Probabilistic Monads, Domains and Classical Information

Michael Mislove[*]

Tulane University
New Orleans, LA 70118

Shannon's classical information theory [18] uses probability theory to analyze channels as mechanisms for information flow. In this paper, we generalize results from [14] for binary channels to show how some more modern tools — probabilistic monads and domain theory in particular — can be used to model classical channels. As initiated in [14], the point of departure is to consider the family of channels with fixed inputs and outputs, rather than trying to analyze channels one at a time. The results show that domain theory has a role to play in the capacity of channels; in particular, the $n \times n$-stochastic matrices, which are the classical channels having the same sized input as output, admit a quotient compact ordered space which is a domain, and the capacity map factors through this quotient via a Scott-continuous map that measures the quotient domain. We also comment on how some of our results relate to recent discoveries about quantum channels and free affine monoids.

1 Introduction

Classical information theory has its foundations in the seminal work of Claude Shannon [18], who first conceived of analyzing the behavior of channels using *entropy* and deriving a formula for *channel capacity* based on mutual information (cf. [6] for a modern presentation of the basic results). Recent work of Martin, et al. [14] reveals that the theory of compact, affine monoids and domain theory can be used to analyze the family of binary channels. In this paper, our goal is to generalize the results in [14] to the case of $n \times n$-channels — channels that have n input ports and n output ports. Our approach also uses the monadic properties of probability distributions to give an abstract presentation of how channels arise, and that clarifies the role of the doubly stochastic matrices, which are special channels. While our work focuses on the classical case, the situation around quantum information and quantum channels is also a concern, and we point out how our results relate to some recent work [7, 15] on quantum qubit channels and free affine monoids. While most of the ingredients we piece together are not new, we believe the approach we present does represent a new way in which to understand families of channels and some of their important features.

The rest of the paper is structured as follows. In the next section, we describe three monads based on the probability measures over compact spaces, compact monoids and compact groups. Each of these is used to present some aspect of the classical channels. We then introduce topology, and show how the capacity of a channel can be viewed from a topological perspective. The main result here is that capacity is the maximum distance from the surface determined by the entropy function and the underlying polytope generated by the rows of a channel matrix, viewed as vectors in $\mathbb{R}^n$ for appropriate n. This leads to a generalization of Jensen's Lemma that characterizes strictly concave functions. Domain theory is then introduced, as applied to the finitely-generated polytopes residing in a compact convex set, ordered by reverse inclusion. Here we characterize when proper maps measure a domain, in the sense of Martin [13]; a closely related result can be found in [16]. Finally, we return to the compact monoid of $n \times n$-stochastic matrices and show that it has a natural, algebraically-defined pre-order relative to which

capacity measures the quotient partial order, which is a compact ordered space. The capacity mapping is also shown to be strictly monotone with respect to this pre-order, which means that strictly smaller channel matrices have strictly smaller capacity. We close with a summary and comments about future work.

2   Three probabilistic monads

The categorical presentation of classical information relies on three monads, each of which has the family $\mathsf{Prob}(X)$ of probability distributions over a set X as the object-level of the left adjoint. The first of these starts with compact Hausdorff spaces, and uses several results from functional analysis: standard references for this material are [4, 17]. We present these monads in turn:

2.1   A spatial monad

We begin with the probability measure monad over topological spaces. If X is a compact Hausdorff space, then $C(X,\mathbb{R})$, the family of continuous, real-valued functions defined on X, is a Banach space (complete, normed linear space) in the sup-norm. The Banach space dual of $C(X,\mathbb{R})$, denoted $C(X,\mathbb{R})^*$ consists of all continuous linear functionals from $C(X,\mathbb{R})$ into $\mathbb{R}$. $C(X,\mathbb{R})^*$ is another Banach space, and the Riesz Representation Theorem implies this is the Banach space of Radon measures on X (those that are both inner- and outer regular). The unit sphere of $C(X,\mathbb{R})^*$ is the family $\mathsf{Prob}(X)$ of probability measures over X. If we endow $\mathsf{Prob}(X)$ with the weak* topology (the weakest topology making all continuous linear functionals into $\mathbb{R}$ continuous), then $\mathsf{Prob}(X)$ becomes a compact, Hausdorff space, by the Banach-Alaoglu Theorem. Prob extends to a functor $\mathsf{Prob}_S\colon \mathsf{Comp} \to \mathsf{CompConv}_{LC}$ from the category of compact Hausdorff spaces and continuous maps, to the category of compact, convex, locally convex spaces and continuous affine maps, via $\mathsf{Prob}_S(X) = \mathsf{Prob}(X)$ and $f\colon X \to Y$ maps to $\mathsf{Prob}_S(f)\colon \mathsf{Prob}_S(X) \to \mathsf{Prob}_S(Y)$ by $\mathsf{Prob}_S(f)(\mu)(A) = \mu(f^{-1}(A))$, for each Borel set $A \subseteq Y$.

Moreover, if the mapping $x \mapsto \delta_x\colon X \to C(X,\mathbb{R})^*$ sending a point to the Dirac measure it defines, is a continuous mapping into the weak* topology. Since X is compact Hausdorff, Urysohn's Lemma implies $C(X,\mathbb{R})$ separates the points of X, and so $x \mapsto \delta_x$ is a homeomorphism onto its image. Another application of Urysohn's Lemma shows each Dirac measure is an extreme point $\mathsf{Prob}(X)$ and in fact the Dirac measures form the set of extreme points of $\mathsf{Prob}(X)$.

A *simple measure* is a finite, convex combination of Dirac measures, i.e., one of the form $\sum_{i \leq n} r_i \delta_{x_i}$, where $r_i \geq 0$, $\sum_i r_i = 1$, and $x_i \in X$ for each i. We let $\mathsf{Prob}_{sim}(X)$ denote this family. The Krein-Milman Theorem implies that $\mathsf{Prob}_{sim}(X)$ is weak* dense among the probability measures. So, if $f\colon X \to C$ is a continuous function from X into a compact subset of a locally convex vector space, then the function $\widehat{f}(\delta_x) = f(x)$ extends uniquely to continuous function $\widehat{f}(\sum_{i \leq n} r_i \delta_{x_i}) = \sum_{i \leq n} r_i f(x_i)$, and then to all of $\mathsf{Prob}(X)$, by the density of the simple measures. Obviously, $\widehat{f}(\delta_x) = f(x)$.

We conclude that the functor Prob_S is left adjoint to the forgetful functor. In fact, Prob_S defines a monad, where the unit of the adjunction is the mapping $\eta_X(x) = \delta_x$ and the multiplication $\mu\colon \mathsf{Prob}(\mathsf{Prob}(X)) \to \mathsf{Prob}(X)$ is integration.

Theorem 1. *The functor Prob_S sending a compact space to its family of probability measure in the weak* topology defines a monad on the category Comp. The unit of the monad sends a point $x \in X$ to the Dirac measure δ_x, and the image of the unit is the set of extreme points in $\mathsf{Prob}(X)$.*

Definition 1. *Let X and Y be compact Hausdorff spaces. A* (lossless) noisy channel *from X to Y is a mapping $f\colon X \to \mathsf{Prob}(Y)$.*

Since Prob_S is a monad, each channel $f\colon X \to \mathsf{Prob}(Y)$ corresponds uniquely to a continuous, affine mapping $\mathsf{Prob}_S(f)\colon \mathsf{Prob}(X) \to \mathsf{Prob}(Y)$ in the Kleisli category $\mathscr{K}_{\mathsf{Prob}_S}$ of Prob_S.

Example 1. *Let $n \geq 1$ and let $\overline{n} = \{0,\dots,n-1\}$ be the discrete, compact space. Then $Prob(\overline{n})$ is the family of probability distributions on n points, and given $m \geq 1$, a channel $f\colon \overline{m} \to Prob(\overline{n})$ is an $m \times n$-stochastic matrix. The family $ST(m,n)$ of $m \times n$-stochastic matrices is then the family of lossless, noisy channels from $\overline{m}$ to $\overline{n}$. Moreover, from our comment about the Kleisli category $\mathscr{K}_{Prob_S}$, we conclude that the family of morphisms $\mathscr{K}_{Prob_S}(\overline{m}, Prob(\overline{n}))$ is $ST(m,n) \hookrightarrow Aff(Prob(\overline{m}), Prob(\overline{n}))$, where $Aff(Prob(\overline{m}), Prob(\overline{n}))$ is the family of continuous affine maps from $Prob(\overline{m})$ to $Prob(\overline{n})$.*

This first probabilistic monad shows that classical channels correspond to mappings in the Kleisli category of the "spatial" monad Prob_S on the category Comp. If we let $m = n$, then $\mathsf{ST}(n) \stackrel{\text{def}}{=} \mathsf{ST}(n,n)$ is also a monoid using composition in the Kleisli category: if $f, g \in \mathsf{ST}(n)$, and $\widehat{g}\colon \mathsf{Prob}(\overline{n}) \to \mathsf{Prob}(\overline{n})$ is the extension of g, then $g \circ f ::= \widehat{g} \circ f \in \mathsf{ST}(n)$. We next present a second monad that gives another account of this special case.

2.2 A monad on monoids

The second monad we define is based on the category CMon of compact monoids and compact monoid homomorphisms. More precisely, a *compact monoid* is a monoid S — a non-empty set endowed with an associative binary operation $(x,y) \mapsto xy\colon S \times S \to S$ that also has an identity element, 1_S — that also is a compact Hausdorff space for which the multiplication is continuous. We can apply the probability functor to such an S to obtain the compact convex (Hausdorff) space $\mathsf{Prob}(S)$ of probability measures on S. If we denote multiplication on S by $\cdot$, then $\mathsf{Prob}(\cdot)\colon \mathsf{Prob}(S \times S) \to \mathsf{Prob}(S)$, and since $\iota_S\colon \mathsf{Prob}(S) \times \mathsf{Prob}(S) \hookrightarrow \mathsf{Prob}(S \times S)$ is an embedding, we have a continuous affine map $\mathsf{Prob}(\cdot) \circ \iota_S\colon \mathsf{Prob}(S) \times \mathsf{Prob}(S) \to \mathsf{Prob}(S)$. This map is called *convolution*, and we denote $(\mathsf{Prob}(\cdot) \circ \iota_S)(\mu,\nu) = \mu *_S \nu$. It is routine to show convolution is associative, so $\mathsf{Prob}(S)$ is a compact affine monoid.

If $\phi\colon S \to T$ is a morphism of compact semigroups, then $\mathsf{Prob}(\phi)\colon \mathsf{Prob}(S) \to \mathsf{Prob}(T)$ is defined by $\mathsf{Prob}(\phi)(\mu)(f) = \int f \circ \phi \, d\mu$ for any $f\colon T \to \mathbb{R}$. If $\mu, \nu \in \mathsf{Prob}(S)$ and $f \in C(T, \mathbb{R})$, then

$$
\begin{aligned}
\mathsf{Prob}(\mu *_S \nu)(f) = \int_S (f \circ \phi) \, d(\mu *_S \nu) \;\; &= \;\; \int_S \int_S f \circ \phi \circ m_S \, d\mu \, d\nu \\
&\stackrel{1}{=} \;\; \int_S \int_S f \circ m_T \circ (\phi \times \phi) \, d\mu \, d\nu \\
&= \;\; (\mathsf{Prob}(\phi)(\mu) *_T \mathsf{Prob}(\phi)(\nu))(f),
\end{aligned}
$$

where $m_S\colon S \times S \to S$, $m_T\colon T \times T \to T$ are the semigroup operations, $*_S, *_T$ denote convolution, and where $\stackrel{1}{=}$ follows from the fact that ϕ is a homomorphism. Thus $\mathsf{Prob}_S\colon (\mathsf{Prob}(S), *_S) \to (\mathsf{Prob}(T), *_T)$ is a semigroup homomorphism. Finally, the fact that $\mathsf{Prob}(\phi)$ preserves the identity follows from the observation that $\delta_x *_S \delta_y = \delta_{xy}$, which implies that δ_1 is an identity for the simple measures, and consequently for all measures since the simple measures are dense.

It follows that restricting Prob_S to the subcategory CMon of Comp yields a functor $\mathsf{Prob}_M\colon \mathsf{CMon} \to \mathsf{CAM}$ into the category of locally convex compact affine monoids and continuous affine monoid maps.

Theorem 2. *The restriction of $Prob_S$ to Mon induces a monad $Prob_M$ whose target is CAM, the category of locally convex compact affine monoids and continuous, affine monoid homomorphisms. The unit of the monad is again the Dirac map, and its image is again the set of extreme points of $Prob(S)$.*

Applying the same reasoning as for Prob_S, we see that each continuous monoid homomorphism $\phi\colon S \to \mathsf{Prob}(T)$ corresponds to a unique morphism of compact affine monoids, $\widehat{\phi}\colon \mathsf{Prob}(S) \to \mathsf{Prob}(T)$ in the Kleisli category $\mathscr{K}_{\mathsf{Prob}_M}$. However, in relation to classical channels, our interest is in the object level of Prob_M:

Example 2. *We return to the example $ST(n)$ of stochastic $n \times n$-matrices. These arise as channels on a discrete, n-element set. For such a set $\overline{n}$, the selfmaps of $\overline{n}$ form a finite — hence compact — monoid. If we denote this monoid by $[\overline{n} \to \overline{n}]$, then applying Prob_M we obtain a compact affine monoid $\mathsf{Prob}_M([\overline{n} \to \overline{n}])$.*

Now, $[\overline{n} \to \overline{n}] \hookrightarrow [\overline{n} \to \mathsf{Prob}(\overline{n})]$ by $f \mapsto \eta_{\overline{n}} \circ f$, where $\eta_{\overline{n}}$ is the unit for Prob_S. Since $ST(n) = [\overline{n} \to \mathsf{Prob}(\overline{n})]$ is a compact affine monoid, this mapping extends to a morphism of compact affine monoids

$$\sum_{i \leq k} r_i \delta_{f_i} \mapsto \sum_{i \leq k} r_i \eta_{\overline{n}} \circ f_i \colon \mathsf{Prob}_M([\overline{n}, \overline{n}]) \to ST(n).$$

Since $\{\eta_{\overline{n}} \circ f \mid f \in [\overline{n} \to \overline{n}]\}$ is the set of extreme points of $ST(n)$, this morphism is surjective. In fact this map is an isomorphism. Thus $ST(n)$ is the free compact affine monoid over $[\overline{n} \to \overline{n}]$.

2.3   A monad over compact groups

Our final use of Prob to define a monad starts with CGrp, the category of compact groups and continuous group homomorphisms. Since CGrp is a subcategory of CMon, we know that applying Prob_M to a compact group yields a compact affine monoid. However, $\mathsf{Prob}_M(G)$ is not a group in general, so the forgetful functor from CMon does not take $\mathsf{Prob}_M(G)$ to a group, but instead yields a compact monoid.

But, when applied to a compact group G *qua* compact monoid, the unit of the monad Prob_M sends each $g \in G$ to $\delta_g \in \mathsf{Prob}(G)$, and this is a monoid — hence group — homomorphism. So, we define a new functor $H\colon \mathsf{CMon} \to \mathsf{CGrp}$ by $H(S) = H(1_S)$, the group of units[1] of the compact monoid S. If $\phi\colon S \to T$ is a morphism of compact affine monoids, then $\phi|_{H(1_S)}\colon H(1_S) \to H(1_T)$ is a morphism of compact groups, so H defines a functor.

Theorem 3. *The functor $H\colon \mathsf{CAM} \to \mathsf{CGrp}$ is right adjoint to the functor $\mathsf{Prob}_G\colon \mathsf{CGrp} \to \mathsf{CAM}$. In fact, the composition $H \circ \mathsf{Prob}_G$ defines a monad on CGrp. Moreover, for any compact group G, we have $H(\mathsf{Prob}_G G) \simeq G$. Again, the unit of the monad is the Dirac map, and the image of G in $\mathsf{Prob}_G(G)$ is the set of extreme points.*

Example 3. *We again consider the case of classical channels. Here, given $n \geq 1$, $ST(n)$ has for its group of units the permutation group $S(n)$. Applying Prob_G, we find that $\mathsf{Prob}_G(S(n))$ is the free compact affine monoid over the group $S(n)$. But this is just the family $DT(n)$ of doubly stochastic $n \times n$-matrices.*

We can use information about $\mathsf{Prob}_G(S(n))$ to conclude information about $DT(n)$. Wendel's Theorem [20] states that, for a compact group G, the compact monoid $\mathsf{Prob}(G)$ has $\{\delta_g \mid g \in G\}$ as its group of units, and the minimal ideal (every compact monoid has one — cf. [10]) is a zero, which in fact is Haar measure on G. In the case of $S(n)$, this reaffirms that the units of $DT(n)$ are the permutations of $\overline{n}$, and that $DT(n)$ has a zero, which is the equidistribution $\sum_{i \leq n} \frac{1}{n} \delta_i$.

Corollary 1. *If G is a finite group, then $\mathsf{Prob}_G(G)$ is the free affine monoid over G, as well as the being the free compact affine monoid over G.*

---

[1] A *unit* of a monoid is an element that has a two-sided inverse with respect to the identity 1_S. The set of units $H(1_S) = \{x \in S \mid (\exists y \in S)\ xy = yx = 1_S\}$ forms the largest subgroup of S that has 1_S as the identity; if S is compact, then so is the group of units.

Proof. If G is a finite group, then $\mathsf{Prob}(G) = \{\sum_{i \leq k} r_i \delta_{g_i} \mid k \in \mathbb{N}, r_i \in [0,1], \sum_i r_i = 1 \wedge g_i \in G\}$ consists of simple measures. If S is an affine monoid and $\phi \colon G \to S$ is a monoid homomorphism, then $\phi(G) \subseteq H(1_S)$, and so $\phi \colon G \to H(1_S)$ is a group homomorphism. Then $\widehat{\phi}(\sum_{i \leq k} r_i \delta_{g_i}) = \sum_{i \leq k} r_i \phi(g_i)$ is easily seen to be a morphism of affine monoids that satisfies $\widehat{\phi}(\delta_g) = \phi(g)$ for each $g \in G$, and $\widehat{\phi}$ is the unique such since $\mathsf{Prob}_G(G)$ consists of simple measures. This shows $\mathsf{Prob}_G(G)$ is the free affine monoid over G, and the Theorem implies it also is the free compact affine monoid over G since G is finite, and hence compact. □

Remark 1. *In [7,15], the free affine monoid over a finite group is employed to deduce properties of quantum channels. The Corollary shows that the free affine monoid over G is nothing other than $\mathsf{Prob}_G(G)$, which implies it is compact, as well as telling us that it has a zero — the uniform distribution on G. We believe other useful properties about quantum channels over finite groups can be deduced from this observation.*

3 Capacity as a topological concept

In this section we develop a new approach to understanding the capacity of a classical channel. Our idea is to analyze capacity from a topological perspective, rather from the usual perspective of inequalities prevalent in information theory. We begin with a brief reprise of the basics of Shannon information; the standard reference for this material is [6].

If $X \colon \mathscr{X} \to \mathbb{R}$ is a random variable on a finite probability space $(\mathscr{X}, p)$, then the *entropy*[2] of X is defined as $H(X) = -\sum_{x \in \mathscr{X}} p(x) \log_2 p(x)$. If $Y \colon \mathscr{Y} \to \mathbb{R}$ is another finite random variable, then the *conditional entropy* of Y given X is

$$H(Y|X) = \sum_{x \in \mathscr{X}} p(x)H(Y|X = x) = \sum_{x \in \mathscr{X}} p(x) \sum_{y \in \mathscr{Y}} p(y|x) \log_2 \frac{1}{p(y|x)}, \tag{1}$$

and the *mutual information in X and Y* is

$$\mathscr{I}(Y,X) = H(Y) - H(Y|X) = H(X) - H(X|Y).$$

If $C \colon \mathscr{X} \to \mathscr{Y}$ is a channel from inputs $\mathscr{X}$ to outputs $\mathscr{Y}$, then C is an $\mathscr{X} \times \mathscr{Y}$-matrix whose (x,y)-entry is the *conditional probability* of output y occurring, given that the input was x. Each distribution p on the inputs $\mathscr{X}$ then produces a corresponding distribution $p \cdot C$ on $\mathscr{Y}$. The *capacity of a channel* is given by

$$\mathsf{Cap} \colon [\mathsf{Prob}(\mathscr{X}) \to \mathsf{Prob}(\mathscr{Y})] \to [0,1] \quad \text{by} \quad \mathsf{Cap}(C) = \sup_{p \in Pr(\mathscr{X})} H(p \cdot C) - H(p \cdot C \mid p),$$

i.e., $\mathsf{Cap}(C)$ is the supremum of the possible mutual information values $\mathscr{I}(p \cdot C \mid p)$ as p ranges over the distributions on $\mathscr{X}$, the set of inputs.

If we let $\mathscr{X} = \mathscr{Y} = \overline{n}$ and $C \colon \overline{n} \to \mathsf{Prob}(\overline{n})$ is a channel, then

$$\mathsf{Cap}(C) = \sup_{\sum_{i \leq n} r_i \delta_i} \left[H(\sum_{j \leq n} r_j C(j|1), \ldots, \sum_{j \leq n} r_j C(j|n)) - \sum_{i \leq n} r_i H(C(i|1), \ldots, C(i|n)) \right]. \tag{2}$$

This formula requires some interpretation.

---

[2]We use $H(X)$ to denote the entropy of a random variable X; this overloads our notation for the maximal subgroups of a monoid S, but we believe the context will be sufficient to make the meaning clear.

1. First, the term to which H is first applied — $(\sum_{j\leq n} r_j C(j,1), \ldots, \sum_{j\leq n} r_j C(j,n))$ — represents a distribution on $\mathscr{Y} = \overline{n}$ obtained from pC, where $p = \sum_i r_i \delta_i$ is a distribution on $\mathscr{X} = \overline{n}$. This is the p-convex combination of the n vectors $(C(1,1), \ldots, C(1,n)), \ldots, (C(n,1), \ldots, C(n,n)) \in [0,1]^n$ comprising the rows of the channel C, where $C(i,j)$ denotes the i, j-entry of C, interpreted as a conditional probability. Since C is a channel, each of these rows is a probability distribution on $\overline{n}$. (As a sanity check, we see that applying H to the convex combination $(\sum_{j\leq n} r_j C(j,1), \ldots, \sum_{j\leq n} r_j C(j,n))$ thus makes sense, since a convex combination of probability distributions is another such, and H applies to probability distributions.)

2. Now, the convex combination $pC = (\sum_{j\leq n} r_j C(j,1), \ldots, \sum_{j\leq n} r_j C(j,n))$ is a point on the polytope $K \subseteq [0,1]^n$ the rows of C generate, so

$$(\sum_{j\leq n} r_j C(j,1), \ldots, \sum_{j\leq n} r_j C(j,n)), H(\sum_{j\leq n} r_j C(j,1), \ldots, \sum_{j\leq n} r_j C(j,n))) \in [0,1]^n \times \mathbb{R}$$

   represents the point on the surface H generates over the polytope K.

3. Likewise the second term, $\sum_{i\leq n} r_i H(C(i|1), \ldots, C(i|n))$ of Equation 2 is a p-convex combination, $p = \sum_{i\leq n} r_i \delta_i$, of the terms $H(C(i,1), \ldots, C(i,n))$, each of which is obtained by applying H to a row of C, regarded as an elemnt of $[0,1]^n$. We can regard each of the points $H(C(i,1), \ldots, C(i,n))$ as being the $n+1$-coordinate of a tuple $(C(i,1), \ldots, C(i,n), H(C(i,1), \ldots, C(i,n)) \in \mathbb{R}^{n+1}$, and hence the p-convex combination of these points lies on the polytope these points generate.

4. Finally, the difference $H(\sum_{j\leq n} r_j C(j|1), \ldots, \sum_{j\leq n} r_j C(j|n)) - \sum_{i\leq n} r_i H(C(i|1), \ldots, C(i|n))$ is the difference in the $n+1$-coordinates described under 2. and 3., so it is height of the vertical line between the point pC in 3. and the corresponding point

$$(C(1,1), \ldots, C(1,n), H(C(1,1), \ldots, C(1,n)), \ldots, (C(n,1), \ldots, C(n,n), H(C(n,1), \ldots, C(n,n))))$$

   on the surface H generates over K.

Thus, $\mathsf{Cap}(C)$ as presented by Equation 2 takes the supremum of the differences between the value of H at a convex combination of the rows of C and the same convex combination of H applied to the rows of C. It is well-known that entropy H is a strictly concave function, and we now take advantage of this to formulate a result about Cap.

Definition 2. *Let $K \subseteq \mathbb{R}^n$ be a convex set. A function $f: K \to \mathbb{R}$ is* strictly concave *if*

$$f(r\vec{x} + (1-r)\vec{y}) > rf(\vec{x}) + (1-r)f(\vec{y})$$

for all $r \in (0,1)$ and all $\vec{x}, \vec{y} \in K$.

We next recall *Jensen's Inequality:*

Theorem 4 (Jensen (cf. [9])). *If $f: K \to \mathbb{R}$ is a convex function defined on a convex subset K of a vector space V, then $Ef(X) \geq f(E(X))$ for a finite random variable $X: \mathscr{X} \to K$, where E denotes expectation. Moreover, if f is strictly convex, then $E(f(X)) = f(E(X))$ implies X is constant.*

Jensen's Inequality is a fundamental result of information theory; for example, it is crucial for proving mutual information is non-negative, that the mutual information in a pair of random variables is 0 iff the random variables are independent, and that entropy itself is strictly concave (cf. [6], Chapter 2). Since f is (strictly) concave iff $-f$ is (strictly) convex, the following generalizes Jensen's Inequality by strengthening the result in case f is strictly convex.

Lemma 1. *If $f: K \to \mathbb{R}$ be defined from a convex subset K to $\mathbb{R}$. Then the following are equivalent:*

1. *f is strictly concave.*

2. *For all $r_1, \ldots, r_m \in (0,1)$ and all $\vec{x}_1, \ldots, \vec{x}_m \in K$,*

$$\sum_{i \leq m} r_i = 1 \quad \Rightarrow \quad f\left(\sum_{i \leq m} r_i \vec{x}_i\right) > \sum_{i \leq m} r_i f(\vec{x}_i).$$

Proof. $(ii) \Rightarrow (i)$ is obvious. For the reverse direction, we proceed by induction on m. The base case, $m = 2$, is just the definition of strict concavity. So suppose (ii) holds for some m, and consider a family $r_1, \ldots, r_{m+1} \in [0,1]$ and $\vec{x}_1, \ldots, \vec{x}_{m+1} \in K$. Since $r_i \in (0,1)$ for each i,

$$
\begin{aligned}
f\left(\sum_{i \leq m+1} r_i \vec{x}_i\right) &= f\left(\sum_{i \leq m-1} r_i \vec{x}_i + (r_m + r_{m+1})\left(\frac{r_m}{r_m + r_{m+1}} \vec{x}_m + \frac{r_{m+1}}{r_m + r_{m+1}} \vec{x}_{m+1}\right)\right) \\
&> \sum_{i \leq m-1} r_i f(\vec{x}_i) + (r_m + r_{m+1}) f\left(\frac{r_m}{r_m + r_{m+1}} \vec{x}_m + \frac{r_{m+1}}{r_m + r_{m+1}} \vec{x}_{m+1}\right) \\
&> \sum_{i \leq m-1} r_i f(\vec{x}_i) + r_m f(\vec{x}_m) + r_{m+1} f(\vec{x}_{m+1}) \\
&= \sum_{i \leq m+1} r_i f(\vec{x}_i).
\end{aligned}
$$

$\square$

Notation: If $K \subseteq \mathbb{R}^n$ be a compact convex set, then $\mathsf{Con}_n(K)$ denotes the family of convex polytopes $\overline{\mathrm{conv}}(\{x_1, \ldots, x_k\})$ generated by finite subsets $\{x_1, \ldots, x_k\} \subseteq K$, where $k \leq n$.

Also note that $P_n \overset{\text{def}}{=} \{x \in [0,1]^n \mid \sum_i x_i = 1\}$ is a compact, convex subset of $[0,1]^n$, which we identify with the family $\mathsf{Prob}(\overline{n})$ of probability distributions on $\overline{n}$.

Proposition 1. *Let $K \subseteq \mathbb{R}^n$ be a compact, convex set, and let $f: K \to \mathbb{R}$ be continuous and strictly concave. Define*

$$\widehat{f}: \mathsf{Con}_n(K) \to \mathbb{R}^{op} \quad \text{by} \quad \widehat{f}(\overline{\mathrm{conv}}(\{x_1, \ldots, x_k\})) = \sup_{(r_1, \ldots, r_k) \in [0,1]^k} f\left(\sum_{i \leq k} r_i x_i\right) - \sum_{i \leq k} r_i f(x_i).$$

Then $\widehat{f}$ is continuous and monotone with respect to reverse inclusion.

Proof. The compactness of K implies that the family $\mathsf{Con}_n(K)$ is closed under filtered intersections in the hyperspace of non-empty, closed subsets of K, and then the continuity of $\widehat{f}$ follows from the continuity of f. This map is clearly monotone. To show it is strictly monotone, let $\overline{\mathrm{conv}}(\{x_1, \ldots, x_k\}), \overline{\mathrm{conv}}(\{y_1, \ldots, y_m\}) \in \mathsf{Con}_n(K)$ with $\overline{\mathrm{conv}}(\{x_1, \ldots, x_k\}) \subsetneq \overline{\mathrm{conv}}(\{y_1, \ldots, y_m\})$. Since f is continuous, $\widehat{f}(\overline{\mathrm{conv}}(\{x_1, \ldots, x_k\}))$ assumes its value at some point in $\overline{\mathrm{conv}}(\{x_1, \ldots, x_k\})$, and since f is strictly concave, this value is not assumed at x_i for any index i. Thus, there is a k-tuple $(r_1, \ldots, r_k) \in (0,1)^k$ with

$$\widehat{f}(\overline{\mathrm{conv}}(\{x_1, \ldots, x_k\})) = f\left(\sum_{i \leq k} r_i x_i\right) - \sum_{i \leq k} r_i f(x_i).$$

Because $\overline{\mathrm{conv}}(\{x_1,\ldots,x_k\}) \subsetneq \overline{\mathrm{conv}}(\{y_1,\ldots,y_n\})$, for each $i \leq k$, there is $(s_{i,1},\ldots,s_{i,m}) \in [0,1]^m$ with $x_i = \sum_{j\leq m} s_{i,j} y_j$, and at least one of the families $(s_{i,j})_{j\leq m} \in (0,1)^m$. For this index i, we have $f(x_i) = f(\sum_{j\leq m} s_{i,j} y_j) > \sum_{j\leq m} s_{i,j} f(y_j)$. Lemma 2(ii) then implies that

$$\sum_{i\leq k} r_i f(x_i) = \sum_{i\leq k} r_i f\left(\sum_{j\leq m} s_{i,j} y_j\right) > \sum_{i\leq k}\sum_{j\leq m} r_i s_{i,j} f(y_j),$$

and so

$$
\begin{aligned}
\widehat{f}(\overline{\mathrm{conv}}(\{x_1,\ldots,x_k\})) &= f(\sum_{i\leq k} r_i x_i) - \sum_{i\leq k} r_i f(x_i) \\
&\leq f\left(\sum_{i\leq k} r_i \left(\sum_{j\leq m} s_{i,j} y_j\right)\right) - \sum_{1\leq k}\sum_{j\leq m} r_i s_{i,j} f(y_j) \\
&\leq \sup_{(s_1,\ldots,s_m)\in[0,1]^m} f\left(\sum_j s_j y_j\right) - \sum_j s_j f(y_j) \\
&= \widehat{f}(\overline{\mathrm{conv}}(\{y_1,\ldots,y_m\})).
\end{aligned}
$$

$\square$

4   Domains

In this section, we introduce domains, which are the next ingredient in our analysis of classical channels. For details about these structures, a standard reference is [1] or [8]. A *partial order* is a non-empty set P endowed with a reflexive, antisymmetric and transitive relation. A subset $D \subseteq P$ is *directed* is every finite subset of D has an upper bound in D; P is *directed complete* if every directed subset of P has a least upper bound in P. We denote directed complete partial orders as dcpos.

If P and Q are dcpos, then $f\colon P \to Q$ is *Scott continuous* if f is monotone and preserves suprema of directed sets. An equivalent definition is available using topology: a subset $U \subseteq P$ is *Scott open* if $U = \mathord{\uparrow} U = \{x \in P \mid (\exists u \in U)\, u \leq x\}$ is an upper set, and for any directed subset $D \subseteq P$, if $\sup D \in U$, then $D \cap U \neq \emptyset$. The Scott-open sets form a topology on P, called the Scott topology, and the functions $f\colon P \to Q$ that are continuous with respect to this topology are exactly those that are Scott continuous, as defined above.

Example 4. *Let K be a compact convex subset of a topological vector space, and let $Con(K)$ denote the compact convex subsets of K. We can order these by reverse inclusion: $C \sqsubseteq C' \Leftrightarrow C' \subseteq C$. A directed family $D \subseteq Con(K)$ is simply a filterbasis, and since K is compact and each set in D is convex, the set $\bigcap D \in Con(K)$. Thus $Con(K)$ is a dcpo.*

We can say more. If $C, C' \in Con(K)$ and $C' \subseteq C^\circ$, the interior of C, then given any directed set D with $\bigcap D \subseteq C'$, there is some $E \in D$ with $E \subseteq C^\circ$, and hence $E \subseteq C$. In this case we say C is way-below C', and we write $C \ll C'$. In fact, if the ambient topological vector space is locally convex, then each $C' \in Con(K)$ is the filtered intersection of those C satisfying $C \ll C'$: this follows from the fact that in any compact Hausdorff space, each compact subset is the filtered intersection of its compact neighborhoods, and the same applies to compact, convex sets in a locally convex topological vector space.

A *domain* is a dcpo satisfying $\{y \in P \mid y \ll x\}$ is directed and $x = \sup\{y \in P \mid y \ll x\}$ for each $x \in P$. The original motivation for domains was to provide semantic models for high-level programming languages, where the fact that any Scott-continuous selfmap on a domain with least element has a least fixed point gives a canonical model for recursion.

Motivated by examples of selfmaps that are not Scott continuous, Martin [13] devised another approach to guaranteeing fixed points, using the concept of a *measurement*: A Scott-continuous function $m: P \to [0, \infty)^{op}$ is said to *measure the content at $x \in P$* if, given $U \subseteq P$ Scott open,

$$x \in U \ \Rightarrow \ (\exists \varepsilon > 0) \ m_\varepsilon(x) \subseteq U,$$

where $m_\varepsilon(x) = \{y \leq x \mid m(y) - m(x) < \varepsilon\}$. We say that *$m$ measures P* if m measures the content at x for each $x \in P$.

For our next result, we need some notions from topology. Recall that s subset $A \subseteq X$ of a topological space is *saturated* if $A = \bigcap\{U \mid A \subseteq U \text{ open}\}$. The *saturation* of a subset A is $\bigcap\{U \mid A \subseteq U \text{ open}\}$, so a set is saturated iff it is equal to its saturation. Moreover, a subset is compact iff its saturation is compact.

Definition 3. *A continuous function $f: X \to Y$ between topological spaces is* proper *if $f^{-1}(K)$ is compact for each saturated, compact subset $K \subseteq Y$.*

A Scott-continuous map $m: P \to Q$ between continuous posets is proper iff $m^{-1}(\uparrow y)$ is Scott compact for each $y \in Q$. We say that a Scott-continuous mapping $m: P \to Q$ between posets is *proper at $x \in P$* if $\downarrow x \cap m^{-1}(\uparrow y)$ is Scott compact in $\downarrow x$ for each $y \in Q$. We use $[0, \infty)^{op}$ to denote the non-negative real numbers in the *dual order*; the following result is from [10]:

Proposition 2. *Let P be a domain and let $m: P \to [0, \infty)^{op}$ be Scott continuous. If m is proper at $x \in P$, then the following are equivalent:*

1. *m measures the content at x.*

2. *m is strictly monotone at $x \in P$: i.e., $y \leq x$ & $m(y) = m(x) \ \Rightarrow \ y = x$.*

In particular, a Scott-continuous, proper map $m: P \to [0, \infty)^{op}$ measures P iff m is strictly monotone at each $x \in P$.

Corollary 2. *Let $P_n \stackrel{\text{def}}{=} \{x \in [0,1]^n \mid \sum_i x_i = 1\}$ be the compact, convex set of distributions on $\overline{n}$. Then $(Con_n(P_n), \supseteq)$ is a domain, and the mapping $cap: Con_n(P_n) \to [0, \infty)^{op}$ by*

$$cap(\overline{conv}(F)) = \sup \left\{ H(\sum_{x \in F} r_x \cdot x) - (\sum_{x \in F} r_x H(x)) \mid r_x \geq 0, \sum_{x \in F} r_x = 1 \right\}$$

measures $(Con_n(P_n), \supseteq)$.

Proof. The discussion in Example 4 applies to $K = P_n$ to show that $Con(P_n)$ is a domain, and $Con_n(P_n)$ is closed in $Con(P_n)$ under filtered intersections. Since $\mathbb{R}^n$ is locally convex, it's easy to show that each $\overline{conv}(F)$ is the intersection of sets $\overline{conv}(G)$, where $\overline{conv}(F) \subseteq \overline{conv}(G)^\circ$ and $|G| \leq n$ if $|F| \leq n$. Hence $(Con_n(P_n), \supseteq)$ is a domain.

A compact, saturated subset of $[0, \infty)^{op}$ has the form $A = [0, r]$ for some r, and then $cap^{-1}([0, r]) = \{\overline{conv}(F) \mid cap(\overline{conv}(F) \leq r\}$ is closed, and hence compact, since cap is continuous and $Con(P_n)$ — and hence also $Con_n(P_n)$ — are compact in the Lawson topology (cf. [1,8]). But $cap^{-1}([0, r])$ also is an upper set, so it is Scott compact. Thus cap is proper, and so it measures $Con_n(P_n)$ iff it is strictly monotone. But the latter follows from Proposition 1, since entropy is strictly concave. $\square$

Remark: As we will see in a moment, the real import of this last result is not so much that cap measures $Con_n(P_n)$, *per se*, but rather that this implies the mapping cap is strictly monotone. This will tell us that under an appropriate (pre-)order, having one channel strictly below another implies that the capacity of the lower channel is strictly less than that of the larger one.

We also note that Martin and Panangaden have obtained results in [16] that can be used to derive the Proposition 2 and Corollary 2.

5   A domain-like structure of $\mathsf{ST}(n)$

We have seen that $\mathsf{ST}(n)$ is a compact affine monoid, and we already commented that every compact semigroup has a unique smallest ideal: i.e., a non-empty subset $I \subseteq S$ satisfying $IS \cup SI \subseteq I$. This *minimal ideal* is denoted $\mathscr{M}(S)$, and it is closed, hence compact. For example, we noted that $\mathscr{M}(\mathsf{DT}(n))$ is a point, which is the equidistribution on $\overline{n}$. A reference for much of the material in this section is [10], where basic results about compact affine monoids are laid out. A good reference for results about transformation semigroups can be found in [5].

Proposition 3. *If $C \in \mathsf{ST}(n)$, then $C(\mathsf{Prob}(\overline{n})) = \overline{\mathrm{conv}}(\{\delta_i \mid 1 \in \overline{n}\})$ is a convex polytope in $[0,1]^n$.*

Proof. If $C \in \mathsf{ST}(n)$, then $C\colon \mathsf{Prob}(\overline{n}) \to \mathsf{Prob}(\overline{n})$ is an affine mapping, so it preserves the convex structure of $\mathsf{Prob}(\overline{n})$. It follows that $C(\mathsf{Prob}(\overline{n})) = \overline{\mathrm{conv}}(\{\delta_i C \mid i \leq n\})$, where $\delta_i \in \mathsf{Prob}(\overline{n})$ is the Dirac measure on $i \in \overline{n}$. Thus, $C(\mathsf{Prob}(\overline{n}))$ is a convex polytope in $\mathsf{Prob}(\overline{n})$.   □

As a result of the Proposition, we can define a relation on $\mathsf{ST}(n)$ by

$$C \equiv C' \;\Leftrightarrow\; C(\mathsf{Prob}(\overline{n})) = C'(\mathsf{Prob}(\overline{n})). \tag{3}$$

This is clearly a closed equivalence relation, and because channels are affine maps.

$$C \equiv C' \;\Leftrightarrow\; C(\{\delta_i \mid i \in \overline{n}\}) = C'(\{\delta_i \mid i \in \overline{n}\}) \tag{4}$$

Now, $C(\delta_i) = \delta_i C = C(i)$, the ith row of C, so $C \equiv C'$ iff C and C' have the same set of rows vectors. Hence,

$$C \equiv C' \;\Leftrightarrow\; (\exists \pi \in S(n))\, M_\pi C = C', \tag{5}$$

where M_π is the stochastic matrix representing the permutation $\pi \in S(n)$.

We also can obtain an algebraic representation of the relation $\equiv$ using the monoid structure of $\mathsf{ST}(n)$.

Definition 4. *If X is a set, then the* full transformation semigroup $T(X)$ *on X is the family of all selfmaps of X under composition. A* transformation semigroup *is a subsemigroup of $T(X)$ for some set X.*

Notation: If $S \subseteq T(X)$ is a transformation semigroup, then for $s, s' \in T(X)$ and $x \in X$, the element $ss' \in S$ denotes the function $ss'(x) = s'(s(x))$ — i.e., we use the "algebraic notation" for function application, which agrees with our representation matrix multiplication as composition of functions.

Here are some simple observations about $T(X)$; the proofs are all straightforward:

1. $T(X)$ is a monoid whose group of units is the family of bijections of X; if X is finite, this is just $S(|X|)$, the group of permutations of $|X|$-many letters.

2. Each constant map $f_x\colon X \to X$ by $f_x(y) = x$ is s *left zero* in $T(X)$: if $g \in T(X)$, then $gf_x = f_x$. It follows from general semigroup theory that $\mathscr{M}(T(X)) = \{f_x \mid x \in X\}$.

3. If S is a transformation semigroup on X and $S \cap \mathscr{M}(T(X)) \neq \emptyset$, then $\mathscr{M}(S) = S \cap \mathscr{M}(T(X))$. Thus, if S contains a constant map, then $\mathscr{M}(S)$ consists of constant maps. This follows from the fact that gf is a constant map if either f or g is one, so the constant maps in S form an ideal.

4. If S is a transformation semigroup, then for each $s \in S$ and each $x \in X$, $f_x \in S^1 s \;\Rightarrow\; x \in s(X)$.[3] Indeed, if $f_x \in Sx \cup \{s\}$, then there is some $s' \in S$ with $f_x = s's$, so $x = s(s'(y)) \in S(X)$.

---

[3] If S is a semigroup, then S^1 denotes the semigroup S with an identity element adjoined.

5. Conversely, if $M(S) = \{f_x \mid x \in X\}$, then $x \in s(X) \Rightarrow f_x \in S^!s$. This follows since $x \in s(X)$ implies $x = s(y)$, for some $y \in X$; if $s \neq f_x$, then $f_x = f_y s \in Ss$.

Definition 5. *Let S be a monoid. We define the relations $\equiv_M$ and $\leq_M$ on S by:*

$$s \equiv_M s' \quad \Leftrightarrow \quad s\mathcal{M}(S) = s'\mathcal{M}(S), \text{ and} \tag{6}$$
$$s \leq_M s' \quad \Leftrightarrow \quad s\mathcal{M}(S) \subseteq s'\mathcal{M}(S).$$

It is routine to show that $\equiv_M$ and $\leq_M$ are both (topologically) closed relations on any compact monoid S.

Combining properties 4. and 5. above on transformation semigroups with equivalences 3—5 yields:

Proposition 4. *Let $n \geq 1$ and let $C, C' \in ST(n)$. Then*

$$C(Prob(\overline{n})) = C'(Prob(\overline{n})) \quad \Leftrightarrow \quad \mathcal{M}(ST(n))C = \mathcal{M}(ST(n))C' \quad \Leftrightarrow \quad (\exists \pi \in S(n)) \, M_\pi C = C'. \tag{7}$$

Since $\equiv_M = \leq_M \cap (\leq_M)^{-1}$, we can form the relation $\leq_M / \equiv_M$, which is a closed partial order on $ST(n)/\equiv_M$.

Theorem 5. *Let $n \geq 1$. Then*

   1. *The relation $\equiv_M$ is a left congruence on $ST(n)$.*

   2. *$(ST(n)/\equiv_M, \leq_M / \equiv_M)$ is a compact ordered space, and the quotient map $\pi \colon ST(n) \to ST(n)/\equiv_M$ is an monotone map.*

   3. *As an ordered space, $ST(n)/\equiv_M \simeq Con_n(P_n)$.*

   4. *Thus cap $\colon (ST(n)/\equiv_M, \leq_m / \equiv_M) \to [0,\infty)^{op}$ measures $(ST(n)/\equiv_M, \leq_m / \equiv_M)$.*

   5. *If $C \in ST(n)$, then $Cap(C) = cap(\overline{conv}(C(1), \ldots, C(n)))$, where $C(i)$ is the ith row of C.*

Proof. For 1., it is clear from Equation 6 that $C \equiv_M C'$ implies $C''C \equiv_M C''C'$. Thus, $\equiv_M$ is a left-congruence.

Because $\equiv_M$ is a closed relation and $ST(n)$ is compact, the quotient space is compact and the quotient map is closed and continuous. The definition of the pre-order $\leq_M$ and the quotient order $\leq_M / \equiv_M$ implies the quotient map is monotone.

3. follows from Proposition 4, from which 4. and 5. are clear. $\qquad \square$

So we see that $ST(n))$ has a natural pre-order $\leq_M$ defined by its algebraic structure as a compact monoid, and if $C \leq_M C'$ then $Cap(C) \leq Cap(C')$. Moreover, this pre-order turns into a partial order on $ST(n)/\equiv_M$, and here capacity is the mapping cap. Importantly, cap — and hence Cap on $ST(n)$ — is strictly monotone with respect to this partial (pre-) order. Moreover, $C \equiv_M C'$ iff $S(n)C = S(n)C'$, so each is a permutation of the rows of the other.

Example 5. *Here's an example of what our results tell us about $ST(n)$. Recall that a Z-channel is a binary channel of the form*

$$Z_p = \begin{pmatrix} 1-p & p \\ 0 & 1 \end{pmatrix} = (1-p)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + p\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix},$$

showing that each lies on a one-parameter semigroup $\psi \colon ([0,1], \cdot) \to ST(2)$ by $\psi(p) = Z_p$. Now, ψ is a homomorphism, so $p < p' \Rightarrow Z_p = Z_{\frac{p}{p'}} Z_{p'}$, while obviously, $p \neq p' \Rightarrow S(2)Z_p \cap S(2)Z_{p'} = \emptyset$. It follows

that $p < p' \Rightarrow \pi(Z_p) < \pi(Z_{p'}) \Rightarrow Cap(Z_p) = cap(\pi(Z_p)) < cap(\pi(Z_{p'})) = Cap(Z_{p'})$, so the Z-channels $\{Z_p \mid 0 \leq p \leq 1\}$ all have distinct capacities.

Similarly, the matrices

$$Z'_p = \begin{pmatrix} 1 & 0 \\ 1-p & p \end{pmatrix} = p \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + (1-p) \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix},$$

form a one parameter semigroup from I_2 to $\mathscr{M}(ST(2))$, along which Cap is strictly decreasing. Now $\mathscr{M}(ST(2)) = \{r \cdot O_1 + (1-r)O_2 \mid 0 \leq r \leq 1\}$, where O_i is the matrix both of whose rows are $(\delta_{1i}\ \delta_{2i})$, for $i = 1, 2$. For each fixed $r \in [0,1]$, there is a one-parameter semigroup $p \mapsto p \cdot I_2 + (1-p) \cdot (r \cdot Z_p + (1-r) \cdot Z'_p)$, and combining the earlier results, we conclude that along this one-parameter semigroup, Cap is strictly decreasing. Note as well that $\overline{\mathrm{conv}}(\{I_2\} \cup \mathscr{M}(ST(2)))$ is equal to the union of these one-parameter semigroups.

We can generalize this verbatim *to $ST(n)$: define a Z-channel in $ST(n)$ to be one of the form $Z_p = p \cdot I_n + (1-p) \cdot O_k$, where $p \in [0,1]$ and O_k is the channel in $\mathscr{M}(ST(n))$ all of whose rows are $(0, \ldots, 0, 1, 0, \ldots)$, where the unique 1 appears in the kth entry. As in the binary case, $p \leq p' \Rightarrow Z_p = Z_{\frac{p}{p'}} Z_{p'}$, and $p \neq p' \Rightarrow S(n)Z_p \cap S(n)Z_{p'} = \emptyset$. So we can again conclude that Cap is strictly decreasing along this one-parameter semigroup. As in the case of $n = 2$, $\mathscr{M}(ST(n)) = \{\sum_{i \leq n} r_k O_k \mid \sum_k r_k = 1\}$, so the result extends* verbatim *to these one-parameter semigroups.*

6   Summary and future work

We have used an array of tools to analyze the capacity map on the set of classical channels. In Section 3, we gave a topological interpretation of capacity of a channel: it is the maximum distance between the surface generated by the entropy function applied to the rows of the channel, and the polytope generated by the entropy function applied to each individual row. This suggests an method for computing capacity: the Generalized Mean Value Theorem implies the capacity is achieved at the unique place where gradient of the capacity function is 0 — this point is unique because entropy is strongly concave. Moreover, this produces the input distribution where capacity is achieved — the celebrated Arimoto–Blahut Algorithm [2, 3] commonly used to compute the capacity of a discrete memoryless channel is an iterative procedure that approximates the capacity, not the input distribution where its value is assumed. An algorithm more closely related to our results can be found in [19], where the iteration scheme follows the concavity of the capacity function using Newton's Method.

We also applied our topological result to derive a domain-theoretic interpretation of capacity, again using the strong concavity of entropy: the family $Con_n(P_n)$ of polytopes with at most n vertices is a domain, and capacity measures this domain. The important point is that this implies capacity is strictly monotone with respect to the partial order. We found that $ST(n)$ has a natural, algebraically-defined pre-order whose associated equivalence $\equiv_M$ defines a closed congruence on $ST(n)$, and modulo which we obtain a copy of $Conv_n(P_n)$. This implies that capacity is strictly monotone with respect to the pre-order on $ST(n)$.

In addition, we used the probabilistic measures on compact spaces to define three monads, each of which tells us something about classical channels. The first realizes $ST(m,n)$ as the morphisms on the Kleisli category of the monad $Prob_S$. The second shows that $ST(n)$ is the free compact, affine monoid $Prob_M[\overline{n} \to \overline{n}]$, while the third shows that $DT(n)$ is the free compact monoid over $S(n)$, the group of permutations on n letters. This last also implies $DT(n)$ has a zero, which is Haar measure on $S(n)$.

   The work discussed here concerns classical channels, but we believe that much of it could be generalized to the quantum setting. We pointed out one connection to existing work on the roll of free affine monoids in analyzing quantum qubit channels. In any case, we have shown that the basic results of [14] generalize from the binary case, where the capacity function of a binary channel was studied in terms of the subinterval of $[0,1]$ it determines. The analog here is the polytope the rows of an $n \times n$-stochastic channel determine.

Acknowledgment

References

[1] Abramsky, S. and A. Jung, Domain theory, in: S. Abramsky, D. M. Gabbay, and T. S. E. Maibaum, editors, Handbook of Logic in Computer Science, volume 3, pages 1–168. Clarendon Press, 1994

[2] Arimoto, S., *An algorithm for computing the capacity of arbitrary DMCs*, IEEE Transaction on Information Theory **18** (1972), pp. 14–22.

[3] Blahut, R., *Computation of channel capacity and rate distortion functions,* IEEE Transaction on Information Theory **18** (1972), pp. 460–473.

[4] Bourbaki, N., "Espaces vectoriels topologiques", Chap. 1 à 5, Masson Paris etc., 1981, vii+368 pp.

[5] Clifford, A. H., and G. B. Preston, "The Algebraic Theory of Semigroups, I," Math. Surveys **7**, Amer. Math. Soc., Providence, R. I., 1962, 224 pp.

[6] Cover, T. and J. Thomas, "Elements of Information Theory", Wiley (2006), xiii + 748 pp.

[7] Crowder, T., *A quantum representation for involution groups*, Proceedings of MFPS 27, ENTCS 276, doi:`10.1016/j.entcs.2011.09.019`

[8] Gierz, G., K. H. Hofmann, K. Keimel, J. D. Lawson, M. Mislove and D. Scott, Continuous Lattices and Domains, Encyclopedia of Mathematics and its Applications **93**, Cambridge University Press, 2003, doi:`10.1017/CBO9780511542725`

[9] Hardy, G. H., J. E. Littlewood and G. Pólya, "Inequalities", 2nd ed., Cambridge University Press, Cambridge, UK, pp. 83-84, 1988.

[10] Hofmann, K. H. and M. Mislove, *Compact affine monoids, harmonic analysis and information theory*, preprint.

[11] Högnäs, G. and A. Mukherjea, "Probability Measures on Semigroups," Springer-Verlag, 2011, 430 pp., doi:`10.1007/978-0-387-77548-7`

[12] Jones, C., Probabilistic non-determinism, University of Edinburgh, Edinburgh, Scotland, 1992

[13] Martin, K., A Foundation for Computation, Tulane University PhD Thesis, New Orleans, LA, 2000.

[14] Martin, K., G. Allwein and I. Moskowitz, *Algebraic information theory for binary channels*, Theor. Comput. Sci. **411** (2010), 1918–1927, doi:`10.1016/j.tcs.2010.01.022`

[15] Martin, K., J. Feng, and S. Krishnan. *A Free Object in Quantum Information Theory*. Electronic Notes in Theoretical Computer Science **267** (2010), 35–47, doi:`10.1016/j.entcs.2010.08.004`

[16] Martin, K. and P. Panangaden, A technique for verifying measurements, Electronic Notes Theoretical Computer Science **218** (2008), pp 261-273, doi:`10.1016/j.entcs.2008.10.016`

[17] Rudin, W. (1991). "Functional Analysis" (2nd ed.). Boston, MA: McGraw-Hill

[18] Shannon, C. E., *A mathematical theory of communication*, Bell Systems Technical Journal **27** (1948), 379–423, 623–656, doi:10.1145/584091.584093

[19] Watanabe, Y., *Iteration procedure for the capacity of discrete memoryless channel,* Transaction of the IECE of Japan **67** (1984), pp. 440–444.

[20] Wendel, J. G., Haar measure and the semigroup of measures on a compact group, Proceedings of the American Mathematical Society **5** (1954), pp. 923-929,