polish space X and a continuous surjective function $f\colon X \to \Gamma$. A measurable space is called **Souslin measurable space** if it is, as a measurable space, isomorphic to a Souslin space. An HMM is called **Souslin HMM** if both the internal space Γ and the output space Δ are Souslin spaces.

[3] A Polish space is a separable, completely metrisable topological space.

Remark. a) Most authors do not require Souslin spaces to be metrisable, but only Hausdorff. We use Bourbaki's definition from [Bou89].

b) Every Souslin space Δ is separable and $\mathcal{P}(\Delta^{\mathbb{Z}})$ is also a Souslin space.

c) Souslin measurable spaces are also called **analytic spaces**. For their definition, it is irrelevant if Souslin spaces are assumed to be metrisable or not. Every non-metrisable Souslin space is Borel isomorphic to a metrisable Souslin space.

d) If we prove a measurable space Γ to be Souslin measurable, we may use it as internal space of a Souslin HMM, implicitly assuming that a compatible Souslin topology is chosen.

For a summary of the (for our purposes) most important properties of Souslin spaces, we refer to Appendix A.1. In particular, the Borel σ-algebra is countably generated, all probability measures are Radon measures and regular versions of conditional probability exist. Further, if a subset of a metrisable space is the measurable image of a Souslin space, it is Souslin as well. For convenience, in the sequel, the term HMM shall always imply that the spaces are Souslin and the HMM is transition-emitting.

Definition 2.11. We use the term **HMM** as synonym for transition-emitting Souslin HMM (see Definitions 2.4 and 2.10).

Given a Souslin space Γ of internal states and a Markov kernel T from Γ to $\Delta \times \Gamma$, which output distribution of HMMs (T, μ) can be achieved by choosing different initial distributions $\mu \in \mathcal{P}(\Gamma)$? In particular, we can start the HMM in an internal state $g \in \Gamma$. Denote the corresponding output process of the HMM (T, δ_g) by $O_T(g) \in \mathcal{P}(\Delta^{\mathbb{N}})$. Because O_T is measurable, it is a Markov kernel from the internal space Γ to the space $\Delta^{\mathbb{N}}$ of future trajectories. The output distribution $O_T^\mu \in \mathcal{P}(\Delta^{\mathbb{N}})$ of the HMM (T, μ) for a general initial distribution $\mu \in \mathcal{P}(\Gamma)$ is given by

$$O_T^\mu \;=\; \int O_T \,\mathrm{d}\mu.$$

Because O_T^μ is linear in μ, the set of achievable output distributions is convex and the extreme points are included in (but not necessarily equal to) the image $\mathrm{Im}(O_T) = \{\, O_T(g) \mid g \in \Gamma \,\}$ of O_T. These considerations also appear in [AC05].

In order to analyse HMMs, we need some further notation. In the rest of this section, we restrict ourselves to countable output spaces Δ (with discrete topology), but allow the internal space Γ to be an arbitrary Souslin space.

Definition 2.12. Let Δ be countable and (T, μ) an HMM. Let $g \in \Gamma$, $d \in \Delta$, and $\nu \in \mathcal{P}(\Gamma)$.

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

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

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

Remark. a) K_g is the distribution of the next output symbol given that the internal state is g, i.e. $K_g = \mathbb{P}(X_1 \mid W_0 = g)$ a.s. Further, K_μ is the distribution 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. In the case of countable Γ and using the definition of T_d from Section 2.1.3, $L_d(\nu)K_\nu(d) = \nu T_d$. For Dirac measures, we obtain

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

Note that L_d is *not* induced by a kernel in the following sense. There is no kernel $l_d\colon \Gamma \to \mathcal{P}(\Gamma)$ 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 $\iota(g) = \delta_g$, because $L_d(\nu)$ is normalised outside the integral as opposed to an individual normalisation of the $L_d(\delta_g)$ inside the integral on the right-hand side. Thus, L_d is not even linear.

It follows from the definition of $(X_{\mathbb{N}}, W_{\mathbb{N}_0})$ by a Markov kernel that the conditional probability given an internal state $W_0 = g$ is obtained by starting the HMM in g. In other words, it is generated by the HMM (T, δ_g). 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.

Lemma 2.13. *Let Δ be countable, (T, μ) an HMM with internal and output processes $W_{\mathbb{N}_0}$, $X_{\mathbb{N}}$. Then $(T, \delta_{W_0(\omega)})$ is a.s. an HMM of $\mathbb{P}(X_{\mathbb{N}} \mid W_0)(\omega)$, and $\big(T, L_{X_1(\omega)}(\mu)\big)$ is a.s. an HMM of $\mathbb{P}(X_{[2,\infty[} \mid X_1)(\omega)$.*

Proof. We first prove that (T, δ_{W_0}) is an HMM of $\mathbb{P}(X_{\mathbb{N}} \mid W_0)$. For $g \in \Gamma$, recall that $O_T(g) \in \mathcal{P}(\Delta^{\mathbb{N}})$ is the distribution of the output process of (T, δ_g). Because O_T is measurable, $O_T \circ W_0$ is $\sigma(W_0)$-measurable. From the definition of $(W_{\mathbb{N}_0}, X_{\mathbb{N}})$, it follows for measurable $G \subseteq \Gamma,\; A \subseteq \Delta^{\mathbb{N}}$ that

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

where the second equality holds because W_0 is distributed according to μ. Thus $O_T \circ W_0$ 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

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

□

2.1.5 Doubly infinite time and size of HMMs

So far we were concerned with HMMs for processes $X_{\mathbb{N}}$ in semi infinite time. The main focus of this work, however, is on stationary processes with time set $\mathbb{Z}$. Every stationary process with time set $\mathbb{N}$ can be uniquely extended to a stationary process in doubly infinite time. Thus, an HMM (T, μ) of the future part $X_{\mathbb{N}}$ of a stationary process $X_{\mathbb{Z}}$ identifies the distribution of the whole process. Nevertheless, we are not satisfied with this model of $X_{\mathbb{N}}$ as model of $X_{\mathbb{Z}}$ for the following reason. The internal process $W_{\mathbb{N}_0}$ is not necessarily stationary and, in general, there exists no extension to a process $W_{\mathbb{Z}}$, such that T is the conditional probability of X_k, W_k given W_{k-1}. Therefore, (T, μ) is not a valid possibility for the generation of the doubly infinite process $X_{\mathbb{Z}}$. To the contrary, we require that the internal process of a model of $X_{\mathbb{Z}}$ has to be stationary. This requirement is equivalent to μ being T-invariant in the sense that

$$\mu(G) \;=\; \int T(\,\cdot\,;\, \Delta \times G) \,\mathrm{d}\mu \qquad \forall G \in \Gamma. \tag{2.3}$$

On the other hand, given an HMM (T, μ), T-invariance of μ ensures that both $X_{\mathbb{N}}$ and $W_{\mathbb{N}_0}$ are stationary and we extend them to processes $W_{\mathbb{Z}}$, $X_{\mathbb{Z}}$ in doubly infinite time. Most of the rest of this thesis is concerned with stationary processes and invariant representations.

Definition 2.14. An HMM (T, μ) is called **invariant**, if μ is T-invariant (i.e. (2.3) is satisfied).

If an HMM is invariant, one may ask whether the generated processes $W_{\mathbb{Z}}$ and $X_{\mathbb{Z}}$ are ergodic. Because $W_{\mathbb{Z}}$ is a Markov process, ergodicity of $W_{\mathbb{Z}}$ can be verified easily, at least if Γ is countable (see [KSK76]). The situation for $X_{\mathbb{Z}}$ is more complicated. A simple sufficient criterion is given by the fact that ergodicity of $W_{\mathbb{Z}}$ implies ergodicity of $X_{\mathbb{Z}}$ (but not vice versa). This criterion is enough for our purposes and we prove it in the following. A complete characterisation of HMMs with ergodic output processes was obtained by Schönhuth in [SJ09].

Proposition 2.15. *Let (T, μ) be an invariant HMM with ergodic internal process $W_{\mathbb{Z}}$. Then the output process $X_{\mathbb{Z}}$ is ergodic as well.*

Proof. Let $A \in \mathfrak{B}(\Delta^{\mathbb{Z}})$ be shift-invariant. Due to stationarity of $P = \mathbb{P}_{X_{\mathbb{Z}}}$, A is measurable w.r.t. the P-completion of the tail σ-algebra on $\Delta^{\mathbb{Z}}$ (e.g. [Dęb09, Lem. 3]). For $x \in \Gamma^{\mathbb{Z}}$, the conditional process $P'_x := \mathbb{P}(X_{\mathbb{Z}} \mid W_{\mathbb{Z}} = x)$ is independently distributed. Thus, we can apply the Kolmogorov 0-1-law to obtain that $P'_x(A) \in \{0, 1\}$ for almost all x. Let $B := \{ x \in \Gamma^{\mathbb{Z}} \mid P'_x(A) = 1 \}$. Then B is shift-invariant (modulo $\mathbb{P}_{W_{\mathbb{Z}}}$), because joint stationarity and shift-invariance of A lead to

$$P'_{\sigma(x)}(A) \;=\; \mathbb{P}\big(\{ X_{\mathbb{Z}} \in A \} \,\big|\, W_{\mathbb{Z}} = \sigma(x)\big) \;=\; \mathbb{P}\big(\{ X_{\mathbb{Z}} \in \sigma^{-1}(A) \} \,\big|\, W_{\mathbb{Z}} = x\big) \;=\; P'_x(A).$$

Consequently, we obtain

$$\mathbb{P}\big(\{ X_{\mathbb{Z}} \in A \}\big) \;=\; \int P\big(\{ X_{\mathbb{Z}} \in A \} \,\big|\, W_{\mathbb{Z}}\big)\, \mathrm{d}\mathbb{P} \;=\; \mathbb{P}\big(\{ W_{\mathbb{Z}} \in B \}\big) \;\in\; \{0, 1\}. \qquad \square$$

If Δ is countable and $X_{\mathbb{Z}}$ is a Δ-valued, stationary process, then there exists a countable HMM generating the future part, $X_{\mathbb{N}}$, of the process. This is not the case for the whole process. There may not exist any invariant, countable HMM of $X_{\mathbb{Z}}$.

Example 2.16. Let $X_{\mathbb{Z}}$ be any stationary process with distribution $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ and uncountably many ergodic components (we define ergodic components in Section 4.1.2). For instance, let $\Delta = \{0, 1\}$ and $P = \int_0^1 P_p \,\mathrm{d}p$, where $P_p \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ is the Bernoulli process with parameter p (i.e. P_p is i.i.d. with $P_p\big([1]\big) = p$) and the integration is w.r.t. Lebesgue measure. Then there is no invariant, countable HMM of $X_{\mathbb{Z}}$, because the output process of such an HMM (T, μ) has only countably many ergodic components. Indeed, if Γ is countable, it follows from the theory of countable Markov chains (e.g. [KSK76]) that the stationary internal process $W_{\mathbb{Z}}$ has a countable number of ergodic components $W^k_{\mathbb{Z}}$, and there are disjoint sets $\Gamma_k \subseteq \Gamma$ such that W^k_n assumes for any $n \in \mathbb{N}$ only values in Γ_k. Because ergodicity of the internal process of an HMM implies ergodicity of the output process, the output of (T, μ) can have at most as many ergodic components as $W_{\mathbb{Z}}$ does. In particular, the number of components is countable. $\lozenge$

There is a natural notion of isomorphism of invariant HMMs.

Definition 2.17. An invariant HMM (T, μ) with space Γ of internal states is called **isomorphic** to an invariant HMM (T', μ') with space Γ' of internal states if they share a common output space Δ and there is a measurable function $\iota\colon \Gamma \to \Gamma'$ that is μ-a.s. injective (i.e. injective on a set of μ-measure one) and satisfies

$$\mu' \;=\; \mu \circ \iota^{-1} \qquad \text{and} \qquad T'(\iota(\,\cdot\,);\, D \times G') \;=\; T\big(\,\cdot\,;\, D \times \iota^{-1}(G')\big) \quad \mu\text{-a.s.}$$

for all $D \in \mathcal{D}$ and $G' \in \mathcal{G}'$. ι is called **isomorphism**.

Remark. a) $\mu' = \mu \circ \iota^{-1}$ implies that ι is essentially surjective in the sense that its image has full measure[4] w.r.t. μ'. Thus an isomorphism is essentially bijective.

b) Obviously, isomorphic HMMs generate the same output process.

The following proposition justifies the name isomorphism for the function ι of Definition 2.17. For every isomorphism, there is an inverse isomorphism.

Proposition 2.18. *Let (T, μ) and (T', μ') be isomorphic invariant HMMs and $\iota\colon \Gamma \to \Gamma'$ an isomorphism. Then there exists an inverse isomorphism $\iota'\colon \Gamma' \to \Gamma$ with*

$$\iota \circ \iota' \;=\; \mathrm{id}_{\Gamma'} \quad \mu'\text{-a.s.} \qquad \text{and} \qquad \iota' \circ \iota \;=\; \mathrm{id}_{\Gamma} \quad \mu\text{-a.s.}$$

Proof. Let ι be injective on $\Lambda \in \mathcal{G}$ with $\mu(\Lambda) = 1$ and define $\Lambda' = \iota(\Lambda)$. Then the restriction $\iota\restriction_\Lambda$ of ι to Λ is a Borel isomorphism between the Souslin spaces Λ and Λ' ([Coh80, Prop. 8.6.2]). Let $j = \iota\restriction_\Lambda^{-1}$. Because Λ' is a Souslin set, it is universally measurable and there is a Borel map $\iota'\colon \Gamma' \to \Gamma$ that coincides with j on a measurable set $A' \in \mathcal{G}'$ with $A' \subseteq \Lambda'$ and $\mu'(A') = 1$ ([Bog07, Cor. 6.5.6]). On A', ι' is injective and $\iota \circ \iota' = \mathrm{id}_{\Gamma'}$. Let $A = \iota^{-1}(A') \cap \Lambda$. Because $\mu' = \mu \circ \iota^{-1}$, we have $\mu(A) = 1$ and obtain $\iota' \circ \iota = \mathrm{id}_\Gamma$ on A. We show that ι' is an isomorphism:

1. $\mu' \circ \iota'^{-1} = \mu \circ \mathrm{id}_\Gamma^{-1} = \mu$ holds because $\iota' \circ \iota = \mathrm{id}_\Gamma$ a.s.

2. Let $D \in \mathcal{D}$ and $G \in \mathcal{G}$. Because μ is T-invariant, $T(\,\cdot\,;\, \Delta \times G) = T\big(\,\cdot\,;\, \Delta \times (\iota' \circ \iota)^{-1}(G)\big)$ holds μ-a.s. Therefore, using that ι is an isomorphism, we obtain μ'-a.s.

$$T(\iota'(\,\cdot\,);\, D \times G) \;=\; T'\big(\iota \circ \iota'(\,\cdot\,);\, D \times \iota'^{-1}(G)\big) \;=\; T'\big(\,\cdot\,;\, D \times \iota'^{-1}(G)\big). \qquad \square$$

The question about a *minimal* HMM of a given process $X_\mathbb{Z}$ suggests itself. In order to make this more precise, we have to define the "size" of an HMM. Because we consider $X_\mathbb{Z}$, and thus Δ, to be fixed, we use the "size" of the internal component. One possibility to define it is the cardinality $|\Gamma|$ of its set of internal states. In the case of invariant HMMs of stationary processes, there is a second possibility which we consider more appropriate for the definition of complexity measures. It is the entropy $H(\mu) = H^{\mathbb{P}}(W_0)$ of the invariant initial distribution. Note that these two possible definitions of size lead to a different ordering of HMMs. We demonstrate in the following example that a very natural looking HMM with the minimal number of internal states can have higher internal entropy than an HMM with more internal states.

[4]The image need not be Borel measurable. It is, however, a Souslin set and thus universally measurable. In particular, it is μ'-measurable.

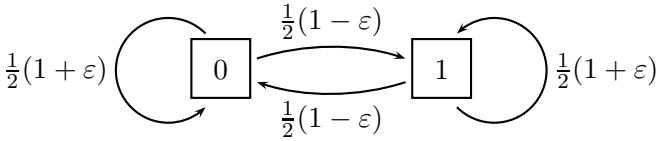


Figure 2.1: "Nearly i.i.d." Markov process used in Example 2.19

Remark (visualisation of generators). We visualise the generator T of a finite HMM as transition graph in the following way. The nodes of the directed graph correspond to the internal states of the HMM and are drawn as circles labeled with the internal state. From node g to node $\hat{g}$, there may be up to $|\Delta|$ edges, labeled by output symbol d and transition probability $p = T(g;\, d, \hat{g})$. Edges are present if and only if $p > 0$. Similarly, we draw transition kernels of Markov processes. Here, the nodes are the states of the process and we draw them as square boxes. There are no output symbols in the edge labels and there is at most one edge from one state to another.
Note that these kinds of visualisation are different from the visualisations of dependence structures as graphical models, which we already used. There, the nodes correspond to random variables instead of states. To distinguish these visualisations, we do not draw circles around nodes in graphical models.

Example 2.19. Let $\Delta := \{\, 0, 1 \,\}$ and, for $\varepsilon \in [0, 1]$, consider the stationary Markov process $X^\varepsilon_{\mathbb{Z}}$ defined by

$$\mathbb{P}\big(\{\, X^\varepsilon_0 = d \,\}\big) \;:=\; \tfrac{1}{2} \qquad \text{and} \qquad \mathbb{P}\big(\{\, X^\varepsilon_{n+1} = \hat{d} \,\} \;\big|\; X^\varepsilon_n = d\big) \;:=\; \begin{cases} \frac{1}{2}(1+\varepsilon) & \text{if } d = \hat{d}, \\ \frac{1}{2}(1-\varepsilon) & \text{if } d \neq \hat{d}. \end{cases}$$

The transition kernel is visualised in Figure 2.1. $X^\varepsilon_{\mathbb{Z}}$ is a disturbed i.i.d. process with disturbance of magnitude ε. For $\varepsilon = 0$, it is i.i.d., and for $\varepsilon = 1$ it is constantly 0 or 1, each with equal probability. It is obvious that there is a stationary HMM of $X^\varepsilon_{\mathbb{Z}}$ with two internal states and $W_k = X_k$, because $X^\varepsilon_{\mathbb{Z}}$ is already Markovian. The internal state entropy of this HMM is $\log(2)$.

No HMM can do with less than two internal states, but we can construct an HMM with lower internal state entropy on three states for sufficiently small ε. The idea is to have one state corresponding to the i.i.d. process and getting most of the invariant measure if ε is small. The other two states correspond to the disturbances towards constantly 0 and 1 respectively. More precisely, let $\Gamma := \{\, 0, 1, 2 \,\}$ and consider the stationary HMM $(T^\varepsilon, \mu^\varepsilon)$ given by

$$T^\varepsilon(g) \;:=\; \begin{cases} \varepsilon\delta_{(g,g)} + (1-\varepsilon)\delta_{(g,2)} & \text{if } g \in \{\, 0, 1 \,\}, \\ \frac{1}{2}\big(\varepsilon\delta_{(0,0)} + \varepsilon\delta_{(1,1)} + (1-\varepsilon)\delta_{(0,2)} + (1-\varepsilon)\delta_{(1,2)}\big) & \text{if } g = 2 \end{cases}$$

(see Figure 2.2) together with the invariant initial distribution $\mu^\varepsilon = \frac{\varepsilon}{2}\delta_0 + \frac{\varepsilon}{2}\delta_1 + (1-\varepsilon)\delta_2$. We verify that this HMM generates $X^\varepsilon_{\mathbb{Z}}$, using the terminology of Section 2.1.3. For every vector ν, νT^ε_0 is a multiple of $(\varepsilon, 0, 1-\varepsilon)$ and νT^ε_1 is a multiple of $(0, \varepsilon, 1-\varepsilon)$. Thus the output process is Markovian and the conditional probability of the next output, given that the last output was 0, is $\varepsilon\delta_0 + (1-\varepsilon)(\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1) = \mathbb{P}(X^\varepsilon_1 \mid X^\varepsilon_0 = 0)$. Because the same holds for the output 1 and the marginals coincide, $(T^\varepsilon, \mu^\varepsilon)$ is an invariant HMM of $X^\varepsilon_{\mathbb{Z}}$ with internal state entropy given by

$$H(\mu^\varepsilon) \;=\; -(1-\varepsilon)\log(1-\varepsilon) - \varepsilon\log(\tfrac{\varepsilon}{2}) \quad \stackrel{\varepsilon\to 0}{\longrightarrow} \quad 0.$$

Thus it is smaller than $\log(2)$ for sufficiently small ε. $\Diamond$

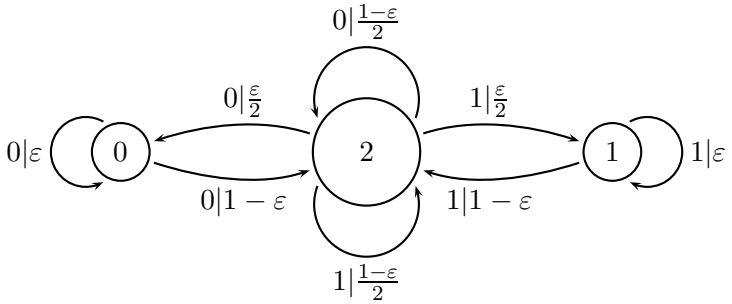


Figure 2.2: HMM used in Example 2.19. The circled nodes are internal states. The edges are transitions, labeled with output symbol d and transition probability p in the form "$d|p$". The HMM generates the Markov process shown in Figure 2.1, but with lower internal state entropy.

2.1.6 Internal expectation process

In this section, we consider only countable Δ. We claimed that the internal operator L_d describes the update of knowledge of the internal state. Now, we look at the process of "knowledge of the internal state," more precisely at the process $Y_{\mathbb{Z}}$ of conditional probabilities of the internal state given the past of the output process. We justify our interpretation of L_d in Lemma 2.21 and show that the internal expectation process $Y_{\mathbb{Z}}$ is a Markov process. These results are in particular needed in the following section to clarify the structure of partially deterministic HMMs.

Definition 2.20 ($Y_{\mathbb{Z}}$ and $H_{\mathbb{Z}}$). Given an invariant HMM, let $Y_{\mathbb{Z}}$ be the $\mathcal{P}(\Gamma)$-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. $H_k(\omega) := H\big(Y_k(\omega)\big)$. We call an HMM **state observable** if $H_k = 0$ a.s. for all k.

Remark. a) 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 $H^{\mathbb{P}}(Y_k)$ of the *random variable* Y_k.

b) An HMM is state observable if and only if there is a measurable function $h\colon \Delta^{-\mathbb{N}_0} \to \Gamma$ such that $W_0 = h \circ X_{-\mathbb{N}_0}$ a.s. This means that the current internal state can always be inferred by an observer.

The following lemma justifies the idea of the internal operator L_d modelling the 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 2.21. *Let (T, μ) be an invariant HMM, Δ countable and $d \in \Delta$. Then*

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

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

$$\int T \,\mathrm{d}Y_0 \;=\; \int \mathbb{P}(X_1,\, W_1 \mid W_0) \,\mathrm{d}\mathbb{P}(\,\cdot \mid X_{-\mathbb{N}_0}) \;=\; \mathbb{P}(X_1,\, W_1 \mid X_{-\mathbb{N}_0}). \tag{2.4}$$

a) Let $d = X_1(\omega)$ and for $G \in \mathcal{G}$ set $F_G := \{ X_1 = d,\ W_1 \in G \}$. We obtain a.s.

$$L_d(Y_0)(G) \overset{(2.4)}{=} \frac{\mathbb{P}(F_G \mid X_{-\mathbb{N}_0})}{\mathbb{P}(F_\Gamma \mid X_{-\mathbb{N}_0})} \overset{(d = X_1(\omega))}{=} \mathbb{P}\big(\{ W_1 \in G \} \bigm| X_{-\mathbb{N}_0},\, X_1\big) \;=\; Y_1(\,\cdot\,)(G).$$

b) The second equality follows directly from (2.4). The first follows because, due to the second equality, $\mathbb{P}\big(\{ X_1 = d \} \bigm| X_{-\mathbb{N}_0}\big)$ is $\sigma(Y_0)$-measurable modulo $\mathbb{P}$. □

Using the previous lemma, we can 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.

Proposition 2.22. *Let Δ be countable. For an invariant HMM, $Y_\mathbb{Z}$ and $H_\mathbb{Z}$ are stationary. $Y_\mathbb{Z}$ is a Markov process with transition kernel*

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

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

$$\begin{aligned} \mathbb{P}(Y_{k+1} \mid Y_{[0,k]} = \nu_{[0,k]}) &\overset{\text{(Lem. 2.21a)}}{=} \mathbb{P}\big(L_{X_{k+1}(\,\cdot\,)}(\nu) \bigm| Y_{[0,k]} = \nu_{[0,k]}\big) \\ &= \sum_{d \in \Delta} \mathbb{P}\big(\{ X_{k+1} = d \} \bigm| Y_{[0,k]} = \nu_{[0,k]}\big) \cdot \delta_{L_d(\nu)}. \end{aligned}$$

$\sigma(Y_{[0,k]})$ is nested between $\sigma(Y_k)$ and $\sigma(X_{]-\infty,k]})$, i.e. $\sigma(Y_k) \subseteq \sigma(Y_{[0,k]}) \subseteq \sigma(X_{]-\infty,k]})$. Therefore, Lemma 2.21 b. implies that we have $\mathbb{P}\big(\{ X_{k+1} = d \} \bigm| Y_{[0,k]} = \nu_{[0,k]}\big) = K_{\nu_k} = K_\nu$ and hence the claim follows. □

2.1.7 Partial determinism

If the generator 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. An example is the shift HMM of Example 2.7. Determinism 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 $f(g, d)$ of the last internal state g and the new output symbol d. In the visualisation of T as transition graph, this means that for every internal state g and output symbol d, there is at most one edge labeled with d and leaving the node g. An example of such an HMM is Example 2.9, where the internal state coincides with the past output and $f(g, d) = gd$.

If the internal space Γ and the output space Δ are finite, partially deterministic HMMs are stochastic versions of *deterministic finite state automata (DFAs)*, an important concept of theoretical computer science (see [HU79, 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 2.23. An invariant HMM (T, μ) is called **partially deterministic** if there is a measurable function $f\colon \Gamma \times \Delta \to \Gamma$, called **transition function**, such that for μ-almost all $g \in \Gamma$, we have $T(g) = K_g \otimes \delta_{f(g,\,\cdot\,)}$, i.e.

$$T(g;\, D \times G) \;=\; K_g\big(D \cap f_g^{-1}(G)\big) \qquad \forall D \in \mathcal{D},\; G \in \mathcal{G},$$

where $f_g(d) := f(g, d)$. We also use the notation $\hat{f}_d(g) := f_g(d)$.

The isomorphisms between partially deterministic HMMs are precisely the essentially bijective maps that "preserve" output kernel and transition function.

Lemma 2.24. *Let (T, μ) and (T', μ') be invariant, partially deterministic HMMs with output kernels K, K', transition functions f, f' and spaces Γ, Γ' of internal states. Let $\iota\colon \Gamma \to \Gamma'$ be a μ-a.s. injective map with $\mu' = \mu \circ \iota^{-1}$. Then ι is an isomorphism if and only if*

$$K'_{\iota(g)} \;=\; K_g \qquad \textit{and} \qquad f'_{\iota(g)}(d) \;=\; \iota \circ f_g(d) \qquad \mu \otimes K\textit{-a.s.}$$

Proof. "if": Obvious from the definitions.

"only if": Let ι be an isomorphism. With $D = \Delta$, we obtain $K'_{\iota(g)} = K_g$ a.s. For all g where this holds and all $G' \in \mathcal{G}'$, we have

$$K_g\big(D \cap {f'_{\iota(g)}}^{-1}(G')\big) \;=\; T'\big(\iota(g);\, D \times G'\big) \;=\; T\big(g;\, D \times \iota^{-1}(G')\big) \;=\; K_g\big(D \cap \iota^{-1} \circ f_g^{-1}(G')\big)$$

for all $D \in \mathcal{D}$, which implies

$$1_{G'} \circ f'_{\iota(g)} \;=\; 1_{G'} \circ (\iota \circ f_g) \qquad K_g\text{-a.s.} \qquad \forall G' \in \mathcal{G}'.$$

The equality holds for μ-almost all g. Because $\mathcal{G}'$ is countably generated, this implies $f'_{\iota(g)} = \iota \circ f_g$ K_g-a.s. for μ-almost all $g \in \Gamma$. □

If Δ is countable, we obtain for partially deterministic HMMs that

$$L_d(\nu)(G) \;=\; \frac{1}{K_\nu(d)} \int_{\hat{f}_d^{-1}(G)} \widehat{K}_d \,\mathrm{d}\nu \qquad \text{and} \qquad L_d(\delta_g) \;=\; \delta_{f_g(d)}. \tag{2.5}$$

The second equation implies $W_k = f_{W_{k-1}}(X_k)$ a.s., justifying the name transition function for f. We obtain this result also for more general spaces. Recall that $K_g(D) = \mathbb{P}\big(\{\, X_1 \in D \,\} \;\big|\; W_0 = g\big)$ for $g \in \Gamma$ and $D \in \mathcal{D}$.

Proposition 2.25. *An invariant HMM (T, μ) is partially deterministic with transition function f if and only if $W_1 = f(W_0, X_1)$ a.s.*

Proof. $W_1 = f(W_0, X_1)$ a.s. is equivalent to $\delta_{f(W_0,X_1)}$ being a version of the conditional probability $\mathbb{P}(W_1 \mid W_0, X_1)$ (see Appendix A.3). We show that this is the case if and only

if the HMM is partially deterministic with transition function f. $\delta_{f(W_0,X_1)}$ is $\sigma(W_0,X_1)$-measurable. Let $A \in \sigma(W_0)$, $G \in \mathcal{G}$, and $B = X_1^{-1}(D) \in \sigma(X_1)$ with $D \in \mathcal{D}$. Using $\delta_{f(W_0,X_1)} = \mathbb{P}\big(f(W_0,X_1) \bigm| W_0, X_1\big)$, we obtain

$$\begin{aligned}\int_{A\cap B} \delta_{f(W_0,X_1)}(G)\,\mathrm{d}\mathbb{P} &= \int_A \mathbb{P}\big(B \cap \{\, f(W_0,X_1) \in G \,\} \bigm| W_0, X_1\big)\,\mathrm{d}\mathbb{P} \\ &\overset{(A\in\sigma(W_0))}{=} \int_A \mathbb{P}\big(\{\, X_1 \in D,\ f(W_0,X_1) \in G \,\} \bigm| W_0\big)\,\mathrm{d}\mathbb{P} \\ &= \int_A K_{W_0}\big(D \cap f_{W_0}^{-1}(G)\big)\,\mathrm{d}\mathbb{P}\end{aligned}$$

Partial determinism means that $K_{W_0}\big(D \cap f_{W_0}^{-1}(G)\big) = \mathbb{P}\big(\{\, X_1 \in D,\ W_1 \in G \,\} \bigm| W_0\big)$ a.s., or equivalently that the integrals of these two functions over all $\sigma(W_0)$-measurable sets A coincide. Therefore, it is equivalent to

$$\int_A K_{W_0}\big(D \cap f_{W_0}^{-1}(G)\big)\,\mathrm{d}\mathbb{P} = \mathbb{P}\big(A \cap \{\, X_1 \in D,\ W_1 \in G \,\}\big) \qquad \forall A \in \sigma(W_0),\ D \in \mathcal{D},\ G \in \mathcal{G}$$

and hence to $\delta_{f(W_0,X_1)} = \mathbb{P}(W_1 \mid W_0, X_1)$ a.s. □

Corollary 2.26. *Every invariant, state observable HMM is partially deterministic.*

Proof. Let the HMM be state observable, i.e. $\mathbb{P}(W_0 \mid X_{-\mathbb{N}_0}) = \delta_{W_0}$ a.s. Then

$$\delta_{W_1} = \mathbb{P}(W_1 \mid X_{-\mathbb{N}_0}, X_1) = \mathbb{P}\big(W_1 \bigm| \mathbb{P}(W_0 \mid X_{-\mathbb{N}_0}), X_{-\mathbb{N}_0}, X_1\big) = \mathbb{P}(W_1 \mid W_0, X_1)$$

a.s. Thus W_1 is a function of W_0, X_1 and the HMM is partially deterministic. □

The converse of this corollary is obviously not true. It may well happen that the internal state of a partially deterministic HMM cannot be derived from the output at any point in time. Assuming countable Δ, we see in the following proposition that the uncertainty $H_k = H(Y_k)$ of the internal state given the past output stays constant over time. This property is crucial for understanding partially deterministic HMMs. Furthermore, in the case of finite entropy $H(\mu)$, the next output symbol is independent of the internal state if the past output is known. The proof is along the following lines. If we know the internal state at one point in time, we can maintain this knowledge due to partial determinism. More generally, the uncertainty H_k of the internal state cannot increase on average and thus $H_{\mathbb{Z}}$ is a supermartingale. But because it is also stationary, the trajectories have to be constant. If two possible internal states led 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 $H_{\mathbb{Z}}$.

Theorem 2.27. *Let Δ be countable and (T,μ) 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 $\mathrm{supp}(Y_0) \subseteq \Gamma$ of the random measure Y_0 is a.s. a constant kernel, i.e.*

$$K_g = K_{\hat{g}} \qquad \forall g, \hat{g} \in \mathrm{supp}\big(Y_0(\omega)\big) \qquad \text{a.s.} \tag{2.6}$$

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 Γ is countable. Note that $\varphi(x) = -x\log(x)$ satisfies $\varphi(\sum x_i) \leq \sum \varphi(x_i)$. We obtain

$$H\big(L_d(\nu)\big) \overset{(2.5)}{=} \sum_{\hat{g}\in\Gamma} \varphi\left(\sum_{g\in \hat{f}_d^{-1}(\hat{g})} \nu(g)\frac{K_g(d)}{K_\nu(d)}\right) \leq \sum_{g\in \hat{f}_d^{-1}(\Gamma)=\Gamma} \varphi\left(\nu(g)\frac{K_g(d)}{K_\nu(d)}\right).$$

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

$$\begin{aligned}\mathrm{E}\big(H_{k+1} \,\big|\, Y_k = \nu\big) &\overset{\text{(Prop. 2.22)}}{=} \sum_{d\in\Delta} K_\nu(d)\cdot H\big(L_d(\nu)\big) \leq -\sum_{d,g} \nu(g)K_g(d)\cdot\log\left(\nu(g)\frac{K_g(d)}{K_\nu(d)}\right)\\ &= H^{\mathbb{P}}(W_k \mid X_{k+1}, Y_k = \nu) \leq H^{\mathbb{P}}(W_k \mid Y_k = \nu) = H(\nu), \quad (2.7)\end{aligned}$$

where the second equality holds because $\mathbb{P}\big(\{\, W_k = g,\ X_{k+1} = d\,\} \,\big|\, Y_k = \nu\big) = \nu(g)K_g(d)$ 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 (2.7) 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. □

Note that the finite-entropy assumption is indeed necessary for the second statement of Theorem 2.27. The shift HMM defined in Example 2.7, for example, is a deterministic HMM that does not (in general) satisfy (2.6).

Example 2.28. Let $(T^\sigma, P_{\mathbb{N}})$ be the (one-sided) shift HMM of the stationary process $X_{\mathbb{Z}}$. The HMM is invariant and deterministic, thus in particular partially deterministic. Let $g = (g_k)_{k\in\mathbb{N}} \in \Gamma$ and note that $K_g = \delta_{g_1}$. Thus, $K_g = K_{\hat{g}}$ implies $g_1 = \hat{g}_1$. Furthermore, $Y_0 = \mathbb{P}(W_0 \mid X_{-\mathbb{N}_0}) = \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0})$ because $W_0 = X_{\mathbb{N}}$ a.s. If (2.6) holds, this means that X_1 is a.s. determined by $X_{-\mathbb{N}_0}$, which is generically not true. ◊

Theorem 2.27 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 2.29. *Let Δ be countable and (T, μ) a partially deterministic, invariant HMM with $H(\mu) < \infty$. Then*

$$\mathbb{P}(X_{\mathbb{N}} \mid W_0 = g) = \mathbb{P}(X_{\mathbb{N}} \mid W_0 = \hat{g}) \qquad \forall g, \hat{g} \in \mathrm{supp}(Y_0) \qquad \text{a.s.}$$

Proof. According to Theorem 2.27, $\mathbb{P}(X_1 \mid W_0 = \cdot) = K$ is constant on $\mathrm{supp}(Y_0)$. To obtain the statement for $X_{[1,n]}$, we consider the n-tuple HMM defined as follows: Its output space is Δ^n, its internal space is Γ and 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

$\widehat{T}\colon \Gamma \to \mathcal{P}(\Delta^n \times \Gamma)$, $\widehat{T}(g) = \mathbb{P}(X_{[1,n]}, W_n \mid W_0 = g)$. The HMM is obviously partially deterministic with state update function $f_{d_n} \circ \cdots \circ f_{d_1}$ and invariant. Thus Theorem 2.27 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. □

2.2 Algebraic representations

There are also more algebraic models of stochastic processes, which dismiss the conception of the internal dynamics being described by a stochastic process. Instead, some vector space replaces the internal states and the "dynamics" is described by linear maps (instead of Markov kernels). These models are proper generalisations of HMMs. They were introduced and termed *stochastic S-modules* by Alex Heller in the very concise paper [Hel65]. Later, in [Jae00], Herbert Jaeger made the construction more explicit and transparent for readers not familiar with module theory. He introduced the name *observable operator model* (*OOM*), provided ways of interpreting them and extended the theory by learning algorithms. Ergodic theory for OOMs was developed in [FS07, SJ09]. In [LSS01], the same model class was also obtained starting from a somewhat different intuition (internal states are constructed as predictions for certain tests) as linear non-controlled *predictive state representations* (*PSRs*).[5] The equivalence of linear non-controlled PSRs and OOMs is shown in [SJR04]. Another name for the same class of models, *generalised HMMs* (*GHMMs*), is used in [Upp89].

Not only do OOMs provide more compact representations of some stochastic processes than HMMs do, but they also turn out to be a useful tool for studying HMMs (see Section 2.2.3).

2.2.1 Observable operator models

In this section, let Δ be countable. OOMs exist also for processes with values in arbitrary spaces (see [Jae99] and Section 2.2.4), but nearly all of the literature assumes that the output space Δ is finite.

We saw in Section 2.1.3 that countable HMMs can be reformulated in terms of a vector corresponding to the initial distribution and a family $(T_d)_{d\in\Delta}$ of sub-stochastic matrices. This idea is generalised in OOM theory, where the matrices are replaced by linear maps on some vector space. This means that the positivity constraint is relaxed. Of course, the probabilities associated to the output process have to be positive, which is required explicitly in the OOM definition. In practice, this condition (3. in the following definition) is a problem for learning algorithms, because it cannot be checked in general. In [Wie08], it is proven that the condition is undecidable in the sense of computation theory, i.e. there cannot exist a general algorithm for deciding if a given structure is a valid OOM.

Definition 2.30. An **observable operator model** (**OOM**) with countable output space Δ is a quadruple $\big(V, (T_d)_{d\in\Delta}, v, l\big)$, where V is a real vector space, $T_d\colon V \to V$ are linear maps, $v \in V$, and l is a linear form on V, such that for $n \in \mathbb{N}$ and $d_1, \ldots, d_n \in \Delta$,

1. $l(v) = 1$, 2. $l \circ \sum_{d\in\Delta} T_d = l$, 3. $P_{d_1,\ldots,d_n} := l \circ T_{d_n} \circ \cdots \circ T_{d_1}(v) \geq 0$.

[5] PSRs exist also for controlled systems, thus including actions of the observer. In principle, they can be non-linear, but most of the theory considers the linear case.

v is called **initial vector**, the T_d are called **observable operators** and l is called **evaluation form**. The process $P \in \mathcal{P}(\Delta^{\mathbb{N}})$, defined by $P\big([d_1, \ldots, d_n]\big) := P_{d_1,\ldots,d_n}$, is called **generated** by the OOM and the dimension $\dim(V)$ of V is called **dimension** of the OOM.

Remark. a) Jaeger fixes a basis of V instead of an evaluation form and defines l to be the sum of coefficients in the basis expansion. This corresponds to the vector $\mathbb{1}$ used in Section 2.1.3.

b) It is straightforward to verify (see [Jae00]) that the P generated by an OOM is really a well-defined probability measure on $\Delta^{\mathbb{N}}$.

OOMs are generalisations of HMMs in the sense that for any HMM, there is a naturally associated OOM generating the same process. Given an HMM (T, μ), we introduced in Definition 2.12 the internal operator L_d which describes the update of the knowledge about the internal state when the symbol d is observed. The normalisation of L_d was necessary to map probability measures onto probability measures but makes it non-linear. If we leave out this normalisation, we obtain a linear operator $T_d(\nu) = K_\nu(d)L_d(\nu)$ from the space $\mathcal{M}(\Gamma)$ of signed measures of bounded variation to itself. The operators T_d incorporate also information about the a priori probability of having observed d and yield an OOM generating the same process.

Definition 2.31. Let (T, μ) be a countable HMM and $V := \mathcal{M}(\Gamma)$ the set of signed measures of bounded variation on Γ. For $d \in \Delta$, we define the linear maps $T_d \colon V \to V$ by

$$T_d(\nu)(G) \;:=\; \int T\big(\,\cdot\,;\,\{\,d\,\} \times G\big)\,\mathrm{d}\nu \qquad \forall \nu \in V = \mathcal{M}(\Gamma),\; G \in \mathcal{G}.$$

Then $\big(V, (T_d)_{d\in\Delta}, \mu, l\big)$ with $l(\nu) := \nu(\Gamma)$ is called the **associated OOM** of (T, μ).

Remark. Let (T, μ) be a countable HMM with assocaited OOM $\big(\mathcal{M}(\Gamma), (T_d), \mu, l\big)$.

a) We have $T_d(\nu) = K_\nu(d)L_d(\nu)$. In particular, $K_\nu(d) = 0$ implies $T_d(\nu) = 0$.

b) If the HMM is finite, the number $|\Gamma|$ of internal states coincides with the dimension of the associated OOM. The internal states correspond to the basis $\{\,\delta_g \mid g \in \Gamma\,\}$ of the associated OOM vector space $\mathcal{M}(\Gamma)$.

c) It is shown in [Hel65] and [Jae00] that finite-dimensional OOMs can exist for some processes that do not allow for an HMM with finitely many internal states.

Lemma 2.32. *The OOM associated to an HMM is a valid OOM and generates the same process as the HMM.*

Proof. This is a special case of Lemma 2.39 below. □

2.2.2 Canonical OOM

It is in general a difficult task to construct an HMM with the minimal number of internal states. Existence of an HMM with finitely many internal states and the necessary number of states depend on an intricate geometrical condition specified by Heller in [Hel65]. Moreover, the HMM with the minimal number of internal states is not unique. The situation for OOMs

is much more pleasant. There is a unique (up to isomorphism) OOM with minimal dimension, and it is obtained by a canonical construction on the space $V := \mathcal{M}(\Delta^{\mathbb{N}})$ of signed measures with bounded variation on $\Delta^{\mathbb{N}}$. Note that $\mathcal{M}(\Delta^{\mathbb{N}}) = \operatorname{span}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big)$, where span denotes the linear hull. The canonical observable operators $\tau_d \colon \mathcal{M}(\Delta^{\mathbb{N}}) \to \mathcal{M}(\Delta^{\mathbb{N}})$ are defined by

$$\tau_d(z) \;=\; z\big([d] \cap \sigma^{-1}(\,\cdot\,)\big),$$

where σ is the left-shift on $\Delta^{\mathbb{N}}$. Further define $l_\Delta \colon \mathcal{M}(\Delta^{\mathbb{N}}) \to \mathbb{R}$ by $l_\Delta(z) = z(\Delta^{\mathbb{N}})$, i.e. the evaluation form l_Δ associates to a measure its total mass. For convenience, we define

$$\tau_{d_1 \cdots d_n} \;:=\; \tau_{d_n} \circ \cdots \circ \tau_{d_1}.$$

Definition 2.33. Let Δ be countable and for $P \in \mathcal{P}(\Delta^{\mathbb{N}})$ let

$$Q_P \;:=\; \big\{\, \tau_{d_1 \cdots d_n}(P) \;\big|\; n \in \mathbb{N}_0,\; d_1, \ldots, d_n \in \Delta \,\big\} \qquad \text{and} \qquad V_P \;:=\; \operatorname{span}(Q_P).$$

For $d \in \Delta$, denote the function $V_P \to V_P$, $z \mapsto \tau_d(z)$ with a slight abuse of notation again by τ_d. Then $\big(V_P, (\tau_d)_{d\in\Delta}, P, l_\Delta\big)$ is called **canonical OOM** of P. If $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$, the canonical OOM of P is the canonical OOM of $P_{\mathbb{N}}$, i.e. $(V_P, (\tau_d), P_{\mathbb{N}}, l_\Delta)$ with $V_P := V_{P_{\mathbb{N}}}$.

Lemma 2.34. *Let Δ be countable and $P \in \mathcal{P}(\Delta^{\mathbb{N}})$. The corresponding canonical OOM is a valid OOM and generates P.*

Proof. This is a special case of Lemma 2.41 below. □

Remark. a) It is straightforward to verify that the canonical OOM has the minimal dimension amongst all OOMs generating P (see [Jae00]). In particular, the dimension of V_P is not bigger (but may be essentially smaller) than the minimal number of internal states required for any HMM generating P.

b) If A is a finite-dimensional cylinder set, the same holds for $[d] \cap \sigma^{-1}(A)$. Therefore, τ_d is weak-$*$ continuous and, consequently, τ_d maps the weak-$*$-closure $\overline{V_P}^{\,w*}$ of V_P to itself. Thus $\big(\overline{V_P}^{\,w*}, (\tau_d)_{d\in\Delta}, P, l_\Delta\big)$ is an OOM of P. The space $\overline{V_P}^{\,w*}$ turns out to be important when we compare the canonical OOM to the causal states in Section 3.4.3.

The dimension of the canonical OOM is an important characteristic of the process. It is called *process dimension* in [Jae00] and *minimum effective degree of freedom* in [IAK92]. Note that we consider it an $\mathbb{N} \cup \{\infty\}$-valued quantity, i.e. it may be infinite, but we do not distinguish between different levels of infinity.

Definition 2.35. The **process dimension** $\dim(P)$ of $P \in \mathcal{P}(\Delta^{\mathbb{N}})$ is the dimension of its canonical OOM,

$$\dim(P) \;:=\; \dim(V_P) \;\in\; \mathbb{N} \cup \{\infty\}.$$

A process is called finite-dimensional (in [Hel65] the term *finitary* is used) if its process dimension is finite. The canonical OOM is closely related to the shift HMM introduced in Example 2.7.

Example 2.36. The (one-sided) shift HMM (T^σ, P) of $P \in \mathcal{P}(\Delta^{\mathbb{N}})$ is a deterministic HMM with set $\Gamma := \Delta^{\mathbb{N}}$ of internal states. It is in general by no means a minimal HMM, and it is not possible to restrict T^σ to a smaller subset of Γ such that it still generates P. If we interpret

the shift HMM as OOM, the internal vector space is $V = \mathcal{M}(\Delta^{\mathbb{N}})$, the associated operators T_d^σ are equal to the canonical ones, i.e. $T_d^\sigma = \tau_d$, the initial vector is the initial distribution of the shift HMM, i.e. $v = P$ and the evaluation form is $z \mapsto z(\Delta^{\mathbb{N}})$. Now it is obvious that we can reduce every OOM to a "cyclic" version by restricting V to $\operatorname{span}\big\{ T^\sigma_{d_1\dots d_n}(v) \;\big|\; n \in \mathbb{N}_0,\ d_1,\dots,d_n \in \Delta \big\}$. This reduced shift OOM is just the canonical OOM. Thus its dimension is minimal, but it can in general not be interpreted as an HMM. ◊

2.2.3 Identifiability problem and existence of finite HMMs

In 1957, the so-called *identifiability problem* was posed by Blackwell and Koopmans ([BK57]). Can we obtain a necessary and sufficient criterion for two different, invariant functional HMMs to generate the same process $X_{\mathbb{Z}}$? This problem gave rise to a sequence of papers obtaining various partial solutions (e.g. [Gil59, Dha63a, Dha63b, Dha65, FR68]). A related problem considered in many of these papers is to find conditions for the existence of finite functional HMMs of a given process $P \in \mathcal{P}(\Delta^{\mathbb{N}})$. This second problem was solved by Heller in [Hel65] with the help of stochastic modules which are equivalent to OOMs. Heller's non-constructive, geometrical condition is the following. There exists a finite HMM of P if and only if there exists a polyhedral convex cone K (i.e. the convex hull of finitely many rays) in the canonical OOM vector space V_P with $P \in K \subseteq \mathcal{M}_+(\Delta^{\mathbb{N}}) \cap V_P$ and $\tau_d(K) \subseteq K$ for all $d \in \Delta$.

The identifiability problem turns out to be easier to solve more generally for OOMs instead of HMMs. Although [Hel65] provided the necessary tools, it was not until 1992 that the identifiability problem was explicitly solved by Ito et al. in [IAK92]. The algorithm given in [IAK92] to check whether two given OOMs generate the same output process requires a computing time that is exponential in dimension of the OOM. In particular, if HMMs are given, the time is exponential in the number of internal states. A more efficient, polynomial time algorithm is given by Schönhuth in [Sch08].

2.2.4 OOMs of Souslin space valued processes

In this section, we extend the definitions of OOMs, associated OOMs and canonical OOMs to the case of processes with values in an arbitrary Souslin space Δ. If Δ is uncountable, we have to index the observable operators with measurable subsets $D \in \mathcal{D}$ instead of elements $d \in \Delta$. In the countable case, the operator T_D is given by $\sum_{d\in D} T_d$ and in general we need a σ-additivity condition. In the vector space V, (countably) infinite sums are not defined, and therefore we cannot require $\tau_{\bigcup_k D_k} = \sum_k \tau_{D_k}$ for disjoint D_k. It is, however, enough to assume the corresponding equality after applying the evaluation form l. We require that Δ is a Souslin space (instead of an arbitrary measurable space) in order to be able to use the Kolmogorov extension theorem. Thus, specifying the finite-dimensional marginals (consistently) is enough to obtain a well-defined process $P \in \mathcal{P}(\Delta^{\mathbb{N}})$.

Definition 2.37. An **OOM** with Souslin output space Δ is a quadruple $\big(V, (T_D)_{D\in\mathcal{D}}, v, l\big)$, where V is a real vector space, $T_D \colon V \to V$ are linear maps, $v \in V$, and l is a linear form on V, such that for $n \in \mathbb{N}$ and $D_1, D_2 \in \mathcal{D}$,

$$1.\ l(v) = 1, \qquad 2.\ l \circ T_\Delta = l, \qquad 3.\ P_{D_1,\dots,D_n} := l \circ T_{D_n} \circ \cdots \circ T_{D_1}(v) \geq 0,$$
$$4.\ l \circ T_{\bigcup_{k\in\mathbb{N}} D_k}(w) = \sum_{k\in\mathbb{N}} l \circ T_{D_k}(w) \qquad \forall w \in V,\ \text{disjoint } D_k \in \mathcal{D}.$$

The process $P \in \mathcal{P}(\Delta^{\mathbb{N}})$, defined by $P\big([D_1 \times \cdots \times D_n]\big) := P_{D_1,\ldots,D_n}$, is called **generated** by the OOM and the OOM is called OOM of P.

Definition 2.38. Let (T, μ) be an HMM and $V := \mathcal{M}(\Gamma)$ the set of signed measures of bounded variation on Γ. For $D \in \mathcal{D}$ we define the linear maps $T_D \colon V \to V$ by

$$T_D(\nu)(G) \;:=\; \int T\big(\,\cdot\,;\, D \times G\big)\,\mathrm{d}\nu \qquad \forall \nu \in V = \mathcal{M}(\Gamma),\; G \in \mathcal{G}.$$

Then $\big(V, (T_D)_{D\in\mathcal{D}}, \mu, l\big)$ with $l(\nu) := \nu(\Gamma)$ is called the **associated OOM** of (T, μ).

Remark. With this notation, we can express invariance of an HMM (T, μ) simply as $\mu = T_\Delta(\mu)$.

Lemma 2.39. *Let (T, μ) be an HMM of $P \in \mathcal{P}(\Delta^{\mathbb{N}})$. Then the associated OOM is an OOM generating P.*

Proof. The OOM properties are obvious. To see that the associated OOM generates the same process, recall that $O_T(g)$ denotes the output process of the HMM (T, δ_g). For $D_1, \ldots, D_n \in \mathcal{D}$ we obtain via induction over n that

$$\begin{aligned} P\big([D_1 \times \cdots \times D_n]\big) \quad &= \quad \int_{g_0\in\Gamma} \int_{D_1\times\Gamma} O_T(\,\cdot\,;\, D_2 \times \cdots \times D_n)\,\mathrm{d}T(g_0)\,\mathrm{d}\mu \\ &= \quad \int O_T(\,\cdot\,;\, D_2 \times \cdots \times D_n)\,\mathrm{d}T_{D_1}(\mu) \\ &\overset{\text{(induction)}}{=} \quad l \circ T_{D_n} \circ \cdots \circ T_{D_2}\big(T_{D_1}(\mu)\big), \end{aligned}$$

which implies the claimed identity of processes. □

The construction of the canonical OOM can also be extended to the case of Souslin spaces in the obvious way. In analogy to Definition 2.33, we define

Definition 2.40. Let Δ be a Souslin space and for $D \in \mathcal{D}$ define

$$\tau_D \colon \mathcal{M}(\Delta^{\mathbb{N}}) \to \mathcal{M}(\Delta^{\mathbb{N}}), \qquad z \mapsto \tau_D(z) \;:=\; z\big([D] \cap \sigma^{-1}(\,\cdot\,)\big)$$

and $l_\Delta(z) := z(\Delta^{\mathbb{N}})$. For $P \in \mathcal{P}(\Delta^{\mathbb{N}})$ let

$$Q_P \;:=\; \big\{\, \tau_{D_n} \circ \cdots \circ \tau_{D_1}(P) \;\big|\; n \in \mathbb{N}_0,\; D_1, \ldots, D_n \in \mathcal{D} \,\big\} \qquad \text{and} \qquad V_P \;:=\; \operatorname{span}(Q_P).$$

Denote the function $V_P \to V_P$, $z \mapsto \tau_D(z)$ with a slight abuse of notation again by τ_D. Then $\big(V_P, (\tau_D)_{D\in\mathcal{D}}, P, l_\Delta\big)$ is called **canonical OOM** of P.

Lemma 2.41. *Let Δ be countable and $P \in \mathcal{P}(\Delta^{\mathbb{N}})$. The corresponding canonical OOM is an OOM of P.*

Proof. Obviously, $\tau_D(V_P) \subseteq V_P$ and therefore the canonical OOM is well-defined. We see that it generates P as follows.

$$\begin{aligned} l_\Delta \circ \tau_{D_n} \circ \cdots \circ \tau_{D_1}(P) \;&=\; \tau_{D_n}\big(\tau_{D_{n-1}} \circ \cdots \circ \tau_{D_1}(P)\big)(\Delta^{\mathbb{N}}) \;=\; \tau_{D_{n-1}} \circ \cdots \circ \tau_{D_1}(P)(D_n) \\ &=\; \cdots \;=\; P\big(D_1 \cap \sigma^{-1}(D_2) \cap \cdots \cap \sigma^{-n+1}(D_n)\big) \\ &=\; P\big([D_1 \times \cdots \times D_n]\big) \end{aligned}$$

The OOM properties are now obvious. □

Chapter 3

Predictive models

So far we considered the task of generating a process or representing its distribution by different kinds of models. With a generative model, we describe the statistics of a process and are able to simulate it. In this chapter, we shift the focus to the related but not identical task of predicting a stationary stochastic process $X_{\mathbb{Z}}$. We interpret $X_{-\mathbb{N}_0}$ as the observed past, and $X_{\mathbb{N}}$ as the future, which we want to predict. In this chapter, $X_{\mathbb{Z}}$ *is always assumed to be stationary* and Δ *is assumed to be a Souslin space.*

3.1 Some information theory

For the subsequent discussion, we need some information theoretic quantities. In most of the literature about information theory (e.g. in the standard reference [CT91]), the underlying spaces are assumed to be either discrete or $\mathbb{R}^n$. For our purposes, considering general measurable spaces is advantageous, and an excellent treatment of this more general case is given in [Kak99].

3.1.1 Entropy and mutual information

The most basic quantity is (Shannon) entropy, $H(\mu)$, of a probability measure μ on a measurable space Γ. It describes how "random" or "diverse" the measure is. It can also be used as measure of the "size" of a probability space, mainly justified by its role in coding theory. This interpretation is important when entropy is used for the definition of various complexity measures. If Γ is finite, the entropy is defined by

$$H(\mu) \;:=\; \sum_{g\in\Gamma} \varphi\big(\mu(g)\big), \qquad \text{where} \qquad \varphi(x) \;:=\; -x\log(x).$$

It satisfies $0 \le H(\mu) \le \log\big(|\Gamma|\big)$. If Γ is not finite, $H(\mu)$ is the supremum of the entropy of finite partitions.

Definition 3.1. Let Γ be a measurable space and $\mu \in \mathcal{P}(\Gamma)$. The **entropy** $H(\mu) \in \overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{\infty\}$ of μ is defined by

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

If X is a random variable, its **entropy** $H^{\mathbb{P}}(X)$ is defined as the entropy of its distribution,

$$H^{\mathbb{P}}(X) \;:=\; H(\mathbb{P}_X).$$

Notation. It is common practice to write $H(X)$ for $H^{\mathbb{P}}(X)$, assuming that it is clear from the context if the argument is a random variable or a measure. We do not follow this convention, because below we investigate measure valued random variables. By distinguishing between H and $H^{\mathbb{P}}$, we avoid any confusion.

Assume Γ is a separable, metrisable space, and $\mu \in \mathcal{P}(\Gamma)$ has finite entropy. Then μ must be supported by a countable set A. In this case,

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

Consequently, the entropy $H^{\mathbb{P}}(X_{\mathbb{Z}})$ of a stochastic process is usually infinite, even for finite Δ. But if $X_{\mathbb{Z}}$ is stationary, it has a well-defined entropy rate.

Definition 3.2. Let $X_{\mathbb{Z}}$ be a stationary process with distribution $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$. Then the **entropy rate** is defined by

$$h_{\mathbb{P}}(X_{\mathbb{Z}}) \;:=\; h(P) \;:=\; \lim_{n \to \infty} \tfrac{1}{n} H^P(X'_{[1,n]}).$$

We occasionally need the conditional entropy of a random variable Y given knowledge of another random variable X. In the case of finite range spaces, it is defined as $H^{\mathbb{P}}(Y \mid X) := H^{\mathbb{P}}(X,Y) - H^{\mathbb{P}}(X)$. Information theory also provides a quantity measuring the total amount of information contained in X about the random variable Y. It is called mutual information and, if both range spaces are finite, it is defined as the reduction of the entropy of Y achieved by the knowledge of X, i.e. $I(X : Y) := H^{\mathbb{P}}(Y) - H^{\mathbb{P}}(Y \mid X)$. In the more general case of uncountable range spaces, conditional entropy and mutual information are, like entropy, defined as a supremum over finite approximations in the obvious way. Also random variables with infinite entropy have a well-defined mutual information that may (or may not) be finite. Mutual information can also be expressed in terms of the Kullback-Leibler divergence.

Definition 3.3. Let $\mu, \nu \in \mathcal{P}(\Gamma)$ for a measurable space Γ. Then the **Kullback-Leibler divergence** is defined by

$$D_{\mathrm{KL}}(\mu \parallel \nu) \;:=\; \begin{cases} \displaystyle\int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu & \text{if } \mu \ll \nu, \\ \infty & \text{otherwise,} \end{cases}$$

where $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ is the Radon-Nikodym derivative of μ w.r.t. ν and $\mu \ll \nu$ means that μ is absolutely continuous w.r.t. ν, i.e. that the Radon-Nikodym derivative exists.

Kullback-Leibler divergence is commonly interpreted as a distance measure, although it is neither symmetric nor satisfies the triangle inequality. The mutual information between random variables X and Y can be expressed as the "distance" from the joint distribution $\mathbb{P}_{X,Y}$ to the product distribution of the marginals,

$$I(X : Y) \;=\; D_{\mathrm{KL}}(\mathbb{P}_{X,Y} \parallel \mathbb{P}_X \otimes \mathbb{P}_Y),$$

in other words the "distance" to being independent. We summarise a few well-known properties of the mutual information. Let X, Y and Z be random variables.

a) $0 \leq I(X:Y) = I(Y:X) \leq H(X)$

b) $I(X:Y) = 0$ if and only if X and Y are independent.

c) If X is conditionally independent of Y given Z, then $I(X:Y) \leq I(Z:Y)$.

d) Let X_n, Y_n be random variables with $\sigma(X_n) \subseteq \sigma(X_{n+1})$, $\sigma(Y_n) \subseteq \sigma(Y_{n+1})$. Further assume $\sigma(X) = \sigma(X_n,\ n \in \mathbb{N})$ and $\sigma(Y) = \sigma(Y_n,\ n \in \mathbb{N})$. Then

$$I(X:Y) \ = \ \sup_{n\in\mathbb{N}} I(X_n:Y_n) \ = \ \lim_{n\to\infty} I(X_n:Y_n).$$

3.1.2 Excess entropy

An important question related to prediction is the following. How much information about the future is contained in the past? The answer is given by the mutual information between $X_{-\mathbb{N}_0}$ and $X_{\mathbb{N}}$. This quantity arises in a number of different contexts, for instance it is used in the discussion of sufficient memories below. It is also a well-accepted complexity measure on its own, studied by Grassberger under the name of *effective measure complexity* ([Gra86]) and by Bialek et al. under the name of *predictive information* ([BNT01]). It is called *excess entropy* by Crutchfield and Feldman ([CF03][1]), because of the following well-known identity.

Proposition 3.4. *Let $X_{\mathbb{Z}}$ be a stationary process with distribution $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ and finite marginal entropy $H^{\mathbb{P}}(X_1) < \infty$. Then*

$$I(X_{-\mathbb{N}_0} : X_{\mathbb{N}}) \ = \ \lim_{n\to\infty} H^{\mathbb{P}}(X_{[1,n]}) - n \cdot h(P).$$

Proof. Let $H_n := H^{\mathbb{P}}(X_{[1,n]})$ and note that $I(X_{-\mathbb{N}_0} : X_{\mathbb{N}}) = \sup_{n,m\in\mathbb{N}} I(X_{[-m,0]} : X_{[1,n]})$. Due to stationarity of $X_{\mathbb{Z}}$,

$$\begin{aligned} I(X_{-\mathbb{N}_0} : X_{\mathbb{N}}) \ &= \ \sup_{n,m\in\mathbb{N}} H_n - H^{\mathbb{P}}(X_{[1,n]} \mid X_{[-m,0]}) \ = \ \sup_n H_n - n \cdot \inf_m H^{\mathbb{P}}(X_1 \mid X_{[-m,0]}) \\ &= \ \sup_n H_n - nh(P) \ = \ \lim_{n\to\infty} H_n - nh(P), \end{aligned}$$

where we used the well-known identity $h(P) = H^{\mathbb{P}}(X_1 \mid X_{-\mathbb{N}_0})$. □

The above representation of predictive information supports its interpretation as complexity measure. It quantifies the amount of apparent randomness in the positive time part $X_{\mathbb{N}}$ of the process that can be "explained" by the past $X_{-\mathbb{N}_0}$. Thus it is a measure of structure of the process. We use the name excess entropy, because this term is used in computational mechanics and also by Dębowski for his ergodic decomposition result (see Chapter 4).

Definition 3.5. Let $X_{\mathbb{Z}}$ be a stationary process with distribution $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$. We call $E(X_{\mathbb{Z}}) := E(P) := I^{\mathbb{P}}(X_{-\mathbb{N}_0} : X_{\mathbb{N}})$ **excess entropy** of $X_{\mathbb{Z}}$ or of P.

It is not difficult to prove but an important fact that the excess entropy of a stationary process is bounded by the internal state entropy of any generative HMM.

Proposition 3.6. *Let (T, μ) be an invariant HMM of $X_{\mathbb{Z}}$. Then $E(X_{\mathbb{Z}}) \leq H(\mu)$.*

Proof. Let $W_{\mathbb{Z}}$ be the internal process. Conditional independence of $X_{\mathbb{N}}$ and $X_{-\mathbb{N}_0}$ given W_0 yields

$$E(X_{\mathbb{Z}}) = I(X_{-\mathbb{N}_0} : X_{\mathbb{N}}) \ \leq \ I(W_0 : X_{\mathbb{N}}) \ \leq \ H^{\mathbb{P}}(W_0) = H(\mu).$$ □

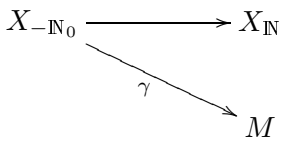
[1]The name *excess entropy* was already used for a related but different quantity in [CP83]

3.2 Computational mechanics

One approach to predictive models of stochastic processes is a theory called computational mechanics. It is based on the fundamental concept of sufficient statistics.

3.2.1 Memories and sufficiency

Not all information of the past $X_{-\mathbb{N}_0}$ is necessary for predicting the future $X_{\mathbb{N}}$. Therefore, one tries to compress the relevant information in a memory variable M via a memory kernel (transition probability) γ. The memory variable assumes values in a measurable space Γ of memory states. This is illustrated as



Of course, the memory variable M has to be conditionally independent of the future $X_{\mathbb{N}}$ given the past $X_{-\mathbb{N}_0}$, and the conditional distribution is given by

$$\mathbb{P}(M \mid X_{\mathbb{Z}}) \;=\; \gamma \circ X_{-\mathbb{N}_0}. \tag{3.1}$$

For technical reasons, we assume that Γ is a Souslin space. Sometimes, we call both the memory variable M and the memory kernel γ simply *memory*. No confusion arises, as one determines the other.

Definition 3.7. A **memory kernel** is a Markov kernel $\gamma\colon \Delta^{-\mathbb{N}_0} \to \mathcal{P}(\Gamma)$ from the past to a Souslin space Γ of **memory states**. The associated random variable M defined by (3.1) is called **memory variable**.

In general, γ reduces the information about the future. Particularly important are memories that avoid this potential reduction and capture all information about the future that is available in the past. In the case of finite excess entropy, we can formalise this requirement in terms of mutual information (Proposition 3.10 below), but more generally this means that the future is conditionally independent of the past given the memory variable. Using the language of statistics, we call such memories *sufficient* for the future.

Definition 3.8. a) Let X, Y, Z be random variables. X is **conditionally independent** of Y given Z, written as $X \perp\!\!\!\perp Y \mid Z$, if

$$\mathbb{P}\big(\{\, X \in A,\; Y \in B \,\} \,\big|\, Z\big) \;=\; \mathbb{P}\big(\{\, X \in A \,\} \,\big|\, Z\big) \cdot \mathbb{P}\big(\{\, Y \in B \,\} \,\big|\, Z\big) \qquad \text{a.s.}$$

for all measurable sets A and B in the range space of X and Y respectively.

b) A memory kernel γ and its associated memory variable M are called **sufficient** if

$$X_{-\mathbb{N}_0} \perp\!\!\!\perp X_{\mathbb{N}} \mid M.$$

The following characterisation of conditional independence is well-known.

Lemma 3.9. *For random variables X, Y, Z with values in Souslin spaces the following holds.*

$$X \perp\!\!\!\perp Y \mid Z \quad \Leftrightarrow \quad \mathbb{P}(Y \mid X, Z) \;=\; \mathbb{P}(Y \mid Z) \;\; a.s.$$

If $Z \perp\!\!\!\perp Y \mid X$, then $I(Z:Y) \leq I(X:Y)$, and if additionally $I(X:Y) < \infty$, then

$$X \perp\!\!\!\perp Y \mid Z \quad \Leftrightarrow \quad I(Z:Y) \;=\; I(X:Y).$$

Let M be a memory variable. We note that $M \perp\!\!\!\perp X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}$ holds by definition of memory, and therefore $I(M:X_{\mathbb{N}}) \leq I(X_{-\mathbb{N}_0}:X_{\mathbb{N}}) = E(X_{\mathbb{Z}})$. Recall that we can interpret $\mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0})$ as measurable function from Ω to $\mathcal{P}(\Delta^{\mathbb{N}})$ and abbreviate

$$\mathbb{P}_{X_{\mathbb{N}}}^{X_{-\mathbb{N}_0}} \;:=\; \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \colon \Omega \to \mathcal{P}(\Delta^{\mathbb{N}}).$$

In the following, we use this notation in particular when we want to emphasise this interpretation as measurable function. Then it is clear what the generated σ-algebra $\sigma(\mathbb{P}_{X_{\mathbb{N}}}^{X_{-\mathbb{N}_0}})$ means and obviously $\sigma\big(\mathbb{P}_{X_{\mathbb{N}}}^{X_{-\mathbb{N}_0}}\big) = \sigma\big(\mathbb{P}\big(\{\, X_{\mathbb{N}} \in A \,\} \;\big|\; X_{-\mathbb{N}_0}\big),\; A \in \mathfrak{B}(\Delta^{\mathbb{N}})\big)$. We now see that a memory is sufficient if and only if the σ-algebra generated by M, $\sigma(M)$, is a refinement of $\sigma(\mathbb{P}_{X_{\mathbb{N}}}^{X_{-\mathbb{N}_0}})$.

Proposition 3.10. *Let $X_{\mathbb{Z}}$ be a stationary process and M a memory variable. Then the following properties are equivalent.*

1. M is sufficient. *2.* $\mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \;=\; \mathbb{P}(X_{\mathbb{N}} \mid M)$ *a.s.*

3. $\sigma(M) \;\supseteq\; \sigma\big(\mathbb{P}_{X_{\mathbb{N}}}^{X_{-\mathbb{N}_0}}\big)$ *modulo* $\mathbb{P}$.

If the excess entropy is finite, $E(X_{\mathbb{Z}}) < \infty$, then the following property is also equivalent

4. $I(M:X_{\mathbb{N}}) \;=\; E(X_{\mathbb{Z}})$.

Proof. *"1. $\Leftrightarrow$ 2.":* We apply Lemma 3.9 twice. Sufficiency is equivalent to $\mathbb{P}(X_{\mathbb{N}} \mid M) = \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}, M)$. Due to $M \perp\!\!\!\perp X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}$, we have for every memory (sufficient or not) that $\mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}, M) = \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0})$.

"2. $\Leftrightarrow$ 3.": Let $\mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) = \mathbb{P}(X_{\mathbb{N}} \mid M)$ a.s. Then $\sigma(M) \supseteq \sigma\big(\mathbb{P}(X_{\mathbb{N}} \mid M)\big) = \sigma(\mathbb{P}_{X_{\mathbb{N}}}^{X_{-\mathbb{N}_0}})$ modulo $\mathbb{P}$. Conversely, if $\sigma(\mathbb{P}_{X_{\mathbb{N}}}^{X_{-\mathbb{N}_0}}) \subseteq \sigma(M)$ modulo $\mathbb{P}$, then

$$\mathbb{P}(X_{\mathbb{N}} \mid M) \;=\; \mathbb{P}(X_{\mathbb{N}} \mid M,\, \mathbb{P}_{X_{\mathbb{N}}}^{X_{-\mathbb{N}_0}}) \;=\; \mathbb{P}(X_{\mathbb{N}} \mid M, X_{-\mathbb{N}_0}) \qquad \text{a.s.}$$

The last equality follows from Lemma A.4 (with $X = Z = X_{\mathbb{N}}$, $Y = X_{-\mathbb{N}_0}$). Finally, we have $\mathbb{P}(X_{\mathbb{N}} \mid M, X_{-\mathbb{N}_0}) = \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0})$.

"1. $\Leftrightarrow$ 4.": From the second statement of Lemma 3.9, we see that sufficiency of the memory is equivalent to $I(M:X_{\mathbb{N}}) = E(X_{\mathbb{Z}})$, provided that $E(X_{\mathbb{Z}}) < \infty$. □

In the following, we frequently use the first equivalence of Proposition 3.10 without further notice. One might take it as alternative definition of sufficient memory.

Remark. The sufficiency property is called *prescient* in [SC01]. It is the central requirement in computational mechanics and sufficient memories are the candidates for predictive models proposed by computational mechanics.

A memory kernel γ does not only induce a memory variable $M_0 := M$ at time zero, but a whole stationary **memory process** $M_{\mathbb{Z}}$ produced by application of γ in each time step. The conditional distribution of $M_{\mathbb{Z}}$ is the product distribution given by

$$\mathbb{P}(M_{\mathbb{Z}} \mid X_{\mathbb{Z}})(\omega) \;=\; \bigotimes_{k\in\mathbb{Z}} \gamma\big(X_{]-\infty,k]}(\omega)\big).$$

Note that the memory process $M_{\mathbb{Z}}$ of a sufficient memory is not necessarily Markovian, as we see in the following simple example.

Example 3.11. Let $\Delta = \{\,0,1\,\}$ and P the Δ-valued i.i.d. process with uniformly distributed marginals. Then every memory kernel is sufficient. Let $\gamma\colon \Delta^{-\mathbb{N}_0} \to \mathcal{P}\big(\{\,0,1\,\}\big)$, $x \mapsto \delta_{h(x)}$ with $h(x) = x_0 \cdot x_{-42}$, where $x = (x_k)_{k\in-\mathbb{N}_0} \in \Delta^{-\mathbb{N}_0}$. That is, $M_0 = 1$ if and only if $X_{-42} = X_0 = 1$. Then M_k is obviously independent of M_{k-1}, but it does depend on M_{k-42}, because $\mathbb{P}\big(\{\,M_k = 1\,\} \,\big|\, M_{k-42} = 1\big) = \frac{1}{2} > \mathbb{P}\big(\{\,M_k = 1\,\}\big)$. ♢

Sufficient memories contain all information about the future that is available in the past. How do we actually extract this information and justify the term "model" for sufficient memories? In the following proposition, we see that sufficient memories induce generative HMMs. In general, the process of internal states of the associated HMM cannot have the same distribution as the memory process $M_{\mathbb{Z}}$, because the latter need not be Markovian. The (first order) Markov approximation of the joint process $(M_{\mathbb{Z}}, X_{\mathbb{Z}})$, however, yields the desired HMM.

Proposition 3.12 (sufficient memories induce generative HMMs). *Let $X_{\mathbb{Z}}$ be a stationary process and γ a sufficient memory kernel with space Γ of memory states. Let $M_{\mathbb{Z}}$ be the corresponding memory process and define $T^\gamma\colon \Gamma \to \mathcal{P}(\Delta\times\Gamma)$ by*

$$T^\gamma(g) \;:=\; \mathbb{P}(X_1,\, M_1 \mid M_0 = g).$$

Then for $x \in \Delta^{-\mathbb{N}_0}$, an HMM of the conditional process $\mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = x) \in \mathcal{P}(\Delta^{\mathbb{N}})$ is a.s. given by $\big(T^\gamma,\, \gamma(x)\big)$. An invariant HMM of $X_{\mathbb{Z}}$ is given by (T^γ, μ_γ) with $\mu_\gamma := \mathbb{P}\circ M_0^{-1}$.

Proof. 1. We abbreviate $T := T^\gamma$. Note that μ_γ is T-invariant, because $M_{\mathbb{Z}}$ is stationary and $T_\Delta(\mu_\gamma) = \int_{g\in\Gamma} T(g;\, \Delta\times\,\cdot\,)\,\mathrm{d}\mu_\gamma$ is the distribution of M_1. Because $\mu_\gamma = \int \gamma\circ X_{-\mathbb{N}_0}\,\mathrm{d}\mathbb{P}$, the second claim follows from the first. Hence it is sufficient to prove $O_T^{\gamma(x)} = \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = x)$ a.s. Recall that O_T^μ denotes the output process of the HMM (T,μ), and $O_T(g) = O_T^{\delta_g}$.

2. We claim $O_T(g) = \mathbb{P}(X_{\mathbb{N}} \mid M_0 = g)$ for $\mathbb{P}_{M_0}$-a.a. $g \in \Gamma$. It is enough to prove it for cylinder sets $A := [D_1 \times \cdots \times D_n]$, and we do this by induction over n. Let σ be the left-shift on $\Delta^{\mathbb{N}}$ and set $B := [D_2 \times \cdots \times D_n]$. Then $A = [D_1] \cap \sigma^{-1}(B)$. The induction hypothesis, together with stationarity and sufficiency of the memory, yields a.s.

$$\begin{aligned} O_T(M_1(\omega);\, B) \;&=\; \mathbb{P}\big(\{\,X_{\mathbb{N}} \in B\,\} \,\big|\, M_0 = M_1(\omega)\big) \;=\; \mathbb{P}\big(\{\,X_{[2,\infty[} \in B\,\} \,\big|\, M_1\big)(\omega) \\ &=\; \mathbb{P}\big(\{\,X_{[2,\infty[} \in B\,\} \,\big|\, X_{]-\infty,1]}\big)(\omega). \end{aligned} \tag{3.2}$$

Now, using $T\circ M_0 = \mathbb{P}(X_1, M_1 \mid M_0)$, we obtain a.s.

$$\begin{aligned} O_T(M_0;\, A) \;&=\; \int_{(d,m)\in D_1\times\Gamma} O_T(m;\, B)\,\mathrm{d}\mathbb{P}(X_1, M_1 \mid M_0) \\ &\overset{(3.2)}{=}\; \int 1_{\{X_1\in D_1\}}\,\mathbb{P}\big(\{\,X_{[2,\infty[} \in B\,\} \,\big|\, X_{]-\infty,1]}\big)\,\mathrm{d}\mathbb{P}(\,\cdot\mid M_0) \\ &=\; \int \mathbb{P}\big(\{\,X_{\mathbb{N}} \in A\,\} \,\big|\, X_{-\mathbb{N}_0}, X_1\big)\,\mathrm{d}\mathbb{P}(\,\cdot\mid M_0) \;=\; \mathbb{P}\big(\{\,X_{\mathbb{N}} \in A\,\} \,\big|\, M_0\big), \end{aligned}$$

where the last equality is due to $M_0 \perp\!\!\!\perp X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}, X_1$. This finishes the induction.

3. Let $M = M_0$. Using sufficiency of M and step 2., we obtain a.s.

$$\begin{aligned} O_T^{\gamma(X_{-\mathbb{N}_0})} &= \int O_T \circ M \, \mathrm{d}\mathbb{P}(\,\cdot \mid X_{-\mathbb{N}_0}) \overset{\text{(step 2.)}}{=} \int \mathbb{P}(X_{\mathbb{N}} \mid M) \, \mathrm{d}\mathbb{P}(\,\cdot \mid X_{-\mathbb{N}_0}) \\ &= \int \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \, \mathrm{d}\mathbb{P}(\,\cdot \mid X_{-\mathbb{N}_0}) \;=\; \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}). \end{aligned}$$ □

Definition 3.13. Let $X_{\mathbb{Z}}$ be a stationary process and γ a sufficient memory kernel. Then we call the HMM (T^γ, μ_γ) of $X_{\mathbb{Z}}$ constructed in Proposition 3.12 the **HMM induced by** γ.

3.2.2 Deterministic memories, partitions and σ-algebras

An important special case of memory kernels are deterministic maps. Because Γ is embedded in $\mathcal{P}(\Gamma)$ via Dirac measures, a measurable function $h\colon \Delta^{-\mathbb{N}_0} \to \Gamma$ induces a memory kernel $\gamma_h(x) := \delta_{h(x)}$, where δ_g is the Dirac measure in g. We call a memory **deterministic** if the memory kernel is induced by a function h in this way. Restricting to deterministic memories is the usual approach in computational mechanics, although an extension to stochastic maps has been considered by Still et al. ([SCE07]). For a deterministic memory γ_h, we may assume that $M = h \circ X_{-\mathbb{N}_0}$ and it is sufficient if and only if h is measurable w.r.t. the σ-algebra generated by the kernel $\mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = \,\cdot\,)$, up to a set of measure zero (Proposition 3.10).

One might be tempted to think that the HMM induced by a sufficient deterministic memory is always state observable (i.e. W_0 is determined by $X_{-\mathbb{N}_0}$), because the memory state is a function of the past. This is, however, not always true. In the following proposition, we provide equivalent conditions to state observability of the induced HMM. In particular, it is equivalent to partial determinism, which is a strictly weaker property for general HMMs (see Section 2.1.7). If $x = (x_k)_{k\in-\mathbb{N}_0} \in \Delta^{-\mathbb{N}_0}$ is the past trajectory and $d \in \Delta$, we denote by $xd \in \Delta^{-\mathbb{N}_0}$ the resulting past when d is observed, i.e. $xd = (y_k)_{k\in-\mathbb{N}_0}$ with $y_0 = d$ and $y_k = x_{k+1}$ for $k \in -\mathbb{N}$.

Proposition 3.14. *Let $h\colon \Delta^{-\mathbb{N}_0} \to \Gamma$ be measurable and such that the deterministic memory $\gamma = \gamma_h$ is sufficient. Let $M_{\mathbb{Z}} = (h \circ X_{]-\infty,k]})_{k\in\mathbb{Z}}$ be the process of memory states, and (T^γ, μ_γ) the induced HMM with Γ-valued internal process $W_{\mathbb{Z}}$. The following properties are equivalent.*

1. *(T^γ, μ_γ) is state observable.*
2. *(T^γ, μ_γ) is partially deterministic.*
3. *$M_1 = f(M_0, X_1)$ a.s. for some measurable $f\colon \Gamma \times \Delta \to \Gamma$.*
4. *$f\big(h(x), d\big) := h(xd)$ is a.s. (w.r.t. $\mathbb{P}_{X_{]-\infty,1]}}$) well-defined for $x \in \Delta^{-\mathbb{N}_0}$, $d \in \Delta$.*
5. *$(W_{\mathbb{Z}}, X_{\mathbb{Z}})$ has the same joint distribution as $(M_{\mathbb{Z}}, X_{\mathbb{Z}})$.*

The functions f in 4. and 3. a.s. coincide with the transition function of the induced HMM.

Proof. *"1. ⇒ 2.":* Corollary 2.26
"2. ⇔ 3.": By definition of the induced HMM, the triples (M_0, X_1, M_1) and (W_0, X_1, W_1) have the same joint distribution. Partial determinism is equivalent to $W_1 = f(W_0, X_1)$ a.s., where f is the transition function (Proposition 2.25).

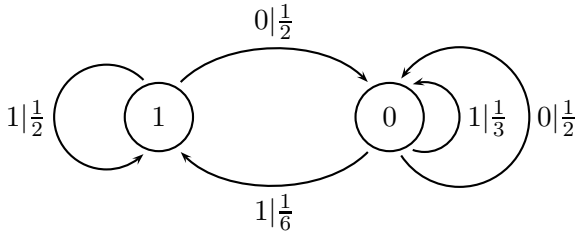


Figure 3.1: Not partially deterministic HMM induced by a sufficient deterministic memory (Example 3.15).

"3. ⇔ 4.": Obvious from $M_1(\omega) = h(xd)$, $M_0(\omega) = h(x)$ if $X_{-\mathbb{N}_0}(\omega) = x$, $X_1(\omega) = d$.
"3. ⇒ 5.": Due to 3., M_1 is $\sigma(M_0, X_1)$-measurable modulo $\mathbb{P}$. Sufficiency implies that $X_{-\mathbb{N}_0} \perp\!\!\!\perp M_0, X_1 \mid M_0$. Thus also $X_{-\mathbb{N}_0} \perp\!\!\!\perp M_1 \mid M_0$, and we obtain

$$\mathbb{P}(X_1, M_1 \mid M_0) \;=\; \mathbb{P}(X_1, M_1 \mid M_0, X_{-\mathbb{N}_0}) \;=\; \mathbb{P}(X_1, M_1 \mid M_{-\mathbb{N}_0}, X_{-\mathbb{N}_0}).$$

This means that the joint process $(M_{\mathbb{Z}}, X_{\mathbb{Z}})$ satisfies the correct Markov property and the Markov approximation $(W_{\mathbb{Z}}, X_{\mathbb{Z}})$ has the same distribution.
"5. ⇒ 1.": Using 5., $M_0 = h(X_{-\mathbb{N}_0})$ implies $W_0 = h(X_{-\mathbb{N}_0})$, i.e. state observability. □

Example 3.15. Let P be the uniform i.i.d. process with values in $\Delta = \{0, 1\}$, and $\gamma(x) = \gamma_h(x)$ with $h(x) = x_{-42} \cdot x_0$, just as in Example 3.11. We have seen there that the memory process is not Markovian, and thus Proposition 3.14 implies that the induced HMM cannot be partially deterministic. The induced HMM is shown in Figure 3.1. Note, however, that Markovianity of the memory process is not sufficient for the equivalent properties of Proposition 3.14. Consider the i.i.d. process on $\Delta = \{0, 1, 2\}$ with uniform marginals and the deterministic memory γ_h defined by

$$h\colon \Delta^{-\mathbb{N}_0} \to \{m_1, m_2, m_3\}, \qquad h(x) \;:=\; \begin{cases} m_1 & \text{if } x_0 \in \{0, 1\}, \\ m_2 & \text{if } x_0 = 2 \text{ and } x_{-1} = 1, \\ m_3 & \text{if } x_0 = 2 \text{ and } x_{-1} \in \{0, 2\}. \end{cases}$$

It is easy to verify that the memory process is Markovian, because the memory state identifies the last observation unless it is m_1, and in that case past memory states do not provide additional information. On the other hand, the induced HMM is not partially deterministic because if the internal state is m_1 and the emitted symbol is 2, the next internal state can either be m_2 or m_3. ◊

Instead of specifying a deterministic memory by a measurable function $h\colon \Delta^{-\mathbb{N}_0} \to \Gamma$, we can define it by a sub-σ-algebra $\mathfrak{R}$ of the Borel σ-algebra on $\Delta^{-\mathbb{N}_0}$. Then the space $\Gamma_{\mathfrak{R}}$ of memory states is the set of atoms of $\mathfrak{R}$, equipped with the final σ-algebra of the canonical projection. More explicitly, the atom $[x]_{\mathfrak{R}}$ of $\mathfrak{R}$ containing $x \in \Delta^{-\mathbb{N}_0}$ is the intersection of all $\mathfrak{R}$-measurable sets containing x, i.e.

$$[x]_{\mathfrak{R}} \;:=\; \bigcap \{ R \in \mathfrak{R} \mid x \in R \}, \qquad x \in \Delta^{-\mathbb{N}_0}.$$

Note that the atoms of $\mathfrak{R}$ are $\mathfrak{R}$-measurable if $\mathfrak{R}$ is countably generated, but otherwise this need not be the case. In fact, they can even be non-measurable w.r.t. $\mathfrak{B}(\Delta^{-\mathbb{N}_0})$. The measurable space $(\Gamma_{\mathfrak{R}}, \mathcal{G}_{\mathfrak{R}})$ of memory states is defined as the quotient space $\Delta^{-\mathbb{N}_0}/\mathfrak{R}$, i.e.

$$\Gamma_{\mathfrak{R}} \;:=\; \{\, [x]_{\mathfrak{R}} \mid x \in \Delta^{-\mathbb{N}_0} \,\} \qquad \text{and} \qquad \mathcal{G}_{\mathfrak{R}} \;:=\; \{\, G \subseteq \Gamma_{\mathfrak{R}} \mid \bigcup G \in \mathfrak{R} \,\}.$$

Because G is a set of atoms, $\bigcup G$ is the union $\bigcup_{[x]_{\mathfrak{R}} \in G} [x]_{\mathfrak{R}}$ of these atoms and at the same time the pre-image ${[\,\cdot\,]_{\mathfrak{R}}}^{-1}(G)$ of G under the canonical projection $x \mapsto [x]_{\mathfrak{R}}$. The canonical projection also defines the memory kernel by $\gamma_{\mathfrak{R}}(x) := \delta_{[x]_{\mathfrak{R}}}$. In other words, the memory variable defined by $\mathfrak{R}$ is equal to the atom of $\mathfrak{R}$ containing the past. We still have to take care that the space $\Gamma_{\mathfrak{R}}$ is a Souslin measurable space.

Proposition 3.16. *Let $\mathfrak{R}$ be a sub-σ-algebra of $\mathfrak{B}(\Delta^{-\mathbb{N}_0})$. Then $\Gamma_{\mathfrak{R}}$ is a Souslin measurable space if and only if $\mathfrak{R}$ is countably generated.*

Proof. Obviously, $\mathfrak{R}$ is countably generated if and only if $\mathcal{G}_{\mathfrak{R}}$ is countably generated (the canonical projection $[\,\cdot\,]_{\mathfrak{R}}$ induces an isomorphism of set algebras). If $\Gamma_{\mathfrak{R}}$ is a Souslin measurable space, it has a countably generated σ-algebra. Conversely, assume that $\mathcal{G}_{\mathfrak{R}}$ is countably generated. Then the atoms of $\mathfrak{R}$ are $\mathfrak{R}$-measurable and thus the singletons in $\Gamma_{\mathfrak{R}}$ are measurable. Furthermore, the canonical projection $[\,\cdot\,]_{\mathfrak{R}}$ is a surjective measurable map from the Souslin space $\Delta^{-\mathbb{N}_0}$ onto $\Gamma_{\mathfrak{R}}$. But if a measurable space with measurable singletons and countably generated σ-algebra is the image of a Souslin space under a measurable map, then it is a Souslin measurable space ([Coh80, Prop. 8.6.5]). □

If $\mathfrak{R} = \sigma(h)$ for a measurable function $h\colon \Delta^{-\mathbb{N}_0} \to \Gamma$, then the quotient space $\Gamma_{\mathfrak{R}} = \Delta^{-\mathbb{N}_0}/\mathfrak{R}$ is as a measurable space isomorphic to the image $\mathrm{Im}(h) \subseteq \Gamma$ of h. The canonical isomorphism ι satisfies $h = \iota \circ [\,\cdot\,]_{\mathfrak{R}}$, and thus considering the memory kernel γ_h induced by h is equivalent to considering the memory kernel $\gamma_{\mathfrak{R}}$ induced by $\mathfrak{R}$. Also note that $\mathfrak{R}$ is countably generated because Γ is a Souslin space. On the other hand, every $\mathfrak{R} \subseteq \mathfrak{B}(\Delta^{-\mathbb{N}_0})$ is generated by a measurable function, namely $\mathfrak{R} = \sigma\big([\,\cdot\,]_{\mathfrak{R}}\big)$. Thus analysing deterministic memories amounts to the same thing as analysing countably generated sub-σ-algebras of $\Delta^{-\mathbb{N}_0}$.

The set of atoms of a sub-σ-algebra $\mathfrak{R} \subseteq \mathfrak{B}(\Delta^{-\mathbb{N}_0})$ is a partition of $\Delta^{-\mathbb{N}_0}$. If $\mathfrak{R}$ is countably generated, the partition is measurable in the sense that the partition elements are measurable subsets of $\Delta^{-\mathbb{N}_0}$. In [SC01], measurable partitions were used instead of σ-algebras. Mainly countable partitions, i.e. partitions into countably many sets, were considered, and in this case the partition uniquely determines the corresponding σ-algebra. More generally, however, there may be many σ-algebras with the same set of atoms. Luckily, it turns out that at most one of them is countably generated. Indeed, according to the Blackwell theorem ([Coh80, Thm. 8.6.7]), every countably generated sub-σ-algebra $\mathfrak{R}$ of the Souslin space $\Delta^{-\mathbb{N}_0}$ consists of all Borel measurable unions of its atoms, i.e.

$$\mathfrak{R} \;=\; \Big\{ A \subseteq \Delta^{-\mathbb{N}_0} \;\Big|\; A \in \mathfrak{B}(\Delta^{-\mathbb{N}_0}),\; A = \bigcup_{x \in A} [x]_{\mathfrak{R}} \Big\}. \tag{3.3}$$

In particular, the σ-algebra is uniquely determined by the partition given by its atoms, together with the fact that it is countably generated. Therefore, we can specify a deterministic memory by defining a (not necessarily countable) measurable partition of $\Delta^{-\mathbb{N}_0}$. But we still have to make sure that the σ-algebra defined by (3.3) (with $[x]_{\mathfrak{R}}$ replaced by the partition element containing x) is countably generated.

Remark. Every measurable partition of $\Delta^{-\mathbb{N}_0}$ defines a sub-σ-algebra $\mathfrak{R}$ of $\mathfrak{B}(\Delta^{-\mathbb{N}_0})$ by (3.3). This σ-algebra is not the one generated by the partition elements. It is the largest (rather than the smallest) sub-σ-algebra with the given atoms. Note that $\mathfrak{R}$ is not countably generated in general. Even for $\Delta = \{\,0, 1\,\}$, there is a measurable partition of $\Delta^{-\mathbb{N}_0}$ that is not the set of atoms of any countably generated sub-σ-algebra of $\mathfrak{B}(\Delta^{-\mathbb{N}_0})$ (see Appendix A.4).

3.2.3 Minimal sufficient memory: Causal states

It is natural to ask how big a sufficient memory has to be and how to obtain a minimal one. There are mainly two possibilities to measure the size of a memory: cardinality $|\Gamma|$ of the set of memory states and Shannon entropy $H^{\mathbb{P}}(M)$ of the memory variable. But unlike in the situation of generative HMMs, both notions of size yield the same notion of minimality and the minimal sufficient memory is essentially unique. Furthermore, it turns out to be deterministic with partition given by the following equivalence relation. Two past trajectories, $x, \hat{x} \in \Delta^{-\mathbb{N}_0}$, are identified if they induce the same conditional probability on the future, i.e.

$$x \sim \hat{x} \quad :\Leftrightarrow \quad \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = x) \;=\; \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = \hat{x}). \tag{3.4}$$

The corresponding equivalence classes are called causal states ([CY89, SC01]).

Definition 3.17. Let $X_{\mathbb{Z}}$ be a stationary stochastic process. The equivalence classes $\mathfrak{C}(x) := \{\, \hat{x} \in X_{-\mathbb{N}_0} \mid \hat{x} \sim x \,\}$ of the relation defined in (3.4) are called **causal states** of $X_{\mathbb{Z}}$. The set

$$\Gamma_{\mathfrak{C}} \;:=\; \mathrm{Im}(\mathfrak{C}) \;=\; \big\{\, \mathfrak{C}(x) \;\big|\; x \in \Delta^{-\mathbb{N}_0} \,\big\}$$

of causal states is equipped with the σ-algebra $\mathcal{G}_{\mathfrak{C}} := \big\{\, A \subseteq \Gamma_{\mathfrak{C}} \;\big|\; \bigcup A \in \mathfrak{B}(\Delta^{-\mathbb{N}_0}) \,\big\}$.

Remark. It is important to note that the causal states depend on the version of conditional probability. As always, we assume that a regular version is chosen. We say that the number of causal states is countable (respectively finite) if there exists a version of conditional probability such that there are only countably (resp. finitely) many equivalence classes. In this case, we assume that the version is chosen such that $\Gamma_{\mathfrak{C}}$ is countable (resp. finite). Note that every other version yields countably many causal states with non-zero probability and the total mass of the remaining ones is zero.

In the following lemma, we show that the causal states induce a deterministic memory in the sense of Section 3.2.2. It is obvious that they partition the space $\Delta^{-\mathbb{N}_0}$, but we have to prove that this partition is measurable, and that the corresponding σ-algebra $\mathfrak{R}_{\mathfrak{C}} \subseteq \mathfrak{B}(\Delta^{-\mathbb{N}_0})$ defined by (3.3) is countably generated. We denote the induced memory kernel $x \mapsto \delta_{\mathfrak{C}(x)}$ by $\gamma_{\mathfrak{C}}$ and the corresponding memory variable by $M_{\mathfrak{C}} = \mathfrak{C} \circ X_{-\mathbb{N}_0}$.

Lemma 3.18. *The causal states are measurable sets and the atoms of a unique countably generated sub-σ-algebra of $\mathfrak{B}(\Delta^{-\mathbb{N}_0})$. In particular, $\Gamma_{\mathfrak{C}}$ is a Souslin measurable space. Furthermore, the induced deterministic memory is sufficient.*

Proof. $f := \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = \cdot\,)$ is a measurable map from $\Delta^{-\mathbb{N}_0}$ into $\mathcal{P}(\Delta^{\mathbb{N}})$. Because the singletons in $\mathcal{P}(\Delta^{\mathbb{N}})$ are measurable, $\mathfrak{C}(x) = f^{-1}\big(f(x)\big)$ is a measurable set for every $x \in \Delta^{-\mathbb{N}_0}$, and $\Gamma_{\mathfrak{C}}$ is the set of atoms of $\sigma(f)$. Because the σ-algebra of $\mathcal{P}(\Delta^{\mathbb{N}})$ is countably generated, $\sigma(f)$ is countably generated as well and according to Blackwell's theorem (see Section 3.2.2), there can be no other countably generated σ-algebra with the same set of atoms. $\Gamma_{\mathfrak{C}}$ is a Souslin measurable space according to Proposition 3.16. Proposition 3.10 yields sufficiency of the induced memory, because $\sigma(M_{\mathfrak{C}}) = \sigma(\mathfrak{C} \circ X_{-\mathbb{N}_0}) = \sigma(f \circ X_{-\mathbb{N}_0}) = \sigma(\mathbb{P}^{X_{-\mathbb{N}_0}}_{X_{\mathbb{N}}})$. □

Definition 3.19. We call the countably generated σ-algebra with set $\Gamma_{\mathfrak{C}}$ of atoms **causal state σ-algebra** and denote it by $\mathfrak{R}_{\mathfrak{C}}$. The induced deterministic memory with memory kernel $\gamma_{\mathfrak{C}}$ and memory variable $M_{\mathfrak{C}}$ is called **causal state memory**.

One of the basic facts of computational mechanics is that the causal state memory is the unique minimal sufficient deterministic memory ([SC01]). This property can easily be extended to our more general measure-theoretic setting and non-deterministic sufficient memories. If two histories $x, y \in \Delta^{-\mathbb{N}_0}$ are in different causal states, every sufficient memory γ assigns orthogonal probability measures to them, i.e. $\gamma(x) \perp \gamma(y)$.[2] Thus the causal state can a.s. be recovered from the value of the memory variable.

Proposition 3.20. *Let γ be a memory kernel with set Γ of memory states. γ is sufficient if and only if there exist disjoint measurable subsets $\Lambda_c \subseteq \Gamma$, $c \in \Gamma_{\mathfrak{C}}$, such that*

$$\gamma(x;\, \Lambda_{\mathfrak{C}(x)}) \;=\; 1 \qquad \text{a.s.}$$

Proof. Let M be the memory variable induced by γ.
"if": Let $f(g) = c$ if $g \in \Lambda_c$. Then $M_{\mathfrak{C}} = \mathfrak{C} \circ X_{-\mathbb{N}_0} = f \circ M$ a.s. and sufficiency of $M_{\mathfrak{C}}$ yields sufficiency of M.
"only if": Define

$$\Lambda_{\mathfrak{C}(x)} \;:=\; \big\{\, g \in \Gamma \;\big|\; \mathbb{P}(X_{\mathbb{N}} \mid M = g) \;=\; \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = x) \,\big\}.$$

By definition of the causal states, the sets Λ_c, $c \in \Gamma_{\mathfrak{C}}$, are well defined and disjoint. They are obviously measurable and we obtain a.s.

$$\begin{aligned} \gamma(x;\, \Lambda_{\mathfrak{C}(x)}) \;&=\; P\big(\{\, M \in \Lambda_{\mathfrak{C}(x)} \,\} \;\big|\; X_{-\mathbb{N}_0} = x\big) \\ &=\; P\Big(\big\{\, \mathbb{P}(X_{\mathbb{N}} \mid M) = \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \,\big\} \;\Big|\; X_{-\mathbb{N}_0} = x\Big) \;=\; 1. \end{aligned}$$

The last equality holds due to sufficiency of M. □

Corollary 3.21. *The causal state memory is the minimal sufficient memory in the sense that for every sufficient memory with memory variable M and set Γ of memory states both*

$$|\Gamma| \;\geq\; |\Gamma_{\mathfrak{C}}| \qquad \text{and} \qquad H^{\mathbb{P}}(M) \;\geq\; H^{\mathbb{P}}(M_{\mathfrak{C}}).$$

Corollary 3.22. *A deterministic memory $\gamma_{\mathfrak{R}}$, specified by a countably generated σ-algebra $\mathfrak{R} \subseteq \mathfrak{B}(\Delta^{-\mathbb{N}_0})$, is sufficient if and only if $\mathfrak{R} \supseteq \mathfrak{R}_{\mathfrak{C}}$ modulo $\mathbb{P}_{X_{-\mathbb{N}_0}}$.*

Due to the minimality of the causal states, their entropy is an important complexity measure called statistical complexity. We analyse its properties in more detail in Section 4.3.

Definition 3.23. Let $X_{\mathbb{Z}}$ be a stationary process with distribution $P \in \mathcal{P}_{\mathfrak{s}}(\Delta^{\mathbb{Z}})$. The entropy of the causal states,

$$C_{\mathfrak{C}}(P) \;:=\; C_{\mathfrak{C}}(X_{\mathbb{Z}}) \;:=\; H^{\mathbb{P}}(M_{\mathfrak{C}}),$$

is called **statistical complexity** of $X_{\mathbb{Z}}$ or of P.

Remark. Because $M_{\mathfrak{C}}$ is a sufficient memory, $I(M_{\mathfrak{C}} : X_{\mathbb{N}}) = E(X_{\mathbb{Z}})$. In particular, the statistical complexity is lower bounded by the excess entropy. That is,

$$C_{\mathfrak{C}}(X_{\mathbb{Z}}) \;\geq\; E(X_{\mathbb{Z}}).$$

In the following section, we see that the difference $C_{\mathfrak{C}}(X_{\mathbb{Z}}) - E(X_{\mathbb{Z}})$, which is defined for $E(X_{\mathbb{Z}}) < \infty$, can be arbitrarily large.

[2] $\mu \perp \nu$ means that there is a measurable set A with $\mu(A) = 1$ and $\nu(A) = 0$

3.2.4 The ε-machine and its non-minimality

Besides the causal states, the second central concept of computational mechanics is the so-called ε-machine. It is the generative HMM induced by the sufficient memory $\gamma_{\mathfrak{C}}$ of causal states and encodes the mechanisms of prediction.

Definition 3.24. The HMM $(T^{\gamma_{\mathfrak{C}}}, \mu_{\gamma_{\mathfrak{C}}})$ induced by the causal state memory is called **ε-machine**.

The following nice properties of the ε-machines and the causal states were obtained in [SC01] for countable state spaces Δ and countably many causal states. Now we can generalise them to Souslin spaces Δ and possibly uncountably many causal states.

Proposition 3.25. *The ε-machine is partially deterministic and state observable. Its internal process has the same distribution as the process $(M_{\mathfrak{C}})_{\mathbb{Z}}$ of causal states. In particular, the process of causal states is Markovian.*

Proof. We have to show that the equivalent properties of Proposition 3.14 are satisfied. To obtain that $(M_{\mathfrak{C}})_1$ is a function of $(M_{\mathfrak{C}})_0$ and X_1, it is sufficient to show that $\mathbb{P}(X_{[2,\infty[} \mid X_{-\mathbb{N}_0}, X_1)$ is a function of $\mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0})$ and X_1. This follows from the fact that with $\mathbb{P}^{X_{-\mathbb{N}_0}} := \mathbb{P}(\,\cdot \mid X_{-\mathbb{N}_0})$ the equality $\mathbb{P}(\,\cdot \mid X_{-\mathbb{N}_0}, X_1)(\omega) = \big(\mathbb{P}^{X_{-\mathbb{N}_0}}(\omega)\big)(\,\cdot \mid X_1)(\omega)$ is a.s. satisfied (see Appendix A.3). In other words, we obtain the conditional probability given $X_{-\mathbb{N}_0}$ and X_1 by first conditioning on $X_{-\mathbb{N}_0}$ and then conditioning the resulting probability measure on X_1. □

The causal states provide the minimal sufficient memory and induce the ε-machine. But is the latter also the minimal generative HMM? In general, the answer is "no". The ε-machine may be arbitrarily much bigger than the minimal HMM. It can be infinite or even uncountable, while there is a generative HMM with only two internal states. This was already mentioned (although not rigorously proven) by Crutchfield in [Cru94], but not everyone who applies computational mechanics seems to be aware of the fact. In the following, we give two examples of this phenomenon. In the first one, the set of causal states is uncountable. In the second one, we demonstrate that restricting to processes with countably many causal states and finite statistical complexity does not solve the problem.

Example 3.26 (uncountable ε-machine). The following HMM (T, μ) with $\Delta := \Gamma := \{\,0, 1\,\}$ turns out to have uncountably many causal states and thus an uncountable ε-machine. Let μ be the uniform distribution on Γ. With a parameter $0 < p < \frac{1}{4}$, we define the generator $T\colon \Gamma \to \mathcal{P}(\Delta \times \Gamma)$ by

$$T(g;\, d, \hat{g}) \;:=\; \begin{cases} 1 - 2p & \text{if } \hat{g} = d = g, \\ p & \text{if } d \neq g, \\ 0 & \text{otherwise.} \end{cases}$$

See Figure 3.2 for an illustration of the transition graph. It is easy to check that μ is T-invariant. Recall that the internal operator $L_d\colon \mathcal{P}(\Gamma) \to \mathcal{P}(\Gamma)$ (Definition 2.12) describes the update of the knowledge about the internal state when the output symbol $d \in \Delta$ is observed. We parametrise $\mathcal{P}(\Gamma)$ by the unit interval with $\iota(y) := y\delta_1 + (1-y)\delta_0$, $y \in [0,1]$, and obtain two "update functions" $f_d\colon [0,1] \to [0,1]$ by

$$f_d(y) \;:=\; \iota^{-1} \circ L_d \circ \iota(y) \;=\; L_d\big(y\delta_1 + (1-y)\delta_0\big)(1), \qquad d \in \Delta,\ y \in [0,1].$$

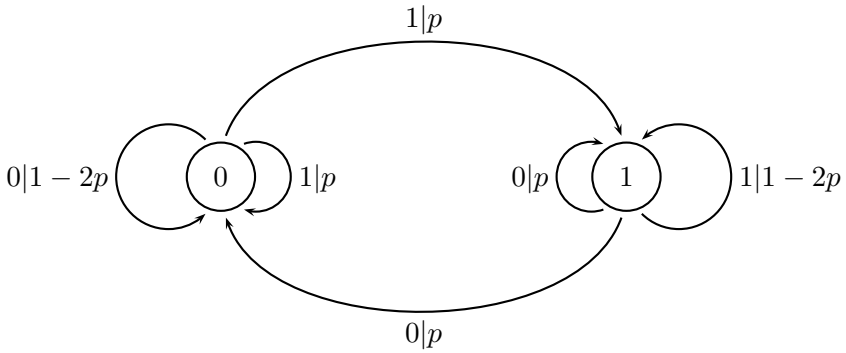


Figure 3.2: HMM with uncountably many causal states used in Example 3.26.

Then, by Lemma 2.13, we obtain for $n \in \mathbb{N}$ and $x = (x_1, \dots, x_n) \in \Delta^n$

$$\mathbb{P}\big(\{\, W_0 = 1 \,\} \bigm| X_{[-n+1,0]} = x\big) \;=\; f_{x_n} \circ \cdots \circ f_{x_1}\big(\mu(1)\big) \;=\; f_{x_n} \circ \cdots \circ f_{x_1}(\tfrac{1}{2}).$$

Because $\mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) = \lim_{n\to\infty} \mathbb{P}(X_{\mathbb{N}} \mid X_{[-n+1,0]})$ a.s. and different internal states lead to different expectations of $X_{\mathbb{N}}$, we obtain on a set $A \subseteq \Delta^{-\mathbb{N}_0}$ of measure one that

$$\mathfrak{C}(x) \;=\; \mathfrak{C}(\hat{x}) \quad\Leftrightarrow\quad \lim_{n\to\infty} f_{x_n} \circ \cdots \circ f_{x_1}(\tfrac{1}{2}) \;=\; \lim_{n\to\infty} f_{\hat{x}_n} \circ \cdots \circ f_{\hat{x}_1}(\tfrac{1}{2})$$

The definitions of f_d and T yield

$$f_0(y) \;=\; \frac{yp}{1-2p+y(4p-1)} \qquad \text{and} \qquad f_1(y) \;=\; \frac{p+y(1-3p)}{2p+y(1-4p)}$$

and we observe that both f_0 and f_1 are strictly increasing. Further, $f_0\big([0,1]\big) = [0, \frac{1}{2}]$ and $f_1\big([0,1]\big) = [\frac{1}{2}, 1]$. Thus, $A \cap \mathfrak{C}(x)$ can contain at most two histories (if $y \in \mathfrak{C}(x) \cap A$ and $y \neq x$, then there is an $n \in \mathbb{N}$ such that $x_k = y_k$ for $k > -n$ and $x_k = y_{-n} = 1 - x_{-n} = 1 - y_k$ for all $k < -n$). In particular, the number of causal states is uncountable for every version of conditional probability. Also note that the statistical complexity $C_{\mathfrak{C}}(X_{\mathbb{Z}})$ is infinite, while the excess entropy $E(X_{\mathbb{Z}})$ is bounded by $H(\mu) = \log(2)$. ◊

In the second example, we use the following two simple technical lemmata.

Lemma 3.27. *Let Γ be a Souslin space and $\mu \in \mathcal{P}(\Gamma)$. Then*

$$H(\mu) \;\geq\; -\log\big(\sup_{g\in\Gamma} \mu(\{\, g \,\})\big)$$

Lemma 3.28. *Let $f\colon \mathbb{R} \to \mathbb{R}$ be strictly decreasing and $x \in \mathbb{R}$ with $x < f^2(x) < f(x)$. Then the points $f^k(x)$, $k \in \mathbb{N}_0$, are distinct.*

Proof. By induction we obtain that $f^{2n}(x)$ is strictly increasing in n, $f^{2n+1}(x)$ is strictly decreasing and $f^{2n}(x) < f^{2k+1}(x)$ for all $n, k \in \mathbb{N}_0$. In particular, the $f^k(x)$ are distinct. □

Example 3.29. In this example, we show that HMMs with arbitrarily small internal state entropy can generate processes with arbitrarily high statistical complexity, even if the latter

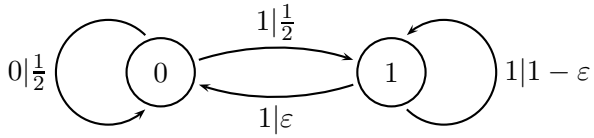


Figure 3.3: HMM of Example 3.29

one is assumed to be finite. Consider the following HMM $(T_\varepsilon, \mu_\varepsilon)$ with parameter $0 < \varepsilon < 1$ and $\Gamma := \Delta := \{0, 1\}$.

$$T_\varepsilon(0) \;:=\; \tfrac{1}{2}(\delta_{(0,0)} + \delta_{(1,1)}), \qquad T_\varepsilon(1) \;:=\; (1-\varepsilon)\delta_{(1,1)} + \varepsilon\delta_{(1,0)},$$

as is illustrated in Figure 3.3. The stationary probability μ_ε is given by $\mu_\varepsilon(0) = \frac{2\varepsilon}{1+2\varepsilon}$. We denote the output process by $X^\varepsilon_{\mathbb{Z}}$ and see from $\mu_\varepsilon(0) \stackrel{\varepsilon\to 0}{\longrightarrow} 0$ that the internal state entropy, the excess entropy and the entropy rate all tend to zero,

$$E(X^\varepsilon_{\mathbb{Z}}) \;\le\; H(\mu_\varepsilon) \stackrel{\varepsilon\to 0}{\longrightarrow} 0 \qquad \text{and} \qquad h(X^\varepsilon_{\mathbb{Z}}) \;\le\; H^{\mathbb{P}}(X^\varepsilon_1) \stackrel{\varepsilon\to 0}{\longrightarrow} 0.$$

Now we determine the causal states of $X^\varepsilon_{\mathbb{Z}}$. It is evident that output symbols preceding the last occurrence of 0 do not influence the prediction (because the internal state is known to be 0 at that time). Thus causal states are unions of sets $[01^k] := \{\, x \in \Delta^{-\mathbb{N}_0} \mid x_{-k} = 0,\ x_l = 1 \text{ for } l = -k+1, \ldots, 0 \,\}$ with $k \in \mathbb{N}_0$. We now claim that no further identification is possible, i.e. the causal states are precisely the sets $[01^k]$. To see this, consider the update function $f_\varepsilon = f_{\varepsilon,1}$ corresponding to L_1 like in Example 3.26. Then

$$f_\varepsilon(x) \;=\; \frac{x(1-2\varepsilon)+1}{x+1}$$

and we obtain

$$\mathbb{P}\big(\{\, W^\varepsilon_1 = 1 \,\} \;\big|\; X^\varepsilon_{-\mathbb{N}_0} \in [01^k]\big) \;=\; f^k_\varepsilon(0).$$

Observe that the function f_ε is strictly decreasing and $0 < f^2_\varepsilon(0) = f_\varepsilon(1) = 1-\varepsilon < 1 = f_\varepsilon(0)$, hence the $f^k_\varepsilon(0)$ are distinct. Therefore, the causal states are precisely the sets $[01^k]$ and

$$C_{\mathfrak{C}}(X^\varepsilon_{\mathbb{Z}}) \;\ge\; -\log\Big(\sup_{k\in\mathbb{N}_0} \mathbb{P}\big(\{\, X^\varepsilon_{-\mathbb{N}_0} \in [01^k] \,\}\big)\Big) \;=\; -\log\big(\mathbb{P}\big(\{\, X^\varepsilon_0 = 0 \,\}\big)\big) \;=\; -\log\big(\tfrac{1}{2}\mu_\varepsilon(0)\big).$$

Consequently, $C_{\mathfrak{C}}(X^\varepsilon_{\mathbb{Z}}) \stackrel{\varepsilon\to 0}{\longrightarrow} \infty$ because $\mu_\varepsilon(0)$ tends to zero. ◊

In Section 3.3.3 below, we see that the ε-machine actually has some minimality properties. Namely, it is minimal in the sub-class of partially deterministic HMMs.

3.2.5 Finite-history computational mechanics

Usually, computational mechanics considers (in theory) past trajectories of infinite length. We consider this viewpoint appropriate for theoretical investigations and the definition of complexity measures. Therefore, the previous and subsequent parts deal only with the infinite-history case. In this section, however, we briefly consider the case of finite, varying observation lengths. This case has also been considered in [FC98a]. We ask the following questions. How can we define sufficient memories and causal states when the observations have arbitrary but only finite length? Is the entropy of a finite-history version of the causal states a good approximation of statistical complexity?

A **finite-history memory kernel** γ assigns to every history $x \in \Delta^n$ of arbitrary but finite length n a probability distribution on the Souslin space Γ of memory states.[3] More precisely,

$$\gamma\colon \Delta^* \;\to\; \mathcal{P}(\Gamma), \qquad \text{where } \Delta^* \;:=\; \bigcup_{n\in\mathbb{N}_0} \Delta^n.$$

Note that Δ^* contains the "empty history," which corresponds to not having observed anything. Because the length of the observed history may vary, a finite-history memory does not only induce a single memory variable at time zero but rather for any history length n a different memory variable M^n with conditional distribution

$$\mathbb{P}(M^n \mid X_{[-n+1,0]}) \;=\; \gamma \circ X_{[-n+1,0]}.$$

We now want to define sufficiency for finite-history memories. Simply assuming the conditional independence property for every length n separately, i.e.

$$X_{\mathbb{N}} \perp\!\!\!\perp X_{[-n+1,0]} \mid M^n \qquad \forall n \in \mathbb{N}, \tag{3.5}$$

is a weak requirement and does not provide the correct definition of sufficiency in the context of finite but varying observation lengths. If a memory kernel satisfies (3.5), the information about the future need not be contained in the memory state alone but may require knowledge of the particular observation length n. The same memory state g can have a completely different implication on the future if it results from different history lengths (see Example 3.31). Therefore, we have to assume that the memory keeps all information about the future without the additional knowledge of n. To this end, imagine that n is determined randomly by an $\mathbb{N}$-valued random variable τ that is assumed to be independent of all other variables. We call such a variable τ **random time**. Combining the family of memory variables M^n, $n \in \mathbb{N}$, with a random time τ we get a new variable M^τ with $M^\tau(\omega) := M^{\tau(\omega)}(\omega)$. Similarly, a past of random length τ is given by the Δ^*-valued random variable $X_{[-\tau+1,0]}$ defined by $X_{[-\tau+1,0]}(\omega) = X_{[-\tau(\omega)+1,0]}(\omega)$. We require that, for all random times τ, the corresponding M^τ contains maximal information about the future, even if the value of τ is not known.

Definition 3.30. We call a finite-history memory **sufficient** if it satisfies

$$X_{\mathbb{N}} \perp\!\!\!\perp X_{[-\tau+1,0]} \mid M^\tau \qquad \forall \text{ random times } \tau.$$

Note that a sufficient memory satisfies (3.5), because the random time can be constant. We illustrate the difference between (3.5) and sufficiency by the following example.

Example 3.31 (why random times?). Let $X_{\mathbb{Z}}$ be a non-i.i.d. Markov process on $\Delta := \{0, 1\}$. Define

$$M^n \;:=\; X_0 \qquad \text{and} \qquad \widehat{M}^n \;:=\; \begin{cases} X_0 & \text{if } n \text{ odd,} \\ 1 - X_0 & \text{if } n \text{ even.} \end{cases}$$

Then both $M = (M^n)_{n\in\mathbb{N}_0}$ and $\widehat{M} = (\widehat{M}^n)_{n\in\mathbb{N}_0}$ are obviously induced by finite-history memories and satisfy (3.5). M is also sufficient but $\widehat{M}$ is not, because the information $\widehat{M}^\tau = g$ is useless if we do not know whether τ is odd or even. $\Diamond$

[3] For simplicity of notation we write $x \in \Delta^n$ instead of $x \in \Delta^{[-n+1,0]}$. The symbol x is reserved for histories in this section.

In the following lemma, we provide a characterisation of sufficiency. Instead of using random times, our equivalent condition explicitly requires that the conditional probabilities of the future given a particular memory state are the same for different history lengths. More precisely, $\mathbb{P}(X_{\mathbb{N}} \mid M^n = g) = \mathbb{P}(X_{\mathbb{N}} \mid M^t = g) =: \Psi(g)$ for appropriately chosen versions of conditional probability and all $n, t \in \mathbb{N}$, $g \in \Gamma$.

Proposition 3.32. *A finite-history memory is sufficient if and only if there is a kernel Ψ from Γ to $\Delta^{\mathbb{N}}$ with*

$$\mathbb{P}(X_{\mathbb{N}} \mid X_{[-n+1,0]}) \;=\; \Psi \circ M^n \qquad \forall n \in \mathbb{N} \;\; a.s.$$

In this case, $\Psi \circ M^n = \mathbb{P}(X_{\mathbb{N}} \mid M^n)$ a.s. for all n.

Proof. We use that sufficiency is equivalent to $\mathbb{P}(X_{\mathbb{N}} \mid X_{[-\tau+1,0]}) = \mathbb{P}(X_{\mathbb{N}} \mid M^\tau)$ a.s. for all random times τ. Note that τ is a function of $X_{[-\tau+1,0]}$, because $\tau = n$ if and only if $X_{[-\tau+1,0]} \in \Delta^n$.
"if": Fix a random time τ. Using the assumption and that τ is independent of $X_{\mathbb{Z}}$, we obtain

$$\begin{aligned}\mathbb{P}(X_{\mathbb{N}} \mid X_{[-\tau+1,0]}) &= \mathbb{P}(X_{\mathbb{N}} \mid X_{[-\tau+1,0]}, \tau) \;=\; \sum_{n\in\mathbb{N}} 1_{\{\tau=n\}} \mathbb{P}(X_{\mathbb{N}} \mid X_{[-n+1,0]}) \\ &= \sum_{n\in\mathbb{N}} 1_{\{\tau=n\}} \Psi \circ M^n \;=\; \Psi \circ M^\tau.\end{aligned}$$

Now we prove that $\Psi \circ M^\tau$ is a version of the conditional probability $\mathbb{P}(X_{\mathbb{N}} \mid M^\tau)$. Indeed, the $\sigma(M^\tau)$-measurability is obvious and for $G \in \mathcal{G}$ we obtain

$$\begin{aligned}\int_{\{M^\tau \in G\}} \Psi \circ M^\tau \, \mathrm{d}\mathbb{P} &= \sum_{n\in\mathbb{N}} \int_{\{\tau=n,\, M^n \in G\}} \Psi \circ M^n \, \mathrm{d}\mathbb{P} \\ &= \sum_{n\in\mathbb{N}} \int_{\{\tau=n,\, M^n \in G\}} \mathbb{P}(X_{\mathbb{N}} \mid X_{[-n+1,0]}) \, \mathrm{d}\mathbb{P}\end{aligned}$$

using $\tau, M^n \perp\!\!\!\perp X_{\mathbb{N}} \mid X_{[-n+1,0]}$ we continue

$$\begin{aligned}&= \sum_{n\in\mathbb{N}} \mathbb{P}\big(\{\, X_{\mathbb{N}} \in \cdot \,\} \cap \{\, \tau = n,\; M^n \in G \,\}\big) \\ &= \mathbb{P}\big(\{\, X_{\mathbb{N}} \in \cdot \,\} \cap \{\, M^\tau \in G \,\}\big).\end{aligned}$$

"only if": Assume that the memory is sufficient. Choose a random time τ with $\mathbb{P}\big(\{\, \tau = n \,\}\big) > 0$ for all $n \in \mathbb{N}$. Define $\Psi := \mathbb{P}(X_{\mathbb{N}} \mid M^\tau = \cdot)$. Due to sufficiency and $\mathbb{P}(X_{\mathbb{N}} \mid X_{[-\tau+1,0]}) = \mathbb{P}(X_{\mathbb{N}} \mid X_{[-\tau+1,0]}, \tau)$, we also obtain $\mathbb{P}(X_{\mathbb{N}} \mid M^\tau) = \mathbb{P}(X_{\mathbb{N}} \mid M^\tau, \tau)$ a.s. Thus, for all $n \in \mathbb{N}$ and $\mathbb{P}_{M^n}$-almost all $g \in \Gamma$,

$$\mathbb{P}(X_{\mathbb{N}} \mid M^n = g) \;=\; \mathbb{P}(X_{\mathbb{N}} \mid M^\tau = g,\, \tau = n) \;=\; \mathbb{P}(X_{\mathbb{N}} \mid M^\tau = g).$$

Sufficiency implies in particular that $\mathbb{P}(X_{\mathbb{N}} \mid M^n) = \mathbb{P}(X_{\mathbb{N}} \mid X_{[-n+1,0]})$. Thus

$$\mathbb{P}(X_{\mathbb{N}} \mid X_{[-n+1,0]}) \;=\; \mathbb{P}(X_{\mathbb{N}} \mid M^n) \;=\; \Psi \circ M^n \qquad \text{a.s.}$$

□

The finite-history causal states are defined analogously to the infinite-history case as a partition of Δ^*. The identified pasts may have different lengths.

Definition 3.33. For $x \in \Delta^*$, let $l(x) = n$ if $x \in \Delta^n$. Define the equivalence relation $\sim$ on Δ^* by

$$x \sim \hat{x} \quad \Leftrightarrow \quad \mathbb{P}(X_{\mathbb{N}} \mid X_{[-l(x)+1,0]} = x) \;=\; P(X_{\mathbb{N}} \mid X_{[-l(\hat{x})+1,0]} = \hat{x}).$$

Then the equivalence classes of $\sim$ are the **finite-history causal states**. The corresponding finite-history memory variables are denoted by $M_{\mathfrak{C}}^n$.

Let Δ be countable. Then, in contrast to the infinite-history case, the set of finite-history causal states is always countable. But, as we see in the following example, it is not true that the finite-history causal states are always "less" than the causal states. Furthermore, the "length n statistical complexity" $H^{\mathbb{P}}(M_{\mathfrak{C}}^n)$ is not always a good approximation of $C_{\mathfrak{C}}(X_{\mathbb{Z}})$. In a sense, the transition from finite to infinite history lengths is discontinuous at infinity.

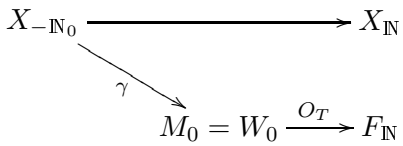
Example 3.34. Let $\Delta := \{0, 1\}$ and for $p \in [0, 1]$ let $P_p \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ be the i.i.d. process with $P_p\big([1]\big) = p$. Define $P := \frac{1}{2}P_p + \frac{1}{2}P_q$ with $0 < p < q < 1$. Then there are two causal states corresponding to the two ergodic components, and the statistical complexity is $C_{\mathfrak{C}}(P) = \log(2)$. Finite histories of the same length, however, are only identified if they have the same number of ones. Therefore, the number of finite-history causal states is infinite. It is straight-forward to see that also $H^{\mathbb{P}}(M_{\mathfrak{C}}^n)$ tends to infinity, roughly like $\log(n)$. $\Diamond$

3.3 The generative nature of prediction

3.3.1 Predictive interpretation of HMMs

We have seen that there can be a huge discrepancy between the minimal sufficient memory and the minimal generative HMM. The requirement of sufficiency is based on a certain understanding of "prediction". Here, we propose an alternative, weaker notion of prediction that allows for a predictive interpretation of all HMMs.

For this interpretation, we model prediction by two steps. First, the past $X_{-\mathbb{N}_0}$ is processed by a memory kernel γ, like in Section 3.2.1 but without the sufficiency assumption. Then the actual prediction is done by generating a predicted future $F_{\mathbb{N}}$. To this end, we assume a generator $T\colon \Gamma \to \mathcal{P}(\Delta \times \Gamma)$, which uses the space Γ of memory states as internal state space. It is initialized by the random memory state M_0 produced by γ. Thus the (non-invariant) HMM $\big(T, \gamma(X_{-\mathbb{N}_0})\big)$ with random initial distribution $\gamma \circ X_{-\mathbb{N}_0}(\omega)$ generates the prediction $F_{\mathbb{N}}$. The situation is illustrated as



where O_T is the kernel from Γ to $\Delta^{\mathbb{N}}$ associating to an initial state g the output distribution of the HMM (T, δ_g) (see Section 2.1.4). Of course, the generated future $F_{\mathbb{N}}$ and the corresponding process $W_{\mathbb{N}_0}$ of internal states is conditionally independent of the real future $X_{\mathbb{N}}$ given the past $X_{-\mathbb{N}_0}$. Thus, we cannot expect the prediction $F_{\mathbb{N}}$ and the future $X_{\mathbb{N}}$ to coincide. But we require that the distributions, conditioned on the known past $X_{-\mathbb{N}_0}$, are identical. This is the

best one can possibly do and means that actual and predicted future cannot be distinguished statistically, based on the observed past.

Definition 3.35. The pair (γ, T) is called **predictive model** of $X_{\mathbb{Z}}$ if Γ is a Souslin space, $\gamma\colon \Delta^{-\mathbb{N}_0} \to \mathcal{P}(\Gamma)$ and $T\colon \Gamma \to \mathcal{P}(\Delta \times \Gamma)$ are measurable, and the process $F_{\mathbb{N}}$ generated by the HMM $\big(T, \gamma(X_{-\mathbb{N}_0})\big)$ satisfies

$$\mathbb{P}(F_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \;=\; \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \qquad \text{a.s.}$$

A memory kernel γ (resp. generator T) is called **predictive** if there exists a generator T (resp. memory γ) such that (γ, T) is a predictive model.

We already know from Proposition 3.12 that sufficient memory kernels induce generative HMMs. In our new terminology, Proposition 3.12 states that every sufficient memory is predictive. As we see in the following, the converse is not true. In fact, predictive memories can be much smaller than any sufficient memory. The following proposition states that the generator of any invariant HMM is predictive, and we know from Section 3.2.4 that generative HMMs can (for some processes) do with fewer internal states and less internal state entropy than sufficient memories.

Proposition 3.36 (generative HMMs are predictive). *Let (T, μ) be an invariant HMM of $X_{\mathbb{Z}}$. Then T is predictive, i.e. there is a memory kernel γ_T, such that (γ_T, T) is a predictive model of $X_{\mathbb{Z}}$. More specifically, we can choose*

$$\gamma_T(x) \;:=\; \mathbb{P}(W_0 \mid X_{-\mathbb{N}_0} = x), \qquad x \in \Delta^{-\mathbb{N}_0}.$$

Proof. We denote the internal processes of the HMM (T, μ) by $W_{\mathbb{Z}}$ and obtain

$$\begin{aligned} \mathbb{P}(F_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \;&=\; \int O_T \,\mathrm{d}\gamma_T(X_{-\mathbb{N}_0}) \;=\; \int \mathbb{P}(X_{\mathbb{N}} \mid W_0 = \,\cdot\,)\,\mathrm{d}\mathbb{P}(W_0 \mid X_{-\mathbb{N}_0}) \\ &=\; \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \qquad \text{a.s.} \end{aligned}$$

Thus (γ_T, T) is a predictive model. □

Definition 3.37. Let (T, μ) be an invariant HMM. Then the memory γ_T constructed in Proposition 3.36 is called **canonical memory kernel** of (T, μ).

Remark. The canonical memory kernel γ_T is nearly the same as the marginal Y_0 of the internal expectation process defined in Section 2.1.6. More precisely, $Y_0 = \gamma_T \circ X_{-\mathbb{N}_0}$.

The ε-machine $(T^{\gamma_{\mathfrak{C}}}, \mu_{\gamma_{\mathfrak{C}}})$ is the HMM induced by the causal state memory $\gamma_{\mathfrak{C}}$. The canonical memory $\gamma_{T^{\gamma_{\mathfrak{C}}}}$ of the ε-machine recovers the causal state memory, i.e. $\gamma_{T^{\gamma_{\mathfrak{C}}}} = \gamma_{\mathfrak{C}}$. This follows from the fact that the internal process of the ε-machine has the same distribution as the causal state process (Proposition 3.25). We emphasise that for a general sufficient memory kernel, the canonical memory of the induced HMM need not coincide with original memory. Assume, for instance, that the initial memory is deterministic but does not satisfy the equivalent conditions of Proposition 3.14 (see Example 3.15 for an example of such memories). Then the induced HMM is not state observable, in other words its canonical memory kernel is not deterministic. In particular, it is different from the original deterministic one.

While a given predictive memory kernel γ need neither be sufficient nor deterministic, it is closely related to a deterministic sufficient memory kernel γ' in the following way. Consider

the set $\Gamma' := \{\, \gamma(x) \mid x \in \Delta^{-\mathbb{N}_0} \,\} \subseteq \mathcal{P}(\Gamma)$ of so-called *information states* and define the Γ'-valued deterministic memory kernel $\gamma'(x) := \delta_{\gamma(x)}$. The associated memory variable satisfies $M' = \mathbb{P}(W_0 \mid X_{-\mathbb{N}_0})$ a.s., where $W_0 = M_0$ is the memory variable of γ. Then, using that $F_{\mathbb{N}}$ is conditionally independent of $X_{-\mathbb{N}_0}$ given W_0, we obtain

$$\mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \;=\; \mathbb{P}(F_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \overset{\text{(Lem. A.4)}}{=} \mathbb{P}\big(F_{\mathbb{N}} \;\big|\; \mathbb{P}(W_0 \mid X_{-\mathbb{N}_0})\big) \;=\; \mathbb{P}(X_{\mathbb{N}} \mid M').$$

In the last equality, we could replace $F_{\mathbb{N}}$ by $X_{\mathbb{N}}$ because $\sigma(M') \subseteq \sigma(X_{-\mathbb{N}_0})$. Thus, γ' is sufficient, and in particular the cardinality of Γ' is lower bounded by the number of causal states. Note that Γ' may have much more elements than Γ.

Remark. We have to point out that the above notion of predictive models does not capture all aspects of prediction.

a) A predictive memory may not allow for an iterative update of the memory state after observing additional output symbols.

b) Given a sufficient memory, the complete conditional future distribution corresponding to a past x is encoded in a single memory state $m \in \Gamma$. This is no longer the case for predictive memories. Assume that we want to use a predictive model for sampling the future distribution given the past x. We choose a memory state m according to $\gamma(x)$ and initialize T with m for generating a prediction. We repeat this procedure and obtain the correct future distribution $\mathbb{P}(F_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = x) = \mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0} = x)$. But if we "forget" the past x and, instead of sampling new memory states, initialize T always with the same m, the resulting distribution can be different. Thus, we have to memorize the *distribution* (the information state) $\gamma(x)$.

3.3.2 Generative complexity

Statistical complexity is a widely used complexity measure, and it is the entropy of the minimal sufficient memory. At the same time, it is also the internal state entropy of the ε-machine. Because the ε-machine is not the minimal generative HMM, and there is some predictive interpretation of these HMMs, we suggest to study the minimal internal entropy of a generative HMM, analogously to statistical complexity, as a complexity measure.

Definition 3.38. The **generative complexity** of a stationary stochastic process $X_{\mathbb{Z}}$ with distribution $P \in \mathcal{P}_{\mathbf{s}}(\Delta^{\mathbb{Z}})$ is

$$C_{\mathrm{HMM}}(X_{\mathbb{Z}}) \;:=\; C_{\mathrm{HMM}}(P) \;:=\; \inf\big\{\, H(\mu) \;\big|\; (T,\mu) \text{ is an invariant HMM of } P \,\big\}.$$

In contrast to the causal states, we do not have a constructive method to obtain the minimal generative HMM. Therefore, the generative complexity is difficult to compute. We know from Proposition 3.6 that it is lower bounded by excess entropy. Because the ε-machine is a generative HMM, it is upper bounded by statistical complexity, i.e.

$$E(X_{\mathbb{Z}}) \;\leq\; C_{\mathrm{HMM}}(X_{\mathbb{Z}}) \;\leq\; C_{\mathfrak{C}}(X_{\mathbb{Z}}).$$

Both inequalities can be strict. We saw in Examples 3.26 and 3.29 that $C_{\mathrm{HMM}}(X_{\mathbb{Z}}) < C_{\mathfrak{C}}(X_{\mathbb{Z}})$ is possible. We now show that the strict inequality $C_{\mathrm{HMM}}(X_{\mathbb{Z}}) < C_{\mathfrak{C}}(X_{\mathbb{Z}})$ also implies that the other inequality, $E(X_{\mathbb{Z}}) < C_{\mathrm{HMM}}(X_{\mathbb{Z}})$, is strict. For let $E(X_{\mathbb{Z}}) = C_{\mathrm{HMM}}(X_{\mathbb{Z}}) < \infty$

and (T, μ) an HMM of $X_{\mathbb{Z}}$ with $H(\mu) = C_{\mathrm{HMM}}(X_{\mathbb{Z}})$. That such an HMM exists is shown in Section 4.4 below. We have to prove $C_{\mathfrak{C}}(X_{\mathbb{Z}}) = C_{\mathrm{HMM}}(X_{\mathbb{Z}})$. Let M be the memory variable of the canonical memory γ_T. $H(\mu) = E(X_{\mathbb{Z}})$ in particular implies $I(W_0 : X_{\mathbb{N}}) = I(X_{-\mathbb{N}_0} : X_{\mathbb{N}})$, which means that W_0 is conditionally independent of $X_{\mathbb{N}}$ given $X_{-\mathbb{N}_0}$. Therefore, $(W_0, X_{\mathbb{Z}})$ has the same distribution as $(M, X_{\mathbb{Z}})$ and thus $I(M : X_{\mathbb{N}}) = I(X_{-\mathbb{N}_0} : X_{\mathbb{N}})$. This means that the canonical memory is sufficient and $C_{\mathrm{HMM}}(X_{\mathbb{Z}}) = H^{\mathbb{P}}(M) \geq C_{\mathfrak{C}}(X_{\mathbb{Z}})$ by Corollary 3.21.

3.3.3 Minimality of the ε-machine

We have seen that there are predictive models smaller than the ε-machine. We now show that this can happen only if the memory kernel is not deterministic. More precisely, determinism of a predictive memory implies sufficiency. This means, in particular, that the ε-machine is minimal among the state observable HMMs of a given process, because state observability is equivalent to determinism of the canonical memory. Below, we also obtain a stronger property for countable Δ. Namely, the ε-machine is also minimal among the partially deterministic HMMs, which is a larger class.

Proposition 3.39. *If a memory map is deterministic and predictive, then it is sufficient.*

Proof. Let (γ, T) be a predictive model and γ deterministic, i.e. $W_0 = M_0 = f \circ X_{-\mathbb{N}_0}$ for some measurable function $f \colon \Delta^{-\mathbb{N}_0} \to \Gamma$. Then a.s.

$$\mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \;=\; \mathbb{P}(F_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) \;=\; \mathbb{P}(F_{\mathbb{N}} \mid X_{-\mathbb{N}_0}, f \circ X_{-\mathbb{N}_0}) \;=\; \mathbb{P}(F_{\mathbb{N}} \mid W_0).$$

Thus, $\mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0})$ is $\sigma(W_0)$-measurable modulo $\mathbb{P}$ and $\mathbb{P}(X_{\mathbb{N}} \mid X_{-\mathbb{N}_0}) = \mathbb{P}(X_{\mathbb{N}} \mid W_0)$ a.s. Because $W_0 = M_0$, this means that the memory is sufficient. □

Corollary 3.40. *The causal state memory is the minimal predictive deterministic memory and the ε-machine is the minimal state observable generative HMM.*

In the case of countable Δ, we can use Proposition 3.39 together with the results of Section 2.1.7 about partially deterministic HMMs to see that the canonical memory of a partially deterministic HMM with finite internal state entropy is sufficient.

Theorem 3.41. *Let Δ be countable and (T, μ) a partially deterministic HMM of P with $H(\mu) < \infty$. Then the canonical memory kernel γ is sufficient. In particular,*

$$H(\mu) \;\geq\; C_{\mathfrak{C}}(P).$$

Proof. Define the relation $g \sim \hat{g} \;:\Leftrightarrow\; \mathbb{P}(X_{\mathbb{N}} \mid W_0 = g) = \mathbb{P}(X_{\mathbb{N}} \mid W_0 = \hat{g})$ and fix one representative of every equivalence class. Let $h \colon \Gamma \to \Gamma$ map g to the representative of its equivalence class. Note that measurability of h is not an issue, because Γ is essentially countable due to $H(\mu) < \infty$. Then $M' := h \circ M$ is a memory with memory kernel $\gamma'(x) = \gamma(x) \circ h^{-1}$, and it is predictive due to the same generator T. From Corollary 2.29, we see that it is deterministic. Thus, it is sufficient due to Proposition 3.39. Because M' is a function of M, we conclude from the data processing inequality that γ must be sufficient as well. □

Together with Proposition 3.20, we directly obtain that the ε-machine is the minimal partially deterministic HMM generating a given stationary process $X_{\mathbb{Z}}$.

Corollary 3.42. *If Δ is countable, the ε-machine is the minimal partially deterministic HMM of $X_{\mathbb{Z}}$.*

Note that the finite-entropy assumption in Theorem 3.41 cannot be dropped. Similarly to Example 2.28, it is straightforward to see that the canonical memory kernel of the shift HMM is in general not sufficient.

3.4 Prediction space

In this section, we represent causal states and ε-machine on the space $\mathcal{P}(\Delta^{\mathbb{N}})$ of probability measures on the future. This representation allows us to show the close relation to the prediction process introduced by Frank Knight in [Kni75] and compare the concepts of computational mechanics, namely causal states, ε-machine and statistical complexity, to other concepts such as the canonical OOM, process dimension and excess entropy. Properties of the discrete prediction process developed in Section 3.4.5 are also helpful to prove lower semi-continuity of statistical complexity later in Section 4.3.

3.4.1 Discrete-time version of Knight's prediction process

Given a Polish space[4] valued, measurable stochastic process with continuous 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 [Kni75] and was developed in [Mey76, Kni81, Kni92]. 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 leads to a lot of technical difficulties. In our simpler, discrete-time 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 characteristic that turns out to have a strong relation to the causal states, and its Markov transition kernel is related to the ε-machine transition. Let $X_{\mathbb{Z}}$ be a stationary process with distribution $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ and recall that the canonical projections from $\Delta^{\mathbb{Z}}$ to Δ are denoted by X'_k.

Definition 3.43. The $\mathcal{P}(\Delta^{\mathbb{N}})$-valued stochastic process $Z_{\mathbb{Z}} = Z^P_{\mathbb{Z}}$ of conditional probabilities, defined on the probability space $\big(\Delta^{\mathbb{Z}}, \mathfrak{B}(\Delta^{\mathbb{Z}}), P\big)$ by

$$Z_k \;:=\; P(X'_{[k+1,\infty[} \mid X'_{]-\infty,k]}), \qquad k \in \mathbb{Z},$$

is called **prediction process** of P or of $X_{\mathbb{Z}}$. $\mathcal{P}(\Delta^{\mathbb{N}})$ is called **prediction space**.

Remark. Frank Knight denotes a more sophisticated construction with the name "prediction space" ([Kni92, Sec. 2.3]). The complicated construction, however, is only necessary because of the continuous time set. The definition in [Kni92] refers to the space of paths of the prediction process, thus corresponding rather to $\mathcal{P}(\Delta^{\mathbb{N}})^{\mathbb{N}}$ than to $\mathcal{P}(\Delta^{\mathbb{N}})$. Nevertheless, we feel that in our simple setting it is appropriate to call $\mathcal{P}(\Delta^{\mathbb{N}})$ prediction space, because we do not need restrictions on the set of possible paths.

[4] Knight actually considers Lusin spaces, which are Borel subsets of compact metrisable spaces. Every Polish space is Lusin and every Lusin space is Borel isomorphic to a Polish 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 illustrative and much easier in our discrete-time setting. The corresponding transition kernel works as follows. Assume that 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. Recall that $\sigma\colon \Delta^{\mathbb{N}} \to \Delta^{\mathbb{N}}$ denotes the left shift and the symbol X'_k is also used for the canonical projections on $\Delta^{\mathbb{N}}$ instead of $\Delta^{\mathbb{Z}}$. There is one technical point involved, namely let for $z \in \mathcal{P}(\Delta^{\mathbb{N}})$

$$\phi_z\colon \Delta^{\mathbb{N}} \to \mathcal{P}(\Delta^{\mathbb{N}}), \qquad x \mapsto \phi_z(x) \;:=\; z\big(\sigma^{-1}(\,\cdot\,) \,\big|\, X'_1\big)(x).$$

Then we have to ensure that $\phi_z(x)$ is jointly measurable in z and x, which amounts to choosing versions of regular conditional probability that depend measurably on the probability measure. In [Kni92, Thm. 1.5], it is proven that such jointly measurable versions exist in Polish spaces, provided that the σ-algebra we are conditioning on is countably generated. We show in Appendix A.3 that this result remains true in Souslin spaces. Because $\sigma(X'_1)$ is countably generated, we may assume in the following that ϕ is jointly measurable.

Proposition 3.44. *Let $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$. The prediction process $Z_{\mathbb{Z}}$ of P is a stationary Markov process. The kernel S from $\mathcal{P}(\Delta^{\mathbb{N}})$ to $\mathcal{P}(\Delta^{\mathbb{N}})$ with $S(z) = z \circ \phi_z^{-1}$, i.e.*

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

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.

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

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

$$P\big(\{\, Z_1 \in B \,\} \,\big|\, Z_0\big) = P\big(\{\, Z_1 \in B \,\} \,\big|\, X'_{-\mathbb{N}_0}\big) = S(Z_0;\, B), \tag{3.6}$$

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 (3.6). □

Definition 3.45. 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. All P-dependent aspects of $Z_{\mathbb{Z}}$ are encoded in its (stationary) marginal distribution.

3.4.2 Prediction space representation of causal states and ε-machine

Recall that the causal states are equivalence classes of histories inducing the same conditional probability distribution on the future. Let ι_P be the function associating to an equivalence class the common distribution on the future, i.e.

$$\iota_P \colon \Gamma_{\mathfrak{C}} \to \mathcal{P}(\Delta^{\mathbb{N}}), \qquad \mathfrak{C}(x) \mapsto P(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0} = x). \tag{3.7}$$

By definition of $\mathfrak{C}$, ι_P is well-defined and injective. It is also measurable, because $P(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0} = \cdot\,)$ is measurable and $\Gamma_{\mathfrak{C}}$ is equipped with the final σ-algebra of $\mathfrak{C}$. Because both $\Gamma_{\mathfrak{C}}$ and $\mathcal{P}(\Delta^{\mathbb{N}})$ are Souslin spaces, ι_P is an isomorphism of measurable spaces onto its image. Note that even if we restrict Δ to be a Polish space, which implies that $\mathcal{P}(\Delta^{\mathbb{N}})$ is Polish, the image of ι_P does not need to be Polish or even measurable. It is, however, a Souslin space. A causal state $g = \mathfrak{C}(x)$ is called "causal" because it captures the part of the past that is relevant for the future of the process. To keep this intuition, we call the measure $\iota_P(g)$ "effect" corresponding to the "cause" g.

Definition 3.46. For $P \in \mathcal{P}_{\mathtt{s}}(\Delta^{\mathbb{Z}})$, the **effect distribution** $\mu_{\mathfrak{C}}(P)$ of P is the marginal distribution of the prediction process. Its topological support is denoted by $\mathfrak{S}_P$ and called **effect space**. In formulas,

$$\mu_{\mathfrak{C}}(P) \;:=\; P \circ Z_0^{-1} \;\in\; \mathcal{P}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big) \qquad \text{and} \qquad \mathfrak{S}_P \;:=\; \operatorname{supp}\big(\mu_{\mathfrak{C}}(P)\big) \;\subseteq\; \mathcal{P}(\Delta^{\mathbb{N}}).$$

Obviously, the effect distribution corresponds to the distribution $\mu_{\gamma_{\mathfrak{C}}} = \mathbb{P}_{M_{\mathfrak{C}}}$ of the causal state memory and the statistical complexity can be computed as the entropy of the effect distribution, i.e.

$$\mu_{\mathfrak{C}}(P) \;=\; \mu_{\gamma_{\mathfrak{C}}} \circ \iota_P^{-1} \qquad \text{and} \qquad C_{\mathfrak{C}}(P) \;=\; H\big(\mu_{\mathfrak{C}}(P)\big).$$

We use this interpretation for proving properties about the statistical complexity as function on $\mathcal{P}_{\mathtt{s}}(\Delta^{\mathbb{Z}})$ in Section 4.3.

Remark. The space of causal states directly corresponds to the image of ι_P, but both of these spaces depend on the chosen version of conditional probability. The effect space $\mathfrak{S}_P$ is free of this defect and uniquely determined by P, because different versions of Z_0 coincide P-a.s. and the effect distribution is the push-forward of P under Z_0. $\mathfrak{S}_P$ can be considered a representation of the causal states in a form independent of the version of conditional probability. It would not be easy to obtain such a representation in the original formulation as equivalence classes on $\Delta^{-\mathbb{N}_0}$, because there is no canonical topology and the different versions are not embedded in a common larger space. Note, however, that the effect space can be uncountable for a process with countably many causal states (then $\operatorname{Im}(\iota_P)$ is dense in $\mathfrak{S}_P$).

If the effect distribution was continuous, lower semi-continuity of the statistical complexity would follow from lower semi-continuity of the entropy. Unfortunately, this is not the case.

Example 3.47. $\mu_{\mathfrak{C}}$ is *not* continuous. Let P be a non-deterministic i.i.d. process. Obviously, the effect 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 [Par61], periodic measures are dense in the stationary measures, and we find an approximating sequence $P_n \xrightarrow{*} P$ of periodic measures P_n. The past of a periodic process determines its future, thus its effect distribution is supported by the set $\mathfrak{M} = \{\, \delta_x \mid x \in \Delta^{\mathbb{N}} \,\}$ of Dirac measures on $\Delta^{\mathbb{N}}$. Because $\mathfrak{M}$ is closed in $\mathcal{P}(\Delta^{\mathbb{N}})$ and does not contain the topological support $\mathfrak{S}_P = \{\, P_{\mathbb{N}} \,\}$ of $\mu_{\mathfrak{C}}(P)$, $\mu_{\mathfrak{C}}(P_n)$ cannot converge to $\mu_{\mathfrak{C}}(P)$. $\Diamond$

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

$$z \;=\; r(\nu) \;:=\; \int_{\mathcal{P}(\Delta^{\mathbb{N}})} \mathrm{id}_{\mathcal{P}(\Delta^{\mathbb{N}})}\,\mathrm{d}\nu, \tag{3.8}$$

where $r\colon \mathcal{P}(\mathcal{P}(\Delta^{\mathbb{N}})) \to \mathcal{P}(\Delta^{\mathbb{N}})$ is called **resultant** (or **barycentre map**) and id is the identity map. In the case of compact Δ, this is a special case of the situation in [Cho69]. Here, we do not need compactness for existence and continuity of the resultant, because it is given by integration over a continuous kernel from $\mathcal{P}(\Delta^{\mathbb{N}})$ to $\Delta^{\mathbb{N}}$. $z = r(\nu)$ 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 $z \circ \sigma^{-1}$.

Lemma 3.48. *$r(\nu S) = r(\nu) \circ \sigma^{-1}$. 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

$$r(\delta_z S) \;=\; \int \mathrm{id}_{\mathcal{P}(\Delta^{\mathbb{N}})}\,\mathrm{d}S(z) \;=\; \int \phi_z\,\mathrm{d}z \;=\; \int z\big(\sigma^{-1}(\,\cdot\,) \;\big|\; X_1'\big)\,\mathrm{d}z \;=\; z \circ \sigma^{-1}. \qquad \square$$

If ν is S-invariant, we also say that ν represents the stationary extension of $r(\nu)$ to $\Delta^{\mathbb{Z}}$. The effect distribution of P is an important S-invariant representation of P.

Lemma 3.49. *Let $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$. Then $\mu_{\mathfrak{C}}(P)$ is S-invariant and represents P.*

Proof. From Proposition 3.44 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

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

We already represented the causal states on prediction space and saw that they are in close relation to the marginal distribution of the prediction process. Now we represent the ε-machine and it is hardly surprising that it is intimately related to the prediction process. First, we define a *prediction HMM* related to the prediction process and later we show in Proposition 3.52 that it is indeed isomorphic to the ε-machine. The "internal state update" of the transition $T^{\mathfrak{C}}$ of the prediction HMM 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 $\phi_z(x) = z({\sigma^{-1}}_{(\,\cdot\,)} \mid X'_1)(x)$.

Definition 3.50. We define the Markov kernel $T^{\mathfrak{C}}$ from $\mathcal{P}(\Delta^{\mathbb{N}})$ to $\Delta \times \mathcal{P}(\Delta^{\mathbb{N}})$ by

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

and call the HMM $\big(T^{\mathfrak{C}}, \mu_{\mathfrak{C}}(P)\big)$ **prediction HMM** of P.

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 effect distribution (Definition 3.46) of some $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$, 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. This statement will be obvious anyway when we show below in Proposition 3.52 that the HMM is isomorphic to the ε-machine. But even more is true. Namely, if $\mu \in \mathcal{P}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big)$ (not necessarily S-invariant) 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 resultant, defined in (3.8), and associates the represented process to μ.

Proposition 3.51. *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, the prediction HMM $\big(T^{\mathfrak{C}}, \mu_{\mathfrak{C}}(P)\big)$ is an invariant HMM of $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$.*

Proof. The output kernel (Definition 2.12) is $K_z = z \circ {X'_1}^{-1}$. The transition function f given by $f_z \circ X'_1 := \phi_z$ is well-defined due to the $\sigma(X'_1)$-measurability of ϕ_z. We have $T^{\mathfrak{C}}(z;\, D \times B) = K_z\big(D \cap f_z^{-1}(B)\big)$ by definition, thus the HMM is partially deterministic. To show that the output process is $r(\mu)$, assume w.l.o.g. that μ is a Dirac measure (the general claim follows by integration over μ). Thus $\mu = \delta_z$ with $z = r(\mu)$. Let $L_d^{\mathfrak{C}}$ be the internal operator of $T^{\mathfrak{C}}$ and recall that, according to Lemma 2.13, $\big(T^{\mathfrak{C}}, L_d^{\mathfrak{C}}(\delta_z)\big)$ 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

$$r\big(L_d^{\mathfrak{C}}(\delta_z)\big) \overset{(2.5)}{=} r\big(\delta_{f_z(d)}\big) \;=\; f_z(d) \;=\; z\big(\sigma^{-1}(\,\cdot\,) \mid X'_1 = d\big),$$

the claim follows by induction. □

The prediction HMM is a representation of the ε-machine on prediction space.

Proposition 3.52. *Let $X_{\mathbb{Z}}$ be a stationary process with distribution $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$. Then the prediction HMM $\big(T^{\mathfrak{C}}, \mu_{\mathfrak{C}}(P)\big)$ is isomorphic to the ε-machine $(T^{\gamma_{\mathfrak{C}}}, \mu_{\gamma_{\mathfrak{C}}})$.*

Proof. We claim that ι_P, defined by (3.7), is an isomorphism. Indeed, we already know that it is injective, measurable, and $\mu_{\mathfrak{C}}(P) = \mu_{\gamma_{\mathfrak{C}}} \circ \iota_P^{-1}$. According to Lemma 2.24, it is sufficient to prove that ι_P "preserves" the output kernel and the transition function. Let K and f be the output kernel and transition function of the prediction HMM, K^{ε}, f^{ε} those of the ε-machine. We obtain a.s.

$$K_{\iota_P \circ \mathfrak{C}(x)} \;=\; \big(\iota_P \circ \mathfrak{C}(x)\big) \circ {X'_1}^{-1} \;=\; P\big(X'_1 \;\big|\; \mathfrak{C}(X'_{-\mathbb{N}_0}) = \mathfrak{C}(x)\big) \;=\; K^{\varepsilon}_{\mathfrak{C}(x)}$$

and

$$\begin{aligned} f_{\iota_P(\mathfrak{C}(x))}(d) \;&=\; P(X'_{[2,\infty[} \mid X'_{-\mathbb{N}_0} = x,\; X'_1 = d) \;=\; P(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0} = xd) \\ &=\; \iota_P\big(\mathfrak{C}(xd)\big) \overset{(\text{Prop. } 3.14)}{=} \iota_P \circ f^{\varepsilon}_{\mathfrak{C}(x)}(d). \end{aligned}$$

□

There are several advantages of working with prediction space instead of equivalence classes on $\Delta^{-\mathbb{N}_0}$. First, $\mathcal{P}(\Delta^{\mathbb{N}})$ possesses a natural topology, which enabled us to define the effect space in a way independent of the version of conditional probability. It also helps us proving lower semi-continuity of statistical complexity in Section 4.3. Second, $\mathcal{P}(\Delta^{\mathbb{N}})$ has an algebraic structure, which allows us to clarify the relation between causal states and the canonical OOM in Section 3.4.3 below. Third, all Δ-valued processes can be treated in a unified way on the same space with the same transition kernel. In the ε-machine, all components depend on the process. The underlying space of internal states depends on P, while the prediction HMMs are all defined on prediction space $\mathcal{P}(\Delta^{\mathbb{N}})$. Even if the partitions defined by the causal states of two processes coincide, the generator $T^{\gamma_{\mathfrak{C}}}$ is different for them. The generator $T^{\mathfrak{C}}$ of the prediction HMM, on the other hand, is universal for all processes. It is only the initial distribution of the prediction HMM that captures all P-dependent aspects.

Let Δ be countable. Given a process $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$, 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 $r(\nu) = P_{\mathbb{N}}$). We already know from Corollary 3.42 that none of them can have lower entropy than the effect distribution. In the next proposition, we see that even more is true. The effect distribution of P is distinguished as the *only* one that can have finite entropy.

Proposition 3.53. *Let Δ be countable, $\nu \in \mathcal{P}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big)$ S-invariant, and $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ the measure it represents. If $\nu \neq \mu_{\mathfrak{C}}(P)$, then $H(\nu) = \infty$.*

Proof. Recall that $Y_0 = \mathbb{P}(W_0 \mid X_{-\mathbb{N}_0})$ (Definition 2.20). Let $H(\nu) < \infty$. According to Proposition 3.51, $(T^{\mathfrak{C}}, \nu)$ is an invariant, partially deterministic HMM of P and we can apply Corollary 2.29. Let $W_{\mathbb{Z}}$ be the $\Gamma = \mathcal{P}(\Delta^{\mathbb{N}})$-valued internal process of the HMM. For almost all fixed ω, Lemma 2.13 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 Proposition 3.51. Thus, $\mathbb{P}(X_{\mathbb{N}} \mid W_0) = W_0$ and

$$z \;=\; \mathbb{P}(X_{\mathbb{N}} \mid W_0 = z) \;\overset{\text{(Cor. 2.29)}}{=}\; \mathbb{P}(X_{\mathbb{N}} \mid W_0 = \hat{z}) \;=\; \hat{z} \qquad \forall z, \hat{z} \in \operatorname{supp}\big(Y_0(\omega)\big).$$

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 2.7).

Example 3.54. Let $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$, $P_{\mathbb{N}} = P \circ X_{\mathbb{N}}^{-1}$ 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 3.53 because ν is not S-invariant (if P is not i.i.d.) ◊

Example 3.55 (lifted shift). Let $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ 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 isomorphic to the one-sided shift with isomorphism ι. This is no contradiction to Proposition 3.53, because $H(\nu) = \infty$ if P is not concentrated on countably many trajectories. In the latter case, $\nu = \mu_{\mathfrak{C}}(P)$. ◊

3.4.3 From causal states to the canonical OOM

Provided that Δ is countable, the representation of the causal states on prediction space as effects also helps us to clarify their close relation to the canonical OOM (Section 2.2.2). Recall that the canonical OOM vector space is $V_P = \operatorname{span}(Q_P)$ with $Q_P = \big\{\, \tau_{d_1\cdots d_n}(P_{\mathbb{N}}) \;\big|\; n \in \mathbb{N}_0,\ d_1, \ldots, d_n \in \Delta \,\big\}$ and observable operators $\tau_d(z) = z\big([d] \cap \sigma^{-1}(\,\cdot\,)\big)$. The connection to the finite-history version of the causal states can be seen very easily. If we replace τ_d in the definition of the canonical OOM vector space by the normalised, non-linear version $z \mapsto \frac{1}{\|\tau_d(z)\|}\tau_d(z)$, where $\|z\| = z(\Delta^{\mathbb{N}})$ is the variational norm, the generated vector space obviously stays the same. Therefore,

$$V_P \;=\; \operatorname{span}\Big\{\, P\big(\sigma^{-n}(\,\cdot\,) \;\big|\; [d_1, \ldots, d_n]\big) \;\Big|\; n \in \mathbb{N}_0,\ d_1, \ldots, d_n \in \Delta,\ P\big([d_1, \ldots, d_n]\big) > 0 \,\Big\}.$$

In the case of finite process dimension, this relation continues to hold for infinite-history causal states. More precisely, V_P turns out to be the linear hull of the effect space $\mathfrak{S}_P$. Because the OOM vector space is defined with finite-length pasts and infinite pasts are used for the definition of $\mathfrak{S}_P$, we can interpret this result as follows. Unlike the set of causal states, the canonical OOM vector space is the same if we consider finite or infinite pasts, provided it is finite-dimensional. See Example 3.34 for an example of a finite-dimensional process with substantially more finite-history causal states than causal states. In the infinite-dimensional case, the situation is more subtle (see Example 3.58) and only the closures of the spaces coincide. Recall that $\overline{V}^{w*}$ denotes the closure of V w.r.t. the weak-$*$ topology.

Theorem 3.56. *Let Δ be countable and $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$. Then*

$$\overline{V_P}^{\,w*} \;=\; \overline{\operatorname{span}(\mathfrak{S}_P)}^{\,w*}.$$

Proof. "$\subseteq$": Let $z \in Q_P$. Then $z = \tau_{d_1\cdots d_n}(P_{\mathbb{N}})$ for some $d_1, \ldots, d_n \in \Delta$. Define A to be the event that the past is $d_1, \ldots, d_n$, i.e. $A := \sigma^n\big([d_1, \ldots, d_n]\big) \subseteq \Delta^{\mathbb{Z}}$. We assume $P(A) > 0$, as otherwise $z = 0$. Further define the non-normalised measure $\widehat{P} := P(A \cap \,\cdot\,) \in \mathcal{M}_+(\Delta^{\mathbb{Z}})$ and recall that $P_{\mathbb{N}}^{-\mathbb{N}_0} = P(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0})$. Let $\mu = \widehat{P} \circ \big(P_{\mathbb{N}}^{-\mathbb{N}_0}\big)^{-1} \in \mathcal{M}_+\big(\mathcal{P}(\Delta^{\mathbb{N}})\big)$. Note that the conditional probability $P(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0})$ in the definition of μ is w.r.t. to P not $\widehat{P}$. Using stationarity of P, we obtain

$$\begin{aligned} z \;&=\; P_{\mathbb{N}}\big([d_1, \ldots, d_n] \cap \sigma^{-n}(\,\cdot\,)\big) \;=\; \int P\big(A \cap \{\, X'_{\mathbb{N}} \in \,\cdot\, \} \;\big|\; X'_{-\mathbb{N}_0}\big)\, \mathrm{d}P \\ &=\; \int_A P(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0})\, \mathrm{d}P \;=\; \int P_{\mathbb{N}}^{-\mathbb{N}_0}\, \mathrm{d}\widehat{P} \;=\; \int \mathrm{id}_{\mathcal{P}(\Delta^{\mathbb{N}})}\, \mathrm{d}\mu \;=\; \|\mu\| \cdot r\Big(\frac{\mu}{\|\mu\|}\Big), \end{aligned}$$

where id is the identity, $\|\mu\| = \mu\big(\mathcal{P}(\Delta^{\mathbb{N}})\big)$ is the variational norm, and $r\colon \mathcal{P}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big) \to \mathcal{P}(\Delta^{\mathbb{N}})$ is the resultant, defined in (3.8). Because $\widehat{P} \ll P$, and thus $\mu \ll \mu_{\mathfrak{C}}(P)$, the support of μ is contained in $\mathfrak{S}_P$. Metrisability of $\mathfrak{S}_P$ implies that we can approximate μ by measures

on $\mathfrak{S}_P$ with finite support ([AB99, 14.10]). Together with continuity of the resultant, this means that the barycentre lies in the closed convex hull of $\mathfrak{S}_P$, i.e.

$$r\big(\tfrac{1}{\|\mu\|}\mu\big) \in \overline{\operatorname{conv}(\mathfrak{S}_P)}^{w*} \qquad \text{and} \qquad z \in \overline{\operatorname{span}(\mathfrak{S}_P)}^{w*}.$$

"$\supseteq$": We have to show that $\overline{V_P}^{w*} \cap \mathcal{P}(\Delta^{\mathbb{N}})$ has full $\mu_{\mathfrak{C}}(P)$-measure, in other words that $P(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0}) \in \overline{V_P}^{w*}$ P-a.s. By the martingale convergence theorem we have for $B \in \mathfrak{B}(\Delta^{\mathbb{N}})$ and $x = (x_k)_{k\in\mathbb{Z}} \in \Delta^{\mathbb{Z}}$ a.s.

$$P\big(\{\, X'_{\mathbb{N}} \in B \,\} \,\big|\, X'_{-\mathbb{N}_0}\big)(x) \;=\; \lim_{n\to\infty} P\big(\{\, X'_{\mathbb{N}} \in B \,\} \,\big|\, X'_{[-n,0]}\big)(x) \;=\; \lim_{n\to\infty} \frac{\tau_{x_{-n}\cdots x_0}(P_{\mathbb{N}})(B)}{P\big([x_{-n},\dots,x_0]\big)}$$

Because $\mathfrak{B}(\Delta^{\mathbb{N}})$ is countably generated and setwise (pointwise) convergence of a sequence of probability measures implies weak-$*$ convergence, we obtain $P(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0}) \in \overline{\mathbb{R}\cdot Q_P}^{w*} \subseteq \overline{V_P}^{w*}$ a.s. □

Recall that the process dimension $\dim(P)$ of P is the dimension of the canonical OOM vector space V_P (Definition 2.35).

Corollary 3.57. *Let Δ be countable and $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$. Then the process dimension satisfies*

$$\dim(P) \;=\; \dim\big(\operatorname{span}(\mathfrak{S}_P)\big).$$

Proof. Finite-dimensional spaces are closed. Thus, if $\dim(P) = \dim(V_P) = \infty$, $\operatorname{span}(\mathfrak{S}_P)$ must also be infinite-dimensional. Otherwise, $V_P = \operatorname{span}(\mathfrak{S}_P)$ and thus the dimensions coincide. □

Remark. Assume $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ has *finite process dimension*. Instead of considering the prediction HMM $\big(T^{\mathfrak{C}}, \mu_{\mathfrak{C}}(P)\big)$ with the whole space $\mathcal{P}(\Delta^{\mathbb{N}})$ as internal states, we can obviously restrict it to the effect space. This HMM is a representation of the ε-machine. If we consider its associated OOM, the corresponding OOM vector space is $\mathcal{M}(\mathfrak{S}_P)$. Compare this to the canonical OOM vector space $V_P = \operatorname{span}(\mathfrak{S}_P)$. The latter can be much lower dimensional, because it utilises the linear structure of $\mathfrak{S}_P$. The spaces are isomorphic if and only if the elements of $\mathfrak{S}_P$ are linearly independent (then $\mathfrak{S}_P$ is in particular finite).

The closures in Theorem 3.56 are really necessary, as we see in the next example. Although $\mathfrak{S}_P$ is closed, $\operatorname{span}(\mathfrak{S}_P)$ is not (in general). Also, in general, neither does $\operatorname{span}(\mathfrak{S}_P)$ contain V_P nor the other way round.

Example 3.58. Let $\Delta = \{\, 0, 1 \,\}$ and for $p \in [0,1]$ let $P_p \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ be the Bernoulli process with parameter p, i.e. P_p is i.i.d. with $P_p\big([1]\big) = p$. Consider the uncountable mixture $P = \int P_p \,\mathrm{d}p$, where integration is w.r.t. Lebesgue measure. Then $\mathfrak{S}_P = \big\{\, P_p \circ {X'_{\mathbb{N}}}^{-1} \,\big|\, p \in [0,1] \,\big\}$ is the set of i.i.d. processes and $\mu_{\mathfrak{C}}(P)$ is the image of Lebesgue measure under the map $p \mapsto P_p \circ {X'_{\mathbb{N}}}^{-1}$. We make the following observations.

1. $\operatorname{span}(\mathfrak{S}_P) \cap \mathcal{P}(\Delta^{\mathbb{N}})$ is the set of *finite* mixtures of i.i.d. processes, in particular $\operatorname{span}(\mathfrak{S}_P)$ is *not* closed.

2. By definition, V_P has countable algebraic dimension, i.e. it is the linear hull of a countable set. Every basis of $\operatorname{span}(\mathfrak{S}_P)$, on the other hand, has to be uncountable (the family $(P_p)_{p\in[0,1]}$ is linearly independent). Thus, V_P cannot contain $\operatorname{span}(\mathfrak{S}_P)$.

3. All elements of $V_P \cap \mathcal{P}(\Delta^{\mathbb{N}})$ have an uncountable number of ergodic components. Therefore, $\operatorname{span}(\mathfrak{S}_P)$ and V_P are even disjoint (except for 0). ◊

3.4.4 Excess entropy and effect distribution

Statistical complexity of a process $P \in \mathcal{P}_{\mathfrak{s}}(\Delta^{\mathbb{Z}})$ is the entropy of the effect distribution $\mu_{\mathfrak{C}}(P)$. Thus it measures "how many" (in the weighted sense of entropy) different future distributions are possible given different past trajectories. It is insensitive to the internal structure of the elements of $\mathcal{P}(\Delta^{\mathbb{N}})$. The process dimension, on the other hand, uses the linear structure of $\mathcal{P}(\Delta^{\mathbb{N}})$. It is the dimension of the support $\mathfrak{S}_P$ of $\mu_{\mathfrak{C}}(P)$. The number of possible future distributions may well be infinite and still contained in a two-dimensional subspace, which is the situation in Example 3.26. It turns out that excess entropy can also be written as a function of $\mu_{\mathfrak{C}}(P)$ and it depends on both the algebraic structure and the "distance structure" given by the Kullback-Leibler divergence. Recall that excess entropy is the mutual information between past and future. For random variables X and Y with values in countable spaces, it is well-known and easy to prove that the mutual information can be rewritten as average Kullback-Leibler divergence,

$$I(X:Y) \;=\; \sum_x \mathbb{P}_X(x)\, D_{\mathrm{KL}}\big(\mathbb{P}(Y \mid X=x) \,\big\|\, \mathbb{P}_Y\big).$$

We could not find the corresponding formula for more general spaces in the literature and therefore give a proof. We use the well-known identity $I(X:Y) = D_{\mathrm{KL}}(\mathbb{P}_{X,Y} \parallel \mathbb{P}_X \otimes \mathbb{P}_Y)$.

Proposition 3.59. *Let X, Y be random variables with values in a Souslin space Γ. Then*

$$I(X:Y) \;=\; \int D_{\mathrm{KL}}\big(\mathbb{P}(Y \mid X) \,\big\|\, \mathbb{P}_Y\big)\, \mathrm{d}\mathbb{P}.$$

Proof. 1. *Case $\mathbb{P}_{X,Y} \not\ll \mathbb{P}_X \otimes \mathbb{P}_Y$:* In this case, $I(X:Y) = \infty$. To show that the right-hand side is also infinite, we show that $\mathbb{P}(Y \mid X) \not\ll \mathbb{P}_Y$ on a set of positive measure. Choose a measurable set A with $\mathbb{P}_{X,Y}(A) > 0$ and $\mathbb{P}_X \otimes \mathbb{P}_Y(A) = 0$ and decompose it into fibres $A = \bigcup_x \{\, x \,\} \times A_x$. Note that the A_x are measurable and

$$0 \;=\; \mathbb{P}_X \otimes \mathbb{P}_Y(A) \;=\; \int_{x\in\Gamma} \mathbb{P}_Y(A_x)\, \mathrm{d}\mathbb{P}_X$$

implies that $\mathbb{P}_Y(A_x) = 0$ $\mathbb{P}_X$-a.s. On the other hand,

$$0 \;<\; \mathbb{P}_{X,Y}(A) \;=\; \int_{x\in\Gamma} \mathbb{P}\big(\{\, Y \in A_x \,\} \,\big|\, X = x\big)\, \mathrm{d}\mathbb{P}_X$$

implies that $\mathbb{P}\big(\{\, Y \in A_x \,\} \;\big|\; X = x\big) > 0$ on a set of positive measure. For all such x, $\mathbb{P}(Y \mid X = x)$ is not absolutely continuous w.r.t. $\mathbb{P}_Y$.

2. *Case $\mathbb{P}_{X,Y} \ll \mathbb{P}_X \otimes \mathbb{P}_Y$:* Let $f := \frac{\mathrm{d}\mathbb{P}_{X,Y}}{\mathrm{d}(\mathbb{P}_X \otimes \mathbb{P}_Y)}$ be the Radon-Nikodym derivative. Then

$$I(X:Y) \;=\; D_{\mathrm{KL}}(\mathbb{P}_{X,Y} \parallel \mathbb{P}_X \otimes \mathbb{P}_Y) \;=\; \int \log(f)\, \mathrm{d}\mathbb{P}_{X,Y}. \tag{3.9}$$

It is easy to see that $\int_{y\in A} f(x,y)\, \mathrm{d}\mathbb{P}_Y$ is a regular version of the conditional probability $\mathbb{P}\big(\{\, Y \in A \,\} \;\big|\; X = x\big)$. Thus we have $\mathbb{P}(Y \mid X) \ll \mathbb{P}_Y$ a.s. and see that

$$\frac{\mathrm{d}\mathbb{P}(Y \mid X = x)}{\mathrm{d}\mathbb{P}_Y}(y) \;=\; f(x,y) \qquad \mathbb{P}_X \otimes \mathbb{P}_Y\text{-a.s.},$$

hence also $\mathbb{P}_{X,Y}$-a.s. We apply this identity to (3.9) and obtain

$$\begin{aligned} I(X:Y) &= \int_{\omega\in\Omega} \log\left(\frac{\mathrm{d}\mathbb{P}(Y \mid X)(\omega)}{\mathrm{d}\mathbb{P}_Y}\big(Y(\omega)\big)\right) \mathrm{d}\mathbb{P} \\ &= \int_{\omega\in\Omega} \left(\int \log\left(\frac{\mathrm{d}\mathbb{P}(Y \mid X)(\omega)}{\mathrm{d}\mathbb{P}_Y}\right) \mathrm{d}\mathbb{P}(Y \mid X)(\omega)\right) \mathrm{d}\mathbb{P}. \end{aligned}$$

The right-hand side coincides with $\int D_{\mathrm{KL}}\big(\mathbb{P}(Y \mid X) \,\big\|\, \mathbb{P}_Y\big)\,\mathrm{d}\mathbb{P}$, as claimed. □

Specialising the proposition to $X = X'_{-\mathbb{N}_0}$ and $Y = X'_{\mathbb{N}}$, we obtain

$$E(P) \;=\; \int D_{\mathrm{KL}}\big(P(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0}) \,\big\|\, P_{\mathbb{N}}\big)\,\mathrm{d}P \;=\; \int D_{\mathrm{KL}}(\,\cdot\, \|\, P_{\mathbb{N}})\,\mathrm{d}\mu_{\mathfrak{C}}(P).$$

This means that the excess entropy is the average Kullback-Leibler divergence from the conditional distribution of the future to the unconditional one. To make the dependence on $\mu_{\mathfrak{C}}(P)$ even more explicit, we rewrite $P_{\mathbb{N}}$ in terms of $\mu_{\mathfrak{C}}$. Namely, it is the barycentre of $\mu_{\mathfrak{C}}(P)$ (Lemma 3.49). We obtain

$$E(P) \;=\; \int D_{\mathrm{KL}}\Big(\,\cdot\, \Big\|\, r\big(\mu_{\mathfrak{C}}(P)\big)\Big)\,\mathrm{d}\mu_{\mathfrak{C}}(P)$$

and see that $E(P)$ uses also the convex structure of $\mathcal{P}(\Delta^{\mathbb{N}})$ via the resultant r.

3.4.5 Discrete prediction process

In this section, we consider the prediction process in the case where not only time but also the state space Δ is discrete. It turns out that in this situation the prediction dynamic is continuous. This is an interesting result on its own, and we also need it to analyse statistical complexity in Section 4.3. In the case of a general Souslin space Δ, even joint measurability of ϕ is a non-trivial fact. For countable Δ, however, we obtain its essential continuity in an elementary way. This is the main reason for the continuity of the prediction dynamic.

Lemma 3.60. *Let Δ be countable and $z, z_n \in \mathcal{P}(\Delta^{\mathbb{N}})$ with $z_n \overset{*}{\rightharpoonup} z$. 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 $A_x := {X'_1}^{-1}\big(X'_1(x)\big)$ and $\Omega_z := \big\{\, x \in \Delta^{\mathbb{N}} \;\big|\; z(A_x) > 0 \,\big\}$. Because Δ is discrete and countable, Ω_z is clopen with $z(\Omega_z) = 1$. Uniform convergence on compacta is equivalent to $\phi_{z_n}(x_n) \overset{*}{\rightharpoonup} \phi_z(x)$ whenever $x_n \to x$ in Ω_z. For sufficiently large n, $X'_1(x_n) = X'_1(x)$ and because σ^{-1} maps cylinder sets to cylinder sets, $\phi_{z_n}(x_n) = \frac{z_n(A_x \cap \sigma^{-1}(\,\cdot\,))}{z_n(A_x)} \overset{*}{\rightharpoonup} \phi_z(x)$. □

Theorem 3.61. *Let Δ be countable. Then 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 3.60. We have to show

$$\int g\,\mathrm{d}S(z_n) \;=\; \int g \circ \phi_{z_n}\,\mathrm{d}z_n \;\overset{n\to\infty}{\longrightarrow}\; \int g \circ \phi_z\,\mathrm{d}z \;=\; \int g\,\mathrm{d}S(z) \tag{3.10}$$

for any bounded continuous g. According to Prokhorov's theorem, the sequence $(z_n)_{n\in\mathbb{N}}$ is uniformly tight and we can restrict the integrations to compact subsets. Because Ω_z is clopen, we have $\lim_{n\to\infty} z_n(\Omega_z) = z(\Omega_z) = 1$ and can restrict to compact subsets of Ω_z. There, the convergence of ϕ_{z_n} is uniform, thus (3.10) holds. □

Chapter 4

Complexity measures of stochastic processes

So far we have, given a fixed stochastic process P, compared three different complexity measures and their motivation: *excess entropy*, *statistical complexity* and *generative complexity*. In this chapter, we suggest to consider *process dimension* (Definition 2.35) also as complexity measure and interpret these four quantities as functions on the space of Δ-valued stationary processes. Thereby Δ *is always assumed to be a Souslin space* with at least two elements.

The following question arises naturally. Which functions $F\colon \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}}) \to \overline{\mathbb{R}}_+ = \mathbb{R}_1 \cup \{\infty\}$ can reasonably be interpreted as complexity measures, and which can not? It would be desirable to have an axiomatic answer to this question, that is a characterisation of the complexity measure functionals. Although we are, of course, far from having such a characterisation, we argue that 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 complicated ones (and hence the complexity is not necessarily 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 should include lower semi-continuity. For the axiomatisation of Shannon entropy, concavity is crucial. We feel that concavity should also be a property of complexity measures. If we mix processes, the resulting process should not be considered less complex than the average original process.

In this chapter, we show that excess entropy, statistical complexity and generative complexity are indeed lower semi-continuous and concave. Thus, they can be interpreted as complexity measures. To obtain lower semi-continuity of statistical complexity, we have to assume that Δ is countable. We also give ergodic decomposition formulas, that is we analyse how the complexities of the ergodic components of a process have to be combined to obtain its complexity. The definitions of all three complexity measures use Shannon entropy and this reflects in the same ergodic decomposition behaviour. We call complexity measures with this behaviour *entropy-based* and demonstrate a few of their elementary properties in Section 4.1, before we show in the subsequent sections that the three discussed complexity measures actually are entropy-based according to our definition. In the last section of this chapter, we suggest to investigate process dimension as candidate of a complexity measure and show that it is also lower semi-continuous and concave for countable state spaces Δ, although it is not entropy-based.

Before we proceed, we give a simple non-continuity example.

Example 4.1 (non-continuity). Let $\pi_p \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ be the Bernoulli process on $\Delta = \{0, 1\}$ with parameter $0 < p < 1$, i.e. $\pi_p\big([1]\big) = p$. Consider the process of throwing a coin that is either slightly biased to 0 or 1, each with probability $\frac{1}{2}$, i.e. $P_\varepsilon = \frac{1}{2}\pi_{\frac{1}{2}+\varepsilon} + \frac{1}{2}\pi_{\frac{1}{2}-\varepsilon}$ with $0 < \varepsilon < \frac{1}{2}$. Then $P_\varepsilon \xrightarrow{*} P_0 = \pi_{\frac{1}{2}}$ for $\varepsilon \to 0$, but all three complexity measures we considered have a discontinuity at $\varepsilon = 0$: $C_{\mathrm{HMM}}(P_\varepsilon) = C_{\mathfrak{C}}(P_\varepsilon) = E(P_\varepsilon) = \log(2)$ but $C_{\mathrm{HMM}}(P_0) = C_{\mathfrak{C}}(P_0) = E(P_0) = 0$. ◊

4.1 Entropy-based complexity measures

4.1.1 Entropy

First, we summarise a few helpful properties of the entropy. It is weak-$*$ lower semi-continuous and concave in a rather general setting and satisfies a nice decomposition formula for convex combinations of mutually singular measures. These results are probably all well-known, note however that lower semi-continuity of the entropy is most often proven w.r.t. variational topology, which is not sufficient for our purposes. Therefore, we provide a proof of lower semi-continuity in Appendix A.5.

Lemma 4.2. *Let Γ be a separable, metrisable space. Then the entropy $H\colon \mathcal{P}(\Gamma) \to \overline{\mathbb{R}}_+$ is weak-$*$ lower semi-continuous.*

It is also important that the entropy is concave and satisfies a nice formula for convex combinations of mutually singular measures. Mutual singularity of a sequence of measures μ_k means that there is a sequence of disjoint, measurable sets A_k with $\mu_k(A_k) = 1$.

Lemma 4.3. *Let Γ be a separable, metrisable space and $\mu \in \mathcal{P}(\Gamma)$ a countable convex combination $\mu = \sum_{k\in I} \nu(k)\,\mu_k$ of measures $\mu_k \in \mathcal{P}(\Gamma)$. Then*

$$\sum_k \nu(k)\, H(\mu_k) \;\leq\; H(\mu) \;\leq\; H(\nu) + \sum_k \nu(k)\, H(\mu_k).$$

Furthermore, the second inequality is an equality if and only if the μ_k are mutually singular or $H(\mu) = \infty$. In particular, H is concave.

We use this property of entropy several times below, in the proof of ergodic decomposition formulas for the different complexity measures.

4.1.2 Entropy-based complexity measures

In this section, we look at a class of non-linear functionals on $\mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ that is characterised by its behaviour w.r.t. to ergodic decomposition. We consider the ergodic decomposition of $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ to be given as a probability measure ν_P on the space of ergodic measures $\mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}}) \subseteq \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$.

Definition 4.4. Let $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$. The **ergodic decomposition** $\nu_P \in \mathcal{P}\big(\mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}})\big)$ of P is defined by

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

If ν_P is supported by a countable set $\{P_1, P_2, \dots\} \subset \mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}})$, we call the P_k with $\nu_P\big(\{P_k\}\big) > 0$ **ergodic components** of P. Otherwise, we say that P has uncountably many ergodic components and call the elements of $\mathrm{supp}(\nu_P)$ ergodic components.

It is well-known that the ergodic decomposition exists and is uniquely determined by P. We see in the following sections that statistical complexity, generative complexity and excess entropy all satisfy the same type of formula when we decompose the measure P into ergodic components. Namely, the complexity of P is the average complexity of its components plus the entropy of the mixture. We take this formula as definition for a subclass of complexity measures.

Definition 4.5. For $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$, denote the ergodic decomposition by $\nu_P \in \mathcal{P}\big(\mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}})\big)$. We call a function $F\colon \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}}) \to \overline{\mathbb{R}}_+$ **entropy-based** if $F(P) = 0$ for all i.i.d. processes P, and F satisfies

$$F(P) \;=\; H(\nu_P) + \int F \,\mathrm{d}\nu_P \qquad \forall P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}}).$$

Remark. a) We need not worry about measurability of F in the previous definition because $H(\nu_P) = \infty$ whenever P has uncountably many ergodic components. Thus $F(P) = \infty$ if ν is not supported by a countable set, and otherwise the integral is actually a countable sum.

b) The assumption that $F(P)$ is zero for i.i.d. processes P is only needed to ensure that F is finite for enough processes. It is a very natural requirement for complexity measures of stochastic processes and often considered the crucial requirement ([FC98b]). It is obviously satisfied for excess entropy, statistical complexity, generative complexity and process dimension.

All entropy-based functionals are non-continuous and using concavity of entropy, we easily obtain that they are concave. Furthermore, entropy-based complexity measures are generically infinite in the sense that the subset $F^{-1}(\infty)$ of processes with infinite complexity contains a dense $\mathcal{G}_\delta$-set.

Proposition 4.6. *Let $F\colon \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}}) \to \overline{\mathbb{R}}_+$ be an entropy-based functional. Then F is concave. Further, the restriction $F\restriction_{F^{-1}(\mathbb{R})}$ of F to the set where it is finite is non-continuous, even in variational topology. If F is in addition lower semi-continuous, $F^{-1}(\infty)$ is a dense $\mathcal{G}_\delta$-set and F is in particular generically infinite.*

Proof. Concavity: Let P_1, $P_2 \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ and $P = \lambda P_1 + (1-\lambda)P_2$ with $0 < \lambda < 1$. If either P_1 or P_2 has uncountably many ergodic components, the same is true for P and $F(P) = \infty$. Let $P_1 = \sum_k \nu_1(k)\pi_k$ and $P_2 = \sum_k \nu_2(k)\pi_k$ for distinct ergodic $\pi_k \in \mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}})$ (some of the $\nu_i(k)$ may be zero if components occur in only one of the ergodic decompositions). Then

$$\begin{aligned} F(P) &= H\big(\lambda\nu_1 + (1-\lambda)\nu_2\big) + \sum_k \big(\lambda\nu_1(k) + (1-\lambda)\nu_2(k)\big)\cdot F(\pi_k) \\ &\geq \lambda H(\nu_1) + (1-\lambda)H(\nu_2) + \lambda\sum_k \nu_1(k)F(\pi_k) + (1-\lambda)\sum_k \nu_2(k)F(\pi_k) \\ &= F(P_1) + F(P_2). \end{aligned}$$

Non-continuity: Let $P, \pi_n \in F^{-1}(\mathbb{R})$ with $\lim_{n\to\infty} \frac{1}{n}F(\pi_n) \to \infty$. Such π_n exist, because if π_n is a mixture of i.i.d. processes, $F(\pi_n)$ is the entropy of the mixture. Define $P_n := \frac{n-1}{n}P + \frac{1}{n}\pi_n$. Then $P_n \to P$ in variational topology, but $F(P_n) \geq \frac{1}{n}F(\pi_n) \to \infty$ by concavity.

Generic infinity: Due to lower semi-continuity, the sets $F^{-1}\big([0,n]\big)$ are closed and $F^{-1}(\infty)$ is a $\mathcal{G}_\delta$-set. To show that it is dense, let $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ and choose $\pi \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ with $F(\pi) = \infty$ (e.g. an uncountable mixture of i.i.d. processes). Then $F^{-1}(\infty) \ni \frac{n-1}{n}P + \frac{1}{n}\pi \to P$. □

4.2 Properties of excess entropy

Lower semi-continuity of the excess entropy is more or less obvious.

Proposition 4.7 (lower semi-continuity). *The excess entropy $E\colon \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}}) \to \overline{\mathbb{R}}_+$ is weak-$*$ lower semi-continuous.*

Proof. By definition of the mutual information, $E(P) = I(X'_{-\mathbb{N}_0} : X'_{\mathbb{N}})$ is a supremum of continuous functions (in P) and thus lower semi-continuous. □

It is not too difficult to prove concavity of the excess entropy directly. But it also follows from Proposition 4.6 and the ergodic decomposition formula obtained in [Dęb06, Dęb09] by Łukasz Dębowski.

Theorem 4.8 (ergodic decomposition). *The excess entropy is an entropy-based complexity measure, i.e.*

$$E(P) \;=\; \int E \,\mathrm{d}\nu_P + H(\nu_P),$$

where ν_P is the ergodic decomposition of P. In particular, E is concave, non-continuous, and generically infinite.

4.3 Properties of statistical complexity

To analyse statistical complexity $C_{\mathfrak{C}}$, we use the identity $C_{\mathfrak{C}}(P) = H\big(\mu_{\mathfrak{C}}(P)\big)$, where $\mu_{\mathfrak{C}}(P)$ is the effect distribution of P (Section 3.4.2). For the proof of lower semi-continuity, we need a compactness argument. To this end, in the case of infinite Δ, we use the next lemma which guarantees that $\mu_{\mathfrak{C}}$ preserves relative compactness. While our lower semi-continuity result only covers the case of countable Δ, the fact that $\mu_{\mathfrak{C}}$ preserves relative compactness holds for all Polish spaces.

Lemma 4.9. *Let Δ be a Polish space and $\mathfrak{M} \subseteq \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ relatively compact. Then $\mu_{\mathfrak{C}}(\mathfrak{M}) := \big\{\, \mu_{\mathfrak{C}}(P) \;\big|\; P \in \mathfrak{M} \,\big\}$ is relatively compact in $\mathcal{P}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big)$.*

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

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

Because f_n is bounded from above by 1, we have $\int f_n \,\mathrm{d}P \leq 1 - \frac{1}{n} P\big(\{\, f_n < 1 - \frac{1}{n} \,\}\big)$ and obtain

$$P\Big(\bigcup_{n\in\mathbb{N}} \{\, f_n < 1 - \tfrac{1}{n} \,\}\Big) \;\leq\; \sum_{n\in\mathbb{N}} n\Big(1 - \int f_n \,\mathrm{d}P\Big) \;\leq\; \sum_{n\in\mathbb{N}} \varepsilon\, 2^{-n} \;=\; \varepsilon.$$

Consequently,

$$\mu_{\mathfrak{C}}(P)(\widetilde{K}) \;=\; P\big(\{\, Z_0 \in \widetilde{K} \,\}\big) \;=\; P\Big(\bigcap_{n\in\mathbb{N}} \{\, f_n \geq 1 - \tfrac{1}{n} \,\}\Big) \;\geq\; 1 - \varepsilon$$

for all $P \in \mathfrak{M}$. We still have to show compactness of $\widetilde{K}$. It is closed because $z_k \overset{*}{\rightharpoonup} z$ 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 4.10 (lower semi-continuity). *Let Δ be countable. Then the statistical complexity $C_{\mathfrak{C}} \colon \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}}) \to \overline{\mathbb{R}}_+$ is weak-$*$ lower semi-continuous.*

Proof. Let $P_n, P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ with $P_n \overset{*}{\rightharpoonup} P$. According to Lemma 4.9, every subsequence of $\big(\mu_{\mathfrak{C}}(P_n)\big)_{n\in\mathbb{N}}$ has an accumulation point (a.p.). Consequently,

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

Every $\mu_{\mathfrak{C}}(P_n)$ is S-invariant. According to Theorem 3.61, 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 resultant $r \colon \mathcal{P}\big(\mathcal{P}(\Delta^{\mathbb{N}})\big) \to \mathcal{P}(\Delta^{\mathbb{N}})$ is continuous, and thus ν represents P. Therefore, according to Proposition 3.53, $H(\nu) \geq C_{\mathfrak{C}}(P)$. In total we obtain

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

As we see in the next theorem, statistical complexity is, just as excess entropy, an entropy-based complexity measure. To obtain the ergodic decomposition formula, we first show that the effect distribution is the average of the effect distributions of the ergodic components. For the proof of the formula for statistical complexity, it would be sufficient to consider the case of countably many ergodic components, but the identity holds in general. Recall that $\mu_{\mathfrak{C}}(P) = P \circ Z_0^{-1}$, where $Z_0 = P(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0})$.

Lemma 4.11. *Let $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ with ergodic decomposition ν_P. Then*

$$\mu_{\mathfrak{C}}(P) \;=\; \int \mu_{\mathfrak{C}} \,\mathrm{d}\nu_P.$$

Proof. Define the function $\xi \colon \Delta^{\mathbb{Z}} \to \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ by

$$\xi(x) \;=\; \begin{cases} \lim\limits_{n\to\infty} \frac{1}{n} \sum\limits_{k=1}^{n} \delta_{\sigma^{-k}(x)} & \text{if the limit exists in the weak-}* \text{ topology,} \\ P & \text{otherwise.} \end{cases}$$

Because $\mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ is a separable, metrisable space, the function ξ is measurable. Together with the fact that $\mathfrak{B}(\Delta^{\mathbb{Z}})$ is countably generated and setwise convergence implies weak-$*$ convergence, Birkhoff's ergodic theorem yields

$$\xi(x) \;=\; \pi \qquad \pi\text{-a.s.} \qquad \forall \pi \in \mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}}). \tag{4.1}$$

Choose a jointly measurable version of conditional probability given $X'_{-\mathbb{N}_0}$ (i.e. $\pi(\,\cdot \mid X'_{-\mathbb{N}_0})(x)$ is measurable in (π, x)), which is possible according to Lemma A.1. Define $F \colon \Delta^{\mathbb{Z}} \to \mathcal{P}(\Delta^{\mathbb{N}})$ by

$$F(x) \;:=\; \xi(x)\big(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0}\big)(x).$$

ξ is $\sigma(X'_{-\mathbb{N}_0})$-measurable: Fix any $d \in \Delta$ and define $h(x) = y$ by $y_k = x_k$ for $k \leq 0$ and $y_k = d$ for $k > 0$. Then $h \colon \Delta^{\mathbb{Z}} \to \Delta^{\mathbb{Z}}$ is $\sigma(X'_{-\mathbb{N}_0})$-measurable and we claim $\xi = \xi \circ h$. Indeed, for the Kantorovich-Rubinshtein metric d_{KR} on $\mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$, we have $d_{\mathrm{KR}}(\delta_{\sigma^{-k}(x)}, \delta_{\sigma^{-k}(h(x))}) \overset{k\to\infty}{\longrightarrow} 0$ and $\xi(x) = \xi\big(h(x)\big)$ follows.

F is a version of $P(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0})$: F is $\sigma(X'_{-\mathbb{N}_0})$-measurable, because ξ has this measurability and the version of conditional probability is jointly measurable. For $B \in \sigma(X'_{-\mathbb{N}_0})$ we obtain

$$\begin{aligned}\int_B F \,\mathrm{d}P &= \int_{\pi\in\mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}})} \int_B F \,\mathrm{d}\pi \,\mathrm{d}\nu_P \overset{(4.1)}{=} \int_{\pi\in\mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}})} \int_B \pi(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0}) \,\mathrm{d}\pi \,\mathrm{d}\nu_P \\ &= \int_{\pi\in\mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}})} \pi\big(B \cap \{\, X'_{\mathbb{N}} \in \cdot\,\}\big) \,\mathrm{d}\nu_P = P\big(B \cap \{\, X'_{\mathbb{N}} \in \cdot\,\}\big).\end{aligned}$$

$\mu_{\mathfrak{C}}(P) = \int \mu_{\mathfrak{C}} \,\mathrm{d}\nu_P$: From the above, we obtain $\mu_{\mathfrak{C}}(P) = P \circ F^{-1}$. Thus,

$$\mu_{\mathfrak{C}}(P) = \int_{\pi\in\mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}})} \pi \circ F^{-1} \,\mathrm{d}\nu_P \overset{(4.1)}{=} \int_{\pi\in\mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}})} \pi \circ \big(\pi(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0})\big)^{-1} \,\mathrm{d}\nu_P = \int \mu_{\mathfrak{C}} \,\mathrm{d}\nu_P. \qquad \square$$

Theorem 4.12 (ergodic decomposition). *Statistical complexity is an entropy-based complexity measure, i.e.*

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

where ν_P is the ergodic decomposition of P. In particular, $C_{\mathfrak{C}}$ is concave, non-continuous, and generically infinite.

Proof. First note that $\mu_{\mathfrak{C}}(P_1)$ and $\mu_{\mathfrak{C}}(P_2)$ are singular for distinct ergodic $P_1, P_2 \in \mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}})$, because there exist disjoint $A_1, A_2 \in \sigma(X'_{\mathbb{N}})$ with $P_k(A_k) = 1$, $k = 1, 2$. If ν_P is not supported by a countable set, $\mu_{\mathfrak{C}}(P)$ cannot be supported by a countable set and $C_{\mathfrak{C}}(P) = H(\nu_P) = \infty$. 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 Lemma 4.11 implies

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

4.4 Properties of generative complexity

We obtain the corresponding results also for generative complexity which was introduced in Section 3.3.2.

Theorem 4.13 (lower semi-continuity). *The generative complexity $C_{\mathrm{HMM}} \colon \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}}) \to \overline{\mathbb{R}}_+$ is weak-$*$ lower semi-continuous.*

Proof. We have to show $C_{\mathrm{HMM}}(P) \leq \liminf_{n\to\infty} C_{\mathrm{HMM}}(P_n)$ for any convergent sequence $P_n \overset{*}{\rightharpoonup} P$ in $\mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$. Assume w.l.o.g. $C_{\mathrm{HMM}}(P_n) < h < \infty$ for some h and let (T_n, μ_n) be an HMM of P_n with entropy $H(\mu_n) \leq h$. Denote the set of internal states by Γ_n. We construct an HMM (T, μ) of P with $H(\mu) \leq \liminf_{n\to\infty} H(\mu_n)$. We may assume that $\Gamma_n = \mathbb{N}$ and reorder the internal states such that $\mu_n(k+1) \leq \mu_n(k)$ for all $n, k \in \mathbb{N}$. Set $\Gamma := \mathbb{N}$.

1. *Construction of μ:* Let $p_n := \mu_n\big([N, \infty[\big)$. Monotonicity of μ_n yields the entropy estimate

$$H(\mu_n) \geq p_n \cdot \inf_{k\geq N} -\log\big(\mu_n(k)\big) = -p_n \cdot \log\big(\mu_n(N)\big) \geq -p_n \log\left(\frac{1-p_n}{N}\right) \geq p_n \log(N).$$

Consequently, $p_n \leq \frac{h}{\log(N)}$, and for $N \to \infty$, p_n converges to zero uniformly in n. Thus the sequence $(\mu_n)_{n\in\mathbb{N}}$ is uniformly tight and, by passing to a subsequence, we may assume that it is convergent, i.e. there is a $\mu \in \mathcal{P}(\mathbb{N})$ with $\mu_n \overset{*}{\rightharpoonup} \mu$.

2. *Construction of T:* Fix $k \in \operatorname{supp}(\mu)$. We show that the sequence $\big(T_n(k)\big)_{n\in\mathbb{N}}$ is uniformly tight. Then, by passing to a subsequence, we may assume that T_n converges pointwise on $\operatorname{supp}(\mu)$ to some $T\colon \mathbb{N} \to \mathcal{P}(\Delta \times \mathbb{N})$. For sufficiently large n, $\mu_n(k) \geq \frac{1}{2}\mu(k) =: a$. For $\varepsilon > 0$, choose N s.t. $p_n = \mu_n\big([N,\infty[\big) \leq \frac{1}{2}\varepsilon a$ for all n. Because μ_n is T_n-invariant,

$$T_n(k;\, \Delta \times [N,\infty[) \;\leq\; \frac{p_n}{\mu_n(k)} \;\leq\; \tfrac{1}{2}\varepsilon.$$

From $P_n \overset{*}{\rightharpoonup} P$, we deduce that $P_n \circ {X'_1}^{-1}$ is uniformly tight (we do not need completeness of $\Delta^{\mathbb{Z}}$, because P_n is a *sequence* of Radon measures, [Bog07, Thm. 8.6.4]). Thus, there is a compact $D \subseteq \Delta$ with $P_n(\{\, X'_1 \in D \,\}) \geq 1 - \frac{1}{2}\varepsilon a$ for all n. Therefore, for large n,

$$T_n\big(k;\, (\Delta \setminus D) \times \mathbb{N}\big) \;\leq\; \frac{\varepsilon a}{2\mu_n(k)} \;\leq\; \tfrac{1}{2}\varepsilon,$$

and $K := D \times [1, N]$ is the desired compactum with $T_n(k;\, K) \geq 1 - \varepsilon$. Thus, we may assume $T_n(k) \overset{*}{\rightharpoonup} T(k)$ for $k \in \operatorname{supp}(\mu)$.

3. *μ is T-invariant:* Because Γ is countable with discrete topology, weak-$*$ convergence in $\mathcal{P}(\Gamma)$ implies convergence in variational norm. Pointwise convergence of T_n on $\operatorname{supp}(\mu)$ and variational convergence of μ_n, together with T_n-invariance of μ_n, yield for $B \subseteq \Gamma$

$$\mu(B) \;=\; \lim_{n\to\infty} \mu_n(B) \;=\; \lim_{n\to\infty} \int T_n(\,\cdot\,;\, \Delta \times B)\, \mathrm{d}\mu_n \;=\; \int T(\,\cdot\,;\, \Delta \times B)\, \mathrm{d}\mu.$$

4. *(T, μ) is an HMM of P:* We have to show convergence of the output process P_n of (T_n, μ_n) to that of (T, μ). Assume for the moment $\operatorname{supp}(\mu) = \Gamma$ and consider the joint internal and output processes. Because Γ is discrete and the Markov transition kernel depends only on the Γ-component, the conditions of Theorem 5 in [Kar75] are satisfied. The theorem states that the Markov processes converge in weak-$*$ topology. This implies in particular convergence of the output processes. We see from the proof of [Kar75, Thm. 1] that $\operatorname{supp}(\mu) \neq \Gamma$ does not lead to problems: Because $\mu_n\big(\operatorname{supp}(\mu)\big) \to 1$ and the processes are stationary, we can choose the compact set A_ε (used in the cited proof) for large enough n as subset of $\big(\Delta \times \operatorname{supp}(\mu)\big)^k$.

5. *$H(\mu) \leq \liminf_{n\to\infty} H(\mu_n)$:* This follows from lower semi-continuity of H. □

With the same proof, we also obtain that a sequence of invariant HMMs (T_n, μ_n) with converging internal entropy $H(\mu_n)$ and a common output process $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ can be used to construct an HMM (T, μ) of P with entropy $H(\mu) = \lim_{n\to\infty} H(\mu_n)$. This means that the infimum in the definition of generative complexity is actually a minimum, and we obtain the following corollary to the proof of Theorem 4.13.

Corollary 4.14. *Let $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$. Then there exists an invariant HMM (T, μ) of P with*

$$H(\mu) \;=\; C_{\mathrm{HMM}}(P).$$

Theorem 4.15 (ergodic decomposition). *The generative complexity is an entropy-based complexity measure, i.e.*

$$C_{\mathrm{HMM}}(P) \;=\; \int C_{\mathrm{HMM}}\, \mathrm{d}\nu_P + H(\nu_P),$$

where ν_P is the ergodic decomposition of P. In particular, C_{HMM} is concave, non-continuous, and generically infinite.

Proof. "$\leq$": Assume that $\nu := \nu_P$ is supported by a countable set. If not, the right-hand side is infinite and the inequality is trivially satisfied. Let (T_k, μ_k) be an HMM of the ergodic component P_k with set Γ_k of internal states, $k \in I \subseteq \mathbb{N}$. Let $\Gamma := \biguplus_k \Gamma_k$ be the disjoint union. Then we obtain an HMM (T, μ) of P with set Γ of internal states as follows. We identify T_k with a kernel to $\Delta \times \Gamma$ (and support in $\Delta \times \Gamma_k$), and μ_k with a measure on Γ (and support in Γ_k) in the obvious way. Then we can define $T(g) := T_k(g)$ if $g \in \Gamma_k$ and $\mu := \sum_k \nu(k)\mu_k$. Obviously, (T, μ) is an HMM of P with $H(\mu) = \sum_k H(\mu_k) + H(\nu)$.

"$\geq$": Let (T, μ) be an HMM of P with countable set Γ of internal states. If there is no such HMM, the left-hand side is infinite and the inequality is trivially satisfied. We decompose the joint process of internal states and output symbols into ergodic components and note that projections of ergodic components are ergodic as well. Because the internal process is stationary and Markov, the ergodic components correspond to a decomposition of the internal states, $\Gamma = \biguplus_i \Lambda_i$, such that the i^{th} ergodic component visits only internal states in Λ_i. Furthermore, every ergodic component of the joint process projects to one of the ergodic components P_k of P and every P_k is reached by at least one such projection. Let $I(k)$ be the set of indices i, s.t. the i^{th} component is projected to P_k. Let $\Gamma_k = \biguplus_{i \in I(k)} \Lambda_i$ and decompose $\mu = \sum_k \hat{\nu}(k)\mu_k$ with probability measures μ_k on Γ_k, more precisely $\operatorname{supp}(\mu_k) \subseteq \Gamma_k$. It is evident that (T, μ_k) is an HMM of P_k, and $\hat{\nu}(k) = \nu(k)$. We obtain

$$H(\mu) \;=\; H(\nu) + \sum_k \nu(k) \cdot H(\mu_k) \;\geq\; H(\nu) + \sum_k \nu(k) \cdot C_{\text{HMM}}(P_k). \qquad \square$$

4.5 Properties of process dimension

Observable operator models are generative algebraic models, representing stochastic processes. The natural measure of the size of an OOM is its dimension, and we already identified the minimal dimension of an OOM, which is the dimension of the canonical OOM vector space V_P, as a characteristic of the process called process dimension (Definition 2.35). These facts suggest to consider the process dimension as candidate of a complexity measure for stochastic processes. Here, we give a further indication that this might be appropriate, namely we prove weak-$*$ lower semi-continuity and concavity. Let Δ be countable and recall that the canonical OOM vector space of $P \in \mathcal{P}(\Delta^{\mathbb{N}})$ is

$$V_P \;=\; \operatorname{span}(Q_P) \qquad \text{with} \qquad Q_P \;=\; \big\{\, \tau_{d_1 \cdots d_n}(P) \;\big|\; n \in \mathbb{N}_0,\; d_1, \ldots, d_n \in \Delta \,\big\},$$

and, for $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$, $Q_P = Q_{P_{\mathbb{N}}}$ and $V_P = V_{P_{\mathbb{N}}}$.

Theorem 4.16 (lower semi-continuity). *Let Δ be countable. Then the process dimension* $\dim\colon \mathcal{P}(\Delta^{\mathbb{N}}) \to \mathbb{N} \cup \{\infty\}$ *is weak-$*$ lower semi-continuous.*

Proof. Let $P_n \overset{*}{\rightharpoonup} P$ in $\mathcal{P}(\Delta^{\mathbb{N}})$ and $\dim(P) \geq m$. We have to show $\dim(P_n) \geq m$ for sufficiently large n. Let V_P be the vector space of the canonical OOM (Definition 2.33). Because $\dim(V_P) = \dim(P)$ and $V_P = \operatorname{span}(Q_P)$, we can pick linearly independent $v_1, \ldots, v_m \in Q_P$ and $d_{ki} \in \Delta$, $k \in \{1, \ldots, m\}$, $i \in \{1, \ldots, N_k\}$ with $v_k = \tau_{d_{k1} \ldots d_{kN_k}}(P)$. Define

$$v_k^n \;:=\; \tau_{d_{k1} \ldots d_{kN_k}}(P_n) \;\in\; V_{P_n}.$$

Then, due to continuity of τ_d, $v_k^n \overset{*}{\rightharpoonup} v_k$. If $v_1^n, \ldots, v_m^n$ are linearly independent for all sufficiently large n, the proof is finished. Suppose for a contradiction that this is not the case and

w.l.o.g. that they are dependent for all n. Then there are $\lambda_k^n \in [-1,1]$ with $\max_k |\lambda_k^n| = 1$ and $\sum_k \lambda_k^n v_k^n = 0$ for all n. Because $[-1,1]^m$ is compact, we may assume by passing to a subsequence that $\lambda_k^n \overset{n\to\infty}{\longrightarrow} \lambda_k$ for some λ_k. Due to weak-$*$ continuity of addition and scalar multiplication, $\sum_k \lambda_k v_k = 0$ and hence linear independence of the v_k yields $\lambda_k = 0$ for all k. This is a contradiction to $\max_k |\lambda_k^n| = 1$. □

Of course, the process dimension is not entropy-based. If two different processes share the same ergodic components, they also have the same process dimension. In this sense, process dimension is not quantitative. It contains only qualitative information and is insensitive to the probability with which the components occur. The dimension of a process is just the sum of the dimensions of its ergodic components. This is not too surprising, because ergodic measures are mutually singular.

Theorem 4.17 (ergodic decomposition). *Let Δ be countable and $P \in \mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$ with ergodic decomposition $\nu_P \in \mathcal{P}\big(\mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}})\big)$. Then*

$$\dim(P) \;=\; \sum_{\pi \in \operatorname{supp}(\nu_P)} \dim(\pi),$$

where the sum is infinite whenever $\operatorname{supp}(\nu_P)$ is an infinite set. In particular, the process dimension is concave.

Proof. We use that the process dimension is equal to the dimension of the effect space, $\dim(P) = \dim(\mathfrak{S}_P)$, by Corollary 3.57. Here, the dimension of a set is the dimension of the generated vector space. Recall that $\mathfrak{S}_P = \operatorname{supp}\big(\mu_{\mathfrak{C}}(P)\big)$ and, by Lemma 4.11,

$$\mu_{\mathfrak{C}}(P) \;=\; \int \mu_{\mathfrak{C}} \,\mathrm{d}\nu_P. \tag{4.2}$$

1. *Case of finitely many ergodic components:* Let $P_1, \dots, P_n$ be the ergodic components of P. Choose disjoint $A_1, \dots, A_n \in \mathfrak{B}(\Delta^{\mathbb{N}})$ with $P_k\big(\{\, X'_{\mathbb{N}} \in A_k \,\}\big) = 1$ for $k = 1, \dots, n$. Then $P_k\big(\{\, X'_{\mathbb{N}} \in A_k \,\} \bigm| X'_{-\mathbb{N}_0}\big) = 1$ P_k-a.s. Because

$$\mathfrak{M}_k \;:=\; \{\, \pi \in \mathcal{P}(\Delta^{\mathbb{N}}) \mid \pi(A_k) = 1 \,\}$$

is closed, this implies $\mathfrak{S}_{P_k} \subseteq \mathfrak{M}_k$. The family $\operatorname{span}(\mathfrak{M}_k)$, $k = 1, \dots, n$, of vector spaces is obviously linearly independent, and thus the vector spaces $V_k := \operatorname{span}(\mathfrak{S}_{P_k})$ are linearly independent as well. Consequently,

$$\dim(P) \;=\; \dim(\mathfrak{S}_P) \overset{(4.2)}{=} \dim\Big(\bigcup_{k=1}^{n} \mathfrak{S}_{P_k}\Big) \;=\; \sum_k \dim(\mathfrak{S}_{P_k}) \;=\; \sum_k \dim(P_k).$$

2. *Case of infinitely many ergodic components:* In this case, the right-hand side is infinite. $\mathfrak{S}_P = \operatorname{supp}\big(\int \mu_{\mathfrak{C}} \,\mathrm{d}\nu_P\big)$ satisfies $\mu_{\mathfrak{C}}(\pi)(\mathfrak{S}_P) = 1$ for ν_P-almost all π, and thus $\mathfrak{S}_P \supseteq \mathfrak{S}_\pi$ for infinitely many π. Because the vector spaces generated by the $\mathfrak{S}_\pi$ are linearly independent (step 1.), this implies that $\dim(\mathfrak{S}_P) = \infty$. Thus the left-hand side is infinite as well.

3. *Concavity:* Because $P \mapsto \nu_P$ is linear, we obtain for $0 < \lambda < 1$ and $P = \lambda P_1 + (1-\lambda) P_2$ that $\operatorname{supp}(\nu_P) = \operatorname{supp}(\nu_{P_1}) \cup \operatorname{supp}(\nu_{P_2})$ and thus $\dim(P) \geq \max\{\, \dim(P_1), \dim(P_2) \,\}$. □

4.6 Open problems

We have seen that in general $E(P) \leq C_{\mathrm{HMM}}(P) \leq C_{\mathfrak{C}}(P)$. How does the process dimension $\dim(P)$ – or rather $\log\big(\dim(P)\big)$ which is more comparable to the entropy based quantities – fit into this line? On one hand, we see from the ergodic decomposition formulas that the process dimension of a mixture of finitely many i.i.d. processes is the number of components, while the statistical complexity is the entropy of the mixture. Thus, $\log\big(\dim(P)\big)$ can be greater than $C_{\mathfrak{C}}(P)$. On the other hand, there are examples of processes, where the minimal number of internal states of any HMM is substantially larger than the process dimension. It is straightforward to check that in the corresponding example of [Jae00], also the entropy has to be larger than the logarithm of the process dimension, i.e. $\log\big(\dim(P)\big)$ can be smaller than $C_{\mathrm{HMM}}(P)$. This means that $\log\big(\dim(P)\big)$ is neither comparable to $C_{\mathrm{HMM}}(P)$ nor to $C_{\mathfrak{C}}(P)$ in general. The situation might be different for excess entropy. We consider it an interesting problem to clarify the relation between process dimension and excess entropy, in particular to find out whether $E(P) \leq \log\big(\dim(P)\big)$ holds in general.

From a theoretical point of view, it seems unsatisfactory that all of the investigated complexity measures are generically infinite and thus do not distinguish between "most" processes. A future goal would be to obtain modified versions that give meaningful results for all (or a generic set of) processes.

Another possible line of research is to extend the definitions of the complexity measures to non-commutative probability theory. Generalisations of OOMs and HMMs to the setting of states on quasi-local C*-algebras already exist under the names of *finitely correlated states* and *C*-finitely correlated states*, respectively ([FNW92]). Because the definitions of causal states, ε-machine and statistical complexity rely heavily on conditional probabilities, it seems to be much more difficult to obtain corresponding generalisations of these terms.

Appendix A

Technical background

A.1 Souslin spaces

We list a few important properties of Souslin spaces (Definition 2.10).

- Every Polish space is a Souslin space.
- Countable products of Souslin spaces are Souslin spaces.
- Every measurable subspace of a Souslin space is a Souslin space ([Bog07, Cor. 6.6.7]).
- If Δ is a Souslin space, $\mathcal{P}(\Delta)$ is also a Souslin space ([Bog07, Thm. 8.9.6]).
- The image of a Souslin space under a measurable map into a separable, metrisable space is a Souslin space ([Bog07, Thm. 6.7.3]).
- Every Souslin subset of a Hausdorff space Γ is universally measurable, i.e., for every $\mu \in \mathcal{P}(\Gamma)$, it is measurable w.r.t. the μ-completion of $\mathcal{G} = \mathfrak{B}(\Gamma)$ ([Bog07, Thm. 7.4.1]).
- Every Souslin space is separable and the Borel σ-algebra is countably generated.
- Every Souslin space is a Radon space, i.e. every probability measure on it is Radon ([Bog07, Thm. 7.4.3]). In particular, all conditional probabilities have regular versions ([Bog07, Thm. 10.4.5]).
- Every measurable bijection between Souslin spaces is a Borel isomorphism, i.e. the inverse map is measurable ([Coh80, Prop. 8.6.2]).

A.2 Extension theorems

Let $(\Delta, \mathcal{D})$ be a measurable space. When we define the distribution $P \in \mathcal{P}(\Delta^{\mathbb{N}})$ of a Δ-valued stochastic process, we usually specify only the finite-dimensional distributions explicitly. Of course, the finite-dimensional distributions $P_{[1,n]} \in \mathcal{P}(\Delta^n)$ have to be consistent in the sense that

$$P_{[1,n]}(A) = P_{[1,n+1]}(A \times \Delta) \qquad \forall A \in \bigotimes_1^n \mathcal{D}.$$

We then need an extension theorem to ensure existence and uniqueness of a measure $P \in \mathcal{P}(\Delta^{\mathbb{N}})$ with the given finite-dimensional marginals. Because finite-dimensional cylinder sets form an intersection stable generator of the product σ-algebra, uniqueness is always satisfied. To prove existence, however, technical assumptions are required. Here, we recall two extension results. The first one is the celebrated Kolmogorov extension theorem ([AB99, Thm. 14.26], [Bog07, Thm. 7.7.1]), which uses the topological assumption that the measures $P_{[1,n]}$ are Radon measures. In particular, this assumption is satisfied if Δ is a Souslin space. The second extension theorem is Ionescu-Tulcea's ([Nev65, Prop. V.1.1], [Bog07, Thm. 10.7.3]), which is free of topological assumptions. Instead, it requires the existence of Markov kernels T_n from Δ^n to Δ such that

$$P_{[1,n+1]} \;=\; P_{[1,n]} \otimes T_n.$$

This amounts to requiring the existence of regular versions of conditional probability, which is in particular guaranteed if Δ is a Souslin space.

A.3 Conditional probabilities

We provide some useful lemmata about conditional probabilities. They are presumably well-known, but easier to prove than to locate in the literature. The first lemma states that in a Souslin space Γ, it is possible to choose for any countably generated σ-algebra $\mathcal{F}$ regular versions of conditional probability given $\mathcal{F}$ that depend measurably on the probability measure. The case of a Polish space Γ is proven in [Kni92, Thm. 1.5], using the Kuratowski isomorphism form Γ onto $[0,1]$. Our proof is similar to the proof of the case $\Gamma = [0,1]$ in [Kni92], but avoids using the special structure of $[0,1]$ and cumulative distribution functions. Note that it is essential that $\mathcal{F}$ is countably generated.

Lemma A.1. *Let Γ be a Souslin space and $\mathcal{F} \subseteq \mathcal{G} = \mathfrak{B}(\Gamma)$ a countably generated sub-σ-algebra. Then there exists a regular, $\mathfrak{B}\big(\mathcal{P}(\Gamma)\big) \otimes \mathcal{F}$-measurable conditional probability given $\mathcal{F}$, i.e. a measurable $\mathcal{E}\colon \mathcal{P}(\Gamma) \times \Gamma \to \mathcal{P}(\Gamma)$, such that $\mathcal{E}(\mu, \cdot\,)(A)$ is a version of $\mu(A \mid \mathcal{F})$ for all $\mu \in \mathcal{P}(\Gamma)$ and $A \in \mathfrak{B}(\Gamma)$.*

Proof. Let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be an increasing sequence of finite σ-algebras with $\sigma\big(\bigcup \mathcal{F}_n\big) = \mathcal{F}$. Then we can obviously choose jointly measurable, regular versions $\mathcal{E}_n$ of conditional probability given $\mathcal{F}_n$ in an elementary way. Define for $\mu \in \mathcal{P}(\Gamma)$, $x \in \Gamma$

$$\mathcal{E}(\mu, x) \;:=\; \begin{cases} \lim\limits_{n\to\infty} \mathcal{E}_n(\mu, x) & \text{if the limit exists in the weak-}* \text{ topology on } \mathcal{P}(\Gamma), \\ \mu & \text{otherwise.} \end{cases}$$

The set where the limit exists is measurable, and because $\mathcal{P}(\Gamma)$ is a separable, metrisable space, the function $\mathcal{E}$ is $\mathfrak{B}\big(\mathcal{P}(\Gamma)\big) \otimes \mathcal{F}$-measurable. Now fix $\mu \in \mathcal{P}(\Gamma)$ and let $\mu(\cdot \mid \mathcal{F})$ be a regular version of conditional probability (it exists because Γ is a Souslin space). Further let $(G_k)_{k\in\mathbb{N}}$ be a countable, intersection stable generator of $\mathcal{G}$. Then outside a fixed set of μ-measure zero

$$\mu(G \mid \mathcal{F}) \;=\; \lim_{n\to\infty} \mathcal{E}_n(\mu, \cdot\,)(G)$$

for all $G = G_k$, $k \in \mathbb{N}$. But the set of $G \in \mathcal{G}$ for which the equality holds is a Dynkin system (due to regularity of $\mu(\cdot \mid \mathcal{F})$) and thus the equation is valid for all $G \in \mathcal{G}$. In particular, this implies

$$\mu(\cdot \mid \mathcal{F})(x) \;=\; \lim_{n\to\infty} \mathcal{E}_n(\mu, x) \;=\; \mathcal{E}(\mu, x) \qquad \mu\text{-a.s.}$$

in the weak-$*$ topology, and $\mathcal{E}(\mu, \cdot\,)$ is a regular conditional probability of μ given $\mathcal{F}$. □

As a corollary, we obtain that first conditioning on one σ-algebra, and then, after changing ones expectations to the result, conditioning on another σ-algebra is equivalent to conditioning on both at the same time.

Corollary A.2 (iterated conditional probability). *Let Γ be a Souslin space, $\mu \in \mathcal{P}(\Gamma)$ and $\mathcal{F}_1, \mathcal{F}_2 \subseteq \mathcal{G}$ sub-σ-algebras. Further assume that $\mathcal{F}_2$ is countably generated. Let $\mu_{\mathcal{F}_1}(x) := \mu(\,\cdot \mid \mathcal{F}_1)(x)$ be a regular conditional probability. Then*

$$\big(\mu_{\mathcal{F}_1}(x)\big)\big(\,\cdot \;\big|\; \mathcal{F}_2\big)(x) \;=\; \mu\big(\,\cdot \;\big|\; \mathcal{F}_1 \vee \mathcal{F}_2\big)(x) \qquad \mu\text{-a.s.},$$

where $\mathcal{F}_1 \vee \mathcal{F}_2 := \sigma(\mathcal{F}_1 \cup \mathcal{F}_2)$.

Proof. Measurability: Choose a jointly measurable, regular conditional probability given $\mathcal{F}_2$ according to Lemma A.1. Then the left-hand side is $\mathcal{F}_1 \vee \mathcal{F}_2$-measurable.
Mean values: Let $G \in \mathcal{G}$, $F_1 \in \mathcal{F}_1$ and $F_2 \in \mathcal{F}_2$ and denote by $\mathrm{E}^{\mu_{\mathcal{F}_1}(x)}$ expectation w.r.t. the measure $\mu_{\mathcal{F}_1}(x)$. Then

$$\begin{aligned}\int_{x \in F_1 \cap F_2} \mu_{\mathcal{F}_1}(x)\big(G \;\big|\; \mathcal{F}_2\big)(x)\,\mathrm{d}\mu \;&=\; \int_{x \in F_1} \mathrm{E}^{\mu_{\mathcal{F}_1}(x)}\big(1_{F_2} \cdot \mu_{\mathcal{F}_1}(x)\big(G \;\big|\; \mathcal{F}_2\big)\big)\,\mathrm{d}\mu \\ &=\; \int_{F_1} \mathrm{E}^{\mu_{\mathcal{F}_1}}(1_{F_2} 1_G)\,\mathrm{d}\mu \;=\; \int_{F_1 \cap F_2} 1_G\,\mathrm{d}\mu.\end{aligned}$$

Because $\mathcal{G}$ is countably generated and the conditional probabilities are regular versions, the claim follows. □

The following two lemmata show very intuitive properties of conditional probabilities. Namely, the conditional probability of X given Y is zero-one valued if and only if X is a function of Y. And if X depends on Y only through Z, then knowledge of the conditional probability $\mathbb{P}(Z \mid Y)$ is as good as knowledge of Y.

Lemma A.3. *Let X, Y be random variables with values in Souslin spaces Γ and Γ' respectively. The following properties are equivalent.*

1. $\mathbb{P}\big(\{\, X \in G \,\} \;\big|\; Y\big) \in \{\, 0, 1 \,\}$ *a.s.* $\forall G \in \mathcal{G}$ *2.* $\mathbb{P}(X \mid Y)(\omega) = \delta_{X(\omega)}$ *a.s.*

3. $X = f \circ Y$ *a.s. for a measurable map* $f \colon \Gamma' \to \Gamma$

Proof. Assume w.l.o.g. that Ω is a Souslin space (e.g. $\Omega = \Gamma \times \Gamma'$). "3. ⇒ 1." is trivial.
"1. ⇒ 2.": Fix $G \in \mathcal{G}$ and let $A := \big\{\, \mathbb{P}\big(\{\, X \in G \,\} \;\big|\; Y\big) = 1 \,\big\}$. Note that $A \in \sigma(Y)$. We have

$$\mathbb{P}(A) \;=\; \int_A 1_A\,\mathrm{d}\mathbb{P} \;=\; \int_A \mathbb{P}\big(\{\, X \in G \,\} \;\big|\; Y\big)\,\mathrm{d}\mathbb{P} \;=\; \mathbb{P}\big(A \cap X^{-1}(G)\big)\,.$$

Similarly, $\mathbb{P}(A) = \mathbb{P}\big(X^{-1}(G)\big)$ and thus $1_A = 1_G \circ X$ a.s. Because $\mathcal{G}$ is countably generated, we can choose the exception set independently of G, and 2. follows.
"2. ⇒ 3.": Because $1_G \circ X$ is $\sigma(Y)$-measurable modulo $\mathbb{P}$ and the singletons in $\mathcal{G}'$ are measurable, there is a measurable map $f \colon \mathrm{Im}(Y) \to \Gamma$ such that $X = f \circ Y$ a.s. Because $\mathrm{Im}(Y)$ is a Souslin set, we can extend f to a universally measurable map on Γ'. Because $\mathcal{G}$ is countably generated, we can modify f on a set of measure zero (w.r.t. $\mathbb{P}_Y$) to obtain a Borel measurable map from Γ' to Γ ([Bog07, Cor 6.5.6]). □

Lemma A.4. *Let X, Y, Z be random variables with $X \perp\!\!\!\perp Y \mid Z$. Then*

$$\mathbb{P}\big(X \,\big|\, \mathbb{P}(Z \mid Y)\big) \;=\; \mathbb{P}(X \mid Y) \qquad \text{a.s.}$$

Proof. Let $\mathcal{F}$ be the σ-algebra generated by $\mathbb{P}(Z \mid Y)$. Note that $\mathcal{F} \subseteq \sigma(Y)$. Using algebraic induction, we see that $\mathrm{E}^{\mathbb{P}}(f \mid Y)$ is $\mathcal{F}$-measurable for every bounded, $\sigma(Z)$-measurable function f. Due to conditional independence, we obtain a.s.

$$\mathbb{P}(X \mid Y) \;=\; \mathrm{E}^{\mathbb{P}}\big(\mathbb{P}(X \mid Y, Z) \,\big|\, Y\big) \;=\; \mathrm{E}^{\mathbb{P}}\big(\mathbb{P}(X \mid Z) \,\big|\, Y\big),$$

and thus $\mathbb{P}(X \mid Y)$ is $\mathcal{F}$-measurable modulo $\mathbb{P}$. □

A.4 Measurable partitions

We show that $\{0,1\}^{-\mathbb{N}_0}$ admits a measurable partition that is not the set of atoms of any countably generated sub-σ-algebra of the Borel σ-algebra. More generally, this is true for every uncountable Polish space.

Lemma A.5. *Let Γ be an uncountable Polish space. Then there is a partition $\pi = \big\{\, [x] \,\big|\, x \in \Gamma \,\big\}$ of Γ into measurable sets $[x] \ni x$, such that π is not the set of atoms of any countably generated sub-σ-algebra of $\mathcal{G}$.*

Proof. According to the Kuratowski theorem ([DM78, Appendix-III.80]), all uncountable Polish spaces are Borel isomorphic. Thus we may assume w.l.o.g. $\Gamma = \mathbb{R}^{\mathbb{N}}$. Let

$$[x] \;:=\; \big\{\, \hat{x} \in \Gamma \,\big|\, \{\, x_k \mid k \in \mathbb{N} \,\} = \{\, \hat{x}_k \mid k \in \mathbb{N} \,\} \,\big\} \;=\; \bigcap_{k\in\mathbb{N}} \bigcup_{n\in\mathbb{N}} \xi_n^{-1}(x_k) \cap \bigcap_{n\in\mathbb{N}} \bigcup_{k\in\mathbb{N}} \xi_n^{-1}(x_k),$$

where $\xi_n \colon \Gamma \to \mathbb{R}$ is the canonical projection onto the n^{th} component. This defines a measurable partition of Γ. Let $\mathcal{F} \subseteq \mathcal{G}$ be a σ-algebra with atoms $[x]$, μ a non-atomic probability measure on $\mathbb{R}$, and $P = \bigotimes_{\mathbb{N}} \mu$ the product measure on Γ. Now $P\big([x]\big) \leq P\big(\bigcup_{k\in\mathbb{N}} \xi_1^{-1}(x_k)\big) \leq \sum_k \mu\big(\{\, x_k \,\}\big) = 0$, thus P is not supported by an atom of $\mathcal{F}$. On the other hand, each atom $[x]$ is invariant under coordinate permutations and this property extends to all $F \in \mathcal{F}$ (F is a union of atoms). Because P is i.i.d., it follows from the Hewitt-Savage 0-1-law that $P\restriction_{\mathcal{F}}$ is $\{0,1\}$-valued. Every $\{0,1\}$-valued measure on a countably generated σ-algebra is a Dirac measure, hence supported by an atom. Therefore, $\mathcal{F}$ cannot be countably generated. □

A.5 Entropy

Lemma A.6. *Let Γ be a separable, metrisable space. Then the entropy $H \colon \mathcal{P}(\Gamma) \to \overline{\mathbb{R}}_+$ is weak-$*$ lower semi-continuous.*

Proof. Let $\mu_n \xrightarrow{*} \mu$ be a convergent sequence in $\mathcal{P}(\Gamma)$ and $G_1, \ldots, G_m$ a measurable partition of Γ. Define $a_k := \mu(G_k)$ and $h := \sum_k \varphi(a_k)$. Measures on metrisable spaces are inner closed-regular, i.e. we can choose closed sets $F_k \subseteq G_k$ with $\mu(F_k) \geq a_k - \varepsilon$. As metrisable spaces are normal, there are disjoint open sets $U_k \supseteq F_k$. The Alexandrov theorem (also called portmanteau theorem; [Bog07, Thm. 8.2.3]) implies $\liminf_{n\to\infty} \mu_n(U_k) \geq \mu(U_k)$. Thus, there is $n(\varepsilon) \in \mathbb{N}$ such that for $n > n(\varepsilon)$,

$$a_k - 2\varepsilon \;\leq\; \mu_n(U_k) \overset{\left(\sum \mu_n(U_k) \leq \sum a_k\right)}{\leq} a_k + 2\varepsilon m \qquad \forall k.$$

Due to uniform continuity of φ, we find for any $\hat{\epsilon} > 0$ an $\varepsilon > 0$, such that for $n > n(\varepsilon)$

$$H(\mu_n) \;\geq\; \sum_{k=1}^{m} \varphi\big(\mu_n(U_k)\big) \;\geq\; h - m \cdot \sup\big\{\, |\varphi(x) - \varphi(y)| \;\big|\; x, y \in [0,1],\; |x-y| \leq 2m\varepsilon \,\big\} \;\geq\; h - \hat{\epsilon}.$$

Consequently, $\liminf_n H(\mu_n) \geq H(\mu)$. □

Lemma A.7. *Let Γ be separable, metrisable and $\mu \in \mathcal{P}(\Gamma)$ satisfy $H(\mu) < \infty$. Then μ is supported by a countable set A and*

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

Proof. It is sufficient to show $H(\mu) = \infty$ for all μ that vanish on singletons. The support of μ exists. Therefore, as μ vanishes on singletons, it has no atoms. Thus, we can partition Γ into two measurable sets with μ-measure $\frac{1}{2}$ each, and recursively into 2^n measurable sets with measure 2^{-n} each. Consequently, $H(\mu) = \infty$. □

Bibliography

[AB99] Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, second edition, 1999.

[AC05] Nihat Ay and James P. Crutchfield. Reductions of hidden information sources. *Journal of Statistical Physics*, 120:659–684, 2005.

[BK57] David Blackwell and Lambert Koopmans. On the identifiability problem for functions of finite Markov chains. *Annals of Mathematical Statistics*, 28, 1957.

[BNT01] W. Bialek, I. Nemenman, and N. Tishby. Predictability, complexity, and learning. *Neural Computation*, 13:2409–2463, 2001.

[Bog07] V. I. Bogachev. *Measure Theory, Volume II*. Springer, 2007.

[Bou89] N. Bourbaki. *General Topology, Chapters 5-10*. Springer-Verlag, 1989.

[BP66] Leonard E. Baum and Ted P. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 37(6), 1966.

[Buk95] Rais G. Bukharaev. *Theorie der stochastischen Automaten*. B.G. Teubner, 1995.

[CF03] James P. Crutchfield and David P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos*, 13(1):25–54, 2003.

[CFW03] 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, Jan 2003.

[Cho69] Gustave Choquet. *Lectures on Analysis, Volume II (Representation Theory)*. W. A. Benjamin, Inc., 1969.

[CMR05] Oliver Cappé, Eric Moulines, and Tobias Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

[Coh80] Donald L. Cohn. *Measure Theory*. Birkhäuser, 1980.

[CP83] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of noisy chaos. *Physica D*, 7:201–223, 1983.

[Cru94] James P. Crutchfield. The calculi of emergence: Computation, dynamics and induction. *Physica D*, 75:11–54, 1994.

[CT91] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley and Sons, Inc., 1991.

[CY89] James P. Crutchfield and Karl Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105–108, 1989.

[Dęb06] Łukasz Dębowski. Ergodic decomposition of excess entropy and conditional mutual information. IPI PAN Reports, nr 993, 2006.

[Dęb09] Łukasz Dębowski. A general definition of conditional information and its application to ergodic decomposition. *Statistics & Probability Letters*, 79:1260–1268, 2009.

[Dha63a] S. W. Dharmadhikari. Functions of finite Markov chains. *Annals of Mathematical Statistics*, 34, 1963.

[Dha63b] S. W. Dharmadhikari. Sufficient conditions for a stationary process to be a function of a finite Markov chain. *Annals of Mathematical Statistics*, 34, 1963.

[Dha65] S. W. Dharmadhikari. A characterisation of a class of functions of finite Markov chains. *Annals of Mathematical Statistics*, 36, 1965.

[DM78] C. Dellacherie and P. Meyer. *Probabilities and Potential*. North-Holland, 1978.

[FC98a] David P. Feldman and James P. Crutchfield. Discovering noncritical organization: Statistical mechanical, information theoretic, and computational views of patterns in one-dimensional spin systems. Santa Fe Institute Working Paper 98-04-026, 1998.

[FC98b] David P. Feldman and James P. Crutchfield. Measures of statistical complexity: Why? *Physics Letters A*, 238:244–252, 1998.

[FNW92] M. Fannes, B. Nachtergaele, and R. F. Werner. Finitely correlated states on quantum spin chains. *Commun. Math. Phys.*, 144(3):443–490, 1992.

[FR68] M. Fox and H. Rubin. Functions of processes with Markovian states. *Annals of Mathematical Statistics*, 39, 1968.

[FS07] Ulrich Faigle and Alexander Schönhuth. Asymptotic mean stationarity of sources with finite evolution dimension. *IEEE Transactions on Information Theory*, 53(7):2342–2348, 2007.

[Gil59] Edgar J. Gilbert. On the identifiability problem for functions of finite Markov chains. *Annals of Mathematical Statistics*, 30, 1959.

[Gra86] P. Grassberger. Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.*, 25:907–938, 1986.

[Hel65] Alex Heller. On stochastic processes derived from Markov chains. *Annals of Mathematical Statistics*, 36:1286–1291, 1965.

[HSF97] J. Henderson, S. Salzberg, and K. H. Fasman. Finding genes in DNA with hidden Markov model. *Journal of Computational Biology*, 4:127–141, 1997.

[HU79] John Hopcroft and Jeffrey Ullman. *Introduction to Automata Theory, Language, and Computation*. Addison-Wesely, Reading, Massachusetts, 1979.

[IAK92] Hisashi Ito, Shun-Ichi Amari, and Kingo Kobayashi. Identifiability of hidden Markov information sources and their minimum degrees of freedom. *IEEE transactions on information theory*, 38(2):324–333, 1992.

[Jae99] Herbert Jaeger. Characterizing distributions of stochastic processes by linear operators. GMD report 62, German National Research Center for Information Technology, 1999. `http://citeseer.ist.psu.edu/jaeger99characterizing.html`.

[Jae00] Herbert Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6):1371–1398, 2000.

[Jel99] Frederick Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1999.

[JWSK07] 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.

[Kak99] Yûichirô Kakihara. *Abstract Methods in Information Theory*. World Scientific Publishing, 1999.

[Kar75] Alan F. Karr. Weak convergence of a sequence of Markov chains. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 33:41–48, 1975.

[Kel98] Gerhard Keller. *Equilibrium States in Ergodic Theory*. London Mathematical Society, 1998.

[Kni75] Frank Knight. A predictive view of continuous time processes. *The Annals of Probability*, 3:573–96, 1975.

[Kni81] Frank Knight. *Essays on the Prediction Process*, volume 1 of *Lecture Notes Series*. Institute of Mathematical Statistics, Hayward, CA, 1981.

[Kni92] Frank Knight. *Foundations of the Prediction Process*. Oxford Science Publications, 1992.

[Kos01] Timo Koski. *Hidden Markov Models for Bioinformatics*, volume 2 of *Computational Biology Series*. Kluwer Academic Publishers, 2001.

[KSK76] John G. Kemeny, J. Laurie Snell, and Anthony W. Knapp. *Denumerable Markov Chains*. Graduate Texts in Mathematics. Springer-Verlag, 1976.

[LA09a] Wolfgang Löhr and Nihat Ay. Non-sufficient memories that are sufficient for prediction. In *Proceedings of Complex'2009, Shanghai*, volume 4 part I of *LNICST*, pages 265–276. Springer, 2009.

[LA09b] Wolfgang Löhr and Nihat Ay. On the generative nature of prediction. *Advances in Complex Systems*, 12(2):169–194, 2009.

[Lau96] Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[Löh09a] Wolfgang Löhr. Predictive models and generative complexity. Submitted to Journal of System Sciences and Complexity, 2009.

[Löh09b] Wolfgang Löhr. Properties of the statistical complexity functional and partially deterministic HMMs. *Entropy*, 11(3):385–401, 2009.

[LSS01] Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. *Advances in Neural Information Processing Systems*, 14, 2001.

[Mey76] P. Meyer. La théorie de la prédiction de F. Knight. *Seminaire de Probabilités*, X:86–103, 1976.

[Nev65] Jacques Neveu. *Mathematical Foundations of the Calculus of Probability*. Holden-Day, INC, 1965.

[OBAJ08] Eckehard Olbrich, Nils Bertschinger, Nihat Ay, and Jürgen Jost. How should complexity scale with system size? *European Physical Journal B*, 63:407–415, 2008.

[Par61] Parthasarathy. On the category of ergodic measures. *Illinois J. Math.*, 5:648–656, 1961.

[SC01] Cosma R. Shalizi and James P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104:817–879, 2001.

[SCE07] Susanne Still, James P. Crutchfield, and Christopher J. Ellison. Optimal causal inference. Informal publication, `http://arxiv.org/abs/0708.1580`, 2007.

[Sch08] Alexander Schönhuth. A simple and efficient solution of the identifiability problem for hidden Markov sources and quantum random walks. In *Proceedings of the International Symposium on Information Theory and its Applications*, 2008.

[SJ09] Alexander Schönhuth and Herbert Jaeger. Characterization of ergodic hidden Markov sources. *IEEE Transactions on Information Theory*, 55(5):2107–2118, 2009.

[SJR04] Satinder Singh, Michael R. James, and Matthew R. Rudary. Predictive state representations: a new theory for modeling dynamical systems. In *AUAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 512–519, Arlington, Virginia, United States, 2004.

[Upp89] Daniel Ray Upper. *Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models*. PhD thesis, University of California, Berkeley, 1989.

[Wie08] Eric Walter Wiewiora. *Modeling Probability Distributions with Predictive State Representations*. PhD thesis, University of California, San Diego, 2008.

Notation

$[D_1 \times \cdots \times D_n]$	Cylinder set $\{ X'_k \in D_k,\ k = 1, \ldots, n \}$ in $\Delta^{\mathbb{Z}}$ or $\Delta^{\mathbb{N}}$
$[d_1, \ldots, d_n]$	Cylinder set $\{ X'_k = d_k,\ k = 1, \ldots, n \}$ in $\Delta^{\mathbb{Z}}$ or $\Delta^{\mathbb{N}}$
$X \perp\!\!\!\perp Y \mid Z$	X is conditionally independent of Y given Z (Def. 3.8)
1_A	Indicator function of a set A, $1_A(x) = \delta_x(A)$
$\mathfrak{B}(\Gamma)$	Borel σ-algebra of a topological space Γ
$\mathfrak{C}(x)$	Causal state (Def. 3.17) of $x \in \Delta^{-\mathbb{N}_0}$
$C_{\mathfrak{C}}(P)$	Statistical complexity (Def. 3.23) of P, $C_{\mathfrak{C}}(P) = H^{\mathbb{P}}(M_{\mathfrak{C}}) = H\big(\mu_{\mathfrak{C}}(P)\big)$
$C_{\mathrm{HMM}}(P)$	Generative complexity (Def. 3.38) of P
δ_x	Dirac measure in x, $\delta_x(A) = 1_A(x)$
$E(X_{\mathbb{Z}}) = E(P)$	Excess entropy (Def. 3.5) of $X_{\mathbb{Z}}$ and its distribution P
$\mathrm{E}(f) = \mathrm{E}^{\mathbb{P}}(f)$	Expectation value, $\mathrm{E}^{\mathbb{P}}(f) = \int f \,\mathrm{d}\mathbb{P}$
$(\Gamma_{\mathfrak{C}}, \mathcal{G}_{\mathfrak{C}})$	(Measurable) space of causal states (Def. 3.17)
$\gamma_{\mathfrak{C}}$	Causal state memory kernel (Def. 3.19), $\gamma_{\mathfrak{C}}(x) = \delta_{\mathfrak{C}(x)}$
$H(\mu)$, $H^{\mathbb{P}}(X)$	(Shannon) entropy (Def. 3.1) of measure μ, random variable X
K, K_g, $\widehat{K}_d$	Output kernel (Def. 2.12) $K\colon \Gamma \to \mathcal{P}(\Delta)$; $K_g(d) = \widehat{K}_d(g) = K(g; d)$
L_d	Internal operator (Def. 2.12) $L_d\colon \mathcal{P}(\Gamma) \to \mathcal{P}(\Gamma)$ $(d \in \Delta)$
$M_{\mathfrak{C}}$	Causal state memory variable (Def. 3.19), $M_{\mathfrak{C}} = \mathfrak{C} \circ X_{-\mathbb{N}_0}$
$\mathcal{M}(\Gamma)$	Space of signed measures of bounded variation on a measurable space Γ
$\mu_{\mathfrak{C}}(P)$	Effect distribution (Def. 3.46) of P
$O_T(g)$, O_T^{μ}	Output distribution of the HMM (T, δ_g), resp. (T, μ); $O_T^{\mu} = \int O_T \,\mathrm{d}\mu$, $O_T(g) = O_T^{\delta_g}$
$\mathcal{P}(\Gamma)$	Space of probability measures on a measurable space Γ
$\mathcal{P}_{\mathsf{s}}(\Delta^{\mathbb{Z}})$, $\mathcal{P}_{\mathsf{e}}(\Delta^{\mathbb{Z}})$	Space of shift-invariant respectively ergodic probability measures on $\Delta^{\mathbb{Z}}$
$\mathbb{P}_X$	Distribution of a random variable X, $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$
$P_{\mathbb{N}}$	Distribution of the future part of the process, $P_{\mathbb{N}} = P_{X'_{\mathbb{N}}} = P \circ {X'_{\mathbb{N}}}^{-1}$
r	Resultant (barycentre map), $r(\nu) = \int \mathrm{id} \,\mathrm{d}\nu$
σ	Left-shift on a sequence space, usually $\Delta^{\mathbb{Z}}$ or $\Delta^{\mathbb{N}}$; $\sigma\big((x_k)_k\big) = (x_{k+1})_k$
$\sigma(f_i,\ i \in I)$	σ-Algebra generated by a family $(f_i)_{i \in I}$ of functions
S	Prediction dynamic (Def. 3.45), transition kernel of the prediction process
$\mathfrak{S}_P$	Effect space (Def. 3.46), $\mathfrak{S}_P = \mathrm{supp}(\mu_{\mathfrak{C}}(P))$; prediction space version of causal states
τ_d, τ_D	Observable operators of the canonical OOM (Def. 2.33, 2.40)
$(T^{\mathfrak{C}}, \mu_{\mathfrak{C}}(P))$	Prediction HMM (Def. 3.50); prediction space version of ε-machine
X'_k	Canonical projection from $\Delta^{\mathbb{Z}}$ or $\Delta^{\mathbb{N}}$ to Δ
$Y_{\mathbb{Z}}$	Internal expectation process (Def. 2.20) of a given HMM, $Y_0 = \mathbb{P}(W_0 \mid X_{-\mathbb{N}_0})$
$Z_{\mathbb{Z}} = Z_{\mathbb{Z}}^P$	Prediction process (Def. 3.43) of P, $Z_0 = P(X'_{\mathbb{N}} \mid X'_{-\mathbb{N}_0})$

Index

Bibliographische Daten

Models of Discrete-Time Stochastic Processes and Associated Complexity Measures
(Modelle stochastischer Prozesse in diskreter Zeit und zugehörige Komplexitätsmaße)
Löhr, Wolfgang
Universität Leipzig, Dissertation, 2009
81+vi Seiten, 5 Abbildungen, 63 Referenzen

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, den 13. Mai 2009

.....
(Wolfgang Löhr)