IMA Journal of Mathematical Control and Information (2013) **30**, 205–238
doi:10.1093/imamci/dns016
Advance Access publication on September 28, 2012

An algebraic framework for information theory: classical information

Manas K. Patra*
Department of Computer Science, University of York, York YO10 5DD, UK and Laboratoire d'Information Quantique, Universite Libre de Bruxelles, Campus Plaine, Bruxelles 1050, Belgium
*Corresponding author: manas@cs.york.ac.uk

and

Samuel L. Braunstein
Department of Computer Science, University of York, York YO10 5DD, UK

[Received on 4 November 2010; revised on 30 December 2011; accepted on 21 April 2012]

This work proposes a complete algebraic model for classical information theory. As a precursor the essential probabilistic concepts have been defined and analysed in the algebraic setting. Examples from probability and information theory demonstrate that in addition to theoretical insights provided by the algebraic model one obtains new computational and analytical tools. Several important theorems of classical probability and information theory are formulated and proved in the algebraic framework.

Keywords: information theory; operator algebras; algebraic probability; entropy; coding theorems.

1. Introduction

The present paper proposes an algebraic model of classical information theory. We then carry out a detailed investigation of the model. The connection between operator algebras and information theory—both classical and quantum—has appeared in the scientific literature since the beginning of information theory and operator algebras—(see e.g. Segal, 1960; Umegaki, 1962; Lindblad, 1974; Araki, 1975) and (Keyl 2002; Kretschmann & Werner 2006; see also Bêny *et al.*, 2007) respectively. The standard formulation of classical information theory (Ash, 1990; Cover & Thomas, 1999) on the other hand is sometimes seen as an important application of probability theory. Thus probabilistic concepts such as distribution function, conditional expectation and independence are vital for the development of information theory. Most previous work including those mentioned above focus on some aspects of information theory, especially the non-commutative generalizations of the concepts of entropy and for specific probabilistic concepts they often resort to a representation on some Hilbert space. As a consequence, there does not appear to be a unified coherent approach based on intrinsically algebraic notions. The construction of such a model is one of the goals of the paper. As probabilistic concepts play such an important role in the development of information theory, we devote a fairly large section to an algebraic approach to probability. It was Segal (1954), one of the major players in the early development of operator theory, who first proposed such an algebraic approach to probability theory. Although we have mostly restricted ourselves to the discrete case, sufficient for our models of communication and information processes, our proposed model is different from Segal's. There are other workers such as Cam (1986), Whittle (1992), Streater (1995) that use algebraic approaches to statistical problem. Cam uses Banach lattices for the set of observables and expectation values are positive linear functionals. Streater's approach is close to ours in that he uses a C^* structure to define classical observables. However, the ambient probability space is there in the background and is invoked to define notions such

as independence. We show that this is not necessary. Most important notions from probability theory can be defined from an abstract perspective. We emphasize that our algebraic analysis was developed primarily to deal with mathematical communication and information theory. We believe that several aspects of our approach are novel (see the section-wise synopsis below) and yield deeper insights ito information processes.

A strong motivation for this paper is the relatively young field of quantum information theory. It is almost folklore that in quantum mechanics, we are forced to deal with non-commutative entities. Thus, the language of C^* algebras, already known to physicists for decades (Emch, 1984; Haag, 1992) as 'the algebra of observables' on which many extensions of classical probabilistic concepts can be made, became a natural setting for quantum information. As a complex quantum information scheme or protocol has several *classical* components (e.g. classical communication, coin-tosses, etc.), it is important that we have a unified model and a single language for quantum and classical information. Such a formulation will be of great help in the difficult task of protocol analysis. Besides a unified framework will be of significant advantage for theoretical analysis. For example, a deeper study of quantum phenomena like (no) quantum broadcasting (Barnum *et al.*, 2007), quantum Huffman coding (Braunstein *et al.*, 2000), channel capacity (Schumacher, 1996) to name a few would benefit from the investigations of these structures. In this framework, we may view a classical process as a special type of process described by *commuting* elements. Therefore, it seems appropriate to investigate this special case first. As we will see the classical structure is quite rich and sheds new light on some familiar aspects of information theory. There is yet another reason. In quantum mechanics, we have several examples of observables taking only a finite number of values (the spectrum is finite). But, in classical mechanics, all variables take on a continuum of values. Therefore, we often see statements like 'a finite-dimensional operator like spin is a purely quantum phenomenon that has no classical analogue'. However, when we talk about information systems finite-dimensional quantum systems have obvious classical analogues. A two-dimensional quantum 'source' corresponds to a classical binary source. Our investigations raise some questions about the possibility of an alternative formulation of probability theory with a more algebraic flavour (Segal, 1954). This is interesting in itself. But it is a side issue in this paper and will only be briefly commented upon. Since our main concern is the mathematical models of information-processing systems, we will be primarily dealing with discrete systems, thus circumventing some tricky topological issues.

Let us recall a simple model of a communication system proposed by Shannon (1948) and Shannon & Weaver (1949). This model has essentially four components: source, channel, encoder/decoder and receiver. The source could be representing very different kinds of objects: a human speaker, a radar antenna or a distant star. We usually have some model of the source. The coding/decoding operation is required for three basic reasons: (i) the source/receiver alphabet and the channel alphabet may be different, (ii) to maximize the rate of information communication and (ii) to detect and correct errors due to noise and distortion. Some amount of noise affects every stage of the operation. So, the behaviour of components are generally modelled as stochastic processes. This is valid in both the classical and often quantum communication processes. The difference, of course, is in the description of the two processes. As in any stochastic process, we specify the source by a family X_t of *random variables* and the various stages of the communication system are modelled as (stochastic) transformations of these variables. The parameter t can be continuous or discrete. In this work, our primary focus will be on discrete processes corresponding to discrete time. Thus, a discrete source can be viewed as a generator of a countable set of random variables. Let us suppose that the source 'tosses' a coin and sends a 1 if it is 'heads' and a 0 otherwise. We may model this by a pair of random variables $\{X_h, X_T\}$ on the probability space $\{H \text{ (heads)}, T \text{ (tails)}\}$ such that $X_H(H) = X_T(T) = 1$ and

$X_H(T) = X_T(H) = 0$. If the coin is unbiased, we say that the *state* of the source is given by a probability measure $\{\frac{1}{2}, \frac{1}{2}\}$. In general, it is $\{p, q\}$ where $0 \leqslant p = 1 - q \leqslant 1$ is the probability of heads. This simple model can be generalized to more complicated sources. Besides these elementary random variables we encounter functions of these variables. Thus, we are led to study *algebras* of random variables. The usual textbook definition of a random variable is that it is a (measurable) function on a probability space S. Hence, in the standard formulation, we need a probability space or sample space to define our random variables or 'observables'. Recall that a probability space is a *triple* $(S, \mathcal{M}, \mu)$, where S is a set, the set of elementary or atomic events, $\mathcal{M}$ is a σ algebra of subsets and μ is the probability measure. Thus, if $\{A_n\}$ is a sequence of mutually disjoint elements from $\mathcal{M}$, then

$$\mu\left(\bigcup_n A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

Moreover, $\mu(S) = 1$ and $\mu(B) \geqslant 0$ for any $B \in \mathcal{M}$. These are essentially the Kolmogorov axioms. A real- or complex-valued random variable is a measurable function form $S \to \mathbb{R}$ or $S \to \mathbb{C}$. Here measurability is with respect to the Borel σ algebra of $\mathbb{R}$ or $\mathbb{C}$. We recall that the Borel σ algebra of any topological space is generated by its open sets. So, in some sense in this formulation the probability space is fundamental and the notion of random variables is based on the former. However, from an observer/experimenter point of view, the random variables are the basic entities because these are precisely the observables. In statistical theories like information theory it is the set of random variables and their distributions and transformations which are of primary interest. Of course, to compute the probability distributions of the random variables, we have to appeal to the original probability space. But once the distributions have been determined for almost all computations they suffice and the underlying probability or sample space plays little role. The fundamental theorem of Kolmogorov (on existence of processes; Shiryayev, 1984; Billingsley, 1995) guarantees that given a set of distribution functions satisfying certain consistency conditions we can reconstruct a probability space and random variables having these distributions. These observations suggest that we take the algebra of random variables or observables as our primary structure and derive all relevant quantities from this structure. One of the advantages is that we deal with a smaller spaces restricted to quantities of interest. In the modelling of security protocols this is a more realistic approach since different participants have access to different sets of observables and may assign different probability structures on the same set of events. They may even assign different event spaces.

In the quantum case there are more fundamental reasons for working with the algebras of observables. We will not go into these here. The current work is an attempt at formulating (classical) information theory in an algebraic framework. We will mainly focus on C^* and von Neumann algebras. We will see that most interesting spaces of observables do have a C^* structure. As mentioned before, we will be dealing with discrete spaces in this work. We also observe that C^* algebras have been studied intensively since the pioneering works of Murray, von Neumann, Gelfand, Naimark and Segal and others starting from 1930s. As we stated at the beginning of this section, several probabilistic and information theoretic concepts such as conditional expectation, entropy, differential entropy have previously been investigated in the algebraic context. However, to the best of our knowledge, there is no work investigating information and communication theory in a purely algebraic framework. Our investigations indicate that most if not all important concepts and constructs of information theory can be dealt with in the algebraic framework. The paper is structured as follows.

In Section 2, we give the basic definitions of the algebras of interest. This section is fairly detailed as we state several structure theorems for finite-dimensional abelian C^* algebras and their tensor products,

possibly *infinite*. There are two reasons for this. The first is to make the paper as self-contained as possible. The second reason is to demonstrate the power and utility of the algebraic techniques. However, for ease in reading we defer the proofs to the appendix. We believe that in these special cases some of the proofs are new. We also give several examples.

Section 3 gives an account of probabilistic concepts from an algebraic perspective. In particular, we investigate the fundamental notion of independence and demonstrate how it relates to the algebraic structure. We note that there is a very sophisticated theory of non-commutative or 'free probability' (Voiculescu *et al.*, 1992). Our approach in the simpler commutative case is different in several aspects. One important point in which our approach seems novel is the definition of a probability distribution function. The definition we give is algebraic in the sense that it depends on the intrinsic properties of the algebra. Specifically, we define a probability distribution function as the weak limit of a net or sequence of elements in a subalgebra representing an approximate identity of an ideal or a subalgebra. To illustrate the practical use of these techniques, we give some typical examples from standard probability theory. The problem of 'waiting time' shows that the algebraic approach can offer new techniques and insights. Finally, using the definition of distribution function and some other constructs, we formulate and prove some of the basic limit theorems in this framework. These are used later in proving results in information theory.

In Section 4, we give a precise algebraic model of information communication system. The fundamental concept of entropy is introduced as a limiting value of typical sequences of the algebra. The notion of typical sequence comes from the limit theorems. In the conventional approach, the limit is taken in the probability (convergence in measure). In our algebraic case, it corresponds to a weak limit. The point is, that we can do all this in purely algebraic setting. We also define and study the crucial notion of a channel. In particular, the *channel coding theorem* is presented as an approximation result. Stated informally,

Every channel other than the useless ones can be approximated by a lossless channel with appropriate coding.

We also show that the notions of zero-error capacity (Shannon, 1956) is very natural in the algebraic setting. In the brilliant work, Lovâsz (1979) demonstrated the powers of algebraic methods. We contend that the intuition for these methods will become clearer in the algebraic models. In the final section, we summarize our constructions and discuss future work. We summarize this introduction with some of main points of the algebraic approach in the paper.

(1) The algebraic framework brings observables, the empirical quantities, that are measured or observed to the forefront.

(2) Most if not all important concepts of probability theory can be formulated in the algebraic framework. The limit theorems, different types of convergence and the notions of statistical independence can be formulated in this framework and this offers a fresh perspective on these important issues. However, our algebraic approach is restricted to random variables taking real or complex values (or vector-valued random variables at most).

(3) This approach provides additional computational and algorithmic tools. We illustrate this in the example of 'waiting time'.

(4) The algebraic framework we believe is more natural in dealing with information and communication processes. It is the widely accepted model in the quantum processes. We show that even for classical communication and information such framework is possible. It provides new insights.

(5) We prove several important results from classical information theory. Although some of our proofs are closely related to the standard proofs, they give us some new insights. For example, we can see that the commutative nature of the (algebraic) classical model is crucial for some and why they cannot be generalized to the non-commutative quantum case.

(6) It is possible to give intrinsically 'algebraic' proofs for the results where we have adapted the standard proofs. But we aim to develop such proofs in a more general setting of non-commutative quantum case. Further, even in a fully algebraic framework it would be certainly fruitful to supplement algebraic techniques with those burrowed from probability and information theory. It gives us new insights in the purely algebraic structure!

(7) Some of the proofs, however, use algebraic techniques. For example, in the proof of Kraft inequality and source coding theorem, we use a Gram–Schmidt-type orthogonalization.

(8) The channel coding theorem is formulated as an approximation result. We approximate the effect of the given channel on the space of codewords by a lossless channel. The proof is algebraic and shows the significance of commutativity assumption.

(9) Our algebraic model of classical communication can be easily integrated with quantum systems, thus offering a unified approach to *hybrid* systems.

2. Algebraic preliminaries

An algebra $\mathcal{A}$ is vector space over a field $\mathtt{F}$ with an associative bilinear product: $\mathcal{A} \times \mathcal{A} \to \mathcal{A}$. We take $\mathtt{F} = \mathbb{C}$, the field of complex numbers. We deal mostly with *unital* algebras, that is, algebras with a unit $\mathbb{1}$. A Banach algebra is an algebra with a non-negative real function $\|\|$ on $\mathcal{A}$ such that

$$\|x\| \geqslant 0 \text{ and } \|a\| = 0, \text{ iff } a = 0,$$
$$\|x + y\| \leqslant \|x\| + \|y\| \text{ (triangle inequality)},$$
$$\|xy\| \leqslant \|x\|\|y\| \text{ (Banach property)},$$

and $\mathcal{A}$ is *complete* in the topology defined by the norm. A C^* algebra B is a Banach algebra with an anti-linear involution * (a map σ is an involution if $\sigma^2 = 1$, it is antilinear if $\sigma(x + cy) = \sigma(x) + \bar{c}\sigma(y)$, c a complex number and $\bar{c}$ its complex conjugate) such that

$$\|xx^*\| = \|x\|^2 \quad \text{and} \quad (xy)^* = y^*x^* \ \forall x, y \in B.$$

This implies that $\|x\| = \|x^*\|$. The quintessential examples of a C^* algebra are the norm-closed self-adjoint subalgebras of $\mathcal{L}(H)$, the set of bounded operators on a Hilbert space of H. The fundamental Gelfand–Naimark–Segal (GNS) theorem states that every C^* algebra can be isometrically embedded in some $\mathcal{L}(H)$. The notion of the spectrum of an operator has an algebraic analogue without reference to the representation space. The resolvent of an element x in the C^* algebra B is the set $R(x) \subset \mathbb{C}$ such that $\lambda \in R(x)$ implies $\lambda\mathbb{1} - x$ is invertible. The spectrum $\mathtt{sp}(x)$ is the complement of the resolvent. The spectrum is a non-empty closed and bounded subset and hence compact. Define $r(x) = \sup\{|\lambda| : \lambda \in \mathtt{sp}(x)\}$, the spectral radius. A basic result (Kadison & Ringrose, 1997) states that

$$r(x) = \lim_{n\to\infty} \|x^n\|^{1/n}.$$

An element x is self-adjoint if $x = x^*$, normal if $x^*x = xx^*$ and positive (strictly positive) if x is self-adjoint and $\mathrm{sp}(x) \subset [0, \infty)((0, \infty))$. A self-adjoint element has a real spectrum and conversely. Since $x = (x + x^*)/2 + i(x - x^*)/2i$, any element of a C^* algebra can be decomposed into self-adjoint 'real' $((x + x^*)/2)$ and 'imaginary' $((x - x^*)/2i)$ parts. For a self-adjoint element x, $r(x) = \|x\|$. The positive elements define a partial order on A: $x \leqslant y$ iff $y - x \geqslant 0$ (positive). An important property of positive elements is that they have unique positive square-roots: if $a \geqslant 0$, there is a unique element $b \geqslant 0$ such that $b^2 = a$. We write $\sqrt{a}$ or $a^{1/2}$ for the square-root. Since $x^*x \geqslant 0$, it has a unique square-root. If x is normal, we write $|x| = \sqrt{x^*x}$. In particular, if x is self-adjoint, $|x| = \sqrt{x^2}$. A self-adjoint element x has a decomposition $x = x_+ - x_-$ into positive and negative parts where $x_+ = (|x| + x)/2$ and $x_- = (|x| - x)/2)$ are positive. An element $p \in B$ is a projection if p is self-adjoint and $p^2 = p$. Given two C^* algebras A and B a homomorphism F is a linear map preserving the product and * structures. It is continuous iff bounded. A continuous isomorphism of C^* algebras is an isometry (norm preserving). A homomorphism is positive if it maps positive elements to positive elements. A (linear) functional on A is a linear map $A \to \mathbb{C}$. The GNS construction starts with a positive functional (mapping positive elements to non-negative numbers) on B. The details may be found in Kadison & Ringrose (1997) and Takesaki (2002). A positive functional ω such that $\omega(\mathbb{1}) = 1$ is called a *state*. The set of states G is convex. The extreme points are called *pure states* and G is the convex closure of pure states (Krein–Millman theorem). A set $B \subset A$ is called a subalgebra if it is a C^* algebra with the inherited product. That is, it is a subalgebra in the algebraic sense and it is *closed* in the norm topology. A subalgebra B is called unital if it contains the identity of A. Our primary interest will be on *abelian* (also called commutative) algebras. The structure theory is a bit different in this case. Of course, the GNS construction is valid and the elements of the algebra act as multiplication operators on the representing Hilbert space. However, there is an alternative representation in the abelian case due to Gelfand and Naimark which will be of primary interest to us. To motivate it consider an example.

Let X be a compact Hausdorff topological space, for example, a closed and bounded set in $\mathbb{R}^n$. Let $C(X)$ denote the space of continuous complex functions on X. It includes the constant functions. If we define addition and multiplication point-wise

$$(f + g)(x) = f(x) + g(x), \quad (fg)(x) = f(x)g(x) \quad \text{and} \quad \|f\| = \sup_{x \in X} |f(x)| \quad \forall f, g \in C(X), \tag{1}$$

then $C(X)$ becomes a complex Banach algebra. Defining $f^*(x) = \overline{f(x)}$ then $C(X)$ is an abelian C^* algebra. This is a prototype of abelian C^* algebras (Kadison & Ringrose, 1997). One can generalize to (essentially) bounded measurable functions on measure spaces with appropriate norm. However, for the purposes of this paper, it suffices to consider compact spaces with measures defined on Borel σ algebras. We will dwell more on this point in the next section. A complex function (not necessarily continuous) is called simple if its range is finite. For example, the *indicator* function I_S of a subset $S \subset X$, given by $I_S(x) = 1$ if $x \in S$ and 0 otherwise is a simple function. It is not continuous unless $S = X$ or S is a connected component. Simple functions play a crucial role in probability and integration theory. From their definition, it follows that the projections in $C(X)$ are precisely the indicator functions. The constant functions 1 and 0 are both projections corresponding to $S = X$ and $\emptyset$, respectively. These are the only projections in $C(X)$ if X is connected. The basic structure theorem for abelian C^* algebras is the following.

THEOREM 1 An abelian C^* algebra with unity is isomorphic to the algebra $C(X)$ for a compact Hasudorff space X. The isomorphism is an isometry (norm preserving).

The main idea of the proof comes from the following observation. In any function algebra $C(X)$ for $p \in X$ the map $\sigma_p : f \to f(p)$, $f \in C(X)$ is a linear functional on $C(X)$. These are multiplicative functionals in the sense that $\sigma_p(xy) = \sigma_p(x)\sigma_p(y)$. In fact, these are the only possible multiplicative functionals. The Gelfand representation for an abstract abelian C^* algebra A identifies the space X as the set of multiplicative functionals and gives it a topology to make these continuous. The details can be found in Kadison & Ringrose (1997).

Now let $X = \{a_1, \ldots, a_n\}$ be a finite set with discrete topology. Then $A = C(X)$ is the set of all functions $X \to \mathbb{C}$. The algebra $C(X)$ can be considered as the algebra of (complex) random variables on the finite probability space X. Let $x_i(a_j) = \delta_{ij}$, $i, j = 1, \ldots, n$. Here $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. The functions $x_i \in A$ form a basis for A. Their multiplication table is particularly simple: $x_i x_j = \delta_{ij} x_i$. They also satisfy $\sum_i x_i = \mathbb{1}$. These are projections in A. They are orthocomplementary in the sense that $x_i x_j = 0$ for $i \neq j$. We call any basis consisting of elements of norm 1 with distinct elements orthocomplementary *atomic*. A set of linearly independent elements $\{y_i\}$ satisfying $\sum_i y_i = \mathbb{1}$ is said to be complete. The next theorem gives us the general structure of any finite-dimensional algebra.

THEOREM 2 Let A be a finite-dimensional abelian C^* algebra. Then there is a unique (up to permutations) complete atomic basis $\mathcal{B} = \{x_1, \ldots, x_n\}$. That is, the basis elements satisfy

$$x_i^* = x_i, \quad x_i x_j = \delta_{ij} x_i, \quad \|x_i\| = 1 \text{ and } \sum_i x_i = \mathbb{1}, \tag{2}$$

Let $x = \sum_i a_i x_i \in A$. Then $\mathrm{sp}(x) = \{a_i\}$ and hence $\|x\| = \max_i\{|a_i|\}$.

The proof is given in the Appendix. However, note that we could have proved the theorem using the Gelfand representation. But the above proof is more intrinsic depending mostly on the structure of the algebra itself. In fact, the Gelfand–Naimark construction is an easy consequence of the theorem in this case. The theorem can be called a form of spectral theorem. A simple consequence of the theorem is the spectral theorem for normal operators on finite-dimensional spaces. It is useful to define some operations on subalgebras of a C^* algebra. Thus, if $B, C \subset A$ are C^* subalgebras of a C^* algebra A, then

$$B + C = \{b + c : b \in B \text{ and } c \in C\} \quad \text{and} \quad BC = \left\{ \sum_i b_i c_i : b_i \in B \text{ and } c_i \in C \right\}.$$

COROLLARY 1 Let A be an abelian C^* algebra satisfying the following conditions. There are finite-dimensional subalgebras A_k, $k = 0, 1, \ldots$ with

$$A = \bigcup_{k=0}^{\infty} A_k \quad \text{and} \quad A_k \subset A_{k+1} \ \ \forall k$$

and for each k corresponding to A_k there is complementary subalgebra $A'_k \subset A_{k+1}$ such that $A_k A'_k = A_{k+1}$, $A_k \bigcap A'_k = \{0, \mathbb{1}\}$ and for $x \in A_k, y \in A'_k$ implies $xy \neq 0$ unless x or y is 0. Then there is a countable basis for A satisfying the first three equations in (2).

Proof. We prove by induction. The case of A_0 is proved in the theorem. Assume that we have an atomic basis $\{y_1^n, \ldots, y_{k_n}^n\}$ for A_n. There is a (unique) atomic basis $\{x_1^n \ldots, x_{m_n}^n\}$ in A'_n. It is now a routine matter to show that $\{x_i^n y_j^n : 1 \leqslant k_n \text{ and } 1 \leqslant m_n\}$ form a basis in A_{n+1}. □

The conditions in the corollary can be slightly weakened by requiring that there be embeddings (injective algebra homomorphisms) $\alpha_k : A_k \to A_{k+1}$ and $\alpha'_k : A'_k \to A_{k+1}$ such that the images $\alpha_k(A_k)$ and $\alpha'_k(A'_k)$ satisfy the conditions stated. Such a structure will appear in the *tensor product* of algebras to be defined below. They play an important role in our modelling of information and communication systems. Let us also note that the basis structure in Theorem 2 may be used to defined a finite-dimensional C^* algebra abstractly.

2.1 *Tensor products*

We next describe an important construction for C^* algebras. Given two C^* algebras A and B, the tensor product $A \otimes B$ is defined as follows. As a set it consists of all finite linear combinations of symbols of the form $\{x \otimes y : x \in A, y \in B\}$ subject to the conditions that for all $x, u \in A,\ y, z \in B$ and $c \in \mathbb{C}$,

$$\begin{aligned} (cx) \otimes y = x \otimes (cy) = c(x \otimes y), \\ (x+u) \otimes y = x \otimes y + u \otimes y \quad \text{and} \quad x \otimes (y+z) = x \otimes y + x \otimes z. \end{aligned} \tag{3}$$

Thus the tensor product is *bilinear*. There are no other relations. Note that by definition the products of the form $x \otimes y$ span $A \otimes B$. Hence, if $\{x_i\}$ and $\{y_j\}$ are bases for A and B, respectively, then $\{x_i \otimes y_j\}$ is a basis for $A \otimes B$. The linear space $A \otimes B$ becomes an algebra by defining $(x \otimes y)(u \otimes z) = xu \otimes yz$ and extending by bilinearity. Explicitly,

$$\sum_i a_i (x_i \otimes y_i) \sum_j b_j (u_j \otimes z_j) = \sum_{ij} a_i b_j (x_i u_j \otimes y_i z_j).$$

The $*$ is defined by $(x \otimes y)^* = x^* \otimes y^*$ and extending *anti-linearly*. The problem is defining the norm since it is not a linear function. In fact, for general C^* algebras there could be a number of inequivalent norms on different completions of $A \otimes B$. This problem of non-uniqueness, however, does not exist if one of the factors is abelian or finite-dimensional. Since, in this work we will be primarily concerned with abelian algebras this point will not be discussed further. Our basic model will be an *infinite* tensor product of finite dimensional C^* algebras which we present next.

Let A_k, $k = 1, 2, \ldots$, be finite-dimensional abelian C^* algebras with atomic basis $B_k = \{x_{k1}, \ldots, x_{kn_k}\}$. Let B^∞ be the set consisting of all infinite strings of the form $z_{i_1} \otimes z_{i_2} \otimes \cdots$ where all but a finite number (> 0) of z_{i_k}s are equal to $\mathbb{1}$ and if some $z_{i_k} \neq \mathbb{1}$ then $z_{i_k} \in B_k$. Explicitly, B^∞ consists of strings of the form $z_{i_1} \otimes z_{i_2} \otimes \cdots \otimes z_{i_k} \otimes \mathbb{1} \otimes \mathbb{1} \otimes \cdots, k = 1, 2, \ldots$ and $z_{i_k} \in B_{i_k}$. Let $\tilde{\mathfrak{A}} = \otimes_{i=1}^\infty A_i$ be the vector space with basis B^∞ such that $z_{i_1} \otimes z_{i_2} \otimes \cdots \otimes z_{i_k} \otimes \cdots$ is linear in each factor separately:

$$\begin{aligned} z_{1_1} \otimes \cdots \otimes (a z_{i_k} + b z'_{i_k}) \otimes z_{i_{k+1}} \otimes \cdots = a(z_{1_1} \otimes \cdots \otimes z_{i_k} \otimes z_{i_{k+1}} \otimes \cdots) \\ + b(z_{1_1} \otimes \cdots \otimes z'_{i_k} \otimes z_{i_{k+1}} \otimes \cdots). \end{aligned}$$

Clearly every $\alpha \in \tilde{\mathfrak{A}}$ is a finite linear combination of elements in B^∞. We define a product in $\tilde{\mathfrak{A}}$ as follows. First, for elements of B^∞:

$$(z_{i_1} \otimes z_{i_2} \otimes \cdots)(z'_{i_1} \otimes z'_{i_2} \otimes \cdots) = (z_{i_1} z'_{i_1} \otimes z_{i_2} z'_{i_2} \otimes \cdots).$$

We extend the product to whole of $\tilde{\mathfrak{A}}$ by linearity. Next define a norm by

$$\left\| \sum_{i_1,i_2,\ldots} a_{i_1 i_2 \ldots} z_{i_1} \otimes z_{i_2} \otimes \cdots \right\| = \sup\{|a_{i_1 i_2 \ldots}|\}.$$

It is straightforward to show that B^∞ is an atomic basis. It follows that the above function is indeed an algebra norm and that $\tilde{\mathfrak{A}}$ is an abelian normed algebra. We also define $*$-operation by

$$\left(\sum_{i_1,i_2,\ldots} a_{i_1 i_2 \ldots} z_{i_1} \otimes z_{i_2} \otimes \cdots \right)^* = \sum_{i_1,i_2,\ldots} \overline{a_{i_1 i_2 \ldots}} z_{i_1} \otimes z_{i_2} \otimes \cdots.$$

It is routine to check that for $x \in \tilde{\mathfrak{A}}$, $\|xx^*\| = \|x\|^2$. Finally, we complete the norm and call the resulting C^* algebra $\mathfrak{A}$. The completion of a norm is a technical device that uses the fact that any normed algebra X can be isometrically mapped to a norm complete algebra (a Banach algebra) $\hat{X}$ and the image X is dense in $\hat{X}$ (see Kadison & Ringrose, 1997).[1] With these definitions $\mathfrak{A}$ is a C^* algebra. An important special case is when all the factor algebras $A_i = A$. We then write the infinite tensor product C^* algebra as $\bigotimes^\infty A$. Intuitively, the elements of an atomic basis B^∞ of $\bigotimes^\infty A$ correspond to strings from an alphabet (represented by the basis B). A general element of A is a linear combination of elements of $\bigotimes^\infty B$. Of particular interest is the two-dimensional algebra G corresponding to a binary alphabet. Thus we name $\bigotimes^\infty G$ the *binary* algebra. Let us fix some notation. For any finite-dimensional C^* algebra A the atomic basis B^∞ for $\bigotimes^\infty A$ constructed above will be denoted by B_A^∞ to emphasize the association. The algebras $\bigotimes^\infty A$ will be our model of signals from a source/encoder which are strings (of arbitrary length) from some alphabet. We next prove a result that is relevant for coding theory. The proof is in the appendix.

PROPOSITION 1 Let A be an abelian C^* algebra of dimension n with atomic basis $B_A = \{x_0, \ldots, x_{n-1}\}$. Let $B_G = \{y_0, y_1\}$ be the atomic basis of the two-dimensional algebra G defined above. Then there are injective algebra homomorphisms

$$\mathcal{J} : \bigotimes^\infty G \to \bigotimes^\infty A \quad \text{and} \quad \mathcal{J}' : \bigotimes^\infty A \to \bigotimes^\infty G$$

that are isometries.

Let us note that from the injective *set maps* j and j' (see the proof in the Appendix) we can construct a *bijective* correspondence between the bases B_A^∞ and B_G^∞ by a Schroeder–Bernstein-type construction (see Kleene, 1952) and this can be lifted to an algebra isometry. But for us, the isomorphisms induced by maps like j and j' (these are certainly not unique) will be greatest interest. Essentially, what the proposition says is that it is often sufficient to restrict our attention to the special algebra $\bigotimes^\infty G$. This proposition together with the algebraic formulation of the communication model implies that it is sufficient to consider a binary alphabet for theoretical purposes.

The next step is to describe the state space. We recall that states of an algebra A are precisely the positive functionals ω that are normalized: $\omega(\mathbb{1}) = 1$. Given a C^* subalgebra $V \subset A$ the set of states of

[1] There are some delicate convergence issues here. Since $\tilde{\mathfrak{A}}$ consisting of finite sums of tensor products is dense in $\mathfrak{A}$ it often suffices to prove some statement about $\tilde{\mathfrak{A}}$ and extended it to $\mathfrak{A}$ by continuity.

V will be denoted by $\mathscr{S}(V)$. Let $\mathfrak{A} = \bigotimes_{i=1}^{\infty} A_i$ denote the infinite tensor product of finite-dimensional algebras A_i. An infinite product state of $\mathfrak{A}$ is a functional of the form

$$\Omega = \omega_1 \otimes \omega_2 \otimes \cdots \text{ such that } \omega_i \in \mathscr{S}(A_i).$$

This is indeed a state of $\mathfrak{A}$ for if $\alpha_k = z_1 \otimes z_2 \otimes \cdots \otimes z_k \otimes \mathbb{1} \otimes \mathbb{1} \cdots \in \mathfrak{A}$ then

$$\Omega(\alpha) = \omega_1(z_1)\omega_2(z_2)\cdots\omega_k(z_k),$$

a *finite* product. Since an arbitrary element of $\mathfrak{A}$ is the limit of sequence of finite sums of elements of the form $\alpha_k,\ k = 1, 2, \ldots, \Omega$ is bounded by the principle of uniform boundedness (Rudin, 1987). Clearly, it is positive. A general state on $\mathfrak{A}$ is a convex combination of product states like Ω.

2.2 *Analytic functions on C^* algebras*

In this section, we discuss another useful construction. Let A be a C^* algebra. Suppose $f(z)$ is an analytic function whose Taylor series $\sum_{n=0}^{\infty} a_n(z-c)^n$ is convergent in a region $|z-c| < R$. The convergence of the series $\sum \|x - c\mathbb{1}\|^n$ for $\|x - c\mathbb{1}\| < R$ implies that the series $\sum_{n=0}^{\infty}(x - c\mathbb{1})^n$ converges (we need completeness of A for this). Thus it makes sense to talk of analytic functions on a C^* algebra. If we have an atomic basis $\{x_1, x_2, \ldots\}$ in an abelian C^* algebra then the functions are particularly simple in this basis. Thus if $x = \sum_i a_i x_i$ then $f(x) = \sum_i f(a_i)x_i$ provided that $f(a_i)$ are defined in an appropriate domain. We will mostly take this as our definition with the understanding that the constant function c is identified with $c\mathbb{1}$.

3. Algebraic approach to probability

We have observed that discrete signals from a source are modelled by an abelian algebra. The elements of the algebra correspond to random variables representing the output of the source. With random variables we always associate a probability distribution. In the standard treatment of probability theory the probability or sample space is introduced first. Random variables are defined as (measurable) real (or complex) functions on this space. One then finds the probability distributions of the random variables and most important quantities like mean, variance and correlations are based on these distributions. In particular, the mean or expectation values of functions of random variables play a central role. Note that random variables can be added and multiplied making it a real algebra (scalars are the constant random variables). Note also that random variables also represent quantities that are actually measured or observed the voltage across a resistor, the currents in an antenna, the position of a Brownian particle and so on. The probability distribution corresponds to the *state* of the devices that produce these outputs. These are often *inferred* from observations. We will take the alternative view and start with these observables as our basic objects. In this way, we single out the objects which are relevant to a specific problem. In the following paragraphs, we formalize these notions.

3.1 *Basic notions*

A classical observable algebra is an abelian complex C^* algebra A. It is convenient to use complex algebras. We can restrict our attention to real algebras whenever necessary. Recall that a state on A is positive linear functional ω such that $\omega(\mathbb{1}) = 1$. We can identify ω with a *probability measure* as follows. Suppose $(M, \mathcal{S}, P)$ is probability space (M= sample space, $\mathcal{S}$ = σ algebra, P = probability measure). Let $L_\infty(M, \mathcal{S}, P)$ (or simply $L_\infty(M)$ if the measure structure is clear) be the set of essentially

bounded measurable complex functions.[2] We can give it a C^* structure as in the case of $C(X)$, the space of continuous functions on a compact topological space X (see equation (1)), but using the essential supremum instead of the ordinary supremum. If $B \in \mathcal{S}$, then the *indicator* function $I_B \in L_\infty(M, \mathbb{C})$ and

$$\int_M I_B \, \mathrm{d}P = P(B),$$

where the integral is defined in the sense of Lebesgue. Note that $\omega_P(f) \equiv \int f \, \mathrm{d}P$ is a positive linear functional on $L_\infty(M)$. Since $\omega_P(\mathbb{1}) = 1$, it is a state.

DEFINITION 1 A *probability algebra* is a pair (A, S) where A is an observable algebra and $S \subset \mathscr{S}(A)$ is a set of states. A probability algebra is defined to be *fixed* if S contains only one state. A probability algebra $\mathscr{A}_1 = (A_1, S_1)$ is defined to be a *cover* of another $\mathscr{A}_2 = (A_2, S_2)$ if there is an algebra homomorphism $\phi : A_1 \to A_2$ and a one-to-one correspondence $\gamma : S_1 \leftrightarrow S_2$ such that the following conditions hold: (i) ϕ is onto and (ii) for all $x \in A_1$ and $\omega \in S_1$: $\omega(x) = \gamma(\omega)(\phi(x))$.

Our definition of cover is similar to the one defined in terms of *subfields* of probability space (Bahadur, 1955) but is more general since we do not restrict to a fixed a probability space. In order to see the connection with probability space, we give an example. Let $(M_1, \mathcal{S}_1, P_1)$ and $(M_2, \mathcal{S}_2, P_2)$ be two probability spaces. Let A_i be the algebra of bounded random variables on M_i, $i \in \{1, 2\}$. Let S_i consist of a single element ω_i defined by $\omega_i(X_i) = \int X_i \, \mathrm{d}P_i$ with $X_i \in A_i$. Suppose that we have maps ϕ and $\gamma(\omega_1) = \omega_2$ satisfying the conditions above. Then if $X \in \mathrm{Ker}(\phi)$ then $X^*X \in \mathrm{Ker}(\phi)$. This implies that $\int X^*X \, \mathrm{d}P_1 = 0$. Since $X^*X \geqslant 0$, X must vanish a.e. Conversely if X vanishes a.e. then so does $\phi(X)$. So, the kernel consists of 'negligible' functions.

Let ω be a state on an abelian C^* algebra A. Call two elements $x, y \in A$ uncorrelated in the state ω if $\omega(xy) = \omega(x)\omega(y)$. Note that this definition depends crucially on the state: the same two elements can be correlated in some other state ω'. Two natural questions are immediate. Are there any states for which *every* pair of elements of A are uncorrelated? Are there a pair of elements which are uncorrelated in *every* state? Two trivial candidates for the second question are $\mathbb{1}$ and 0. Either of them is uncorrelated to every element. We implicitly exclude these two trivial cases. Concerning the second question the answer is negative in general. On the first question, a state ω is called multiplicative if $\omega(xy) = \omega(x)\omega(y)$ for all $x, y \in A$. Note that the notion of positivity defines a partial order on the space of functionals making it an ordered vector space (Kadison & Ringrose, 1997). The set of states, $\mathscr{S}$, is convex in the usual sense that for numbers $p_i \geqslant 0$, $\sum_{i=1}^k p_i = 1$ and states ω_i, $i = 1, \ldots, k$ the functional $\sum_i p_i \omega_i$ is also a state. The extreme points of $\mathscr{S}$ are called *pure* states. In the case of abelian C^* algebras a state is pure if and only of it is multiplicative (Kadison & Ringrose, 1997). Thus in a pure state any two observables are uncorrelated. This is not generally true in the non-abelian quantum case.

Next we come to the important notion of *independence*. First, given $B \subset A$ let $A(B)$ denote the subalgebra generated by B (the smallest subalgebra of A containing B). Two subsets $B_1, B_2 \subset A$ are defined to be *independent* if all the pairs $\{(x_1, x_2) : x_1 \in A(B_1), x_2 \in A(B_2)\}$ are uncorrelated. As independence and correlation depend on the state we sometimes write ω-independent/uncorrelated to emphasize this. Clearly, independence is much stronger condition than being uncorrelated. It is easy to construct examples in three or more dimensions where a pair of observables x, y are uncorrelated but they are not independent: for example, x^2 and y maybe correlated. However, in two dimensions x and y are uncorrelated if and only if one of them is 0 or $c\mathbb{1}$. Let us note that as in the quantum case two dimensions is

[2] A function f is said to be essentially bounded if there is a constant K such that $|f(x)| \leqslant K$ almost everywhere (a.e.). The essential supremum is the infimum over all such K: ess $\sup(|f|) = \inf\{k : P\{x : |f(x)| > k\} = 0\}$.

an exceptional case. The next theorem (see Appendix for a proof) shows the structural implications of independence.

THEOREM 3 Two sets of observables S_1, S_2 in a finite-dimensional abelian C^* algebra A are independent in a state ω if and only if for the (unital) subalgebras $A(S_1)$ and $A(S_2)$ generated by S_1 and S_2, respectively, there exist states $\omega_1 \in \mathscr{S}(A(S_1)),\ \omega_2 \in \mathscr{S}(A(S_2))$ such that $(A(S_1) \otimes A(S_2), \{\omega_1 \otimes \omega_2\})$ is a cover of $(A(S_1S_2), \omega')$ where $A(S_1S_2)$ is the subalgebra generated by $\{S_1, S_2\}$ and ω' is the restriction of ω to $A(S_1S_2)$.

We can even extend the above theorem (see Appendix for the proof) to infinite tensor product by restricting to finite segments. The next step is to extend the notion of independence to more than two subsets. Let $S_1, \ldots, S_k \subset A$ and ω a state of A. Then the subsets are defined to be ω-independent if for all $x_i \in A(S_i),\ i = 1, \ldots, k$ we have

$$\omega(x_1 \cdots x_k) = \omega(x_1) \cdots \omega(x_k).$$

Here $A(S_i)$ is the subalgebra generated by S_i. We can then show that for states $\omega_i \in (A(S_i))$, the restriction of ω to $A(S_i)$ the pair $(A(S_1) \otimes \cdots \otimes A(S_k), \omega_1 \otimes \cdots \otimes \omega_k)$ is a cover of $A(S_1 \cdots S_k), \omega'$, where ω' is the restriction of ω to $A(S_1 \cdots S_k)$, the algebra generated by $S_1, \ldots, S_k$. We thus see the relation between independence and (tensor) product states in the classical or commutative theory. The non-commutative or quantum case is more delicate and requires careful handling.

3.2 *Probability distribution functions*

In this section, we investigate another important concept of probability theory—the (cumulative) distribution function (d.f)—in the algebraic framework. As the paper's primary concern is an alternative formulation of mathematical models of information and communication we do not undertake an extensive exploration of the algebraic approach to probability concepts. However, the notion of a distribution function underpins large part of probability theory and its applications. One of the advantages of using C^* or more general Banach algebra is that we have both algebraic and analytical methods at our disposal. The textbook definition of the distribution function of a random variable X on a probability space $\{M, P\}$ (here P is the probability measure) is as follows. It is the function

$$F(x) = P\{m \in M : X(m) \leqslant x\} \equiv P\{X \leqslant x\}.$$

Now any set $N \subset M$ is characterized by its indicator function I_N. Thus, $F(x) = E(I_{\{X \leqslant x\}})$, where E is the expectation. The indicator function is not continuous in general, and so, it will not belong to the algebra of continuous functions. However, it can be approximated by continuous functions that vanish at infinity. This is the motivation for our use of a technical device called the approximate unit in C^* algebras to define: given a subalgebra $B \subset A$ of an abelian C^* algebra let $B_a = \{x \in A : xs = 0\ \forall s \in B\}$ be the *annihilator* of S. This is an ideal[3] hence there is an *approximate identity*. An approximate identity in an ideal B is a *net* $\{y_\lambda\}$ with $0 \leqslant y_\lambda \leqslant \mathbb{1}$ such that $xy_\lambda \to x$ (also $y_\lambda x \to x,\ \forall x \in B$ if the algebra is non-abelian). For the details, see Kadison & Ringrose (1997). Obviously, S_a cannot contain the identity of the original algebra unless $S = \{0\}$. We only mention that nets (Kelley, 1975) are generalization of sequences where the indexing set is not required to be countable. However, in the case of separable

[3] An ideal of a algebra A is a subset I of A which is closed under addition and for every $x \in A$, $xI \subset I$. A non-zero proper ideal cannot contain the identity of A.

algebras (algebras with a dense countable set) the reader may substitute 'sequence' for 'net'. In the following it will suffice for our purpose to restrict to the separable case although we often use the language of 'nets'. We can now define the distribution of a set of observables.

DEFINITION 2 Let $S = \{x_1, x_2, \ldots, x_n\}$ be a finite self-adjoint subset of A where (A, ω) is a fixed probability algebra. For $\mathfrak{t} = (t_1, t_2, \ldots, t_n) \in \mathbb{R}^n$ let $S_\mathfrak{t} \subset A$ denote the set of elements $\{(t_i \mathbb{1} - x_i) : i = 1, \ldots, n\}$ and $S_\mathfrak{t}^-$ the set of elements $\{z_- : z \in S_\mathfrak{t}\}$, negative parts of members of $S_\mathfrak{t}$. Let $\{e_\lambda(\mathfrak{t})\}$ be approximations of identity in the annihilator ideal $(S_\mathfrak{t}^-)_a$. Then the ω-distribution of S is defined to be the real function

$$F_S(\mathfrak{t}) = \lim_\lambda \omega(e_\lambda).$$

The rationale for this definition is simple. For convenience, restrict to a single random variable. Suppose X is a bounded random variable on a probability space $\{\Omega, \mathcal{S}, P\}$. For a real number t let $X_t = tI - X$, a random variable. Then the (cumulative) distribution function of X is the probability of the event E_t, where $E_t = \{\alpha \in \Omega : X_t(\alpha) \geqslant 0\}$. For a fixed t, let $X_t = X_{t+} - X_{t-}$, the difference between the two non-negative random variables, the positive and negative parts of X_t. Consider now X_{t-} and let $G_t = \{\alpha : X_t(\alpha) < 0\} = \Omega - E_t$. Then X_{t-} is > 0 on G_t and 0 outside it. If Y is any function on Ω such that $YX_{t-} = 0$ then Y must vanish on G_t. Conversely, any function Y that vanishes on G_t satisfies the equation $YX_{t-} = 0$. In particular, the indicator function $\mathscr{I}_{E_t}$ satisfies it. The function $\mathscr{I}_{E_t}$ is the identity on $(X_{t-})_a$ and its expectation value $\int \mathscr{I}_{E_t}\, dP = P(E_t)$. Although the indicator functions are not generally continuous, we can approximate them by a sequence of continuous functions. This sequence is an approximate identity in the C^* algebra of continuous functions. In most cases, the algebras will be separable. Then the nets can be replaced by sequences. Note that since the net $\{e_\lambda\}$ is bounded and increasing the net $\{\omega(e_\lambda)\}$ converge. Finally, let us observe that even though the approximate identity is not unique the distribution function as defined above is unique. To prove this suppose $\{e_\lambda\}, \{f_\mu\}$ are two approximate identities. Then using the fact $\omega(e_\lambda f_\mu - e_{\lambda'} f \mu') = \omega(f_\mu(e_\lambda - e_{\lambda'}) + e_{\lambda'}(f_\mu - f_{\mu'}))$ is Cauchy as $f_\mu(e_\lambda - e_{\lambda'}) \to (e_\lambda - e_{\lambda'})$ and $e_{\lambda'}(f_\mu - f_{\mu'}) \to f_\mu - f_{\mu'}$, we conclude that the double-net $\{\omega(e_\lambda f_\mu)\}$ converges to the limit $\lim_\lambda \omega(e_\lambda) = \lim_\mu \omega(f_\mu)$. Extending the definition of the d.f to an arbitrary element z in the algebra is simple. Write $z = x + iy$ where x and y are self-adjoint. Let $F_x(t)$ and $F_y(t)$ denote the d.f of x and y, respectively. Then the d.f of z: $F_z(t) = F_x(t) + iF_y(t)$.

THEOREM 4 Let $x_1, \ldots, x_n$ be self-adjoint elements of an abelian C^* algebra A. Let $F(t_1, \ldots, t_n)$ be their joint distribution function. Then $F(t_1, \ldots, t_n)$ is non-negative, left-continuous and non-decreasing in each variable. We also have boundary conditions

$$\lim_{t_1, \ldots, t_n \to \infty} F(t_1, \ldots, t_n) = 1 \quad \text{and} \quad \lim_{t_1, \ldots, t_n \to -\infty} F(t_1, \ldots, t_n) = 0.$$

If the elements are independent and $F(t_i)$ denotes the distribution function of x_i, then

$$F(t_1, \ldots, t_n) = F(t_1)F(t_2) \cdots F(t_n).$$

If a sequence $x_n \to x$ is then the corresponding d.f's $F_{x_n}(t) \to F_x(t)$.

This is of course a standard result in probability theory, but we give an algebraic proof (see Appendix). We see that, starting from a purely algebraic definition of independence and distributions, we can recover their essential properties. In particular, for algebras that are finite or infinite tensor product of finite-dimensional algebras, we have the following explicit characterization.

PROPOSITION 2 Let A be a finite-dimensional abelian C^* algebra. Let $x \in \otimes^\infty A$ and x_a its annihilating ideal. Suppose x is a finite sum. Then there is a unique (up to permutation) decomposition

$$x = \sum a_i P_i \quad \text{such that} \quad P_i P_j = \delta_{ij} P_j \quad \text{and} \quad a_i \neq 0 \text{ distinct.}$$

Further, there exist polynomials without constant term g_i such that $P_i = g_i(x)$. Thus, $x = \sum_i a_i g_i(x)$. Then the ideal x_a has an identity $\mathbb{1} - \sum_i P_i$.

Proof. Since x is finite sum it may be considered as an element of $\otimes^n A$ for some finite n. The space $\otimes^n A$ has a finite atomic basis, say $\{Y_1, \ldots, Y_m\}$ $(m = 2^{\dim(A)})$. Let $x = \sum_{i=1}^m a_i Y_i$ and let $J = \{i : a_i = 0\}$. Then $x = \sum_{i \notin J} a_i Y_i$. Let P_i be the sum of all Y_i for which the coefficients a_i are equal, then $x = \sum_i a_i P_i$ with distinct and non-zero a_i. Next use Lagrange interpolation to obtain polynomials g_i such that $g_i(0) = 0$ and $g_i(a_j) = \delta_{ij}$. To prove uniqueness, let $x = \sum_j b_j Q_j$ be another such decomposition. Then $xP_iQ_j = a_i P_i Q_j = b_j P_i Q_j$. Since $\sum_i P_i x = x$ for a fixed i, there must be at least one j_i with $P_i Q_{j_i} \neq 0$ then $a_i = b_{j_i}$. There cannot be more than one such j_i since the b_j's are distinct. On the contrary, we conclude that $i \leftrightarrow j_i$ is a permutation. The last statement follows trivially. □

Let $x = \sum_i a_i P_i$ be as in the proposition. We call this the spectral decomposition of x. If ω is a state, define

$$\mathscr{J}_\omega(x) = \sum_i \omega(P_i) P_i.$$

The map $\mathscr{J}_\omega(x)$ can be considered as a 'centroid' of the possible outcomes of measurement of x. We can extend the proposition to arbitrary element in $\mathscr{A} = \otimes^\infty A$ by using a sequence of finite-dimensional projections as above to approximate. However, the proposition suffices for most of our requirements. Now let

$$Z = \sum_{k=1}^{\infty} X_k, \quad X_k \in \otimes^k A,$$

Z may *not* be a member of $\mathscr{A}$ in general and we treat the above as a formal sum. However, we suppose that for real t, $(t\mathbb{1} - Z)_+ = (|t\mathbb{1} - Z| + (t\mathbb{1} - Z))/2$ can be expressed as finite sum. We will see an example below. Then the required identity is given as follows. It is clear that for $\delta > 0$ small enough $|(t+\delta)\mathbb{1} - Z| + ((t+\delta)\mathbb{1} - Z) = \sum_k a_k Y_k : a_k > 0$ is finite sum where Y_k constitute an atomic basis. Let $P_\delta = \sum_k Y_k$. Then the required identity is given by $P_0 = \lim_{\delta \to 0} P_\delta$. This is essentially a variant of equation (10) in Theorem 4.

We note that there is another approach to distribution functions, using *characteristic functions*, that can be adapted to our algebraic formulation. Thus given a hermitian element x in a (fixed) probability algebra (A, ω), define the probability density function (p.d.f) of x as

$$f_x(t) = \int_{-\infty}^{\infty} \omega(e^{iv(x - t\mathbb{1})}) \, \mathrm{d}v.$$

This is easily recognized as the inverse Fourier transform of a characteristic function in the standard case (if it exists). However, to prove that it satisfies the properties of p.d.f in the case of a C^* algebra requires some work. This can be done using the operator algebra version of Radon–Nikodym theorem (Sakai, 1971). A more important problem with this approach is that it is quite difficult to extend it to the *non-commutative* or quantum case (see, e.g. Patra & Braunstein, 2011 for more on this).

3.3 *Examples*

In this section, we consider some examples from standard probability theory. It will be demonstrated that the algebraic approach not only gives a different perspective on some familiar situations but also provides additional computational tools. First, we review the correspondence between some concepts from the standard theory with our algebraic model. An event in probability theory is a measurable subset of the probability space. The random variable characterizing any (measurable) subset S is its indicator function I_S. In the algebraic language, it is a projection Q_S. The probability of the event corresponds to the expectation value $\omega(Q_S)$ of the projection. In the cases, we consider the projections will generally exist in the algebra itself. In some cases we consider infinite formal sums which are *not* in the algebra but any finite segment of the sum do belong to the algebra. In the actual computation, we always use a 'cut-off' to restrict to such a finite segment. In the cases where projections are not members of the algebra, we can find a sequence (or net) that 'converges in the mean' to the appropriate projection or indicator function. This situation generally arises in the continuous case which is only touched upon peripherally.

(1) *Binomial distribution.* Consider again infinite sequences of Bernoulli trials as in the second example of the previous section. We can think of coin-tossing with 'heads' signalling success. Let Z be the observable (random variable) corresponding to the number of successes. What is its d.f? Let n, k be a positive integers with $k < n$. We want to find the distribution $F(k:n)$ of Z. Recall that G is the two-dimensional algebra and let $A = \otimes^n G$. Let $\{y_0, y_1\}$ be the atomic basis of G with y_1 corresponding to success. Set

$$\begin{aligned} Z &= \sum_{\mathcal{S}} y_1 \otimes y_0 \otimes \cdots \otimes y_0 + \sum_{\mathcal{S}} 2y_1 \otimes y_1 \otimes y_0 \cdots \otimes y_0 \\ &\quad + \cdots + \sum_{\mathcal{S}} r \underbrace{y_1 \otimes y_1 \otimes \cdots \otimes y_1}_{r} \otimes \underbrace{y_0 \otimes y_0 \otimes \cdots \otimes y_0}_{n-r} \\ &\quad + \cdots + n y_1 \otimes y_1 \otimes \cdots \otimes y_1 \\ &= \sum_{r=1}^{n} r Y_r. \end{aligned}$$

Here $\mathcal{S}$ denotes the distinct permutations of the factors in the tensor product. Thus, the rth term Y_r is the sum of all $\binom{n}{r}$ products with r y_1's. Its value is r. Note that $Y_r Y_s = \delta_{rs}$. We have

$$U = |k\mathbb{1} - Z| - (k\mathbb{1} - Z) = \sum_{r=k+1}^{n} r Y_r.$$

In this case, the identity in the annihilator ideal of U exists and is given by the projection operator $P = \sum_{r=0}^{k} Y_r$. Since

$$F(k:n) = \Omega(P) = \sum_{r=0}^{k} \Omega(Y_r) = \sum_{r=0}^{k} \binom{n}{r} p^r (1-p)^{n-r} \text{ with}$$

$$\Omega = \omega \otimes \omega \otimes \cdots \text{ and } \omega(y_1) = p$$

we get the (cumulative) Bernoulli distribution. Note that we can use the above formula to find the distribution in states where the observables are not independent. Algebraically, the state is not a product state which means the trials are *not* independent.

(2) *Waiting time.* Let us start with a simple version of the problem of waiting time. Suppose that we have a binary source with fixed probability distribution emitting a bit per unit time. The waiting time is the time elapsed before the first appearance of 1. It is a random variable or observable W in our formalism. Using the notation above let

$$W = y_0 \otimes y_1 \otimes \mathbb{1} \otimes \cdots + 2y_0 \otimes y_0 \otimes y_1 \otimes \mathbb{1} \otimes \cdots + 3y_0 \otimes y_0 \otimes y_0 \otimes y_1 \otimes \mathbb{1} \otimes \cdots + \cdots.$$

This is an unbounded infinite sum and does not belong to the algebra. However, for any $t \geqslant 0$,

$$\begin{aligned} F_W(t) &\equiv \frac{|t\mathbb{1} - W| + t\mathbb{1} - W}{2} \\ &= ty_1 \otimes \mathbb{1} + (t-1)y_0 \otimes y_1 \otimes \mathbb{1} + \cdots + (t - \lfloor t \rfloor) \underbrace{y_0 \otimes \cdots \otimes y_0}_{\lfloor t \rfloor \text{ factors}} \otimes y_1 \otimes \mathbb{1} \end{aligned}$$

is finite (of course, $F_W(t) = 0$ for $t < 0$). Here $\lfloor t \rfloor$ is the largest integer $\leqslant t$. Using the trick explained before the examples, we replace t by $t + \delta$ (this is to take into account the case when t is an integer). The required projection (approximate identity) is

$$P_W(t) = y_1 \otimes \mathbb{1} + y_0 \otimes y_1 \otimes \mathbb{1} + \cdots + \underbrace{y_0 \otimes \cdots \otimes y_0}_{\lfloor t \rfloor \text{ factors}} \otimes y_1 \otimes \mathbb{1}.$$

The distribution function in a state Ω is given by $f_W(t) = \Omega(P_W(t))$. If $\Omega = \omega \otimes \omega \otimes \cdots$ is an infinite product state with $\omega(y_1) = p = 1 - \omega(y_0)$, then $F_W(t) = \sum_{k=0}^{\lfloor t \rfloor} p(1-p)^k$.

Next we generalize the problem of waiting time to arbitrary strings. Explicitly, given a string ξ of length n, the waiting time is the time before a contiguous stream of bits matching ξ appears. The preceding case is for $\xi = 1$. We will only construct the observable corresponding to waiting time W in this general case. It gives a nice illustration of the algebraic techniques. Let X be the tensor representation of ξ. Waiting time 0 corresponds to the observable $X \otimes \mathbb{1}$. We use the following notation. Write $\mathbb{1}_1$ for the identity in the two-dimensional space G and $\mathbb{1}_k = \mathbb{1}_1 \otimes \mathbb{1}_1 \otimes \cdots \mathbb{1}_1$, the k-fold tensor product. The symbol $\mathbb{1}$ (without subscripts) will be reserved for the identity in $\otimes^\infty A$. The element $Y_0 = X \otimes \mathbb{1}$ corresponds to waiting time 0: the first n symbols received match the given string. We expect the element corresponding to waiting time 1 will be 'proportional' to $Y_1' = \mathbb{1}_1 \otimes X \otimes \mathbb{1}$. Although Y_0 and Y_1' are projections, they need not be orthocomplementary in the sense $Y_1' Y_0 = 0$. So, they do not correspond to mutually exclusive events. Recall that when interpreted as functions on some measure space projections are indicators of measurable sets (events). We therefore adopt an orthogonalization scheme similar to Gram–Schmidt. The observable $Y_1 = Y_1' - Y_1' Y_0$ is a projection and satisfies $Y_1 Y_0 = 0$. Viewed as a function it takes value 1 only when the input string is of the $\zeta = b_0 \xi \ldots$ and such that the prefix of length n of ζ does not match ξ. It corresponds to waiting time 1. Defining inductively, let

$$\begin{aligned} Y_m &= Y_m' - Y_m'(Y_0 + Y_1 \cdots + Y_{m-1}) \\ &= \mathbb{1}_m \otimes X \otimes \mathbb{1} - \mathbb{1}_m \otimes X \otimes \mathbb{1}(Y_0 + Y_1 \cdots + Y_{m-1}). \end{aligned}$$

It is easily verified that $Y_jY_k = \delta_{jk}Y_k$. The element $W = \sum_{k=0}^{\infty} kY_k$ corresponds to the waiting time in this case. Again it is not an element of the algebra but $Z_t = |t\mathbb{1} - W| + t\mathbb{1} - W$ is. In principle, we can compute the expected waiting time from this.

(3) *Markov chains.* We define a discrete time Markov chain on an observable algebra (A, ω) as a sequence of *positive* and *unital* maps $\{\phi_0, \phi_1, \ldots\}$ and an initial element $z_0 \in A$. Let $\mathcal{A} = \{x_1, x_2, \ldots\}$ be a fixed atomic basis. A configuration is a sequence $\{z_0, z_1, \ldots\}$ where each $z_i \in \mathcal{A}$. The usual term for what we call configuration is simply 'state' but the latter has a very specific meaning in operator algebras. Let $\xi_n = \{z_0, z_1, \cdots, z_n\}$ be a finite segment of the configuration. We are interested in the transition from z_0 to z_n via the path ξ_n. The transition probability is defined recursively as follows.

$$y_1 = \phi_0(z_0), \quad y_k = \phi_{k-1}(z_{k-1}y_{k-1}) \quad \text{and} \quad \text{transition probability } p\left(z_0 \overset{\xi_n}{\to} z_n\right) = \omega(z_n y_n).$$

Let us examine this definition in the special case of stationary Markov chains. A Markov chain is defined to be stationary if all the transition maps are identical: $\phi_0 = \phi_1 = \phi_2 = \cdots$. For a stationary chain

$$\begin{aligned} p(z_0 \overset{\xi_n}{\to} z_n) &= \omega(z_n\phi(z_{n-1}\phi(z_{n-2}\phi(\cdots z_1\phi(z_0))))) \\ &= \omega(z_0)\phi(i_n, i_{n-1})\phi(i_{n-1}, i_{n-2})\cdots\phi(i_1, i_0). \end{aligned}$$

Here $\phi(i,j)$ is the (ij)th matrix element of ϕ with respect to the basis $\mathcal{A}$ and $z_k = x_{i_k}$. This looks very similar to quantum transition probability. In the later case, x_i are projections on a Hilbert space. Further, when we consider transitions over all possible paths, we get an analogue of Feynman's 'sum over paths' for total transition probability.

3.4 *Limit theorems*

The limit theorems of probability theory are important for its theoretical structure as well as its empirical justification. We will be primarily concerned with the bounded case where the proofs are simpler. We state two of these but prove only the *weak law of large numbers*. From information theory perspective, it is perhaps the most useful limit theorem. Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed (i.i.d) random variables on a probability space Ω with probability measure P. Let μ be the mean of X_1 (hence any X_i). We assume that the variance $E(X_1 - \mu)^2$ is bounded. Here, $E(X)$ denotes the expectation value of random variable X.

- *Weak law of large numbers*. Let

$$S_n = \frac{X_1 + \cdots + X_n}{n}.$$

Given $\epsilon > 0$

$$\lim_{n\to\infty} P(|S_n - \mu| > \epsilon) = 0.$$

- *Central limit theorem*. If $0 < E(X_1^2) = \sigma < \infty$, then for any real x as $n \to \infty$

$$P\left(\frac{S_n}{\sqrt{n}\sigma} \leqslant x\right) \to \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-(t-\mu)^2/2\sigma\right) \mathrm{d}x.$$

A few comments about these famous limit theorems. These are statements about different types of convergence (Billingsley, 1995). The theorems can be strengthened but since we are dealing with bounded random variables the above-mentioned formulations suffice. These theorems require assigning of probabilities. All we have at our disposal is the algebra and one or more positive functionals (states) which give us expectation values. But we have already seen how to define probability distribution functions. What we need are appropriate projections or approximations to them. Given a self-adjoint observable x and a real number a write $x-a$ for the element $x-a\mathbb{1}$. Let $A(x-a)_+$ be the (two-sided) ideal generated by the positive part of $x-a$.[4] Let $e_n=(x-a)_+[(x-a)_++\delta_n]^{-1}$, where $0<\delta_n$ such that $\lim_{n\to\infty}\delta_n=0$. Then it can be shown that for any $y\in A(x-a)_+$, $\lim_{n\to\infty}e_ny\to y$ in the norm. Hence, $\{e_n\}$ is an increasing *sequence* approximating identity (see Takesaki, 2002). We write $\mathbb{P}(x>a)$ for this approximate identity in $A(x-a)_+$. It is not unique but that does not matter since all the limits that we use it to define are independent of the particular choice. The probability corresponding to the 'event' $x>a$ is defined to be $P(x>a)=\omega(\mathbb{P}(x>a))=\lim_{n\to\infty}\omega(e_n)$. Similarly, we can define $P(x<a)=\omega(\mathbb{P}(x<a))$, where $\mathbb{P}(x<a)=\{f_n\}$ is an approximate identity in the ideal $A(x-a)_-$ obtained by replacing $(x-a)_+$ by $(x-a)_-$ in e_n. We can define more complicated events by algebraic operations, but it is not necessary for what follows. We also note that although we use probabilistic language in the statements of the results below all the expressions are actually defined in a strictly algebraic setting without reference to any underlying probability space.

LEMMA 1 (Chebyshev inequality) Let $x,y\in A$ be self-adjoint where (A,ω) is an observable algebra and $y\geqslant 0$. For any number $\epsilon>0$ we have

$$P(y>\epsilon)\leqslant\frac{\omega(y)}{\epsilon}\ \text{(Markov)}$$

and

$$P(|x-\omega(x)|>\epsilon)\leqslant\frac{\omega([x-\omega(x)]^2)}{\epsilon^2}\ \text{(Chebyshev)}.$$

Proof. Let $\{e_n\}$ be an approximate identity in the ideal $A(y-\epsilon)_+$. By definition $e_n\leqslant\mathbb{1}$. Hence, $\omega(y)=\omega(ye_n)+\omega(y(\mathbb{1}-e_n))\geqslant\omega(ye_n)$. Since $(y-\epsilon)_-$ annihilates the ideal $A(y-\epsilon)_+$, $\omega(ye_n)=\omega([y-\epsilon]e_n)+\omega(\epsilon e_n)=\omega([y-\epsilon]_+e_n)+\epsilon\omega(e_n)\geqslant\epsilon\omega(e_n)$. Hence, $\omega(y)\geqslant\epsilon\omega(e_n)$. Taking limits we obtain the first inequality. Observe that for any $x\in A$, $P(|x|>\epsilon)=P(|x|^2>\epsilon^2)$ for the ideals $A(|x|-\epsilon)_+$ and $A(|x|^2-\epsilon^2)_+$ coincide. This follows from the identity $|x|^2-\epsilon^2=(|x|+\epsilon)(|x|-\epsilon)$ and hence $(|x|^2-\epsilon^2)_+=(|x|+\epsilon)(|x|-\epsilon)_+$ plus the fact that $|x|+\epsilon$ is invertible. Now, as $P(|x-\omega(x)|>\epsilon)=P(([x-\omega(x)]^2>\epsilon^2)$ for self-adjoint x, the second inequality follows from the first by putting $y=(x-\omega(x))^2$ and using ϵ^2 in place of ϵ. □

We will prove next a convergence result which implies the weak law of large numbers.

THEOREM 5 (Law of large numbers (weak)) If $x_1,\ldots,x_n,\ldots$ are ω-independent self-adjoint elements in an observable algebra and $\omega(x_i^k)=\omega(x_j^k)$ for all positive integers i,j and k (they are i.i.d) then

$$\lim_{n\to\infty}\omega\left(\left|\frac{x_1+\cdots+x_n}{n}-\mu\right|^k\right)=0,\quad \mu=\omega(x_1)\quad\text{and}\quad k>0.$$

[4] The ideal generated by a subset B of a C^* algebra A is the smallest ideal containing B.

Proof. We may assume $\mu = 0$ (by reasoning with $x_i - \omega(x_i)$ instead of x_i). First we prove the statement for $k = 2$. Then $\omega(|(x_1 + \cdots + x_n)/n|^2) = \sum_i \omega(x_i^2)/n^2 = \omega(x_1^2)/n$. The first equality follows from independence ($\omega(x_i x_j) = \omega(x_i)\omega(x_j) = 0$ for $i \neq j$) and the second from the fact that they are i.i.d. The case $k = 2$ is now trivial. Now let $k = 2m$. Then $|x_1 + \cdots + x_n|^k = (x_1 + \cdots + x_n)^k$. Put $s_n = (x_1 + \cdots + x_n)/n$. Expanding s_n^k in a multinomial series we note that independence and the fact that $\omega(x_i) = 0$ implies that all the terms in which at least one of the x_i has power 1 do not contribute to $\omega(s_n^k)$. The total number of the remaining terms is $O(n^{2m} - 1)$. Since the denominator is n^{2m}, we see that $\omega(s_n^k) \to 0$. Since for any $x \in A$, $|x| = (x^2)^{1/2}$ can be approximated by polynomials in x^2 we conclude that $\omega(|s_n|) \to 0$. Finally, using the Cauchy–Schwartz-type inequality $\omega(|s_n|^{2r+1}) \leqslant \omega(s_n^2)\omega(s_n^{2r})$, we see that the theorem is true for all k. □

COROLLARY 2 Let $x_1, \ldots, x_n$ and μ be as in the theorem and set $s_n = (x_1 + \cdots + x_n)/n$. Then for any $\epsilon > 0$ there exist n_0 such that for all $n > n_0$

$$P(|s_n - \mu| > \epsilon) < \epsilon.$$

Proof. Using Chebyshev inequality, we have $P(|s_n - \mu| > \epsilon) = P(|s_n - \omega(s_n)| > \epsilon) \leqslant \omega(|s_n - \mu|^2)/\epsilon^2$. As $\omega(|s_n - \mu|^2) \to 0$ (Theorem 5), there is n_0 such that $\omega(|s_n - \mu|^2) < \epsilon^3$ for $n > n_0$. □

4. Communication and information

We now come to our original theme: an algebraic framework for communication and information processes. We can view information as a measure of our state of ignorance or uncertainty in the following sense. We are uncertain about the outcomes of a certain experiment. After the observation process, this uncertainty is removed and we we have gained some information, the more the uncertainty higher the gain in information. Mathematically, we associate a numerical value with a probability distribution of some physical quantity which we identify with an observable. Any manipulation of the quantity, for example, transmitting it or measuring it is given by some operation on the observable. Since our primary goal is the modelling of information processes, we refer to the simple model of communication in Section 1 and recast it in the algebraic framework.

4.1 *Source and coding*

DEFINITION 3 A source is a pair $\mathscr{S} = (X, S)$ where $X \subset A$, A a C^* algebra and S is a set of states. A source is static if S consists of single state. It is discrete if X is countable.

This definition abstracts the essential properties of a source. A real source could be an animate (human speech, for example) or inanimate object (a radio transmitter, for example). Its output can be considered discrete, for example, a keyboard with a fixed alphabet or continuous like radiation from a star. In this work we will be mainly concerned with discrete sources. Then X will be called the *source alphabet*. We assume that at each instant there is a probability distribution on the letters of the alphabet characterizing the *state* of the source at that instant. Thus a discrete source is a countable set of random variables. In the algebraic view, it is a sequence $X = \{x_n\}$ of elements a C^* algebra. The set of states S, called the states of the source, provide the probability distributions. If this distribution does not change (equivalently S consists of a single element) then we have a static source. We will mostly deal with static sources in this work. When we model transmission of information as a Markov process, the state of the source is identified with the initial probability distribution. There is dual view. Suppose that a source $\mathscr{S}$ emits letters from a finite alphabet. This implies that we can distinguish two distinct elements of X

after an observation. We say that the source is unambiguous and model it by demanding that distinct members of X are orthocomplementary. Then the set X in the above definition is a subset of an atomic basis (corresponding to the alphabet) of the algebra A. We will exclusively deal with unambiguous sources. The reason for this elaborate definition is that in general *quantum* sources do not satisfy this. Henceforth, the term source will mean a discrete, static, unambiguous source unless stated otherwise. For a state ω, define

$$\mathcal{O}_\omega = \sum_{i=1}^{n} \omega(x_i)x_i, \ \{x_1, \ldots, x_n\} \text{ an atomic basis.}$$

We say that $\mathcal{O}_\omega$ is the output of the source in state ω. Intuitively, $\mathcal{O}_\omega$ is a kind of mean 'point' in the space of outputs (compare it with the notion of centre of mass in mechanics). More importantly, it facilitates the calculation of important quantities and has close analogy with the quantum case. The quantum analogue may be pictured as follows. The source outputs 'particles' in definite 'states' x_i with probability $p_i = \omega(x_i)$. Note that here state corresponds to a projection operator. A measurement for x_i means applying the dual operator ω_i ($\omega_i(x_j) = \delta_{ij}$) giving $\omega_i(\mathcal{O}_\omega) = p_i$.

Let $\mathscr{Z} = (X, \omega)$ be a source. Suppose every $x \in X$ belongs to a finite-dimensional subalgebra generated by a (finite) set of ω-independent elements. Then using the Theorem 3 we may assume that $A = \bigotimes^\infty B$ where B is finite-dimensional abelian C^* algebra and ω is an (infinite) product state. In this case, each element of X is a tensor product of elements of an atomic basis $\mathcal{B}$ of B. The basis $\mathcal{B}$ is identified with the alphabet and X with strings from the alphabet. In the rest of the paper, we assume that X is the product basis of atomic elements. For example, if B is the two-dimensional algebra with atomic basis $\{y_0, y_1\}$, then X is the set of elements of the form $z_1 \otimes z_2 \otimes \cdots \otimes z_k \otimes \mathbb{1} \otimes \mathbb{1} \otimes \cdots$, where $z_i \in \{y_0, y_1\}$.

4.2 *Source coding*

Let B be a finite-dimensional C^* algebra and $A = \bigotimes^\infty B$. We consider $\otimes^n B$ as a subalgebra of A via the standard embedding (all 'factors' beyond the nth place equal $\mathbb{1}$). Let X_n be its atomic basis in some fixed ordering and let $X = \bigcup_n X_n$. Let B' be another finite-dimensional C^* algebra and $A' = \bigotimes^\infty B'$. A source coding is a linear map $f : B \to T \subset \sum_{k \geqslant 1}^m \otimes^k B'$. Here T is the linear subspace. It induces a (linear) map

$$\otimes^n f : \otimes^n B \to A' \text{ given by } \otimes^n f(x_1 \otimes \cdots \otimes x_n) = f(x_1) \otimes \cdots \otimes f(x_n),$$

where $\otimes^n f$ extends to a unique map $F : A \to A'$. Note that we first induce a map on $\otimes^n B$, $n = 1, 2, \ldots$, and *then* lift it to A. We allow the map f to take values that are not simple products. Moreover, f need *not* be unital. For classical communication, we require that each atomic basis element $x_i \in B$ be mapped to a tensor product of atomic basis elements. Since we are dealing with classical information, it will be implicitly assumed that all the codes are classical. Let us consider an example to clarify these points.

EXAMPLE Let $\{x_0, x_1, x_2, x_3\}$ be an atomic basis for B. Let $B' = G$ with atomic basis $\{y_0, y_1\}$. Define f_1 by $f_1(x_0) = y_0, f_1(x_1) = y_1, f_1(x_2) = y_0 \otimes y_1$ and $f_1(x_3) = y_1 \otimes y_0$. Denote by $\hat{f}_1$ its extension to tensor products. Since $\hat{f}_1(x_0 \otimes x_1) = y_0 \otimes y_1 = \hat{f}_1(x_2)$, $\hat{f}_1$ is not injective. Hence it cannot be inverted on its range. Consider next the map $f_2(x_0) = y_0, f_2(x_1) = y_0 \otimes y_1, f_2(x_2) = y_0 \otimes y_1 \otimes y_1$ and $f_2(x_3) = y_1 \otimes y_1 \otimes y_1$. This map is invertible but one has to look at the complete product before finding the inverse. It is not *prefix-free*.

Now going back to the general formulation, a code $f : B \to T$ is defined to be prefix-free if for distinct members x_1, x_2 in an atomic basis of B, $f'(x_1)f'(x_2) = 0$ where f' is the map $f' : B \to \bigotimes^\infty B'$

induced by f. That is, distinct elements of the atomic basis of B are mapped to *orthocomplementary* elements. Recall that two elements x, y of an algebra are considered orthocomplementary if their product $xy = 0$. Now, in the standard formulation an alphabet is a finite set and a code is a map from $Y \to Z^+$ where Y, Z are alphabets and Z^+ is the set of non-empty finite strings from Z. The definition of prefix-free in this case is clear. In the algebraic language, the free monoidal structure defined by concatenation is replaced by the tensor structure. Then the 'code-word' $z_1 \otimes z_1 \otimes \cdots \otimes z_k \otimes \mathbb{1} \otimes \mathbb{1} \otimes \cdots$ is not orthogonal to another $z'_1 \otimes z'_1 \otimes \cdots \otimes z'_m \otimes \mathbb{1} \otimes \mathbb{1} \otimes \cdots$ with $k \leqslant m$ if and only if $z_1 = z'_1, \ldots, z_k = z'_k$. We observe that one has to be careful about correspondence between the two approaches. For example, one might be tempted to identify the identity $\mathbb{1}$ with the empty string but the $\mathbb{1}$ is the sum of the members of an atomic basis! The binary operation '+' has a relatively lesser role in the classical formalism, but it is crucial in the quantum framework (via superposition principle). Our first result is a useful and well-known inequality proved using algebraic techniques.

LEMMA 2 (Kraft inequality) Let B be an n-dimensional abelian C^* algebra. Corresponding to a finite sequence $k_1 \leqslant k_2 \leqslant \cdots \leqslant k_m$ of positive integers, let $\alpha_1, \ldots, \alpha_m$ be a set of prefix-free elements in $\sum_{i \geqslant 1} \otimes^i B$ such that $\alpha_i \in \otimes^{k_i} B$. Further, suppose that each α_i is a tensor product of elements from a fixed atomic basis of B. Then

$$\sum_{i=1}^{m} n^{k_m - k_i} \leqslant n^{k_m}. \tag{4}$$

Proof. Let $= \{y_1, \ldots, y_n\}$ be the fixed atomic basis of B and set $k_m = M$. We can then restrict our attention to the finite-dimensional algebra $Z = \sum_{i=1}^{M} \otimes^i B$. Let $\alpha_1 = z_1^1 \otimes \cdots \otimes z_{k_1}^1 \otimes \mathbb{1} \otimes \cdots \otimes \mathbb{1}$ where $z_i^1 \in$. Let $\beta = z_1^1 \otimes \cdots \otimes z_{k_1}^1$ and

$$Z_1 = \{\beta \otimes \gamma : \gamma \in \otimes^{M-k_1} B\}.$$

Then $Z_1 \subset \otimes^M B$ is a subalgebra (without unit) of dimension n^{M-k_1}. The assumption that α_i are prefix-free implies $\alpha_2, \alpha_3, \ldots, \alpha_M$ must be in Z'_1 the 'orthogonal' complement to Z_1 in Z. Dimension of $Z'_1 = n^M - n^{M-k_1}$. Repeating this argument with $\alpha_2, \ldots, \alpha_{k_{m-1}}$ we conclude that α_{k_m} must be in a subspace of dimension $n^M - n^{M-k_1} - n^{M-k_2} - \cdots - n^{M-k_{m-1}}$. Since α_{k_m} is non-zero the $n^M - n^{M-k_1} - n^{M-k_2} - \cdots - n^{M-k_{m-1}} \geqslant 1$. This is equivalent to the relation (4). □

With the notation of the lemma we call the sequence $W = \{\alpha_1, \ldots, \alpha_m\}$ decipherable if the tensor product of any two distinct finite-ordered sequence of elements from W are distinct. The sequences may have repeated elements. The Kraft inequality is valid for decipherable sequences (MacMillan, 1953). However, the proof is essentially combinatorial. The Kraft inequality also provides a sufficiency condition for prefix-free code (Ash, 1990; Cover & Thomas, 1999). Thus the existence of a decipherable code of word lengths $(k_1, k_2, \ldots, k_m)$ implies the existence of a prefix-free code of same word lengths. In the following, we restrict ourselves to prefix-free codes. If $g : A \to \otimes^\infty B$ is a prefix-free code, then it maps orthogonal elements to orthogonal elements. It is therefore an algebra isomorphism (a one-to-one homomorphism). Next we have a technical lemma that is useful in finding bounds.

LEMMA 3 Let f be a continuous real function on $(0, \infty)$ such that $xf(x)$ is convex and $\lim_{x \to 0} xf(x) = 0$. Let A be a finite-dimensional C^* algebra with atomic basis $\{x_1, \ldots, x_n\}$ and ω a state on A. Then for any set of numbers $\{a_i : i = 1, \ldots, n;\ a_i > 0$ and $\sum_i a_i \leqslant 1\}$ we have

$$\omega\left(\sum_i f\left(\frac{\omega(x_i)}{a_i}\right) x_i\right) \geqslant f(1).$$

Proof. Let $\omega(x_i) = p_i$. We have to show that $\sum p_i f(p_i/a_i) \geqslant f(1)$. First assume that all $p_i > 0$ and $\sum_i a_i = 1$. Then

$$\sum_i p_i f(p_i/a_i) = \sum_i a_i \frac{p_i}{a_i} f\left(\frac{p_i}{a_i}\right) \geqslant f\left(\sum p_i\right) = f(1)$$

by convexity of $xf(x)$. The general case can be proved by starting with a_i corresponding to $p_i > 0$ and adding extra a_j's to satisfy $\sum_i a_i = 1$ if necessary. The corresponding p_j is set to 0. Now define a new function $g(x) = xf(x)$, $x > 0$ and $g(0) = 0$. The conclusion of the lemma follows by arguing as above with g. □

Using the lemma for the function $f(x) = \log x$ and Lemma 2, we easily deduce the following.

PROPOSITION 3 (Noiseless coding) Let $\mathscr{S}$ be a source with output $\mathcal{O}_\omega \in A$, a finite-dimensional C^* algebra with atomic basis $\{x_1, \ldots, x_n\}$ (the alphabet). Let g be prefix-free code such that $g(x_i)$ is a tensor product of k_i members of the code basis. Then

$$\omega\left(\sum_i \log nk_i x_i + \log \mathcal{O}_\omega\right) \geqslant 0.$$

Next we give a simple application of Theorem 5. First define a positive functional $\mathtt{Tr}$ on a finite-dimensional abelian C^* algebra A with an atomic basis $\{x_1, \ldots, x_d\}$ by $\mathtt{Tr} = \omega_1 + \cdots + \omega_d$, where ω_i are the dual functionals to the basis: $\omega_i(x_j) = \delta_{ij}$. It is clear that $\mathtt{Tr}$ is independent of the choice of atomic basis. Informally, the function $\mathtt{Tr}$ gives the dimension of a projection.

THEOREM 6 (Asymptotic Equipartition Property (AEP)) Let $\mathscr{S}$ be a source with output $\mathcal{O}_\omega = \sum_{i=1}^d \omega(x_i)x_i$ where ω is a state on the finite-dimensional algebra with atomic basis $\{x_i\}$. Then given $\epsilon > 0$ there is a positive integer n_0 such that for all $n > n_0$ d sign.

$$P(2^{-n(H(\omega)+\epsilon)} \leqslant \otimes^n \mathcal{O}_\omega \leqslant 2^{-n(H(\omega)-\epsilon)}) > 1 - \epsilon,$$

where $H(\omega) = -\omega(\log_2(\mathcal{O}_\omega))$ is the *entropy* of the source and the probability distribution is calculated with respect to the state $\Omega_n = \omega \otimes \cdots \otimes \omega$ (n factors) of $\otimes^n A$. If Q denotes the identity in the subalgebra generated by $(\epsilon I - |\log_2(\otimes^n \mathcal{O}_\omega) + nH|)_+$, then

$$(1 - \epsilon)2^{n(H(\omega)-\epsilon)} \leqslant \mathtt{Tr}(Q) \leqslant 2^{n(H(\omega)+\epsilon)}.$$

Before proving the theorem some explanations are necessary. First $\log_2 x\ (= \ln x / \ln 2)$ is usually defined for strictly positive elements of a C^* algebra.[5] We extend the definition to all non-zero $x \geqslant 0$. The standard method of extending complex functions (continuous or analytic) functions to a C^* algebra is via functional calculus (Kadison & Ringrose, 1997). However, in our case it is simpler. Let $\{y_i\}$ be a atomic basis in an abelian C^* algebra. Let $y = \sum_i a_i y_i$ with $a_i \geqslant 0$. Then define $\log_2 y = \sum_i b_i y_i$ where $b_i = \log a_i$ if $a_i > 0$ and 0 otherwise. This definition implies that some standard properties of log are no longer true (e.g. $2^{\log x} \neq x$). But in the present context it gives the correct result when we take expectation values as in the formulas in the theorem. A somewhat longer but mathematically better justified route is to 'renormalize' the state. Thus if $\omega(x_i) = 0$ for k indices we define $\omega'(x_i) = \delta$, where δ is arbitrarily small but positive and $\omega'(x_j) = \omega(x_j) - k\delta$ where $\omega'(x_j) > k\delta$. If we can prove the theorem now for ω'

[5] Henceforth log will be always with respect to base two unless specified otherwise.

and since the relations are valid in the limit $\delta \to 0$ then we are done. We will not take this path but implicitly assume that the probabilities are positive. Finally, note that the element Q is a projection on the subalgebra generated by $(\epsilon I - |\log_2(\otimes^n \mathcal{O}_\omega) - nH|)_+$. It corresponds to the set of strings whose probabilities are between $2^{-nH-\epsilon}$ and $2^{-nH+\epsilon}$. The integer $\mathtt{Tr}(Q)$ is simply the cardinality of this set.

Proof of the theorem. First note that $\log ab = \log a + \log b$ for elements $a, b \geqslant 0$ in A. We can write $\otimes^n \mathcal{O}_\omega = X_1 X_2 \cdots X_n$ where $X_i = \mathbb{1} \otimes \mathbb{1} \otimes \cdots \otimes \mathcal{O}_\omega \otimes \mathbb{1} \otimes \cdots \otimes \mathbb{1}$ with $\log \mathcal{O}_\omega$ the ith factor. The fact that Ω_n is a product state on $\otimes^n A$ (corresponding to a source whose successive outputs are independent) implies that X_i are independent and i.i.d. We can now apply the corollary to Theorem 5 yielding $P(|\log(\otimes^n \mathcal{O}_\omega) - \Omega_n(\log X_1)| > \epsilon) = P(|\log(\otimes^n \mathcal{O}_\omega) - \omega(\log(\mathcal{O}_\omega))| > \epsilon)$. □

4.3 *Communication channels*

Every form of communication requires channels through which signals are sent and received. It is perhaps the most important component in the mathematical models of communication. We will not be dealing with real channels which are complex physical objects—the atmosphere, a telephone cable, a bus on the mainboard of a computer are some examples. Our object is to give simple mathematical models of a channel which still yield interesting results relevant for concrete channels. The original paper of Shannon characterized channels by a transition probability function. Thus, the channel (precisely a two-way channel) has an input alphabet X and output alphabet Y and a sequence of random functions $\phi_n : X^n \to Y^n$. The latter are characterized by probability distributions $p_n(y^{(n)}|x^{(n)})$, the interpretation being: $\phi_n(x^{(n)}) = y^{(n)}$ with conditional probability $p_n(y^{(n)}|x^{(n)})$. Note that the distribution depends on the entire history. We say that such a channel has (infinite) memory. A channel has finite memory if there is an integer $k \geqslant 0$ such that if $x^{(n)} = x_n x_{n-1} \cdots x_{n-k+1} \cdots x_1$ then $p_n(y^{(n)}|x^{(n)}) = p_n(y^{(n)}|x'^{(n)})$ for any string x'_n of length n such that $x'_n = x_n, \ldots, x'_{n-k+1} = x_{n-k+1}$. That is, the probability distribution depends on the most recent k symbols seen by the channel. A channel is *memoryless* if $k = 1$. Since we will be dealing mostly with discrete memoryless channels (without feedback) (DMC), this property will be tacitly assumed unless stated otherwise. In the memoryless case, it is easy to show the simple form of transition probabilities

$$p_n(y^{(n)}|x^{(n)}) = p_n(y_1 \cdots y_n|x_1 \cdots x_n) = p(y_1|x_1)p(y_2|x_2) \cdots p(y_n|x_n). \tag{5}$$

This motivates us to define the *channel transformation matrix* $C(y_j|x_i)$ with $y_j \in Y$ and $x_i \in X$. As before in this work, X and Y will be finite sets. Since the matrix $C(y_j|x_i)$ is supposed to represent the probability that the channel outputs y_j on input x_i, we must have $\sum_j C(y_j|x_i) = 1$ for all i. In other words, matrix $C(ij) = C(y_j|x_i)$ is *row stochastic*. This is the standard formulation (Khinchin, 1957; Ash, 1990; Cover & Thomas, 1999).[6] We now turn to the algebraic formulation. We restrict ourselves to two-terminal channels here.

Definition 4 A DMC $\mathcal{C} = \{X, Y, C\}$ where X and Y are abelian C^* algebras of dimension m and n, respectively, and $C : Y \to X$ is a unital positive map. The algebras X and Y will be called the input and output algebras of the channel, respectively. Given a state ω on X we say that (X, ω) is the input source for the channel.

We recall that a positive map $C : Y \to X$ is a linear map such that $C(y) \geqslant 0$ if $y \geqslant 0$. Sometimes, we write the entries of C in the more suggestive form $C_{ij} = C(y_j|x_i)$ where $\{y_j\}$ and $\{x_i\}$ are atomic bases for

[6] In this work, we will not deal with channel coding and decoding. Including these concepts is not difficult but complicates the notation.

Y and X, respectively. Thus $C(y_j) = \sum_i C_{ij} x_i = \sum_i C(y_j|x_i)x_i$. Note that in our notation C is an $m \times n$ matrix. Its transpose $C^{\top}_{ji} = C(x_i|y_j)$ is the channel matrix in the standard formulation. We have to deal with the transpose because the channel is a map *from* the output alphabet to the input alphabet. This may be counterintuitive but observe that any map $Y \to X$ defines a unique dual map $\mathcal{S}(X) \to \mathcal{S}(Y)$, on the respective state spaces. Informally, a channel transforms a probability distribution on the input alphabet to a distribution on the output. In other words, given an input source, there is a unique output source determined by the channel. Let us note that in case of abelian algebras every positive map is guaranteed to be *completely positive* (Takesaki, 2002). This is no longer true in the non-abelian case. Hence, for the quantum case, complete positivity has to be explicitly imposed on (quantum) channels.

We characterize a channel by input/output algebras (of observables) and a positive map. Like the source output we now define a useful quantity called *channel output*. Corresponding to the atomic basis $\{y_i\}$ of Y, let $\otimes^k y_{i(k)}$ be an atomic basis in $\otimes^n Y$. Here $i(k) = (i_1 i_2 \cdots i_k)$ is a multi-index. Similarly, we have an atomic basis $\{\otimes^k x_{j(k)}\}$ for $\otimes^k X$. The level-k channel output is defined to be

$$O^k_C = \sum_{i(k)} y_{i(k)} \otimes C^{(k)}(y_{i(k)}). \tag{6}$$

Here $C^{(k)}$ represents the channel transition probability matrix on the k-fold tensor product corresponding to strings of length k. In the DMC case it is simply the k-fold tensor product of the matrix C. The channel output defined here encodes most important features of the communication process. First, given the input source function[7] $\mathcal{I}_{\omega^k} = \sum_i \omega^k(x_{i(k)})x_{i(k)}$, the output source function is defined by

$$\mathcal{O}_{\tilde{\omega}^k} = I \otimes \mathrm{Tr}_{\otimes^k X}((\mathbb{1} \otimes \mathcal{I}_{\omega^k})O^k_c) = \sum_i \sum_j C(y_{i(k)}|x_{j(k)})\omega^k(x_{j(k)})y_{i(k)}. \tag{7}$$

Here, the state $\tilde{\omega}^k$ on the output space $\otimes^k Y$ can be obtained via the dual $\tilde{\omega}^k(y) = \tilde{C}^k(\omega^k)(y) = \omega^k(C^k(y))$. The formula above is an alternative representation which is very similar to the quantum case. The *joint output* of the channel can be considered as the combined output of the two terminals of the channel. This is obtained by *not* tracing out over the input in equation (7). Thus the joint output

$$\begin{aligned} \mathcal{J}_{\tilde{\Omega}^k} &= (\mathbb{1} \otimes \mathcal{I}_{\omega^k})O^k_C = \sum_{ij} \Omega^k(y_{i(k)} \otimes x_{j(k)})y_{i(k)} \otimes x_{j(k)} \text{ with} \\ \Omega^k(y_{i(k)} \otimes x_{j(k)}) &= C(y_{i(k)}|x_{j(k)})\omega(x_{j(k)}). \end{aligned} \tag{8}$$

Let us analyse the algebraic definition of channel given above. For simplicity of notation, we restrict ourselves to level 1. The explicit representation of channel output is

$$\sum_i y_i \otimes \sum_j C(y_i|x_j)x_j.$$

We interpreted this as follows: if on the channel out-terminal y_i is observed then the input could be x_j with probability $C(y_i|x_j)\omega(x_j)/\sum_j C(y_i|x_j)\omega(x_j)$. Now suppose that for a fixed i $C(y_i|x_j) = 0$ for all j except one say, j_i. Then on observing y_i at the output we are certain that the the input is x_{j_i}. If this is true for all values of y, then we have an instance of a lossless channel. It is easy to write the channel matrix

[7] We called this the source output before. But as the channel has two terminals we call it input source function to avoid confusion.

in this case. Thus, given $1 \leqslant j \leqslant n$, let d_j be the set of integers i for which $C(y_i|x_j) > 0$. The lossless property implies that $\{d_j\}$ form a partition of the set $\{1, \ldots, m\}$. The corresponding channel output is

$$O_C = \sum_j \left(\sum_{i \in d_j} C(y_i|x_j) y_i \right) \otimes x_j.$$

Clearly lossless channels are the most desirable for communication of information. At the other extreme is the *useless* channel in which there is no correlation between the input and the output. To define it formally, consider a channel $\mathcal{C} = \{X, Y, C\}$ as above. The map C induces a map $C' : Y \otimes X \to X$ defined by $C'(y \otimes x) = xC(y)$. Given a state ω on X the dual of the map C' defines a state Ω_C on $Y \otimes X$: $\Omega_C(y \otimes x) = \omega(C'(y \otimes x)) = C(y|x)\omega(x)$. We call Ω_C the joint (input–output) state of the channel. A channel is useless if Y and X (identified as $Y \otimes \mathbb{1}$ and $\mathbb{1} \otimes X$, respectively) are Ω_C independent.

LEMMA 4 A channel $\mathcal{C} = \{X, Y, C\}$ with input source (X, ω) is useless iff the matrix $C_{ij} = C(y_j|x_i)$ is of rank 1.

Proof. Suppose $\mathcal{C}$ is useless. Note that $\Omega_C(\mathbb{1} \otimes x) = \omega(x)$ and $\Omega_C(y \otimes \mathbb{1}) = \tilde{\omega}(y)$, where $\tilde{\omega}(y) = \omega(C(y))$ is the image of ω under the dual of the map C. Then Ω_C independence implies $C(y_j|x_i)\omega(x_i) = \omega(x_i)\tilde{\omega}(y_j)$. We may assume that all $\omega(x_i) > 0$ (otherwise we just discard it). Hence, $C(y_j|x_i) = \tilde{\omega}(y_j)$ and this proves necessity. Now if C_{ij} has rank 1, then all the rows are non-zero multiples of any one row, say the first. Since C is a row stochastic matrix, the rows must be identical, that is, $C_{ij} = a_j = \tilde{\omega}(y_j)$ and independence is trivially verified. □

The definition of a useless channel captures the intuition that if there is no correlation between the input and output then we can recover practically nothing. The *channel coding* theorem asserts that apart from this extreme case we can decode the output to recover a large portion of the input with high probability of success. The algebraic version of the channel coding theorem assures that it is possible to approximate, in the long run, an arbitrary channel (excepting the useless case) by a lossless one.

THEOREM 7 (Channel coding) Let $\mathcal{C}$ be a channel with input algebra X and output algebra Y. Let $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^m$ be atomic bases for X and Y, respectively. Given a state ω on X, if the channel is not useless then for each k there are subalgebras $Y_k \subset \otimes^k Y, X_k \subset \otimes^k X$, a map $C_k : Y_k \to X_k$ induced by C and a lossless channel $L_k : Y_k \to X_k$ such that

$$\lim_{k \to \infty} \Omega(|O_{C_k} - O_{L_k}|) = 0 \quad \text{on } T_k = Y_k \otimes X_k.$$

Here $\Omega = \otimes^\infty \Omega_C$ and on $\otimes^k Y \bigotimes \otimes^k Y$ it acts as $\Omega^k = \otimes^k \Omega_C$ where Ω_C is the state induced by the channel and a given input state ω. Moreover, if $r_k = \dim(X_k)$, then $\log r_k/k = R + O(1/k)$. R is called transmission rate. Further, any $R < I(X, Y)$ is achievable where the mutual information

$$I(X, Y) = \Omega(\log O_C - \log O_{\tilde{\omega}}).$$

First let us clarify the meaning of the above statements. The theorem simply states that on the chosen set of codewords the channel output of C_k induced by the given channel can be made arbitrarily close to that of a lossless channel L_k. Since a lossless channel has a definite decision scheme for decoding, the choice of L_k is effectively a decision scheme for decoding the original channel's output when the input is restricted to our 'code-book'. This in turn implies that the probability of error tends to 0.

Proof. From an atomic basis of $\otimes^k X$, choose a subset A_k of cardinality r_k (to be determined). Let X_k be the subalgebra generated by A_k. Write $C^{(k)}$ for the k-fold tensor product of C. Let Q_k be the identity on X_k (it is the sum of all the members of A_k). For an atomic basis B_k of $\otimes^k Y$, let

$$B'_k = \{y \in B_k : Q_k C^{(k)}(y) \neq 0\}.$$

Let Y_k be the subalgebra generated by B'_k and $C_k : Y_k \to X_k$ denote the linear map $C_k(y) = Q_k C^{(k)}(y)$. Informally, if we restrict the messages to observables in A_k then the output algebra is Y_k. The new channel map is C_k. We now have a new channel $\tilde{\mathcal{C}}^k = (X_k, Y_k, C_k)$. Throughout the most of the proof we will assume that we are working in $T_k = Y_K \otimes X_k$ with the appropriate maps. We next define L_k as follows. For $y_i \in B'_k$ let $C_k(y_i) = \sum_j C_k(y_i|x_j)x_j$, $x_j \in A_k$. For fixed y_i let i_r be the index for which $C_k(y_i|x_{i_r})$ is maximum (if there are more than one index equal to this maximum choose one arbitrarily). Let $L_k(y_i) = x_{i_r}$ where L_k is a channel map which defines a lossless channel. As we see below, L_k does approximate O_{C_k} in T_k with small error (in probability). What this means is that with high probability we can correctly associate a unique and correct input to a given channel output.[8] Set $r_k = 2^{kR}$ and let

$$O_{\tilde{\omega}^k} = \sum_{y \in B'_k} \tilde{\omega}(y)(y \otimes \mathbb{1}) \quad \text{and} \quad O_{\omega^k} = \sum_{x \in A_k} \omega(x)(\mathbb{1} \otimes x).$$

Here $O_{\tilde{\omega}^k}$ and O_{ω^k} are, respectively, the input and output source function for the channel $\tilde{\mathcal{C}}^k$. By construction both are strictly positive and hence invertible. Let Z_k be the identity on the ideal generated by $(\log O_{C_k} - \log O_{\tilde{\omega}^k} - k(R+\epsilon))_+ = (\log (O_{C_k} O_{\tilde{\omega}^k}^{-1}) - k(R+\epsilon))_+$, $\epsilon > 0$ in T_k.[9] Note that $O_{C_k} = O_{\Omega^k} O_{\omega^k}^{-1}$ on T_k. Since

$$Z_k |O_{C_k} O_{\tilde{\omega}^k}^{-1} - 2^{k(R+\epsilon)}| = (O_{C_k} O_{\tilde{\omega}^k}^{-1} - 2^{k(R+\epsilon)})_+ = Z_k (O_{\Omega^k} O_{\omega^k}^{-1} O_{\tilde{\omega}^k}^{-1} - 2^{k(R+\epsilon)}) \geqslant 0$$

and $Z_k^2 = Z_k$, we conclude that $Z_k O_{\omega^k} \leqslant Z_k O_{\Omega^k} O_{\tilde{\omega}^k}^{-1} 2^{-k(R+\epsilon)} \leqslant Z_k 2^{-k(R+\epsilon)}$. The last inequality follows from the fact that $O_{\Omega^k} O_{\tilde{\omega}^k}^{-1} \leqslant \mathbb{1}$. We also have $|O_{C_k} - O_{L_k}| \leqslant \mathbb{1}_k$ and $\Omega^k(Z_k) = \mathtt{Tr}(Z_k O_{\Omega^k})$. The last fact is true for any projection as can be verified using an atomic basis. We now have

$$\begin{aligned} \Omega^k(Z_k |O_{C_k} - O_{L_k}|) &\leqslant \Omega^k(Z_k) = \mathtt{Tr}(Z_k O_{\Omega^k}) \leqslant \mathtt{Tr}(Z_k O_{\omega^k}) \\ &\leqslant 2^{-k(R+\epsilon)} \mathtt{Tr}(Z_k) \leqslant 2^{-k(R+\epsilon)} r_k = 2^{-k\epsilon}. \end{aligned}$$

Hence $\Omega^k(Z_k |O_{C_k} - O_{L_k}|) \to 0$ as $k \to \infty$. To complete the proof we look at the complementary part: $(\mathbb{1}_k - Z_k)|O_{C_k} - O_{L_k}|$ where $\mathbb{1}_k$ is the identity in T_k. Consider the projection $\mathbb{1}_k - Z_k$. Z_k is the identity in the annihilating ideal of F_{k-}, where $F_k = (\log O_{C_k} - \log O_{\tilde{\omega}^k} - k(R+\epsilon)\mathbb{1})$. Let $G_k = (\log (\otimes^k O_C O_{\tilde{\omega}}^{-1}) - k(R+\epsilon)\mathbb{1}) \in \otimes^k Y \otimes^k X$. Then since F_k is the restriction of G_k to a subspace $G_k = F_k + F'_k$, there is an $F'_k \in \otimes^k Y \otimes \otimes^k X$ with $F_k F'_k = 0$ (we use the fact the channel is memoryless). Hence $G_{k-} = F_{k-} + F'_{k-}$. It follows that $F_{k-} \leqslant G_{k-}$ and Z'_k, the identity on the annihilating ideal of G_{k-} satisfies $Z'_k \leqslant Z_k$. This implies $\Omega(\mathbb{1}_k - Z_k) \leqslant \Omega(\mathbb{1} - Z'_k)$. By definition $\Omega(\mathbb{1} - Z'_k) =$

---

[8] We have combined two types of decoding scheme: the ideal observer decoding (Ash, 1990) and typical set decoding (Cover & Thomas, 1999).

[9] This ideal is $T_k(\log (O_{C_k} O_{\tilde{\omega}^k}^{-1}) - k(R+\epsilon))_+$. Note that we write the scalar $k(R+\epsilon)$ instead of the more accurate $k(R+\epsilon)\mathbb{1}_k$ where $\mathbb{1}_k$ is the unit in T_k.

$P(\log \otimes^k O_C O_{\tilde{\omega}}^{-1}/k - (R+\epsilon) < 0)$ is the probability that $G_k/k < R+\epsilon$. But

$$\Omega(|(\log \otimes^k O_C O_{\tilde{\omega}}^{-1})/k - \Omega(\log O_C - \log O_{\tilde{\omega}})\mathbb{1}|) \to 0 \quad \text{as } k \to \infty$$

follows from the law of large numbers (see Theorem 5 and its corollary). Since the mutual information $I(X|Y) = \Omega(\log O_C O_{\tilde{\omega}}^{-1}) = \Omega(\log O_C - \log O_{\tilde{\omega}}) = H(Y) - H(Y|X)$ and $H(Y|X)$ is the conditional entropy. Thus if we have $R < I(X,Y)$, say $R \leqslant I(X,Y) - 2\epsilon$, then $\Omega(\mathbb{1} - Z_k') = P(\log \otimes^k O_C O_{\tilde{\omega}^k}^{-1}/k - (R+\epsilon) < 0) \leqslant P(|\log(\otimes^k O_C O_{\tilde{\omega}}^{-1})/k - I(X|Y)\mathbb{1}| > \epsilon) =$ but the latter $\to 0$ (recall that $I(X|Y) = \Omega(\log(\otimes^k O_C O_{\tilde{\omega}}^{-1}))$, the expectation value). Putting it all together we have for any $\epsilon > 0$ and $R \leqslant I(X,Y) - 2\epsilon$

$$\begin{aligned}\Omega((\mathbb{1}_k - Z_k)|O_{C_k} - O_{L_k}|) &\leqslant \Omega(\mathbb{1}_k - Z_k) \leqslant \Omega(\mathbb{1} - Z_k') \\ &= P(|\log(\otimes^k O_C O_{\tilde{\omega}}^{-1})/k - \Omega(\log(\otimes^k O_C O_{\tilde{\omega}}^{-1})\mathbb{1}| > \epsilon)) \to 0 \quad \text{as } k \to \infty.\end{aligned}$$

As we already have $\Omega(Z_k|O_{C_k} - O_{L_k}|) \to 0$, the proof is complete. □

Note that in invoking the law of large numbers we have to lift to the full algebra because Ω is not a state in the code space. The channel coding theorem implies that it is possible to choose a set of 'codewords' which can be transmitted with high reliability. Moreover, from the proof, it is clear that *any* decoding scheme may be used although the convergence rate may be poor for a random decoding scheme. It is easy to see that for a lossless channel the input entropy $H(X)$ is equal to the mutual information. We may think of this as conservation of entropy or information which justifies the term 'lossless'. Since it is always the case that $H(X) - H(X|Y) = I(X,Y)$ the quantity $H(X|Y)$ can be considered the loss due to the channel. The channel coding theorem is perhaps the most celebrated theorem in Shannon's work although his proof was not rigorous. Such proofs were given by others later (see, for example, Cover & Thomas, 1999 and the references therein). The algebraic version of the theorem serves two primary purposes. First, we attempt to make the proof as 'algebraic' as possible. More importantly, it gives us the commutative perspective from which we will seek possible extensions to the non-commutative case. Secondly, the channel map L can be used for a decoding scheme. Thus, we may think of a coding–decoding scheme for a given channel as a sequence of pairs (X_k, L_k) as above. Thus given any positive number $R < I(X,Y)$, there is subalgebra X_k of dimension $\lfloor 2^{Rk} \rfloor$ in $\otimes^k X$ such that the map $C_k : Y_k \to X_k$, where Y_k is as defined above and C_k is the restriction of the channel map $C^{(k)}$ of degree k to Y_k can be approximated by a lossless channel map L_k. So, we can consider $I(X,Y)$ as a bound on the rate R at which messages can be transmitted reliably. Now $I(X,Y)$ depends on the initial probability distribution and hence the channel capacity is defined as

$$\mathrm{C} = \sup_{\omega_X} I(X,Y),$$

where ω_X runs through the set of states in X. There is a converse to the channel coding theory which states that any achievable rate R satisfies $R \leqslant \mathrm{C}$. An algebraic version of this can be formulated and proved but we omit it. An alternative (and perhaps operational) version of the capacity would be $\mathrm{C}' = \sup_{\omega_X} R$. We use this in the next section on proving results on capacities of new channels that are defined by combining some existing channels. The coding theorems can be extended to more complicated scenarios like ergodic sources and channels with finite memory or feedback. We will not pursue

these issues further here. But we are confident that these generalizations can be appropriately formulated and proved in the algebraic framework.

4.3.1 *Zero error capacity and algebra of channels.* In this section, we continue with some further developments of communication channels. We use the notation of previous section. First consider the characterization of a lossless channel by the channel L output vector

$$O_L = \sum_j \left(\sum_{i \in d_j} L(y_i|x_j) y_i \right) \otimes x_j,$$

where the sets $d_j \subset \{1, \ldots, , n\}$ form a partition. Let ω_i be the dual functional or states. Then the above condition can be restated as: L is lossless if the elements

$$\mathtt{Y}_j = I_Y \otimes \omega_i(O_L) = \sum_{i \in d_j} L(y_i|x_j) y_i,$$

are mutually 'orthogonal': $\mathtt{Y_i Y_j} = 0$ if $i \neq j$. In the case of an arbitrary channel $\mathcal{C}$ define similarly

$$\mathtt{Y}_j = I_Y \otimes \omega_i(O_C) = \sum_{i \in \gamma_j} C(y_i|x_j) y_i,$$

where $\gamma_j \subset \{1, \ldots, n\}$ such that $i \in \gamma_j$ if and only if $C(y_i|x_j) > 0$. Now let $\mathtt{N}'$ be subset of an atomic basis of X of maximal cardinality such that for every distinct pair $x_i, x_j \in \mathtt{N}'$ the corresponding 'outputs' $\mathtt{Y}_i$ and $\mathtt{Y_j} = 0$ are orthocomplementary. Let $\mathtt{N}$ be the subalgebra generated by $\mathtt{N}'$. We observe that this notion of orthocomplementarity was stated in Körner & Orlitsky (1998) without the algebraic structure introduced here. We similarly define subalgebras $\mathtt{N}^k$ of $\otimes^k X$ such that the images of atomic basis in $\otimes^k Y$ under the (transposed) channel map of level k are orthocomplementary. Let $N(C, k)$ be the dimension of $\mathtt{N}^k$. Although the subalgebras $\mathtt{N}^k$ need not be unique, the dimensions are (being maximal). The simple observation that if $\{\alpha_1, \ldots, \alpha_r\}$ and $\{\beta_1, \ldots, \beta_s\}$ are two orthocomplementary sets in algebras A and B, respectively, then $\{\alpha_i \otimes \beta_j : 1 \leqslant i \leqslant r \text{ and } 1 \leqslant j \leqslant s\}$ is an orthocomplementary set implies the following useful property:

$$N(C, k + l) \geqslant N(C, k) N(C, l).$$

If the code words are restricted to some subalgebra $\mathtt{N}^k$, then the messages can be decoded with zero-error probability. The notion of zero-error capacity is defined similar to the ordinary capacity as the limit $R_0(C) = \lim_{k \to \infty} \log N(C, k)/k$. The calculation of $R_0(C)$ is a difficult problem and following Shannon (1956) it is translated to problem in graph theory. Curiously, some of the most powerful tools in graph theory used for analysing this problem are *algebraic* (see Körner & Orlitsky, 1998). Indeed, the brilliant work of Lovâsz (1979) uses orthogonality relations (in a scalar product sense) to solve a conjecture of Shannon. However, we will not explore these ideas further as our goal is to lay the foundations of an algebraic framework for classical and quantum information. In this context, let us point out some striking differences between the two. Although we use sums of vectors and tensors, it is the basic atomic vectors which have proper interpretation. For example, we can only interpret an expression like $\alpha = p_1 x_1 + p_2 x_2 + p_3 x_3$ $(p_1 + p_2 + p_3 = 1)$ as a formal sum. In a classical system, we cannot have genuine superposition of observables. Another way of viewing this is that in the classical

case we have only one measurement basis, namely, $x_1, x_2, \ldots$, the atomic basis. But in a quantum system it is possible in principle to observe in any basis which is a combination of the given one.

Let us now look at another construction introduced by Shannon (1956). Suppose we have two channels

$$\mathcal{C} = (X, Y, C) \quad \text{and} \quad \mathcal{C}' = (X', Y', C').$$

We form two new channels: the sum and product channels.

- *Product channel*. The product channel $\mathcal{C} \otimes \mathcal{C}'$ is defined as

$$\mathcal{C} \otimes \mathcal{C}' = (C \otimes C', X \otimes X', Y \otimes Y').$$

- *Sum channel*. The sum channel $\mathcal{C} \oplus \mathcal{C}'$ is defined as

$$\mathcal{C} \oplus \mathcal{C}' = (C \oplus C', X \oplus X', Y \oplus Y').$$

The above definition implies that for the new channel $\bar{\mathcal{C}} = (\bar{C}, \bar{X}, \bar{Y})$, which stands for either the product or sum channel, we have k-level maps $\bar{C}_k = \otimes^k \bar{C}$ on the space (of k-strings) $\otimes^k \bar{X}$ to $\otimes^k \bar{Y}$. After a bit of algebra it is easily shown that for the two operations the level k operation is equivalent to the following:

- *Product channel*:

$$\otimes^k \bar{C} \sim \otimes^k C \bigotimes \otimes^k C', \quad \otimes^k \bar{X} \sim \otimes^k X \bigotimes \otimes^k X', \quad \otimes^k \bar{Y} \sim \otimes^k Y \bigotimes \otimes^k Y'.$$

- *Sum channel*:

$$\otimes^k \bar{C} \sim \otimes^k C \bigoplus \otimes^k C', \quad \otimes^k \bar{X} \sim \otimes^k X \bigoplus \otimes^k X', \quad \otimes^k \bar{Y} \sim \otimes^k Y \bigoplus \otimes^k Y'.$$

The equivalence in these relations refers to statistical equivalence in the sense that the outputs have same value in appropriate product states. There are some implicit assumption; the most important one being that the two channels are *non-interfering*. This is often a reasonable assumption in the classical case but not for quantum channels. We conclude the section with a result on the capacity K of product and sum channels. In the case of interfering channels, the situation is much more complicated.

PROPOSITION 4 The capacity L of product and sum channels are given by the following formulas.

$$K(C \otimes C') = K(C) + K(C') \quad \text{and} \quad K(C \oplus C') = \log\left(\exp\left(K(C)\right) + \exp\left(K(C')\right)\right).$$

Proof (*Sketch*). The formulas follow from the fact that for any two spaces M and N

$$\dim(M \otimes N) = \dim(M)\dim(N) \quad \text{and} \quad \dim(M \oplus N) = \dim(M) + \dim(N).$$

Now from Theorem 7 we get two lossless channel maps L_{k_1}, L'_{k_2} on channels $\mathcal{C}$ and $\mathcal{C}'$ approximating the respective channel maps. Let $X_{k_1} \subset \otimes^{k_1} X$ and $Y_{k_2} \subset \otimes^{k_2} Y$ denote the subalgebras on which L and L' are defined, respectively. Take $k = \max(k_1, k_2)$. If $k_1 \leqslant k = k_2$, then we can extend L_{k_1} to a map on an appropriate subspace $X_k \subset \otimes^k X$ such that the approximation condition holds and $\log \dim(X_{k_1})/k_1 =$

$\log \dim(X_k)/k + O(1/k)$. This will imply that

$$K(C \otimes C') \leqslant K(C) + K(C').$$

The converse is proved by showing first that any approximation lossless channel $\bar{L}$ may be assumed to be a product channel because of the independence of X and X' and hence for any $\delta > 0$

$$K(C \otimes C') > K(C) + K(C') - \delta.$$

Thus, we have the first formula of the proposition. The second formula is slightly more complicated. Informally, we use the fact that, if we have two subspaces X_k and X'_k of 'codewords', then the subspace $X_k \oplus X'_k$ is appropriate codeword subalgebra for the sum channel. If R_1 and R_2 are the respective rates, then we have

$$\dim(X_k \oplus X'_k) \sim m^{kR_1} + m'^{kR_2}, \quad \text{where } \dim X(X') = m(m').$$

Now reasoning as in the product case we get the formula for capacity of sum channels. □

The proof in the above proposition shows that relations between various channel capacities are related to corresponding relations between the dimensions. The point is that information theoretic concepts are mapped on to algebraic concepts. We can do more sophisticated analysis for deeper results combining combinatorial and algebraic techniques. For example, Lovâsz's seminal work (Lovâsz, 1979) can be put in the present algebraic context giving insight to the origins of some of the constructions.

5. Conclusion and preview of the future work

In the preceding sections, we have laid the basic algebraic framework for information theory. This work was devoted to classical parts of information theory corresponding to abelian algebras. Since information theory relies heavily on probabilistic concepts, we devoted a major part of the paper to algebraic probability theory. Although we often confined our discussion to finite-dimensional algebras corresponding to finite sample spaces, it is possible to extend it to infinite-dimensional algebras of continuous sample spaces. In this regard, a natural question is: can the algebraic formulation replace Kolmogorov axiomatics based on measure theory? Naively, the answer is no because the assumption of a norm-compete algebra imposes the restriction that the random variables that they represent must be *bounded*. Moreover, the GNS construction implies that the algebraic framework is essentially equivalent to (almost) bounded random variables on a locally compact space. In order to deal with the unbounded case we have to go beyond the normed algebra structures. A possible course of action is indicated in the examples given in Section 3.3: via the use of a 'cut-off'. A more general approach would be to consider sequences which converge in a topology weaker than the norm topology to elements of a larger algebra. These and other related issues on foundations are deep and merit a separate investigation.

The second major theme of this paper is information theory in the algebraic framework. As some the most important results of information theory concern finite or discrete alphabet, we have primarily dealt with these cases only. In this context, we can treat ergodic sources, channels with finite memory and multi-terminal channels. These topics will be investigated in the future in the non-commutative setting. However, let us recall one of the principal motivations of this paper: the construction of a single framework for dealing with quantum and classical information. We have seen that the algebraic theory in the commutative case already indicates the close analogies between the two cases. We will delve deeper into these analogies and aim to throw light on some basic issues such as quantum Huffman coding

(Braunstein *et al.*, 2000), channel capacities and general no-go theorems among others, once we formulate the appropriate models. In this context, let us mention that many investigators have recognized the importance of the algebraic framework but a comprehensive algebraic model which can be extended to infinite-dimensional case is lacking. We aim to address these important issues in subsequent work.

References

Araki, H. (1975) Relative entropy of states of von Neumann algebras. *Publ. Res. Inst. Math. Sci.*, **11**, 173–192.

Ash, R. B. (1990) *Information Theory*. New York: Dover Publications.

Bahadur, R. R. (1955) Statistics and subfields. *Ann. Math. Stat.*, **26**, 490.

Barnum, H., Barrett, J., Leifer, M. & Wilce, A. (2007) Generalized no-broadcasting theorem. *Phys. Rev. Lett.*, **99**, 240501.

Bêny, C., Kempf, A. & Kribs, D. W. (2007) Quantum error correction of observables. *Phys. Rev. A*, **76**, 042303.

Billingsley, P. (1995) *Probability and Measure*. New York: John Wiley.

Brattelli, O. (2002) *Operator Algebras and Quantum Statistical Mechanics*. Berlin: Springer.

Braunstein, S. L., Fuchs, C. A., Gottesman, D. & Lo, H.-K. (2000) A quantum analog of Huffman coding. *IEEE Trans. Inf. Theory*, **46**, 1545.

Cam, L. L. (1986) *Asymptotic Methods in Statistical Decision Theory*. Berlin: Springer.

Cover, T. M. & Thomas, J. A. (1999) *Elements of Information Theory*. New York: John Wiley.

Emch, G. (1984) *Mathematical and Conceptual Foundations of 20th-Century Physics*. Amsterdam: North-Holland.

Haag, R. (1992) *Local Quantum Physics*. Berlin: Springer.

Kadison, R. V. & Ringrose, J. R. (1997) *Fundamentals of the Theory of Operator Algebras*, vol. I. Providence, RI: American Mathematical Society.

Kelley, J. L. (1975) *General Toplogy*. Berlin: Springer.

Keyl, M. (2002) Fundamentals of quantum information theory. *Phys. Rep.*, **369**, 531–548.

Khinchin, A. Ya. (1957) *Mathematical Foundations of Information Theory*. New York: Dover Publications.

Kleene, S. C. (1952) *Introduction to Metamathematics*. Amsterdam: North-Holland.

Körner, J. & Orlitsky, A. (1998) Zero-error information theory. *IEEE Trans. Inf. Theory*, **44**, 2207.

Kretschmann, D. & Werner, R. F. (2006) Quantum channels with memory. *Phys. Rev. A*, **72**, 062323.

Lindblad, G. (1974) Expectations and entropy inequalities for finite quantum systems. *Commun. Math. Phys.*, **39**, 111–119.

Lovâsz, L. (1979) On the Shannon capacity of graph. *IEEE Trans. Inf. Theory*, **25**, 1.

MacMillan, B. (1953) The basic theorems of information theory. *Ann. Math. Stat.*, **24**, 196–219.

Patra, M. K. & Braunstein, S. L. (2011) Quantum Fourier transform, Heisenberg groups and quasiprobability distributions. *New J. Phys.*, **13**, 063013.

Rudin, W. (1987) *Real and Complex Analysis*, 3rd edn. McGraw-Hill.

Sakai, S. (1971) *C^* and W^* Algebras*. Berlin: Springer.

Schumacher, B. (1996) Sending entanglement through noisy channels. *Phys. Rev. A*, **54**, 2614.

Segal, I. E. (1954) Abstract probability spaces and a theorem of Kolmogoroff. *Am. J. Math.*, **76**, 721–732.

Segal, I. E. (1960) A note on the concept of entropy. *J. Math. Mech.*, **9**, 623–629.

Shannon, C. E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423, 623–656.

Shannon, C. E. (1956) The zero-error capacity of a noisy channel. *IRE Trans. Inf. Theory*, **IT-2**, 8.

Shannon, C. E. & Weaver, W. W. (1949) *The Mathematical Theory of Communication*. Champaign, IL: University of Illinois Press.

Shiryayev, A. N. (1984) *Probability*. Berlin: Springer.

Streater, R. F. (1995) *Statistical Dynamics: A Stochastic Approach to Nonequilibrium Thermodynamics*. London: Imperial College Press.

Takesaki, M. (2002) *Theory of Operator Algebra I*. Berlin: Springer.

Umegaki, H. (1962) Conditional expectation in an operator algebra, IV (Entropy and Information). *Kodai Math. Semin. Rep.*, **14**, 59–85.

VOICULESCU, D., DYKEMA, K. & NICA, A. (1992) *Free Random Variables*. CRM Monograph Series. Providence, RI: AMS.
WHITTLE, P. (1992) *Probability via Expectation*, 3rd edn. Berlin: Springer.

Appendix

Proof of Theorem 2. Let $\{y_1, \ldots, y_n\}$ be a basis for A. Since the self-adjoint elements $(y_i + y_i^*)/2$ and $i(y_i - y_i^*)/2$ span A, we can choose an independent set. Hence, we may assume that the y_i are self-adjoint. Then each y_i^2 is positive and hence possesses a square-root $|y_i|$. Moreover, $|y_i| \geqslant y_i$ (Kadison & Ringrose, 1997; Brattelli, 2002).[10] We can therefore write each $y_i = (|y_i| + y_i)/2 - (|y_i| - y_i)/2$, as the difference of two positive elements. Again, choosing an independent set, we may assume that y_i themselves are positive with norm 1. Let $S = \{z : z \geqslant 0 \text{ and } \|z\| \leqslant 4\}$. S is convex and compact (being closed and bounded) and $y_i \in S$. Hence, by the Krein–Millman theorem (Kadison & Ringrose, 1997), S is the convex closure of its *extreme points*.[11] We may assume that these extreme points have norm 1 (obviously discarding 0). Since each y_i can be written as a finite convex sum of its extreme points, we can pick a basis $x_1, \ldots, x_n$ of extreme points. We complete the proof by showing that the x_i's satisfy equations (2) and that they are unique.

Now $\|x_i\| = 1$ implies that for any $|\lambda| > 1$, $\lambda - x_i = \lambda(\mathbb{1} - \lambda^{-1}x_i)$ is invertible. This can be proved by using the geometric series of $(1 - \lambda^{-1}x_i)^{-1}$. Hence if $a \in \mathrm{sp}(x_i)$, then $0 \leqslant a \leqslant 1$ and $\mathbb{1} - x_i$ is positive. Since $\mathrm{sp}(x_i - x_i^2) = \{a - a^2 : a \in \mathrm{sp}(x_i)\}$ and $a - a^2 \geqslant 0$, it follows that $x_i - x_i^2 \geqslant 0$. As $x_i = (2(x_i - x_i^2) + 2x_i^2)/2$, a convex combination of two positive elements in S and x_i is a non-zero extreme point, we must have $x_i - x_i^2 = 0$ or $x_i - x_i^2 = x_i^2$. The last possibility is ruled out because it would imply $\|x_i\| = 2\|x_i^2\| = 2\|x_i\|^2 = 2$. Hence $x_i^2 = x_i$. To prove that they are orthogonal, observe that $x_i - x_i x_j = x_i(1 - x_j)$ is positive. Thus $x_i = (2x_i(1 - x_j) + 2x_i x_j)/2$ is a convex combination of points in S. Hence, as before either $x_i x_j = 0$ or $x_i = x_i x_j$. With x_j in place of x_i we conclude that $x_i x_j = 0$ or $x_j = x_i x_j$. Thus the only possibility for $x_i \neq x_j$ is that $x_i x_j = 0$.

To prove the decomposition property let $\mathbb{1} = \sum_i a_i x_i$. Squaring and using the orthogonality of x_i's we conclude that $a_i = 1$ or 0. If some $a_k = 0$, then $x_k = x_k \mathbb{1} = x_k \sum_i a_i x_i = 0$. Hence, all $a_k = 1$. Finally, let $\{z_i\}$ be another basis satisfying (2). Let $z_i = \sum_j b_{ij} x_j$. As before, $b_{ij} = 1$ or 0 and the matrix (b_{ij}) is a 0–1 matrix. For fixed i, let T_i be the set of integers j such that $b_{ij} = 1$. Then $x_i x_j = 0$, $i \neq j$ implies T_i and T_j are disjoint. This along with the last condition in (2) implies that T_i's form a partition of the set $\{1, \ldots, n\}$. Thus each T_i is a singleton and the matrix (b_{ij}) has exactly one 1 in each row and column. It is a permutation matrix.

Let $x = \sum_i a_i x_i$ be an element of A. Then $\lambda\mathbb{1} - x = \sum_i (\lambda - a_i)x_i$. This is invertible iff $\lambda \neq a_i$, $i = 1, \ldots, n$ with inverse $\sum_i (\lambda - a_i)^{-1} x_i$. The proof is complete. □

Proof of Proposition 1. We observe that it is sufficient to define an injective *set* map j (respectively j') from B_G^∞ to B_A^∞ (respectively B_A^∞ to B_G^∞). For we can first extend these to linear maps $\mathcal{J}$ (respectively $\mathcal{J}'$) on the appropriate spaces. The fact that the bases are atomic will ensure that these are injective algebra homomorphisms, in fact, isometries. Let

$$j(z_1 \otimes \cdots \otimes z_k \otimes \mathbb{1} \otimes \cdots) = \phi(z_1) \otimes \cdots \otimes \phi(z_k) \otimes \mathbb{1} \otimes \cdots,$$

[10] Bratteli (p. 35) gives a proof which does not use Gelfand representation.
[11] Recall that extreme points of a convex set are those which cannot be written as a non-trivial convex combination of some members of the set

where $z_i \in \{y_0, y_1\}$ and $\phi(y_0) = x_0,\ \phi(y_1) = x_1$. To construct j' let the binary representation of the integer $n-1$ be of length $k+1$ where $k = \lfloor \log_2 n \rfloor$. For $0 \leqslant r \leqslant n-1$ $r = b_0^r + b_1^r 2 + b_2^r 2^2 + \cdots + b_k^r 2^k$ be the binary representation of r of length $k+1$ (pad it with 0's if necessary). Let $\psi : B_A \to B_G^\infty$ be the map defined by

$$\psi(x_r) = y_{b_0^r} \otimes y_{b_1^r} \otimes \cdots \otimes y_{b_k^r}$$

extend it to a map $j' : B_A^\infty \to B_G^\infty$ by

$$j'(z_1 \otimes \cdots \otimes z_k \otimes \mathbb{1} \otimes \cdots) = \phi(z_1) \otimes \cdots \otimes \phi(z_k) \otimes \mathbb{1} \otimes \cdots.$$

The map j' is injective and the proof is complete. □

Proof of Theorem 3. First assume that $S_1 = \{x\}$ and $S_2 = \{y\}$. Let $\{x_1, \ldots, x_n\}$ be an atomic basis of A. Let $x = \sum_i a_i x_i$ and $y = \sum_i b_i x_i$. Some of these coefficients may be 0 and some may be equal. Write

$$x = a_1 P_1 + a_2 P_2 + \cdots + a_k P_k \quad \text{and} \quad y = b_1 Q_1 + b_2 Q_2 + \cdots + b_l Q_l.$$

Here the a_i's are distinct and $P_i = x_{i_1} + x_{i_2} + \cdots + x_{i_r}$ corresponding to all basis elements whose coefficients are equal to a_i. Similarly, for Q_j's. Note that $P_i P_m = \delta_{im}$ and $Q_j Q_s = \delta_{js}$. By Lagrange interpolation there are polynomials $f_i(\lambda),\ i = 1, \ldots, k$ and $g_j,\ j = 1, \ldots, l$ such that $f_i(a_r) = \delta_{ir}$ and $g_j(b_s) = \delta_{js}$. Since x, y are ω independent,

$$\omega(f_i(x) g_j(y)) = \omega(P_i Q_j) = \omega(P_i)\omega(Q_j). \tag{9}$$

The subalgebra $A(S_1)(A(S_2))$ is generated by the P_i's (Q_j's). Clearly $\{P_i : i = 1, \ldots, k\}$ and $\{Q_j : j = 1, \ldots, l\}$ are atomic bases for $A(S_1)$ and $A(S_2)$, respectively. Define states ω_1 and ω_2 of $A(S_1)$ and $A(S_2)$, respectively, by restricting ω to these subalgebras. Let $\phi : X \otimes Y \to A'$ be the natural map $\phi(u \otimes v) = uv$. Using equation (9), it is a routine check that $(A(S_1) \otimes A(S_2), \{\omega_1 \otimes \omega_2\})$ is a cover of $(A(S_1, S_2), \omega')$.

Now for the general case. Since $A(S_1)$ and $A(S_2)$ are subalgebras of A they have atomic bases $\{u_i\}$ and $\{v_j\}$, respectively. As in the previous case we have polynomials $\{p_i\}$ and $\{q_j\}$ in several variables such that $p_i(x_1, \ldots, x_{k_i}) = u_i$ and $q_j(y_1, \ldots, y_{m_j}) = v_j$, where $x_i \in S_1$ and $q_i \in S_2$. We do not have easy interpolating polynomial in this case. By repeating the argument of the singleton case above, we get the appropriate cover and complete the proof. The converse is clear from the definition of a cover and the fact that in a *product state* $\omega_1 \otimes \omega_2(z_1 \otimes z_2) = \omega_1(z_1)\omega_2(z_2)$. □

Proof of Theorem 4. We sketch an algebraic proof in the current setting. The most direct approach is to use the notion of continuous function calculus which essentially asserts that continuous functions on the spectrum can be lifted to define functions on the algebra. More precisely, given an element $x \in A$ there is an isometric algebra homomorphism between the algebra of continuous functions on the spectrum of x, $C(\mathrm{sp}(x))$ and the closed subalgebra $C(x)$ generated by x (Kadison & Ringrose, 1997). Thus for every function $f(u)$ on $\mathrm{sp}(x)$ there is a unique element $f(x)$ in $C(x)$ such that if $f(u) \geqslant 0$ then $f(x) \geqslant 0$. Since for any real c and $\delta > 0$, $|t + \delta - u| - (t + \delta - u) \leqslant |t - u| - (t - u)$ we infer that $|t + \delta - x| - (t + \delta - x) \leqslant |t - x| - (t - x)$ for self-adjoint $x \in A$. Now for any $y \in A$ if $xy = 0$ then $|x|y = 0$ and hence $x_+ y = x_- y = 0$. So if $x \leqslant z$ and $v \in A$, then $zv = 0$ implies $xv = 0$. Thus the annihilator ideal of $|t + \delta - x| - (t + \delta - x)$ contains the annihilator ideal of $|t - x| - (t - x)$. The continuity follows from the following construction which is useful for calculating distributions. Write $x(t) = t\mathbb{1} - x$, $\mathfrak{t} =$

$(t_1, t_2, \ldots, t_n)$ and $\chi(\mathtt{t}) = x_1(t_1)_+ \times x_2(t_2)_+ \times \cdots \times x_n(t_n)_+$. For integer $m > 0$ let

$$e_m(\mathtt{t} + 1/m) = m\chi(\mathtt{t} + 1/m)(1 + m\chi(\mathtt{t} + 1/m))^{-1} \equiv \frac{m\chi(\mathtt{t} + 1/m)}{1 + m\chi(\mathtt{t} + 1/m)}, \tag{10}$$

where $\mathtt{t} + 1/m = (t_1 + 1/m, t_2 + 1/m, \ldots, t_n + 1/m)$. Although $e_m(\mathtt{t} + 1/m)$ is *not* a member of the annihilating ideal $S^-(\mathtt{t})_a$ of $S^-(\mathtt{t})$ it belongs to $S^-(\mathtt{t} + 1/m)_a \supset S^-(\mathtt{t})_a$. Let $e_\lambda(\mathtt{t})$ be an approximate identity in $S^-(\mathtt{t})_a$. One can show using the Gelfand representation that

$$\lim_\lambda \omega(e_\lambda(\mathtt{t})) = \lim_{m\to\infty} \omega(e_m(\mathtt{t})).$$

We omit the details but the reader can convince herself by taking an algebra of functions.

This implies the first part of the theorem. To prove the boundary conditions, we use the fact that the spectrum of any element $x \in A$ is bounded by $\|x\|$. Hence, for $t < -\|x\|$, $t\mathbb{1} - x$ has a strictly negative spectrum. Then $(t\mathbb{1} - x)_- = -(t\mathbb{1} - x)$ is invertible and its annihilator ideal consists of 0 alone. Consequently, $F(t, \ldots,) = 0$ for all $t < -\|x\|$. The other extreme case is proved similarly, $t\mathbb{1} - x$ being strictly positive for $t > \|x\|$. Finally, suppose the elements $\{x_1, x_2, \ldots, x_n\}$ are independent. Since x_+ lies in the closed subalgebra generated by x the definition of independence and equation (10) implies that the joint distribution function is a product. One proves the last statement using a sequence like (10). □