Note on Negative Probabilities and Observable Processes

Ulrich Faigle and Alexander Schönhuth[2]

[1] Mathematisches Institut/ZAIK, Universität zu Köln
Weyertal 80, 50931 Köln, Germany
faigle@zpr.uni-koeln.de
[2] aschoen@zpr.uni-koeln.de

Abstract. A mathematical framework for observable processes is introduced *via* the model of systems whose states may be time dependent and described by possibly "negative probabilities". The model generalizes and includes the linearly dependent models or observable operator models for classical discrete stochastic processes. Within this model a general convergence result for finite-dimensional processes, which generalize finite state (hidden) Markov models, is derived. On the philosophical side, the model furthermore offers an explanation for Bell's inequality in quantum mechanics.

Keywords. Negative Probability, Observable Process, Markov Chain, Stochastic Process, Bell's Inequality, HHM, LDP, OOM,

1 Introduction

Observations that cannot be carried out with absolute exactness are usually modeled by random variables within the probabilistic framework proposed by Kolmogorov [11]. However, the Kolmogorov model of probability may not always be appropriate. This issue is intensively discussed, for example, in quantum theory, where the paradox of Einstein, Podolsky and Rosen [6] has led to fundamental questions about the existence of *hidden states*. Bell's inequality [2], for example, is necessarily satisfied by random variables in the Kolmogorov model. Experimental results, however, have led researchers to doubt its validity in quantum mechanical observations (see, *e.g.*, Aspect *et al.* [1]). Therefore, alternative probabilistic models have been proposed. In particular, it has been argued that "probabilities" should be allowed to be negative (see *e.g.*, Khrennikov [12] for an extensive discussion).

In the present note, we sketch a theory of observations under uncertainties based on linear state descriptions of physical systems that can be interpreted as classical (Kolmogorov) probabilities when they are non-negative (and hence as possibly negative "probabilities" in general). Our framework naturally includes as special cases the models of linearly dependent processes or linear operator models that have been proposed for the analysis of (classical) discrete stochastic processes (see, *e.g.*, Heller [8], Ito *et al.* [9] and Jaeger [10]). It can be viewed as a generalization of the Markov chain model (with possibly hidden states). Purely formally, this model may imply "negative transition probabilities" for hidden states even in the case of classical stochastic processes.

Dagstuhl Seminar Proceedings 05031
Algorithms for Optimization with Incomplete Information
http://drops.dagstuhl.de/opus/volltexte/2005/108

However, our point here is not a philosophical discussion of physical interpretations of "negative probabilities". While also our model offers an explanation for observations that possibly violate Bell's inequality, we generally simply accept the model as mathematically feasible and capable of providing a valid analysis of observation processes. As an example, we prove a convergence result for state vectors that holds in particular for so-called finite-dimensional (classical) stochastic processes and shows that one can do meaningful empirical statistics on such processes.

2 Systems, States and Observables

We consider a (non-empty) set $\Omega = \{\omega_1, \ldots, \omega_n\}$ of *pure states* associated with some *system* $\mathcal{S}$. We want to perform observations on $\mathcal{S}$ and assume that the results we obtain depend on the pure state $\omega \in \Omega$ of the system, on which we have perhaps only partial information.

REMARK. The finiteness assumption on Ω is for simplicity of the exposition. It is not difficult to generalize the model to possibly infinite sets of pure states.

As usual, we call a subset $E \subseteq \Omega$ an *event* and associate with it its *characteristic function*

$$\chi_E : \Omega \to \{0,1\} \quad \text{where} \quad \chi_E(\omega) = 1 \iff \omega \in E \; .$$

The event E is said to be *observable* if χ_E is a random variable, *i.e.*, there is a number $0 \leq \mathbf{p}_E \leq 1$ describing the observation probabilities

$$Pr\{\chi_E = 1\} = \mathbf{p}_E \quad \text{and} \quad Pr\{\chi_E = 0\} = 1 - \mathbf{p}_E \; .$$

We assume that the event Ω is always trivially observable with probability $\mathbf{p}_\Omega = 1$ (and hence $\mathbf{p}_\emptyset = 0$). We do not necessarily assume that all events are observable, but that observability depends on the *preparation state* $\mathbf{w}$ of the system $\mathcal{S}$. We view $\mathbf{w}$ as a function

$$\mathbf{w} : 2^\Omega \to \mathbb{R}$$

on the collection of all events.

2.1 Linear Observation Models

The preparation state $\mathbf{w}$ of $\mathcal{S}$ is called *linear* if there are parameters $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ such that

$$\mathbf{w}(E) = \sum_{\omega_i \in E} \alpha_i \quad \text{for all } E \subseteq \Omega \; .$$

So the linear preparation states $\mathbf{w}$ are represented by the n-dimensional coordinate vectors:

$$\mathbf{w} \quad \longleftrightarrow \quad \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \in \mathbb{R}^n \; .$$

CONVENTION. In the following, we restrict ourselves to *linear observation models*, namely observation models where all preparation states $\mathbf{w}$ are linear. Keeping this in mind, we will refer to a preparation state $\mathbf{w}$ often simply as a *state* of the system $\mathcal{S}$ and think of it as a coordinate vector in $\mathbb{R}^n$. The unit vectors in $\mathbb{R}^n$ thus correspond precisely to the pure states of $\mathcal{S}$.

Under the linearity assumption, each state $\mathbf{w}$ obeys the computational rules

$$\begin{array}{rcll} \mathbf{w}(\Omega) & = & 1 & \\ \mathbf{w}(\Omega \setminus E) & = & 1 - \mathbf{w}(E) & \text{for all } E \subseteq \Omega \\ \mathbf{w}(A \cup B) + \mathbf{w}(A \cap B) & = & \mathbf{w}(A) + \mathbf{w}(B) & \text{for all } A, B \subseteq \Omega. \end{array}$$

REMARK [NEGATIVE PROBABILITIES]. A non-negative (linear) state can of course be interpreted as a classical *probability distribution* on the set Ω. Therefore, one could generally call the state parameter $\mathbf{w}(E)$ the "probability" of the event E – accepting with this terminology "negative probabilities" or "probabilities exceeding 1" come into existence. The possibility of mathematical models for stochastic events on the basis of negative probabilities was already noted by Dirac [4] and Feynman [7].

REMARK [QUANTUM PROBABILITIES]. Our model for observations is in spirit quite similar to the one common in quantum mechanics, where the *state* of a system is also described by a coordinate vector with possibly negative components, from which probabilities for observations on the system are computed. The difference of the two models lies in the way probabilities are derived from the state vector.

In quantum mechanics, the state vector $\mathbf{u} = [u_1, \ldots, u_n]^T$ has unit *length*:

$$\|\mathbf{u}\|^2 = |u_1|^2 + \ldots + |u_n|^2 = 1 \; .$$

So the squared absolute values of the components are the elementary "probabilities". The relevant transformations on quantum systems thus are *unitary* transformations, under which the length of coordinate vectors is invariant.

In our model, coordinate vectors $\mathbf{w} = [\alpha_1, \ldots, \alpha_n]^T$ of states have unit coordinate *sum*:

$$u_1 + \ldots + u_1 = 1 \; .$$

The relevant transformations in our model will therefore be *Markov* type transformations, which retain the coordinate sums of coordinate vectors, as will become clearer in the sequel.

2.2 Information Functions

Generalizing the concept of characteristic functions, we define an *information function* to be a function

$$X : \Omega \to \Sigma$$

into a finite set (or *alphabet*) Σ of symbols. X is *observable (in the state* $\mathbf{w}$*)* if

$$0 \; \leq \; Pr\{X = a\} = \mathbf{w}\{\omega \in \Omega \mid X(\omega) = a\} \; \leq \; 1$$

holds for all $a \in \Sigma$. An observable information functions thus appears like a classical Σ-valued random variable:

$$\sum_{a \in \Sigma} Pr\{X = a\} = 1 \quad \text{and} \quad Pr\{X = a\} \geq 0 \; .$$

We say that k observable information functions $X_i : \Omega \to \Sigma_i$ $(i = 1, \ldots, k)$ are (jointly) *compatible* if the information function

$$X : \Omega \to \Sigma := \Sigma_1 \times \ldots \times \Sigma_k \quad \text{with} \quad X(\omega) = [X_1(\omega), \ldots, X_k(\omega)]$$

is observable (in the state $\mathbf{w}$).

If Σ is a set of real numbers, we define the *expected value* of the observable information function X in the state $\mathbf{w} = [\alpha_1, \ldots, \alpha_n]^T$ as usual *via*

$$E(X) = \sum_{a \in \Sigma} aPr\{X = a\} = \sum_{i=1}^{n} X(\omega_i)\alpha_i \; .$$

If $X, Y : \Omega \to \mathbb{R}$ are compatible information functions, their product XY is readily verified to yield an observable information function so that it is meaningful to define the *covariance* of X and Y (in the state $\mathbf{w}$) as

$$\langle X, Y \rangle := E(XY) = \sum_{i=1}^{n} X(\omega_i)Y(\omega_i)\alpha_i \; .$$

2.3 Bell's Inequality

It is important to note that joint compatibility of a set of information functions is a stronger condition than pairwise compatibility. We illustrate this fact with the inequality of Bell [2], which in our model takes the following form.

Lemma 1 (Bell's Inequality). *Assume that the three information functions $X, Y, Z : \Omega \to \{-1, +1\}$ are compatible in the state $\mathbf{w}$. Then their covariances satisfy the inequality*

$$|\langle X, Y \rangle - \langle Y, Z \rangle| \; \leq \; 1 - \langle X, Z \rangle \; . \tag{1}$$

Proof. Any choice of $x, y, z \in \{-1, +1\}$ satisfies the inequality

$$|xy - yz| \; \leq \; 1 - xz \; .$$

By assumption, all the probabilities $p_{xyz} := Pr\{X = x, Y = y, Z = z\}$ are non-negative. So we conclude

$$\begin{aligned} |\langle X, Y \rangle - \langle Y, Z \rangle| = \Big| \sum_{x,y,z} (xy - yz)p_{xyz} \Big| \; &\leq \; \sum_{x,y,z} |xy - yz| p_{xyz} \\ &\leq \sum_{x,y,z} (1 - xz)p_{xyz} \; = \; 1 - \langle X, Z \rangle \; . \end{aligned}$$

◇

Bell's inequality (1) may not hold if X, Y, Z are not jointly compatible in the state $\mathbf{w}$. In fact, the inequality could already be violated when we only know that X, Y and Z are *pairwise* compatible.

Consider for example the set $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$ and functions X, Y, Z as in the following table:

	ω_1	ω_2	ω_3	ω_4	ω_5
X	-1	$+1$	-1	-1	-1
Y	$+1$	$+1$	-1	$+1$	-1
Z	$+1$	$+1$	$+1$	-1	-1

One can check that X, Y, Z are observable and pairwise compatible in the state

$$\mathbf{w} = [-1/3, 1/3, 1/3, 1/3, 1/3]^T \ .$$

The covariances relative to $\mathbf{w}$ are

$$\langle X, Y\rangle = +1, \ \langle Y, Z\rangle = -1/3, \ \langle X, Z\rangle = +1 \ ,$$

violating Bell's inequality.

2.4 Identically Repeated Observations

Assume that we measure the information function $X : \Omega \to \Sigma$ two times "under identical circumstances", *i.e.*, with the system $\mathcal{S}$ being twice in the same preparation state $\mathbf{w} = [\alpha_1, \ldots, \alpha_n]^T$. We can model this situation with the tensor product

$$\Omega \otimes \Omega = (\Omega \times \Omega, \mathbf{w} \otimes \mathbf{w}) \ , \quad \text{where} \quad \mathbf{w} \otimes \mathbf{w} = [\alpha_i \cdot \alpha_j] \in \mathbb{R}^{n \times n} \ ,$$

and the information functions $X_1, X_2 : \Omega \times \Omega \to \Sigma$ whose values are given as

$$X_1(\omega_i, \omega_j) = X(\omega_i) \quad \text{and} \quad X_2(\omega_i, \omega_j) = X(\omega_j) \ .$$

It is clear how this model generalizes to more than two identical observations. The next section provides a further generalization within the framework of discrete time processes.

REMARK. Bell's inequality can also be studied from the point of view of relative frequencies of observations made under identical conditions. The expected values of the relative frequencies then coincide in the limit with the expected values (see also Section 4.2).

2.5 Observable Processes and Generalized Markov Chains

We now turn to the observation of the information function $X : \Omega \to \Sigma$ over discrete time intervals. We assume that the system $\mathcal{S}$ is in the state

$$\mathbf{w}^t = \begin{bmatrix} p_1^{(t)} \\ \vdots \\ p_n^{(t)} \end{bmatrix} \ \in \ \mathbb{R}^n$$

and that X is observable at each time point $t = 0, 1, \ldots$. Let X_t be the observed value of X at time t. By (X_t) we denote the resulting *(discrete) observable process*.

We call (X_t) a *(generalized) Markov chain* if there exists an $(n \times n)$-matrix $M = [m_{ij}]$ such that

(i) $\sum_{i=1}^{n} m_{ij} = 1$ for all $j = 1, \ldots, n$.
(ii) $\mathbf{w}^t = M\mathbf{w}^{t-1} = \ldots = M^t\mathbf{w}^0$ for all $t = 1, 2, \ldots$.

M is the *transition matrix* of the Markov chain (X_t). M decomposes naturally into a sum of matrices,

$$M = \sum_{a \in \Sigma} T^a \ ,$$

where the $(n \times n)$-matrix T^a retains from M precisely the rows i with $X(\omega_i) = a$ and has zero-rows otherwise. We have

$$Pr\{X_{t+1} = a\} = \sum_{X(\omega_i)=a} p_i^{(t+1)} = \sum_{X(\omega_i)=a} \sum_{j=1}^{n} m_{ij} p_i^{(t)} = \mathbf{1}^T T^a \mathbf{w}^t \ ,$$

where the row vector $\mathbf{1}^T = [1, 1, \ldots, 1]$ effects the coefficient sum of the vector $T^a\mathbf{w}^t$.

Given $v = a_1 a_2 \ldots v_t \in \Sigma^t$, we set

$$T^v := T^{a_t} T^{a_{t-1}} \ldots T^{a_1}$$

and then observe by multinomial expansion of M^t the representation

$$\mathbf{w}^t = M^t\mathbf{w}^0 = \big(\sum_{a \in \Sigma} T^a\big)^t \mathbf{w}^0 = \sum_{v \in \Sigma^t} T^v \mathbf{w}^0 \ . \tag{2}$$

2.6 Random Walks

The generalized Markov chain (X_t) offers an alternative view on the system $\mathcal{S}$. We imagine that $\mathcal{S}$ is in some pure state $\pi_t = \omega_j$ at time t and then moves into the state ω_i with (possibly negative) probability

$$p(\omega_i|\omega_j) = m_{ij} \ .$$

In this sense, $\mathcal{S}$ performs a *random walk* on Ω. At time t, $\mathcal{S}$ is in ω_i with probability

$$Pr\{\pi_t = \omega_i\} = p_i^{(t)} \ .$$

We assume that the symbol $X(\omega_i) \in \Sigma$ is emitted when $\pi_t = \omega_i$. Thus we obtain

$$Pr\{X_t = a\} = \sum_{X(\omega_i)=a} p_i^{(t)} \ .$$

REMARK. While the Markov chain (X_t) is observable (by definition), the underlying random walk (π_t) need not be an observable process.

2.7 LPDs, OOMs and HMMs

It is straightforward to extend the notion of *compatibility* of information functions to observable processes (X_t). In the case of Markov chains compatibility amounts to (X_t) exhibiting classical conditional probabilities

$$Pr(a|v) \geq 0 \ .$$

A compatible Markov process (X_t) may be viewed as a classical (discrete) stochastic processes. Not going into details here we just remark that compatible Markov chains can be shown to be equivalent with the stochastic processes that have been termed *finitary linearly dependent* (so-called *LDPs*) (*cf.* Heller [8] and Ito [9]) or the *finite-dimensional observable operator models* (or *OOMs*) (*cf.* Jaeger [10]).

All these models are, of course, closely related to classical (discrete) *hidden Markov models* (or *HMMs* for short) that are well understood (see, *e.g.*, Elliot *et al.* [5]), which in our context essentially are Markov chains with a non-negative transition matrix M. OOMs are strictly more general than HMMs. Jaeger [10] gives the example of a so-called *probability clock* that can be modeled as an OOM but not as an HMM with a finite number of pure states.

3 Generalized Stochastic Processes

Abstracting Markov chains, we think of a (generalized) *discrete stochastic process* (P_t) *with alphabet* Σ as a function

$$P : \Sigma \times \Sigma^* \to \mathbb{R} \quad \text{such that} \quad \sum_{w \in \Sigma^t} P(a|w) = 1 \ \text{ for all } a \in \Sigma, w \in \Sigma^* \ ,$$

where $\Sigma^* := \bigcup_{t=0}^{\infty} \Sigma^t$ is the collection of all finite *words* (or *strings*) over the alphabet Σ. We denote by $t = |w|$ the *length* of the word $w \in \Sigma^t$ and let $\square \in \Sigma^0$ be the *empty word* of length $|\square| = 0$.

The interpretation is, of course, that the process starts at time $t = 0$ with the empty word and adds, at time $t + 1$ and with probability $P(a|w)$, the letter $P_{t+1} = a$ to the word w already produced.

We define the *transition* probabilities (or *conditional* probabilities) for all strings $v = b_1 \dots b_k \in \Sigma^k$ and $w \in \Sigma^t$ by

$$P(v|w) := P(b_1|w)P(b_2|wb_1) \dots P(b_k|wb_1 \dots b_{k-1})$$

and note for all $w \in \Sigma^*$:

$$\sum_{v \in \Sigma^k} P(v|w) = 1 \quad \text{for all } k \in \mathbb{N}, w \in \Sigma^* \ .$$

$P(w) := P(w|\square)$ is the probability for $w \in \Sigma^t$ to have been produced at time t, yielding the relation

$$P(wv) = P(v|w)P(w) \ .$$

We say that (P_t) is *observable* if for all $a \in \Sigma$ and all $t = 0, 1, \ldots$,

$$Pr\{P_{t+1} = a\} = \sum_{w \in \Sigma^t} P(wa) = \sum_{w \in \Sigma^t} P(a|w)P(w) \geq 0.$$

REMARK [Compatibility]. Call (P_t) *compatible* if all conditional probabilities are classical, *i.e.*, $P(a|w) \geq 0$ holds for all $a \in \Sigma, w \in \Sigma^*$. It follows that a compatible stochastic process is in particular observable. In fact, the compatible stochastic processes are the classical stochastic processes.

The transition probabilities describe (P_t) *via* an infinite matrix $\mathcal{P} = [P(v|w)]$ with rows and columns indexed by the words $v, w \in \Sigma^*$. Letting $\mathbf{g}_w$ be the column of $\mathcal{P}$ corresponding to the word w (and hence having the components $P(v|w)$), $\mathbf{g}_w$ contains all the information on the future of the process, given that the word w has been observed so far.

So $\mathbf{g}_w$ can be understood as a representative for the *state* of the process once the word w has been realized. The expected subsequent state is the probabilistically weighted superposition of all possible subsequent states:

$$\mu(\mathbf{g}_w) = \sum_{a \in \Sigma} P(a|w)\mathbf{g}_{wa} \ .$$

Denoting the expected state of the process at time t by

$$\mathbf{g}^t := \sum_{w \in \Sigma^t} P(w)\mathbf{g}_w$$

and extending μ by linearity, we find

$$\begin{aligned} \mathbf{g}^{t+1} = \sum_{wa \in \Sigma^{t+1}} P(wa)\mathbf{g}_{wa} \ &= \ \sum_{w \in \Sigma^t} \sum_{a \in \Sigma} P(w)P(a|w)\mathbf{g}_{aw} \\ = \sum_{w \in \Sigma^t} P(w)\mu(\mathbf{g}_w) \ &= \ \mu\big(\sum_{w \in \Sigma^t} P(w)\mathbf{g}_w)\big) \ = \ \mu(\mathbf{g}^t) \ . \end{aligned}$$

We call μ the *evolution operator* of the stochastic process (P_t) and refer to

$$\mathbf{g}^t = \mu(\mathbf{g}^{t-1}) = \mu^2(\mathbf{g}^{t-2}) = \ldots = \mu^t\mathbf{g}^0$$

as its *ground state* at time t. $\mathbf{g}^t$ contains all the information about the future of (P_t). Indeed, for any $v = b_1 b_k \ldots b_k \in \Sigma^*$, we obtain the *prediction probability* of v at time t as

$$Pr\{P_{t+1} = b_1, \ldots, P_{t+k} = b_k\} = \sum_{w \in \Sigma^t} P(v|w)P(w) = \mathbf{g}^t(v) \ , \tag{3}$$

where $\mathbf{g}(v)$ denotes the v-component of the ground state vector $\mathbf{g}^t$.

3.1 Stobits

In the case of the binary alphabet $\Sigma = \{0, 1\}$, the words $w \in \Sigma^t$ are the $(0, 1)$-strings of length t, *i.e.*, boolean t-bits. A ground state $\mathbf{g}^t$ thus represents a certain probability distribution on the collection of all t-bits (and thus yields a "preparation state" in the sense of Section 2). In analogy with the terminology of t-qubits in quantum computation, one could call a probability distribution on the collection of t-bits a *t-stobit* ("*stochastic t-bit*").

The evolution operator μ describes the stochastic process (P_t) as an evolution of the "empty bit" (*i.e.*, 0-stobit) over time.

3.2 Finite-Dimensional Processes

We define the *dimension* $\dim(P_t)$ of the (generalized) stochastic process (P_t) as the minimal number m such that the ground state $\mathbf{g}^m = \mu^m \mathbf{g}^0$ is a linear combination of the preceding ground states $\mathbf{g}^0, \ldots, \mathbf{g}^{m-1}$:

$$\mathbf{g}^m = \sum_{i=0}^{m-1} \alpha_i \mathbf{g}^i \quad (\alpha_i \in \mathbb{R}).$$

REMARK. For example, it can be shown that the generalized Markov chains of Section 2 give rise to finite-dimensional observable stochastic processes.

Assume $m = \dim(P_t) < \infty$. Then, by definition, the set

$$\mathcal{B} = \{\mathbf{g}_0, \ldots, \mathbf{g}^{m-1}\}$$

is linearly independent. In fact, $\mathcal{B}$ is a linear basis for the collection of all ground states. To see this, let $M = [p_{ij}]$ be the (unique) matrix such that

$$\mu(\mathbf{g}^j) = \mathbf{g}^{j+1} = \sum_{i=0}^{m-1} p_{ij} \mathbf{g}^i \quad (j = 0, \ldots, m-1).$$

We now find

$$\mathbf{g}^{m+1} = \mu(\mathbf{g}^m) = \sum_{j=0}^{m-1} \alpha_j \mu(\mathbf{g}^j) = \sum_{i=0}^{m-1} \beta_i \mathbf{g}^i \ , \quad \text{where} \quad \beta_i = \sum_{j=0}^{m-1} p_{ij} \alpha_j \ .$$

More generally, every ground state $\mathbf{g}^t$ admits a unique representation with respect to $\mathcal{B}$,

$$\mathbf{g}^t = \sum_{i=0}^{m-1} \alpha_i^{(t)} \mathbf{g}^i \ ,$$

where $[\alpha_1^{(t)}, \ldots, \alpha_{m-1}^{(t)}]^T$ is precisely the first column vector in the matrix M^t. Letting $\mathbf{e}_0$ be the first unit vector in $\mathbb{R}^m$, we have in matrix notation:

$$\mathbf{g}^t = [\mathbf{g}^0, \ldots, \mathbf{g}^{m-1}] M^t \mathbf{e}_0 \ .$$

We note: The evolution operator μ is described by the matrix $M = [p_{ij}]$. We therefore call M the *evolution matrix* of the stochastic process (P_t).

3.3 State Representations

As we have seen, in the case $\dim(P_t) = m < \infty$, every ground state $\mathbf{g}^t$ can be described by a unique coordinate vector $\mathbf{w}^t = [\alpha_0^{(t)}, \ldots, \alpha_{m-1}^{(t)}]^T$ relative to the basis $\mathcal{B}$ such that ground state evolution arises as

$$\mathbf{w}^0 = \mathbf{e}_0 \quad \text{and} \quad \mathbf{w}^t = M\mathbf{w}^{t-1} = M^t\mathbf{w}^0 \quad (t = 1, 2, \ldots).$$

The coordinate vectors $\mathbf{w}^t$ are always (preparation) states in the sense of Section 2. Indeed, we observe:

Lemma 2. *Let $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$ be arbitrary. Then*

$$\mathbf{g}^t = \sum_{j=1}^{k} \alpha_j \mathbf{g}^{t_j} \quad \Longrightarrow \quad \sum_{j=1}^{k} \alpha_j = 1 \; .$$

Proof. $1 = \sum_{a\in\Sigma} Pr\{P_{t+1} = a\} = \sum_{a\in\Sigma} \mathbf{g}^t(a) = \sum_{j=1}^{k} \sum_{a\in\Sigma} \alpha_j \mathbf{g}^{t_j}(a) = \sum_{j=1}^{k} \alpha_j$. ◇

REMARK [REVERSIBILITY]. The evolution matrix M of a finite-dimensional stochastic process (P_t) is non-singular. Hence the initial state $\mathbf{w}^0$ can be recovered from the ground state $\mathbf{w}^t$ at any given time t. In this sense, the process does not lose any information over time and is *reversible*.

REMARK [MARKOV CHAINS]. Markov chains not only furnish examples of finite-dimensional stochastic processes. In fact, one can show that every finite-dimensional (generalized) stochastic process arises from some (generalized) Markov chain.

4 A Law of Large Numbers

Say that the stochastic process (P_t) is *bounded* if the prediction probabilities $\mathbf{g}^t(v)$ are bounded, *i.e.*, there are constants $c_1, c_2 \in \mathbb{R}$ such that

$$c_1 \leq \mathbf{g}^t(v) \leq c_2 \quad \text{for all } v \in \Sigma^* \text{ and } t = 0, 1, \ldots.$$

Every classical stochastic process, for example, is bounded by $c_1 = 0$ and $c_2 = 1$.

We want to show that the ground state vectors of a bounded finite-dimensional stochastic process converge on the average to some limit $\mathbf{g}$:

$$\frac{1}{k}(\mathbf{g}^0 + \ldots + \mathbf{g}^{k-1}) \; \rightarrow \; \mathbf{g} \; .$$

For classical stochastic processes, this result implies that one can do empirical statistics. It order to derive the convergence property, we first establish a general convergence result for sequences of vectors.

4.1 Stability of Sequences

Let V be an n-dimensional vector space over the field $\mathbb{C}$ of complex numbers. Assume V to be equipped with a norm $\|\mathbf{v}\|$ that is derived from a hermitian inner product

$$\langle \mathbf{v}|\mathbf{w}\rangle \quad \text{such that} \quad \|\mathbf{v}\| = \sqrt{\langle \mathbf{v}|\mathbf{v}\rangle}\ .$$

Let $F : V \to V$ be a linear operator and denote by F^k its k-iterate. We say that the vector $\mathbf{a} \in V$ is *F-bounded* if there exists some constant $c \in \mathbb{R}$ such that

$$\|F^k\mathbf{a}\| \leq c \quad \text{for all } k = 1, 2, \ldots.$$

Since F is linear, it follows that arbitrary linear combinations of F-bounded vectors are again F-bounded. So F gives rise to the linear subspace

$$\mathcal{F} = \{\mathbf{a} \in V \mid \mathbf{a} \text{ is } F\text{-bounded}\} \subseteq V\ .$$

Note that $\mathcal{F}$ is F-invariant. Indeed, the definition immediately yields:

$$\mathbf{a} \in \mathcal{F} \quad \Longrightarrow \quad F\mathbf{a} \in \mathcal{F}.$$

REMARK. F-boundedness is closely related to the concept of *matrix stability* in the sense of Brayton and Tong [3], where F is called *stable* if there exists a full-dimensional F-invariant compact set $K \subseteq V$ with $0 \in K$. Here, however, we are not so much interested in the operator F but the behavior of an element $\mathbf{a} \in V$ under the action of F. It is not difficult to see that $\mathcal{F} = V$ holds if F is a stable matrix.

Letting $I = F^0$ be the identity operator, we furthermore associate with F and every $k \geq 1$ the k-th *averaging operator*

$$\overline{F}_k := \frac{1}{k}(I + F + F^2 + \ldots + F^{k-1})\ .$$

and call the vector $\mathbf{a} \in V$ *F-stable* if there exists some $\overline{\mathbf{a}} \in V$ such that

$$\lim_{k\to\infty} \overline{F}_k\mathbf{a} \ \to\ \overline{\mathbf{a}}.$$

Also the averaging operators $\overline{F}_k$ are linear. So the F-stable elements of V form a linear subspace $\overline{\mathcal{F}}$ of V as well. The main result in this section says that $\mathcal{F}$ is a subspace of $\overline{\mathcal{F}}$.

Theorem 1. *Let $F : V \to V$ be a linear operator and $\mathbf{a} \in V$ an F-bounded element. Then the k-averages $F_k\mathbf{a}$ converge to some $\overline{\mathbf{a}} \in V$,*

$$\lim_{k\to\infty} \overline{F}_k\mathbf{a} \ =\ \lim_{k\to\infty} \frac{\mathbf{a} + F\mathbf{a} + \ldots + F^{k-1}\mathbf{a}}{k} \ =\ \overline{\mathbf{a}}\ ,$$

where $\overline{\mathbf{a}}$ is either the null vector $0 \in V$ or an eigenvector of F with eigenvalue $\lambda = 1$.

Proof. We prove the Theorem by induction on the dimension n, assuming it to be true for all dimensions $n' < n$. So there is no loss in generality when we assume $\mathcal{F} = V$ (otherwise, the claim follows from the induction hypothesis already).

If F has at least two distinct eigenvalues $\lambda_1 \neq \lambda_2$, then V decomposes into the direct sum of two F-invariant subspaces:

$$V = V_1 \oplus V_2 \quad \text{with} \quad \dim V_i < n \ .$$

$\mathbf{a} = \mathbf{a}_1 + \mathbf{a}_2$ holds for unique elements $\mathbf{a}_i \in V_i$ and we have

$$\|\mathbf{a}\|^2 = \|\mathbf{a}_1\|^2 + \|\mathbf{a}_2\|^2 \ .$$

Since the V_i are F-invariant, the $\mathbf{a}_i$ are F-bounded. So the induction hypothesis implies that the vectors $\mathbf{a}_i$ are F-stable, which in turn implies that their sum $\mathbf{a}$ is F-stable.

So we are left with the case where F has a unique eigenvalue $\lambda \in \mathbb{C}$. We distinguish two cases. If $\lambda \neq 1$, the linear operator $I - F$ is non-singular. So there is some $\mathbf{v} \in V$ such that

$$\mathbf{a} = (I - F)\mathbf{v} \ .$$

The geometric summation formula for the powers of F yields

$$(I + F + \ldots + F^{k-1})(I - F) = I - F^k$$

and, because $\mathbf{v}$ if F-bounded, the convergence

$$\overline{F}_k\mathbf{a} = \frac{(I - F^k)\mathbf{v}}{k} = \frac{1}{k}\mathbf{v} - \frac{1}{k}F^k\mathbf{v} \ \rightarrow \ 0 \in V \ .$$

It remains to deal with the case where $\lambda = 1$ is the unique eigenvalue of the linear operator F. By the Cayley-Hamilton Theorem, there exists a minimal number N such that

$$(F - I)^N\mathbf{v} = 0 \quad \text{for all } \mathbf{v} \in V \ .$$

CLAIM: $N = 1$ (*i.e.*, $F = I$).

Suppose to the contrary that $N \geq 2$ is true. Then there exists some $\mathbf{v} \in V$ with $(F-I)\mathbf{v} \neq 0$ and $(F - I)^2\mathbf{v} = 0$. Consequently, we have for all $k \geq n$,

$$F^k\mathbf{v} = [(F - I) + I]^k\mathbf{v} = \sum_{s=0}^{k} \binom{k}{s} (F - I)^s\mathbf{v} = \mathbf{v} + k(F - I)\mathbf{v} \ ,$$

which implies that $\mathbf{v}$ is not F-bounded. This contradiction shows $F = I$ and thus $\overline{F}_k\mathbf{a} = \mathbf{a}$. So convergence follows trivially.

$\diamond$

REMARK. An F-stable element $\mathbf{a} \in V$ always yields average convergence to either 0 or an eigenvector $\overline{\mathbf{a}}$ of F with eigenvalue $\lambda = 1$. The point of Theorem 1 is a criterion for guaranteed stability.

REMARK. One can show that the converse of Theorem 1 is also true, *i.e.*, $a \in V$ is F-bounded if and only if a is F-stable.

4.2 Convergence of Ground States

Let $\mathcal{B} = [\mathbf{g}^0, \ldots, \mathbf{g}^{m-1}]$ the basis of the finite-dimensional stochastic process (P_t). Then $\mathcal{B}$ contains some $(m \times m)$-submatrix $\mathbf{B}$ of full rank rk $\mathbf{B} = m$ and any ground state $\mathbf{g}^t$ is already determined by its restriction $\mathbf{g}_t$ of the components corresponding to the rows of $\mathbf{B}$. In particular, the evolution operator μ acts linearly on those restrictions of the ground states.

Hence, if (P_t) is bounded, Theorem 1 guarantees the convergence of the averages of the ground states to some $\mathbf{g}$. Because

$$\sum_{a \in \Sigma} \mathbf{g}^t(a) = 1 \quad \Longrightarrow \quad \sum_{a \in \Sigma} \mathbf{g}(a) = 1 \ ,$$

we conclude $\mathbf{g} \neq 0$. So $\mathbf{g}$ must be an eigenvector of the evolution operator μ with eigenvalue 1.

CLASSICAL STOCHASTIC PROCESSES. Assume that (P_t) is a classical finite-dimensional stochastic process. Then the average convergence of the ground states implies the average convergence of the prediction probabilities for all $a_1, \ldots, a_k \in \Sigma$:

$$Pr\{X_{t+1} = a_1, \ldots X_{t+k} = a_k\} = \mathbf{g}^t(a_1 \ldots a_k) \ \rightsquigarrow \ \mathbf{g}(a_1 \ldots a_k) \ .$$

For example, fixing any $a \in \Sigma$ and defining the *empirical counting function* by

$$Y_t \ := \ \begin{cases} 1 & \text{if } X_t = a \\ 0 & \text{if } X_t \neq a, \end{cases}$$

we see that the expected value of the averaged count converges:

$$\begin{aligned} E\big(\frac{Y_1 + \ldots Y_t}{t}\big) &= \frac{E(Y_1) + \ldots E(Y_t)}{t} \\ &= \frac{\mathbf{g}^0(a) + \ldots + \mathbf{g}^{t-1}(a)}{t} \ \ \rightarrow \ \ \mathbf{g}(a) \ . \end{aligned}$$

Because $\sum_{a \in \Sigma} \mathbf{g}^t(a) = 1$ for all t, we obtain a classical limiting probability distribution on Σ:

$$\sum_{a \in \Sigma} \mathbf{g}(a) = 1 \quad \text{and} \quad \mathbf{g}(a) \geq 0 \ .$$

A similar averaged counting result holds for any $v \in \Sigma^*$. This observation indicates that we can do statistics on finite-dimensional stochastic processes.

References

1. A. ASPECT, J. DALIBARD, G. ROGER: *Experimental tests of Bell's inequalities using time-varying analyzers*, Phys. Rev. Lett. 49, 1804 (1982).

2. J.S. BELL: *On the problem of hidden variables in quantum mechanics*, Rev. Mod. Phys. 38. 447-452 (1966).
3. R.K. BRAYTON AND C.H. TONG: *Stability of dynamical systems: A constructive approach*, IEEE Transactions on Circuits and Systems 26, 224-234 (1979).
4. P.A.M. DIRAC: *The physical interpretation of quantum mechanics*, Proc. Royal Soc. London A 180, 1-39 (1942).
5. R.J. ELLIOT, L. AGGOUN, J.B. MOORE: *Hidden Markov Models*, Springer-Verlag, Heidelberg, 1995.
6. A. EINSTEIN,, B. PODOLSKY, N. ROSEN: *Can quantum mechanical descriptions of physical reality be considered complete?*, Phys. Rev. 47, 777-780 (1935).
7. R.P. FEYNMAN: *Quantum Implications, Essays in Honour of David Bohm*, B.J. Hiley and F.D. Peat eds., Routledge and Kegan Paul, London, 235-246 (1987).
8. A. HELLER: *On stochastic processes for functions of finite Markov chains*, Ann. of Mathematical Statistics 30, 688-697 (1965).
9. H. ITO, S.-I. AMARI, K- KOBAYASHI: *Identifiability of hidden Markov information sources and their minimum degrees of freedom*, IEEE Transactions on Information Theory 38, 324-333 (1992).
10. H. JAEGER: *Observable operator models for discrete stochastic time series*, Neural Computing 12, 1371-1398 (2000).
11. A.N. KOLMOGOROFF: *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer-Verlag, Berlin, 1933.
12. A. KHRENNIKOV: *Interpretations of Probability*, VSP Science, Zeist, 1999.