Preprint

"Channel automorphisms and categories in information theory"

Yury Polyanskiy
(Юрий Полянский)

1 Main category of interest

Recall the definition of morphisms between measurable spaces:

Definition 1 *A category* `RanTra` *of random transformations is defined as follows: objects are measurable spaces; a morphism* $F : \mathbf{A} \to \mathbf{B}$ *is a transition probability kernel between the spaces; composition of morphisms and the totality of all morphisms* $Hom(\mathbf{A}, \mathbf{B})$ *are defined in a natural way. Probability measure* μ *on the space* $\mathbf{A}$ *is simply a morphism* $1 \to \mathbf{A}$ *from the space of cardinality* 1*.*

Given a measure μ on $\mathbf{A}$, we define a pushforward measure $F_*\mu$ via

$$\forall E \subset \mathbf{B} : (F_*\mu)(E) \stackrel{\triangle}{=} \mu(F^{-1}E)\,. \tag{1.1}$$

Some remarks:

1. A discrete space with M elements is denoted $[M]$.
2. Any measurable function $f : \mathbf{A} \to \mathbf{B}$ defines a morphism as follows:

   $$\forall x \in \mathbf{A}, E \in \sigma\mathbf{B} : \quad F_f(E|x) = 1\{f(x) \in E\}\,.$$

   In this sense `RanTra` is an extension of `Meas`.
3. A morphism F obtained from injective function is a monomorphism. Similarly, a morphism obtained from a non-injective function cannot be a monomorphism. Otherwise, the criterion for F to be a monomorphism is easier to describe after applying a functor to $\texttt{Vect}_{\mathbb{R}}$ (see below).
4. Similar discussion applies to epimorphisms and surjective functions. However, for epimorphisms the following additional necessary criterion is frequently useful:

   $$E_1 \neq E_2 \implies \exists x_1, x_2 : \quad F(E_1|x_1) \neq F(E_2|x_2)\,.$$

1.1 Binary Hypothesis Testing

We can modify the category `RanTra` by adding to each space a pair of measures P, Q.

Definition 2 *A category `BinHT` of binary hypothesis testing problems is defined as follows: objects are triplets consisting of a measurable space and a pair of measures P, Q. A morphism $F : (\mathbf{A}, P, Q) \to (\mathbf{B}, P', Q')$ is a transition probability kernel between the spaces satisfying:*

$$F_*(P) = P', \quad F_*(Q) = Q' .$$

Composition of morphisms and the totality of all morphisms $Hom(\mathbf{A}, \mathbf{B})$ are defined as in `RanTra`*.*

We can frequently omit the space $\mathbf{A}$ and simply talk about the binary hypothesis testing problem (P, Q). Some remarks:

- If $P \sim Q$ then (P, Q) is isomorphic to a problem over the space $\mathbb{R}$ via a map:

$$\mathbf{A} \to \mathbb{R} : x \to \frac{dP}{dQ}(x) .$$

- Any f-divergence defines a contravariant functor from `BinHT` to the poset $\mathbb{R}$. In fact any such functor is precisely a g-divergence defined in [12]. Note that $-\beta_\alpha(P, Q)$ (see [1, Chapter 2]) for every α defines a g-divergence which is not a monotone transformation of any f-divergence [12].

- More generally, a family $\{\mathcal{D}_\alpha, \alpha \in A\}$ of g-divergences defines a contravariant functor:

$$\texttt{BinHT} \to \prod_{\alpha \in A} \mathbb{R} ,$$

  where the category on the right is a poset.

  Conjecture: A (covariant) functor $\beta_\alpha(P, Q)$ defines an equivalence of category `BinHT` and a category of convex maps $[0, 1] \to [0, 1]$ understood as a poset.

  Convenience of this identification is obvious since it gives a clearer understanding of the category `BinHT`. However, there is a problem: a natural product in `BinHT`:

$$(\mathbf{A}, P, Q) \times (\mathbf{B}, P', Q') \stackrel{\triangle}{=} (\mathbf{A} \times \mathbf{B}, P \times P', Q \times Q')$$

  is not easily understood in the target category of convex maps. In non-fancy language this means that there is no convenient description of $\beta_\alpha(P \times P', Q \times Q')$ in terms of $\beta_\alpha(P, Q)$ and $\beta_\alpha(P', Q')$.

  Challenge: Find another equivalent representation of category `BinHT` where unlike the β_α-representation the product also has a natural form. (Conjecture: take all Rényi divergences!)

2 Channels

Warning on notation
Notice that here the term "channel" is equivalent to a random transformation in [1], whereas the channel in [1] means a sequence of random transformations.

Definition 3 *A channel $\mathcal{A} = (\mathbf{A}, K, \mathbf{B})$ consists of measurable spaces $\mathbf{A}, \mathbf{B}$ and a morphism between them. The diagram corresponding to the channel is*

$$\begin{array}{c} \mathbf{A} \\ \big\downarrow K \\ \mathbf{B} \end{array} .$$

Examples:

1. Discrete channel (DC) is specified by finite spaces $\mathbf{A}$, $\mathbf{B}$ (with power-set σ-algebras) and a $|\mathbf{A}| \times |\mathbf{B}|$ transition matrix K.

2. A particular discrete channel $\mathcal{I}_M = ([M], 1, [M])$, where the kernel 1 is just the identity map, is of importance for data compression.

3. For example, $BSC(n, \delta)$ is a channel with $\mathbf{A} = \mathbf{B} = \mathbb{Z}_2^n$ and

   $$Y^n = X^n + Z^n \,,$$

   where Z^n is binary i.i.d. noise, independent of X^n and with $\mathbb{P}[Z_j = 1] = \delta$.

4. The (hard-constrained) channel $AWGN(n, P)$ is the channel with input space

   $$\mathbf{A} = \{x \in \mathbb{R}^n : ||x|| \leq nP\}$$

   output space $\mathbf{B} = \mathbb{R}^n$ and the transition kernel K defined via

   $$Y^n = X^n + Z^n \,,$$

   where Z^n is iid gaussian noise $Z_j \sim \mathcal{N}(0, 1)$.

Definition 4 *A morphism $F = (f, g)$ of channels $F : \mathcal{A}_1 \to \mathcal{A}_2$ is a pair of maps making the following diagram commute:*

$$\begin{array}{ccc} \mathbf{A}_1 & \xrightarrow{f} & \mathbf{A}_2 \\ \big\downarrow K_1 & & \big\downarrow K_2 \\ \mathbf{B}_1 & \xleftarrow[g]{} & \mathbf{B}_2 \end{array} .$$

The set of all morphism between two channels is denoted $Hom(\mathcal{A}_1, \mathcal{A}_2)$. If it is non-empty then we say that channel $\mathcal{A}_1$ is noisier than $\mathcal{A}_2$ and write

$$\mathcal{A}_1 \prec \mathcal{A}_2 \,,$$

which in words means that it is possible (via a pair of devices f and g) to simulate channel $\mathcal{A}_1$ over the channel $\mathcal{A}_2$. Clearly $\prec$ defines a partial order on the set of channels. We say that two channels are weakly isomorphic and write $\mathcal{A}_1 \asymp \mathcal{A}_2$ if

$$\mathcal{A}_1 \prec \mathcal{A}_2 \,, \;\; \textit{and} \;\; \mathcal{A}_2 \prec \mathcal{A}_1 \,.$$

Examples:

1. One way to obtain channels with non-empty Hom between them is the following construction, which corresponds to modulation in communication. Take $\mathcal{B} = (\mathbf{A}', K', \mathbf{B}')$ to be any channel, fix some measurable spaces $\mathbf{A}, \mathbf{B}$ and morphisms $f : \mathbf{A} \to \mathbf{A}'$ and $g : \mathbf{B}' \to \mathbf{B}$. The channel $\mathcal{A} = (\mathbf{A}, K, \mathbf{B})$ is defined to have the kernel $K = g \circ K' \circ f$, i.e. the induced dashed line on the following diagram

$$\begin{array}{ccc} \mathbf{A} & \xrightarrow{f} & \mathbf{A}' \\ \vdots\, K & & \downarrow K' \\ \mathbf{B}_1 & \xleftarrow[g]{} & \mathbf{B}' \end{array} \quad .$$

   Obviously, the set $Hom(\mathcal{A}, \mathcal{B})$ contains at least morphism (f, g).

2. A particular example of the above construction is BPSK-modulated $AWGN(1, P)$ channel that leads to the following morphism F of channels:

$$BSC(1, Q(\sqrt{P})) \xrightarrow{F} AWGN(1, P) \,, \tag{2.1}$$

   where $\{0, 1\}$ are mapped to $\{-\sqrt{P}, \sqrt{P}\}$ and decoded via maximum likelihood decoder (signum function).

3. Most channels are incomparable and thus for them $Hom(\mathcal{A}, \mathcal{B})$ is empty. For an example notice that capacity $C(\mathcal{A})$ (as maximal mutual information) satisfies, by data-processing inequality:

$$\mathcal{A} \prec \mathcal{B} \quad \Longrightarrow \quad C(\mathcal{A}) \leq C(\mathcal{B}) \,.$$

   Therefore, for example $Hom(BSC(n, 0), BSC(n, \delta))$ is empty for $\delta \neq 1$.

4. **TODO:** Read reference [9] and see what counter-example it gives to Shannon's conjecture about sufficient condition for having $\mathcal{A} \prec \mathcal{B}$.

5. Finally, there is always one special identity morphism inside $Hom(\mathcal{A}, \mathcal{A})$, namely $I = (1_{\mathbf{A}}, 1_{\mathbf{B}})$.

The example 1 is an instance of a very important constuction, "modulation construction". We give the following definition:

Definition 5 *The totality of all possible morphisms between spaces* $\mathbf{A}, \mathbf{B}$ *through the channel* $\mathcal{B} = (\mathbf{A}', K', \mathbf{B}')$ *is denoted as* $Mod_{\mathcal{B}}(\mathbf{A}, \mathbf{B})$. *Obviously,*

$$Mod_{\mathcal{B}}(\mathbf{A}, \mathbf{B}) \subset Hom(\mathbf{A}, \mathbf{B}) .$$

To each element of $Mod_{\mathcal{B}}(\mathbf{A}, \mathbf{B})$ *corresponds a channel* $\mathcal{A} = (\mathbf{A}, g \circ K' \circ f, \mathbf{B})$ *and a morphism* $F = (f, g) : \mathcal{A} \to \mathcal{B}$. *Clearly we have then* $\mathcal{A} \prec \mathcal{B}$; *we will say that* $\mathcal{A}$ *is* $\mathbf{A} \to \mathbf{B}$ *modulation of* $\mathcal{B}$.

Various information theoretic problems can be restated in the language of the above "modulation construction". **Examples:**

1. Almost lossless and lossy data compression. A source is a measurable space $\mathbf{A}$ with a chosen measure P_X and a distortion function $d(\cdot, \cdot)$. The question is to describe all $\mathbf{A} \to \mathbf{A}$ modulations of $\mathcal{I}_M$ channel (see above). To each $K \in Mod_{\mathcal{I}_M}(\mathbf{A}, \mathbf{A})$ we associate its average distortion as follows:

$$d(K) = \int_{\mathbf{A} \times \mathbf{A}} d(P_X \times K) d(x, \hat{x}) ,$$

   where $d(\cdot, \cdot)$ is a distortion measure; for example, $d(x, \hat{x}) = 1\{x \neq \hat{x}\}$ for almost lossless data compression. The goal is to compute

$$d^*(\mathbf{A}, M) = \inf_{K \in Mod_{\mathcal{I}_M}(\mathbf{A}, \mathbf{A})} d(K) .$$

2. Channel coding. Given a fixed channel $\mathcal{A}$ the goal is to describe all $[M] \to [M]$ modulations of $\mathcal{A}$. More concretely, to each modulation $K \in Mod_{\mathcal{A}}([M], [M])$ we associate its average probability of error:

$$e(K) = 1 - \frac{1}{M} \sum_{j=1}^{M} K(j, j) ,$$

   and the question is to compute for each M the function

$$e^*(\mathcal{A}, M) = \inf_{K \in Mod_{\mathcal{A}}([M], [M])} e(K) .$$

Even more restrictively, the question is to compute the asymptotic properties of

$$e^*(\mathcal{A}^n, 2^{nR}),$$

for each R.

TODO: Note that if we denote modulation by (f, g) then the probability of success is $\operatorname{tr} fKg$ and hence we are trying to solve the problem:

$$\sup_{f,g} \operatorname{tr} fKg. \tag{2.2}$$

However, since $\operatorname{tr} fKg = \operatorname{tr} Kgf = \operatorname{tr} gfK$, we can interpret the problem (2.2) in a completely different way ("trace-duality").

3. Joint source channel coding (JSCC). Note that in the data compression we fixed the channel to be $\mathcal{I}_M$ and in the channel coding we fixed the source to be $[M]$. In the JSCC we do not fix either. The question is to describe for a given source $(\mathbf{A}, P_X, d)$ the smallest possible average distortion achievable over all $\mathbf{A} \to \mathbf{A}$ modulations of a fixed channel $\mathcal{B}$:

$$d^*(\mathbf{A}, \mathcal{B}) = \inf_{K \in Mod_{\mathcal{B}}(\mathbf{A},\mathbf{A})} \int_{\mathbf{A}\times\mathbf{A}} d(P_X \times K) d(x, \hat{x}).$$

We proceed to channel isomorphisms.

Definition 6 *The channels $\mathcal{A}$ and $\mathcal{B}$ are strongly isomorphic if there exists a pair of morphisms $F : \mathcal{A} \to \mathcal{B}$ and $G : \mathcal{B} \to \mathcal{A}$ such that $F \circ G = I_{\mathcal{B}}$ and $G \circ F = I_{\mathcal{A}}$.*

Obviously, if channels are strongly isomorphic then they are weakly isomorphic. The question is now to understand what constitutes the set of classifying invariants.

Examples:

1. Clearly, maximal mutual information afforded by the channel is equal for any weakly isomoprhic channels ("capacity is a weak-invariant").

2. Moreover, the cardinality of input and output spaces must be equal for strongly-isomoprhic channels ("cardinalities are strong-invariants").

3. Of course, not all weakly isomorphic channels are strongly isomorphic. For example, consider a channel $\mathcal{A} = ([1], P_Y, \mathbf{B})$, where P_Y is some fixed measure on $\mathbf{B}$. Then it is weakly isomorphic to $\mathcal{B} = ([2], P_Y, \mathbf{B})$ where the kernel is such that the output $Y \in \mathbf{B}$ is independent of the input and is distributed as P_Y. Clearly, though, that $\mathcal{A}$ and $\mathcal{B}$ are not strongly isomorphic since $|[1]| \neq |[2]|$.

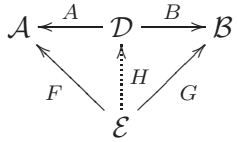
This construction can be easily generalized to show

Proposition 7 *Channels with zero capacity are weakly isomorphic.*

2.1 Non-existence of categorical channel products

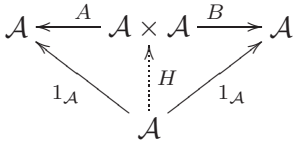
Categorical co-product is what we call channel sum (or a parallel channel). What is a categorical product?

Take a pair of channels $\mathcal{A}, \mathcal{B}$. Then in the usual category theory language, their product is the channel $\mathcal{D} = \mathcal{A} \times \mathcal{B}$ and a pair of (projection) morphisms A, B such that for any other channel $\mathcal{E}$ and morphisms $F : \mathcal{E} \to \mathcal{A}$ and $G : \mathcal{E} \to \mathcal{B}$ there exists a unique morphism $H : \mathcal{E} \to \mathcal{D}$ making the following diagram commute:



(2.3)

It raises the question whether channel products always exist. Here is an outline why $\mathcal{A} \times \mathcal{A}$ does not exist, when $\mathcal{A} = BSC(1, \delta)$. Indeed, first by taking $\mathcal{E} = \mathcal{A}$ and $F = G = 1_{\mathcal{A}}$ and obtaining the $H : \mathcal{A} \to \mathcal{A} \times \mathcal{A}$ which simultaneously extends two identity morphisms:



(2.4)

By writing diagrams for left and right morphisms we see that the output parts of A, B satisfy the property that $\operatorname{supp} a_2(0) \cap \operatorname{supp} a_2(1) =$ and similarly for b_2. Denote by K and T the kernels of the channels $\mathcal{A}$ and $\mathcal{A} \times \mathcal{A}$. Then since $T = a_2 \circ K \circ a_1$ we see that

$$\operatorname{supp} a_2(0) \cup \operatorname{supp} a_2(1) = \operatorname{supp} b_2(0) \cup \operatorname{supp} b_2(1) = \operatorname{supp} T\,,$$

where $\operatorname{supp} T$ is the union of supports of all $T(x), x \in (\mathcal{A} \times \mathcal{A})_1$. On the other hand, looking at the input side: $a_1 \circ h_1 = 1$ – implies that the input space of $\mathcal{A} \times \mathcal{A}$ splits into four disjoint sets $L_{00}, L_{01}, L_{10}, L_{11}$:

$$L_{ij} = \{x \in (\mathcal{A} \times \mathcal{A})_1 : \operatorname{supp} a_1(x) = \{i\}, \operatorname{supp} b_1(x) = \{j\}\}$$

and each subset is non-empty of course.

Now, on one hand for any x in L_{00} we must have

$$T(x) = a_2(0) = b_2(0)\,.$$

(here $a_2(0)$ and $b_2(0)$ are measures on $(\mathcal{A} \times \mathcal{A})_2$ of course) On the other hand, for any $x \in L_{01}$ we have

$$T(x) = a_2(0) = b_2(1)\,,$$

and we can see that since $(\mathcal{A} \times \mathcal{A})_1 = \cup L_{ij}$ we have

$$T = a_2(0) = a_2(1) = b_2(0) = b_2(1) = P_Y \,,$$

for some measure P_Y. This is a contradiction since (2.4) implies $\mathcal{A} \times \mathcal{A}$ and $\mathcal{A}$ are weakly isomorphic and thus must have equal capacities, but capacity of $\mathcal{A} \times \mathcal{A}$ is zero.

Note: another observation is that for any measurable space $\mathbf{A}$ and maps $f_1 : \mathbf{A} \to (\mathcal{A})_1$, $g_1 : \mathbf{A} \to (\mathcal{B})_1$ we can always take a channel $\mathcal{E} = (\mathbf{A}, 1, [1])$ and construct morphisms F, G to $\mathcal{A}, \mathcal{B}$ such that their input parts are f_1, g_1. Thus, if (2.3) holds then input part $(\mathcal{A} \times \mathcal{B})_1$ by the universal property of set-products must be actually equal to $(\mathcal{A})_1 \times (\mathcal{B})_1$[1].

3 Channel automorphisms

The set $Hom(\mathcal{A}, \mathcal{A})$ carries a natural structure of monoid. However, because there is always an element $P_Y \in Hom(\mathcal{A}, \mathcal{A})$ which induces some fixed distribution P_Y regardless of the input, this monoid's Grothendieck group is trivial. Instead, we restrict attention to the invertible elements of $Hom(\mathcal{A}, \mathcal{A})$.

Definition 8 *The automorphism group of the channel $\mathcal{A}$ is defined as follows:*

$$\operatorname{Aut} \mathcal{A} = \{F \in Hom(\mathcal{A}, \mathcal{A}) : \exists G \;\; s.t. \; F \circ G = 1, G \circ F = 1\} \,.$$

Simple remarks:

1. $\operatorname{Aut} \mathcal{A}$ naturally acts on both the input space $\mathbf{A}$ and output space $\mathbf{B}$. In the discrete case we have for each $\phi \in \operatorname{Aut} \mathcal{A}$:

$$P_{Y|X}(y|x) = P_{Y|X}(\phi(y)|\phi(x)) \,. \tag{3.1}$$

   Later addon: This requires some explanation. An element of $Hom(\mathcal{A}, \mathcal{A})$ has two components (f, g) and two elements are composed as

$$(f, g) \cdot (f', g') = (f \circ f', g' \circ g) \,. \tag{3.2}$$

   Therefore, if we just consider a naive action on $\mathbf{A}$ and $\mathbf{B}$ defined for each $\phi = (f, g)$ as

$$\text{wrong action:} \quad \phi(x) = f(x), \phi(y) = g(y)$$

   then because of (3.2) we have $\operatorname{Aut} \mathcal{A}$ acting on $\mathbf{A}$ on the left and on $\mathbf{B}$ on the right. There are two problems defining the action in this way. First, equation (3.1) should be rewritten in much less intuitive form:

$$\text{wrong action:} \quad P_{Y|X}(y|x) = P_{Y|X}(\phi^{-1}(y)|\phi(x)) \,.$$

[1]This is not really true, since f_1 and g_1 are not necessarily deterministic morphisms, and in the category where morphisms are transition probability kernels, one can not gurantee uniqueness of the map to $\mathbf{A} \times \mathbf{B}$.

Second, the map $\phi(x,y) = (f(x), g(y))$ does not define an action on $\mathbf{A} \times \mathbf{B}$ since $\phi \circ \phi_1(x,y)$ in general is not equal to either $(\phi\phi_1)(x,y)$ or $(\phi_1\phi)(x,y)$.

The correct definition of the (left) action of $\operatorname{Aut}\mathcal{A}$ on $\mathbf{A} \times \mathbf{B}$ is given as follows:

$$\begin{aligned} \phi = (f,g) : \mathbf{A} \times \mathbf{B} &\to \mathbf{A} \times \mathbf{B} && (3.3) \\ (x,y) &\mapsto (f(x), g^{-1}(y)) \,. && (3.4) \end{aligned}$$

It is under this action that equation (3.1) makes sense. It also means that all elements (x,y) in the orbit have the same probability $W(y|x)$. Also, for group-noise channels typical ϕ acts on (x,y) as $(g \circ x, g \circ y)$.

All in all: the rule of thumb is: if some transformation $f : \mathbf{A} \to \mathbf{A}$ is "equivalent" to $h : \mathbf{B} \to \mathbf{B}$ in the sense that $P_{Y|X} \circ f = h \circ P_{Y|X}$, then the element of $Hom(\mathcal{A}, \mathcal{A})$ is (f, h^{-1}), but the action of this element is $(x,y) \to (f(x), h(y))$.

2. Let $\mathcal{A} = BSC(n,\delta)$ where n is the blocklength and δ is the crossover probability. For the BSC we have

$$\operatorname{Aut}\Sigma = \mathbb{Z}_2^n \rtimes S_n \tag{3.5}$$

where S_n is the symmetric group on n elements. Indeed, denote a metric space of all $0,1$-strings of length n with Hamming norm $||\cdot||$ as B_n. Take $F = (\phi_1, \phi_2) \in \operatorname{Aut}\Sigma$ and denote. Then we have, by definition of the BSC

$$P_{Y|X}(y|\phi_1(x)) = P_{Y|X}(\phi_2^{-1}(y)|x) \,,$$

or, equivalently,

$$\delta^{||y-\phi_1(x)||}(1-\delta)^{n-||y-\phi_1(x)||} = \sum_{y' \in \phi_2^{-1}(y)} \delta^{||y'-x||}(1-\delta)^{n-||y'-x||} \,. \tag{3.6}$$

Summing over all x we find

$$|\phi_2^{-1}(y)| = 1 \,,$$

meaning that ϕ_2 is a bijection of B_n. Then from (3.6) it follows that

$$\forall x, y : \quad ||y - \phi_1(x)|| = ||\phi_2^{-1}(y) - x|| \,, \tag{3.7}$$

which implies that ϕ_1 is also a bijection. Of course, it is obvious that ϕ_1 and ϕ_2 are bijections from the definition of the automorphism; a direct proof demonstrates that we do not loose generality by that restriction at least in this important case.

Finally, from (3.7) by taking $y = \phi_1(x)$ we find that $\phi_2 = \phi_1$ and that each one is an automorphism of the Hamming space B_n. Hence (3.5) is proved because $\mathbb{Z}_2^n \rtimes S_n$ is the full automorphism group of the Hamming space.

3. For $\mathcal{A} = AWGN(n, P)$ we have

$$\operatorname{Aut}\Sigma = O(n)\,,$$

where $O(n)$ is the orthogonal group. This follows easily since an input component of $F \in \operatorname{Aut}\Sigma$ must preserve the cost function; i.e. if $\phi_1 = \operatorname{In} F$ then

$$x^2 = (\phi_1(x))^2\,,$$

and hence ϕ_1 is an orthogonal transformation.

4. For a general channel $\mathcal{A}^n$ we only know that there is an injection

$$S_n \hookrightarrow \operatorname{Aut}\mathcal{A}^n\,,$$

where S_n acts by permuting coordinates. This is just the expression of the fact that $\mathcal{A}^n$ is a memoryless channel.

5. In general, once a symmetry group is known, we can define generalized types as orbits in the input space under the action of $\operatorname{Aut}\mathcal{A}$. E.g., for the most general DMC of blocklength n we find that the symmetry group is S_n and thus the orbits the action of S_n are exactly the Csiszar-Korner-Marton types. However, for channels with larger $\operatorname{Aut}\mathcal{A}$, such as BSC, the group acts transitively (i.e. there is only one orbit) and therefore, type-splitting becomes not-necessary.

6. It is always true that $\operatorname{Aut}\mathcal{A}\times\mathcal{B}$ contains $\operatorname{Aut}(\mathcal{A})\times\operatorname{Aut}(\mathcal{B})$. Moreover, $\operatorname{Aut}(\mathcal{A}\times\mathcal{A})$ contains $\operatorname{Aut}(\mathcal{A})^2 \rtimes S_2$. However, the precise description might be hard. Indeed for example if $\mathcal{A} = BSC(1, 1/2)$ then

$$\operatorname{Aut}\mathcal{A}^2 = S_4 \neq (\operatorname{Aut}\mathcal{A})^2\,,$$

because of the additional symmetries.

7. **Q:** Aut for the group-noise channel (3.14).

   A: Proved to be nothing nice, unless P_N is of general type (i.e. $P_N(a) = P_N(b)$ implies $a = b$), in which case

$$\operatorname{Aut}\mathcal{A} \sim G\,,$$

because for any element $(g, g') \in \operatorname{Aut}\mathcal{A}$ there must be x_0 s.t. $g(x_0) = 1$, and since $[g(x)]^{-1}g'(y) = x^{-1}y$ for all $x, y \in G$ we must have

$$g'(y) = x_0^{-1}y\,,$$

and also

$$g(x) = x_0^{-1}x\,.$$

Note however, this is not very useful since P_N will not be of general type even for $BSC(n, \delta)$ for $n > 1$.

8. **TODO:** Aut for channel sum.

Theorem 9 *If P_X is a capacity achieving input distribution, then $P_X \circ g^{-1}$ is also capacity achieving for any $g \in \operatorname{Aut}\mathcal{A}$. For channels with $|\mathbf{A}|, |\mathbf{B}| < \infty$, the unique capacity achieving output distribution P_Y is constant on the orbits of $\operatorname{Aut}\mathcal{A}$.*

Remark: In particular if $\mathcal{A} = (\mathcal{A}_1)^n$ (an n-fold product) then the unique capacity achieving n-letter distribution is constant along the Y-types.

Proof: Simple data-processing demonstrates that for any $g \in \operatorname{Aut}\mathcal{A}$ we have

$$I(P_X, P_{Y|X}) = I(P_X \circ g^{-1}, P_{Y|X})\,.$$

Assume $|\mathbf{A}|, |\mathbf{B}| < \infty$, then

$$\begin{aligned} P_Y(y) &= \sum_{x\in\mathbf{A}} P_X(x)P_{Y|X}(y|x) && (3.8)\\ &= \sum_{x\in\mathbf{A}} P_X(x)P_{Y|X}(g(y)|g(x)) && (3.9)\\ &= \sum_{x\in\mathbf{A}} P_X(g(x))P_{Y|X}(g(y)|g(x)) && (3.10)\\ &= \sum_{x\in\mathbf{A}} P_X(x)P_{Y|X}(g(y)|x) && (3.11)\\ &= P_Y(g(y))\,, && (3.12) \end{aligned}$$

where (3.9) is by (3.1), (3.10) is because P_Y is unique and $P_X \circ g$ is also capacity achieving, (3.11) is because g is bijection of $\mathbf{A}$.**QED**.

3.1 Symmetric channels

Definition 10 *A (discrete) channel $\mathcal{A}$ is called Dobrushin-symmetric if every row of $P_{Y|X}$ is a permutation of the first and every column of $P_{Y|X}$ is a permutation of the first.*

Definition 11 *A (discrete) channel $\mathcal{A}$ is called Gallager-symmetric if output alphabet $\mathbf{B}$ can be split into a disjoint union of sub-alphabets such that restricted to each sub-alphabet $P_{Y|X}$ has the Dobrushin property: every row (every column) is a permutation of the first row (column).*

Definition 12 *A channel $\mathcal{A}$ is called input-symmetric (output-symmetric) if $\operatorname{Aut}\mathcal{A}$ acts transitively on the input (output) space.*

Definition 13 *A channel $\mathcal{A}$ is called weakly input-symmetric if there exists an input $x_0 \in \mathbf{A}$ and a collection of random transformations $T_x : \mathbf{B} \to \mathbf{B}, x \in \mathbf{A}$ such that:*

$$T_x \circ P_{Y|X=x_0} = P_{Y|X=x} \tag{3.13}$$

and

$$T_x \circ P_Y^* = P_Y^* \,,$$

where P_Y^* is the capacity-achieving output distribution.

TODO: Hidden assumption on the existence of P_Y^*. What should I do about it?

3.2 Relations between definitions of symmetric channels

1. My previous definition of input-symmetric channels required: for each permutation of inputs there must exist a permutation of outputs "undoing" the first. This is equivalent to requiring that $\operatorname{Aut}\mathcal{A}$ contain $S_{|\mathcal{A}|}$ canonically. This is too hard a requirement. It is not satisfied even for $BSC(2,\delta)$. Thus, this claim automatically rebuffs the conjecture that in the old-definition we have "product of inp.-sym. is inp.-sym."

2. Since $\operatorname{Aut}(\mathcal{A}) \times \operatorname{Aut}(\mathcal{B}) \leq \operatorname{Aut}\mathcal{A} \times \mathcal{B}$ we have trivially: *product of input symmetric channels is input symmetric.*

3. Relation to Dobrushin's definition. Recall that Dobrushin says that DMC W is symmetric if every row is a permutation of the first row and every column is a permutation of the first column. Obviously,

   $$\text{Dobrushin} \not\Longleftrightarrow \text{square}$$

   Easily, for each *group-noise channel*, i.e. channel of the form

   $$Y = X \circ N \tag{3.14}$$

   where $\circ$ is a composition inside some group G and $X, Y, N \in G$; each such channel is Dobrushin-symmetric. The converse is not true. According to [11] the latin squares that are Cayley tables are precisely the ones in which composition of two rows (as permutations) gives another row. An example of the latin square which is not a Cayley table is the following:

   $$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 5 & 4 & 1 & 3 \\ 3 & 1 & 2 & 5 & 4 \\ 4 & 3 & 5 & 2 & 1 \\ 5 & 4 & 1 & 3 & 2 \end{pmatrix}$$

   where numbers $1, 2, 3, 4, 5$ are to be replaced with arbitrary probabilities $p_1, \dots p_5$.

In fact, this channel is not even input-symmetric: exchange of the first and fourth rows is not possible to undo via column permutations. So we have shown:

$$\text{group-noise} \implies \text{Dobrushin, square} \tag{3.15}$$
$$\text{Dobrushin, square} \not\Longrightarrow \text{input-symmetric (let alone group-noise)} \tag{3.16}$$
$$\text{input-symmetric, square} \not\Longrightarrow \text{Dobrushin}\,, \tag{3.17}$$

where the counter-example for the last statement is the following channel:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 3 & 2 & 4 \\ 4 & 2 & 3 & 1 \\ 4 & 3 & 2 & 1 \end{pmatrix} \tag{3.18}$$

4. Channel (3.18) also demonstrates:

$$\text{Gallager-symmetric, square} \not\Longrightarrow \text{Dobrushin}$$

5. Note that it is an easy consequence of the definitions that

$$\text{input-symmetric} \implies \text{every row is a permut-n of the first} \tag{3.19}$$
$$\text{output-symmetric} \implies \text{every column is a permut-n of the first} \tag{3.20}$$

thus we have

$$\text{input-symmetric, output-symmetric} \implies \text{Dobrushin} \tag{3.21}$$

6. By splitting $\mathbf{B}$ into orbits of $\operatorname{Aut}\mathcal{A}$ on $\mathbf{B}$ we see that a subchannel $\mathbf{A} \to \{\text{orbit}\}$ is input and output symmetric. Thus by (3.21) we have:

$$\text{input-symmetric} \Longrightarrow \text{Gallager-symmetric} \tag{3.22}$$

7. Notice that the capacity achieving output distribution is constant on each subalphabet of a Gallager-symmetric channel. Thus:

$$\text{Gallager-symmetric} \implies \text{weakly input-symmetric}\,. \tag{3.23}$$

At the same time

$$\text{weakly input-symmetric} \not\Longrightarrow \text{Gallager-symmetric}\,, \tag{3.24}$$

where the counter-example is the following channel

$$W = \begin{bmatrix} 1/7 & 4/7 & 0 & 2/7 \\ 4/7 & 1/7 & 0 & 2/7 \\ 1/7 & 0 & 4/7 & 2/7 \\ 1/7 & 4/7 & 2/7 & 0 \end{bmatrix} . \tag{3.25}$$

Indeed, $\det W \neq 0$ and hence the capacity achieving input distribution is unique. But $P_X = [1/4, 1/4, 3/8, 1/8]$ achieves uniform P_Y and is thus optimal. Clearly any permutation $T_x : \{1,2,3,4\} \to \{1,2,3,4\}$ fixes a uniform P_Y and thus the channel is weakly input-symmetric. At the same time it is not Gallager-symmetric since no column is a permutation of another.

8. Here is another example of the w.i.s. channel which is not Gallager-symmetric. The importance of this example is that it makes use of the freedom of having T_x being randomized. The channel is:

$$W = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/4 & 1/4 \end{pmatrix}$$

To show it is w.i.s. notice that the $\operatorname{Aut} W$ contains element that flips first two inputs (rows 1,2) and flips last two inputs (columns 2,3). Thus $P_Y^*(2) = P_Y^*(3)$. Now take $x_0 = 1$ (corresp., first row) and $T_2 : \{1,2,3\} \to \{1,3,2\}$ and T_3 is given by the following matrix

$$T_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix}$$

To check that $T_3 \circ W(\cdot|1) = W(\cdot|3)$ simply write

$$\begin{pmatrix} 1/2 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix} = \begin{pmatrix} 1/2 & 1/4 & 1/4 \end{pmatrix}$$

Finally, by symmetry distribution P_Y^* has the form

$$P_Y^* = \begin{pmatrix} a & b & b \end{pmatrix} ,$$

and therefore $T_3 \circ P_Y^* = P_Y^*$ and $T_2 \circ P_Y^* = P_Y^*$ as required.

9. **Q:** Understand Sason's definition of the symmetric channel:

$$P_{Y|X}(y|x) = P_{Y|X}(g_x^{-1} y|0) ,$$

and $g_{x_1} \circ g_{x_2} = g_{x_1+x_2}$, i.e. the input space is a group $\mathbb{Z}_{|\mathbf{A}|}$ and there is a group homomorphism $Z_{|\mathbf{A}|} \to S_Y$ together with the noise distribution $P_N(\cdot) = P_{Y|X}(\cdot|0)$.

A: Sason's definition is just the requirement that once $\mathbf{A}$ is given a structure of $Z_{|\mathbf{A}|}$, there must be a subgroup of $\operatorname{Aut} G$ that acts on $\mathbf{A}$ by addition. Of course, such channel is automatically input-symmetric.

3.3 Sphere-packing and feedback

Theorem 14 *For input-symmetric channel with $|\mathbf{A}| < \infty$, uniform input distribution achieves capacity and thus we have*

$$C = D(W_a||PW), \forall a \in \mathbf{A}\,,$$

where W is the channel kernel, P is any capacity achieving distribution and W_a is the measure $W(\cdot|a)$. Furthermore, the dispersion can be calculated as

$$V \stackrel{\triangle}{=} V(W||PW|P) = V(W_a||PW), \forall a \in \mathbf{A}$$

and the upper bound on $\log M^*$ is improved to

$$\log M^*(n,\epsilon) \le nC - \sqrt{nV}Q^{-1}(\epsilon) + \frac{1}{2}\log n + O(1)\,,$$

and also holds with feedback.

Proof: The first claim follows from concavity of $I(P,W)$ in P and

$$U = \sum_{g \in \operatorname{Aut}\mathcal{A}} \frac{1}{|\operatorname{Aut}\mathcal{A}|} P \circ g^{-1}\,,$$

where U is uniform.

Second claim follows from the previous theorem after noticing that $V(W_a||PW)$ does not depend on a when $\operatorname{Aut}\mathcal{A}$ acts transitively on $\mathbf{A}$:

$$\begin{aligned}
V(W_{g(a)}||PW) + C^2 &= \sum_{y\in\mathbf{B}} W(y|g(a)) \log^2 \frac{W(y|g(a))}{PW(y)} && (3.26)\\
&= \sum_{y\in\mathbf{B}} W(g(y)|g(a)) \log^2 \frac{W(g(y)|g(a))}{PW(g(y))} && (3.27)\\
&= \sum_{y\in\mathbf{B}} W(g(y)|g(a)) \log^2 \frac{W(g(y)|g(a))}{PW(y)} && (3.28)\\
&= \sum_{y\in\mathbf{B}} W(y|a) \log^2 \frac{W(y|a)}{PW(y)} && (3.29)\\
&= V(W_a||PW)\,. && (3.30)
\end{aligned}$$

The last result follows from a fact

$$\beta_\alpha(P_{WY^n}, P_W \times (PW)^n) = \beta_\alpha(P_{Y^n|X^n=x^n}, (PW)^n)\,.$$

QED.

Note: Dobrushin's result is not contained and does not contain this last theorem. His proof works in the additional cases when all rows of W are permutations of each other, and thus the capacity achieving output distribution is uniform (as shown above, this does not imply the channel is input-symmetric). To include Dobrushin's result we need to go to weakly input-symmetric channels. *Input-symmetric differs from weakly input-symmetric in imposing a group structure on transformations T_x (see* (3.13)*). As the next theorem suggests, apparently, such requirement is superficial for the problems of sphere packing with feedback.*

Theorem 15 *Consider a weakly input-symmetric channel $\mathcal{A}$. Assume also that*

$$C' \stackrel{\triangle}{=} \sup_{x\in\mathbf{A}} D(P_{Y|X=x}||P_Y^*) < \infty \tag{3.31}$$

$$\sup_{x\in\mathbf{A}} V(P_{Y|X=x}||P_Y^*) < \infty\,. \tag{3.32}$$

Then all of the following hold:

1. *For $x_0 \in \mathbf{A}$ and P_Y^* as in Definition 13 we have:*

$$C' = D(P_{Y|X=x_0}||P_Y^*)\,,$$

2. *For any distribution P_{X^n} on $\mathbf{A}^n$ we have*

$$\beta_\alpha^n(P_{X^nY^n}, (P_Y^*)^n) \geq \beta_\alpha^n((P_{Y|X=x_0})^n, (P_Y^*)^n)\,. \tag{3.33}$$

3. *For any distribution P_X supported on $\{x \in \mathbf{A} : D(P_{Y|X=x}||P_Y^*) = C'\}$ we have*

$$V(P_X, P_{Y|X}) = V(P_{Y|X=x_0}||P_Y^*) \stackrel{\triangle}{=} V'\,. \tag{3.34}$$

4. *The following (sphere packing) bound holds with and without feedback:*

$$\log M^*(n,\epsilon) \leq -\log \beta_\alpha^n((P_{Y|X=x_0})^n, (P_Y^*)^n)\,,$$

5. *Finally, if*

$$V' > 0 \text{ and } T(P_{Y|X=x_0}||P_Y^*) < \infty$$

then as $n \to \infty$ we have

$$\log M^*(n,\epsilon) \leq nC' - \sqrt{nV'}Q^{-1}(\epsilon) + \frac{1}{2}\log n + O(1)\,.$$

If $V' = 0$ then we have

$$\log M^*(n,\epsilon) \leq nC' - \log(1-\epsilon)\,.$$

Remark: For discrete memoryless channels we know that $C' = C$ and $V' = V$ where (C, V) is the capacity-dispersion pair.

Proof: The first claim follows from the data-processing applied to $(P_{Y|X=x_0}, P_Y^*)$ and any transformation T_x from (3.13). To prove the second claim notice that

$$\begin{aligned} \beta_\alpha(P_{Y|X=x}, P_Y^*) &= \beta_\alpha(T_x \circ P_{Y|X=x_0}, T_x \circ P_Y^*) && (3.35) \\ &\geq \beta_\alpha(P_{Y|X=x_0}, P_Y^*) && (3.36) \end{aligned}$$

where (3.36) follows from the data-processing inequality applied to β_α. A straightforward generalization then shows that for any $x^n \in \mathbf{A}^n$ we have

$$\beta_\alpha^n(P_{Y^n|X^n=x^n}, (P_Y^*)^n) \geq \beta_\alpha^n(P_{Y^n|X^n=x_0^n}, (P_Y^*)^n)\,, \tag{3.37}$$

where x_0^n denotes the string of n letters x_0. Now following the proof of Lemma 29 in [**?**] we get (3.33).

For any $x \in \mathbf{A}$ we have that the second moment of $\log \frac{dP_{Y|X=x}}{dP_Y^*}$ is finite by (3.32) and hence by the CLT:

$$\log \beta_\alpha^n((P_{Y|X=x})^n, (P_Y^*)^n) = -nD(P_{Y|X=x}||P_Y^*) - \sqrt{nV(P_{Y|X=x})}Q^{-1}(\alpha) + o(\sqrt{n})\,.$$

Therefore, from (3.37) we conclude that (3.34) holds (otherwise (3.37) will be violated either for $\alpha < 1/2$ or for $\alpha > 1/2$).

The last two claims are proved as for the input-symmetric case. **QED**

3.4 Functor: $\mathtt{RanTra} \to \mathtt{Vect}_{\mathbb{R}}$.

To any measurable space X we associate a vector space V_X of measurable functions on it. To any morphism $X \stackrel{P_{Y|X}}{\longrightarrow} Y$ we associate a linear map $V_Y \stackrel{F}{\longrightarrow} V_X$ as follows:

$$F(\phi)(x) = \int_Y \phi(y) dF_{Y|X=x}\,.$$

This is a contravariant faithful functor (that is, $F_1 = F_2$ (linear maps) if and only if $P_{Y|X}^1 = P_{Y|X}^2$ (transition kernels)). The functor is not full: its image is the set of all positive linear maps preserving constant 1-function.

Remarks:

1. The map between $[n]$ and $[k]$ is just a $n \times k$ stochastic matrix. This establishes a functor between the subcategory of `RanTra` consisting of discrete spaces and finite-dimensional $\mathbb{R}$-vector spaces with stochastic-matrices as morphisms between them. For example, the image of the $BSC(1, \delta)$ is

$$\begin{pmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{pmatrix}$$

2. Channel sum corresponds to direct sum of vector-spaces and direct-sum of morphisms.

3. Channel product corresponds to tensor product. For example,

$$BSC(n,\delta) \leftrightarrow \begin{pmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{pmatrix}^{\otimes n}$$

Note that the operator on the operator on the right is diagonalizable with $n+1$ eigenspaces corresponding to eigenvalue $(1-\delta)^k, k = 0, \ldots, n$ and dimension $\binom{n}{k}$.

This last point brings an interesting interpretation to Shannon's theorem. For any (positive, 1-preserving) linear map $F : V_Y \to V_X$ between finite-dimensional $\mathbb{R}$-vector spaces there exists an inclusion g^* and a surjection f^* from an 2^k-dimensional space $\mathbb{R}^{2^k}$ such that the overall map

$$\mathbb{R}^{2^k} \xrightarrow{g^*} V_Y^{\otimes n} \xrightarrow{F^{\otimes n}} V_X^{\otimes n} \xrightarrow{f^*} \mathbb{R}^{2^k}$$

satisfies

$$||f^* \circ F^{\otimes n} \circ g^* - I_{2^k}||_\infty \leq \epsilon$$

provided that $k < nC$ and n is sufficiently large (depending on how small ϵ is). This is not possible for $k > nC$.

Thus Shannon's theorem shows that tensor products asymptotically have exponentially large rigid linear subspaces.

Future work:

1. Interpret Shannon's theorem for $BSC(n,\delta)$ in terms of the eigenspace decomposition of $F^{\otimes n}$.

2. describe symmetries and channel equivalences in the target category of finite-dimensional spaces.

3. According to [10], if two doubly stochastic matrices K and K' are similar, then there exist doubly stochastic matrices C and C_1 such that

$$CK = K'C\,, \text{ and } KC_1 = C_1K'\,,$$

or in the diagrammatic way:

$$\begin{array}{ccccc} \mathbf{A} & \xrightarrow{C} & \mathbf{A}_1 & \xrightarrow{C_1} & \mathbf{A} \\ \downarrow{\scriptstyle K} & & \downarrow{\scriptstyle K'} & & \downarrow{\scriptstyle K} \\ \mathbf{B} & \xrightarrow{C} & \mathbf{B}_1 & \xrightarrow{C_1} & \mathbf{B} \end{array}\quad, \tag{3.38}$$

where $|\mathbf{A}| = |\mathbf{B}| = |\mathbf{A}_1| = |\mathbf{B}_1|$ and $K \sim K'$ (as square matrices). Same conclusion is true for stochastic (not doubly) K and K' under some additional conditions (see [10]). Unfortunately, the directions of arrows in (3.38) is not correct, so we can not immediately conclude that $K \prec K'$ or $K \succ K'$.

3.5 Functor: `RanTra` $\to$ `ConVect`

Define a category of *coned vector spaces* as follows:

Definition 16 *A finite dimensional coned vector space V is an $\mathbb{R}$-vector space together with a positive cone C_+ and a unity element $1_V \in C_+$ which satisfy the following properties:There exists a unique (upto reordering) basis $e_\alpha, \alpha \in I$ of V such that*

$$C_+ = \text{hull}\{0, e_\alpha, \alpha \in I\}\,.$$

In this case the unity element should satisfy

$$1_V = \sum_\alpha e_\alpha\,.$$

Morphism of coned spaces is a morphism of vector spaces that preserves a positive cone and the unity.

The following is a contravariant functor from `RanTra`: to a measurable space S we associate a coned space $F(S) = V_S$ of functions $S \to \mathbb{R}$ with C_+ being the set of all positive functions and 1_S being the identity function[1] And to a morphism of spaces $f : S \to R$ we associate a linear map

$$F(f) : v(r) \to \int v(r) f(dr|s)\,.$$

The canonical basis of V_S is naturally indexed by elements of S:

$$\text{basisof } V_S = \{e_s, s \in S\}.$$

We also adopt the following convention: for each subset $R \subset S$ we denote

$$e_R = \sum_{r \in R} e_r \in V_S\,.$$

Remarks:

1. Functor F establishes the (co-)equivalence of categories (the apparent reverse functor is actually an adjoint functor) of finite measurable spaces and finite-dimensional coned spaces.

2. Each coned space has a natural inner product under which the C_+-basis is orthogonal.

3. Taking any subset of basis vectors e_α we can form a subspace $V' \subset V$ such that the orthogonal projection is a morphism of coned spaces. Inclusion $V' \hookrightarrow V$, however, is not a morphism since it does not preserve unities.

[1] Note that this V_S also has a natural ring structure (under pointwise multiplication) but morphisms of measurable spaces do not preserve this structure, so we did not include it in the definition of a coned space.

4. Two interesting subsets are defined on each coned space:

$$K_V = \{x \in V : \forall \alpha < x, e_\alpha > \in [0,1]\}, \tag{3.39}$$
$$P_V = \text{hull}(e_\alpha, \alpha \in I) = \{x \in C_+ :< x, 1_V >= 1\}. \tag{3.40}$$

5. A probability measure P on S corresponds to an element $p \in P_V$:

$$\mathbb{E}_P[f(S)] =< p, f > .$$

6. Note that any (image under the functor) of a binary morphism $S \stackrel{H}{\to} [2]$ is completely specified by a single element $h \in K_{V_S}$. Indeed:

$$V_2 \stackrel{\tilde{H}}{\to} V_S : \quad \tilde{H}(v_0) = 1_S - \tilde{H}(v_1)$$

and thus $\tilde{H}(v_0)$ should belong to K_{V_S} and is otherwise arbitrary.

7. A binary hypothesis testing curve for (S, P, Q) can be found as

$$\beta_\alpha(P,Q) = \inf_{h \in K : < h,p > \geq \alpha} < h, q > \tag{3.41}$$
$$= \alpha \frac{< p, q >}{||p||} + \inf_{v : \alpha \frac{p}{||p||} + v \in K} < v, q > \tag{3.42}$$

Finally, corresponding to a coding diagram

$$[M] \stackrel{f}{\to} X \stackrel{T}{\to} Y \stackrel{g}{\to} [M]$$

we have

$$[M] \stackrel{\tilde{g}}{\to} Y \stackrel{\tilde{T}}{\to} X \stackrel{\tilde{f}}{\to} [M].$$

Note that when the encoder is deterministic and without repeated codewords we have:

$$\tilde{f}(e_x) = \begin{cases} 0, & x \notin f([M]), \\ e_w, & x = f(w), w \in [M]. \end{cases} \tag{3.43}$$

When the decoder is deterministic then we have

$$\tilde{g}(e_{\hat{w}}) = \sum_{y \in D_{\hat{w}}} e_y = e_{D_{\hat{w}}}, \tag{3.44}$$

where $D_{\hat{w}} = g^{-1}(\hat{w})$ is a decoding set.

Below we make the following assumptions about f, g and T:

1. $X = Y = S$ (i.e. spaces of inputs and outputs can be identified). Then the main vector space is denoted V instead of V_S.

2. encoder and decoder are as in (3.43), (3.44).

3. Therefore, we can identify $[M]$ as a subset of S. Then $V_M \hookrightarrow V$ is an (orthogonal) inclusion and $\tilde{f}$ is a projection:

$$\tilde{f} = \mathrm{Proj}(V \to V_M)\,.$$

4. Moreover, the code is "linear", so that the cardinality of each $D_{\hat{w}}$ is the same. This implies that $\tilde{g}$ can be extended from an orthogonal operator $V_M \to V$ to a full orthogonal operator $V \to V$ satisfying

$$\tilde{g}^* \circ \tilde{g} = |D_{\hat{w}}| I = \frac{|S|}{M} I\,.$$

After these assumptions we have the following picture: an endomorphism of a coned space $\tilde{T} : V \to V$ is given. A coned subspace V_M is contained as inclusion in V and an orthogonal operator $\tilde{g} : V \to V$ is given. Together they satisfy:

$$\mathrm{tr}\, \tilde{f}\tilde{T}\tilde{g} = M(1-\epsilon)\,,$$

where ϵ is the probability of error of the original code (f, g). Note that $\tilde{f}\tilde{T}\tilde{g}$ is an $|S|$-by-$|S|$ matrix with block-structure:

$$\tilde{f}\tilde{T}\tilde{g} = \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix}\,,$$

where Λ is a doubly stochastic matrix (because of "linearity" of the code). A code is good if

$$A = \tilde{f}\tilde{T}\tilde{g} - \tilde{f}$$

is almost a zero-operator. More precisely we have the estimates:

$$\epsilon\sqrt{\frac{M}{M-1}} \;\leq\; ||Ae_s|| \leq \sqrt{2}\epsilon \tag{3.45}$$

$$||A|| \;\leq\; \sqrt{2M}\epsilon\,. \tag{3.46}$$

(Note that it is not clear immediately whether even for BSC $\sqrt{M}\epsilon \to 0$ for all rates?)

3.6 Functor: channels $\to$ synchronized channels

Recall that a kernel $P_{Y|X} : \mathbf{A} \to \mathbf{B}$ in the definition of the channel is required to satisfy two conditions:

1. for a fixed $\mathbf{x} \in \mathbf{A}$, $P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\cdot)$ is a probability measure on $(\mathbf{B}, \mathcal{F}_{\mathbf{B}})$, where $\mathcal{F}_{\mathbf{B}}$ is the σ-algebra implicit in the definition of $\mathbf{B}$; and

2. for a fixed $E \in \mathcal{G}$ the function $\mathbf{x} \mapsto P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(E)$ is $\mathcal{F}_{\mathbf{A}}$-measurable.

For problems with feedback the notion of time and causality is of vital importance. We therefore need to add some more structure to the definition of the channel.

Definition 17 *A synchronized channel $\mathcal{A} = (\mathbf{A}, K, \mathbf{B}, \{\mathcal{F}_n, \mathcal{G}_n\})$ is an abstract channel $(\mathbf{A}, \mathbf{B}, K)$ with filtrations $\mathcal{F}_n$ and $\mathcal{G}_n$ on $\mathbf{A}$ and $\mathbf{B}$, resp., and the requirement that K be a transition probability kernel from $(\mathbf{A}, \mathcal{F}_n)$ to $(\mathbf{B}, \mathcal{G}_n)$ for each $n \in \mathbb{Z}$. Naturally we define $Hom(\mathcal{A}, \mathcal{B})$ as morphisms measurable w.r.t. $\mathcal{F}_n, \mathcal{G}_n$. Similarly* $\operatorname{Aut}\mathcal{A}$ *is defined as invertible elements of $Hom(\mathcal{A}, \mathcal{A})$.*

For notational simplicity, we also assume there are two pre-chosen sequences of functions on $\mathbf{A}$ and $\mathbf{B}$, such that

$$\mathcal{F}_n = \sigma\{X_k, k \le n\}\,, \text{ and } \mathcal{G}_n = \sigma\{Y_k, k \le n\}\,.$$

Definition 18 *A time-shift for a channel $\mathcal{A}$ is a pair of maps $\mathbf{A} \stackrel{T_A}{\to} \mathbf{A}$ and $\mathbf{B} \stackrel{T_B}{\to} \mathbf{B}$ such that the following diagram commutes:*

$$\begin{array}{ccc} \mathbf{A} & \xrightarrow{T_A} & \mathbf{A}_2 \\ \downarrow{\scriptstyle K} & & \downarrow{\scriptstyle K} \\ \mathbf{B} & \xrightarrow{T_B} & \mathbf{B}_2 \end{array} \quad .$$

A partition of the channel $(\mathcal{A}, T)$ is a pair of partitions π and σ (of $\mathbf{A}$ and $\mathbf{B}$, resp.) such that

$$\forall S \in \sigma \forall P \in \pi \forall x_1, x_2 \in P : K(S|x_1) = K(S|x_2) \tag{3.47}$$

(this is simply an expression of the usual condition for a non-anticipatory channel).

Note that if (π, σ) is a partition, then $(T_A^{-1}\pi, T_B^{-1}\sigma)$ is a partition too.

Definition 19 *A channel $(\mathcal{A}, T)$ is said to be Bernoulli if there are finite partitions of input and output spaces (satisfying* (3.47)*) such that the action of T on them generates σ-algebras on $\mathbf{A}$ and $\mathbf{B}$ respectively.*

Any channel $\mathcal{A}$ can be extended memorylessly to a synchronized Bernoulli channel $\mathcal{A}'$ by the following construction. We take $\mathbf{A}' = \mathbf{A}^\infty$ and X_j as the usual projections onto

j-th coordinate; similarly we construct $\mathbf{B}' = \mathbf{B}^\infty$ and Y_j. The kernel $P_{\mathbf{Y}|\mathbf{X}}$ is defined as an extension of the following sequence of finite dimensional kernels:

$$P_{Y^n|X^n=(x_1,\dots x_n)} = \prod_{j=1}^{n} P_{Y|X=x_1}\,,$$

where the product is the product of measures on $\mathcal{B}$. A synchronized channel obtained in this way starting from finite spaces $\mathbf{A}$ and $\mathbf{B}$ is called discrete memoryless (DMC). The Bernoulli shift T is obviously defined as a time shift. This description is functorial (i.e. maps of channels $\mathcal{A} \to \mathcal{B}$ induce maps of synchronized channels $\mathcal{A}' \to \mathcal{B}'$).

4 Symmetric codes

Linear codes over $BSC(n,\delta)$ correspond to a particular way of generating constellations in the input space $\mathbb{Z}_2^n$. Namely, there is an injection

$$\mathbb{Z}_2^n \hookrightarrow \operatorname{Aut} BSC(n,\delta)$$

and for any subgroup H of $\mathbb{Z}_2^n$ we define codebook to be all H translates of the fixed vector $x_0 = (0,0,\dots,0)$. Moreover, typically the way to choose $H = \mathbb{Z}_2^k$ is by giving $\mathbb{Z}_2^n$ the structure of a vector space $\mathbb{F}_2^n$ and defining $H = \operatorname{span}(g_1,\dots,g_k)$ for a certain collection $(g_1,\dots,g_k)$ of $\mathbb{F}_2$-independent vectors.

To summarize, a $[k,n]$ linear code is defined by a choice of injection

$$\mathbb{Z}_2^k \hookrightarrow \operatorname{Aut} BSC(n,\delta)$$

and an initial vector x_0.

As a generalization we give the following definition, which encompasses linear codes over BSC and BIAWGN, geometrically uniform (GU) [2] and G-generated codes [8] over AWGN.

Definition 20 *A symmetric code for the channel $\mathcal{A}$ is defined by a pair (x_0, H) where $x_0 \in \mathbf{A}$ is an element of the input space $\mathbf{A}$ of $\mathcal{A}$ and H is a subgroup of* $\operatorname{Aut}\mathcal{A}$. *Explicitly, the encoder is a map $H \to \mathbf{A}$ given by*

$$f(h) = h(x_0)\,.$$

Therefore, the cardinality of the code is $|H|$; the cardinality of the codebook is $[H : H \cap \operatorname{Stab} x_0]$. The decoder is assumed to be a maximum likelihood one.

Although such a choice might look quite restrictive, these codes do achieve capacity of some symmetric channels [7]. At the same time the syndrome decoding generalizes to them and thus they are somewhat easier to decode than a general code.

Although, for BSC the probability of error does not depend on the choice of x_0, it might not be true in general.[1] **TODO:** conjugate subgroups correspond to the same code? Change of basepoint within H-orbit is a permutation of codewords? Cyclic codes corresp. to $\sigma H \sigma^{-1} = H$. More generally, are there normal subgroups of the Coxeter group?

Theorem 21 *Consider channel $\mathcal{A} = (\mathbf{A}, K, \mathbf{B})$ and a symmetric code (H, x_0). Then there exists a randomized maximum likelihood decoder such that with this decoder the average probability of error equals to the maximum probability of error and, of course, the P_e is minimal among all possible decoders.*

Proof: This is an extension of the Theorem 56, Appendix A [**?**]. Denote the kernel of $\mathcal{A}$ as $P_{Y|X}$ for convenience.

For each y define the maximum likelihood function:

$$ml(y) = \max_{h \in H} P_{Y|X}(y|hx_0)$$

and the cardinality of the set of maximum likelihood codewords:

$$N(y) = |\{h \in H : P_{Y|X}(y|hx_0) = ml(y)\}| .$$

Now consider a randomized maximum likelihood decoder:

$$d_r : \mathbf{B} \to H ,$$

where

$$d_r(y) = h \quad \text{w.p. } \tfrac{1}{N(y)} \text{ if } P_{Y|X}(y|hx_0) = ml(y)$$

Denote

$$q(y, h) = \mathbb{P}[d_r(y) = h] .$$

Then by definition $q(y, h) = \frac{1}{N(y)}$ if h is among maximum likelihood codewords and 0 otherwise. Notice,

$$q(h_1 y, h) = q(y, h_1^{-1} h) .$$

Therefore, the probability of error of decoding the codeword h is

$$\begin{aligned} \lambda_h &= \mathbb{P}[d_r(Y) \neq h | X = hx_0] && (4.1)\\ &= \mathbb{E}\left[1 - q(Y, h) | X = hx_0\right] && (4.2)\\ &= \mathbb{E}\left[1 - q(h(Y), 1) | X = x_0\right] && (4.3)\\ &= \lambda_1 . && (4.4) \end{aligned}$$

Therefore, all codewords are equally protected, as claimed.

Remarks:

[1] This property of the BSC is a consequence of the fact that for each $H \leq \mathbb{Z}_2^n$ there is an H' isomorphic to H and such that x_0 and 0 belong to the same H'-orbit.

1. In [8] authors compute the following. Suppose $G \leq \operatorname{Aut} \mathcal{A}$ is abelian subgroup and suppose that it acts regularly (or simply transitively) on $\mathbf{A}$. Then what is the maximal rate achievable by symmetric codes $(H, 0)$ where $H \leq G^n$:

$$C_G = \lim_{\epsilon \to 0} \liminf_{n \to \infty} \max_{H \leq G^n, P_e(H) \leq \epsilon} \frac{1}{n} \log |H| \, .$$

For example, when $G = \mathbb{Z}_{p^r}$ the answer turns out to be

$$C_G = \min_{0 \leq l \leq r} \frac{r}{l} C_l \, ,$$

where C_l is the usual Shannon capacity of the channel $\mathcal{A}_l$ with inputs restricted to $p^{r-l}\mathbb{Z}_{p^r}$ (under obvious identification of $\mathbb{Z}_{p^r}$ and $\mathbf{A}$. Interestingly, authors show example of a channel with $G = \mathbb{Z}_8$ for which C_G less than the Shannon capacity.

4.1 Application to AWGN

Consider an AWGN of blocklength n: it has in input space

$$\mathbf{F}_n = \{\mathbf{x} : ||\mathbf{x}||^2 = nP\} \subset \mathbb{R}^n$$

(it is a standard method to reduce the power constraint $||\mathbf{x}||^2 \leq nP$ to the one with equality by adding an $(n+1)$-st coordinate).

The channel acts from $\mathbf{F}_n$ to $\mathbb{R}^n$ by adding white Gaussian noise:

$$Y^n = X^n + Z^n \, , \qquad Z^n \sim \mathcal{N}(0, I_n) \, .$$

A simple argument shows then

$$\operatorname{Aut} \text{AWGN} = O(n)$$

since $O(n)$ is the totality of isometries of Euclidean space preserving the sphere.

Symmetric codes for AWGN should be, according to this paradigm, be identified with discrete subgroups of $O(n)$, the orthogonal group. Classifying and understanding all such subgroups is a long-standing mathematical problem. However, there is a better-understood subgroup $C_n \leq O(n)$ called the Coxeter subgroup, which is isomorphic to $\mathbb{Z}_2^n \rtimes S_n$. Interestingly, restricting attention to C_n and its subgroups is equivalent to factoring the encoder through (2.1), or equivalently applying the BPSK modulation and then doing linear-coding on top of it.

Note however, that inclusion $G \hookrightarrow O(n)$ is an orthogonal representation of the group G (any finite dimensional representation can be ortogonalized, so there is no loss of generality here). Thus, any and all symmetric codes for the AWGN come from representations of finite groups. *Moreover, the ones that correspond to reducible representations are equivalent to concatenating to shorter codes. Thus it is imperative to focus on irreducible representations.*

Remarks:

1. It might be worthwile to initially focus on faithful representations (i.e. only 1 element of G is represented by an identity operator).

2. If we also require the action of G to be free (which is equivalent to asking that no respective linear operators had an eigenvalue of 1) then finding such action is the same as finding an $(n-1)$-dimensional compact smooth Riemannian manifold with constant positive sectional curvature (equivalently, compact topological manifolds with finite fundamental group G and covered by S^{n-1}).

3. Classically (see [2]) codes from subgroups of $O(n)$ were proposed by Slepian in [3] (a later extensive review is in [4]). However, the problem with their approach is that they restrict attention to commutative groups for which there cannot be real irreducible representation for dimension > 2. Note that geometric uniform codes [2] differ from our treatment in allowing translations in addition to $O(n)$ (to include things like lattice codes), so is not directly interesting. **TODO:** Read [3] and [4].

4. Maybe related: Slepian gave a funny counter-example of a 10-point constellation in $\mathbb{R}^5$ which has transitive symmetry group, but cannot be generated as an orbit code by a subgroup of the symmetry group. A recent example of orbit codes: [5] and [6].

Real and complex representations

Given a real representation ρ_{real} we can form a complex representation by taking $\rho_{real} \otimes \mathbb{C}$. We define any complex representation ρ to be *essentially real* if there exists some ρ_{real} such that $\rho \simeq \rho_{real} \otimes \mathbb{C}$. For a complex representation ρ its realification is denoted by $\rho_{\mathbb{R}}$.

Remarks:

1. ρ_{real}–irred. does not imply $\rho_{real} \otimes \mathbb{C}$ – irred. (c/ex: $\mathbb{Z}_4$ with generator $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$).

2. ρ – irred. does not imply $\rho_{\mathbb{R}}$ irred. (c/ex: any essentially real rep.).

3. For any representation ρ_{real} the space supporting $\rho_{real} \oplus \rho_{real}$ can be given $\mathbb{C}$-structure such that $\rho_{real} \oplus \rho_{real} \simeq_{\mathbb{C}} \rho_{real} \otimes \mathbb{C}$ (as complex reps.).

4. ρ–essentially real $\iff$ $\exists$ basis of V s.t. all matrices of ρ are real.

5. ρ–essentially real $\iff$ $\exists$ equivariant antilinear map $J : V \to V$ satisfying $J^2 = 1$. Antilinearity ensures that the space V splits into $+1$ and -1 eigenspaces evenly. ρ_{real} is taken to be a restriction of ρ to $+1$ eigenspace.

6. Necessary criterion: ρ – essentially real then $\rho_{\mathbb{R}} \simeq \rho_{real} \oplus \rho_{real}$ for some ρ_{real}. Is it sufficient? I.e. do we always have $\rho_{\mathbb{R}} \simeq \rho_{real} \oplus \rho_{real} \implies \rho \simeq \rho_{real} \otimes \mathbb{C}$?

7. Given a $\mathbb{C}$-vector space V we can define its complex-conjugate $\bar{V}$ as follows: $\bar{V} = V$ (as an abelian group) but $\lambda \cdot v \stackrel{\triangle}{=} (\bar{\lambda})v$ (where here and below $\cdot$ denotes multiplication in $\bar{V}$ and absence of $\cdot$ – that of V). The following are simple relations between V and $\bar{V}$:

   - $Hom_{\mathbb{C}}(V,V) = Hom_{\mathbb{C}}(\bar{V},\bar{V}) = \{H \in End_{\mathbb{R}}(V) : H$ commutes with $i : V \to_{\mathbb{C}} V\}$.
   - $Hom_{\mathbb{C}}(V,\bar{V}) = Hom_{\mathbb{C}}(\bar{V},V) = \{H \in End_{\mathbb{R}}(V) : H$–antilinear$\}$.
   - One may construct isomorphism between V and $\bar{V}$ by picking an arbitrary basis $\{e_j\}$ and defining $J : V \to \bar{V}$ via $e_j \to e_j$, $ie_j \to -ie_j$ and extending this by $\mathbb{R}$-linearity. All such isomorphisms (as $\mathbb{R}$-linear maps) are characterized by the condition:
   $$J^2 = 1, J \in End_{\mathbb{R}}(V)\,. \tag{4.5}$$
   - A hermitian form on $V \iff Bilin(V,\bar{V}) \iff V \otimes \bar{V}$. Consequently, choice of any hermitian form induces an isomorphism
   $$V^* \simeq \bar{V}\,.$$
   - Given a representation ρ on V, we can define a representation $\bar{\rho}$ on $\bar{V}$: $\bar{\rho}_g(v) = \rho_g(v)$ (i.e. by definition $\bar{\rho}_{\mathbb{R}} = \rho_{\mathbb{R}}$; the definition makes sense because $Hom_{\mathbb{C}}(V,V) = Hom_{\mathbb{C}}(\bar{V},\bar{V})$).
   - ρ and $\bar{\rho}$ not necessarily isomorphic: e.g. rep. of $\mathbb{Z}_4$ on $\mathbb{C}^1$ with i as a generator.
   - Some useful identities:
   $$\rho \simeq \rho_1 \oplus \rho_2 \implies \bar{\rho} \simeq \bar{\rho}_1 \oplus \bar{\rho}_2 \tag{4.6}$$
   $$\rho_{\mathbb{R}} \otimes \mathbb{C} \simeq \rho \oplus \bar{\rho}\,. \tag{4.7}$$
   - $\rho \simeq \rho_{real} \otimes \mathbb{C} \iff \rho \simeq \bar{\rho}$ with the isomorphism given by a map satisfying (4.5) $\iff$ there exists a basis of V s.t. all matrices of ρ are real.
   - If $\rho \simeq \bar{\rho}$, then $\rho_{\mathbb{R}}$ cannot be irreducible. Indeed, isomorphism $H : V \to \bar{V}$ cannot be identity and clearly it satisfies
   $$H\rho_g = \rho_g H \quad \text{as elements of } End_{\mathbb{R}}(V)\,.$$
   At the same time we cannot claim that H will satisfy (4.5). The only thing we can guarantee is that in any $\mathbb{R}$-basis obtained from $\mathbb{C}$-basis we have:
   $$H \leftrightarrow \begin{pmatrix} A & B \\ B & -A \end{pmatrix}$$
   TODO: explicit counter-example of $\rho \simeq \bar{\rho}$ and H not satisfying (4.5).

4.2 Application to noncoherent fast fading channel: SISO

Noncoherent fading channel acts on complex input letters by multiplication and addition, independently on a per-letter basis, as follows:

$$Y = HX + Z\,,$$

where $H \sim \mathcal{N}_c(0,1)$ and $Z \sim \mathcal{N}_c(0,1)$ (independent of each other).

Consequently, for blocklength n we have channel acting between

$$\mathcal{F}_n \stackrel{\triangle}{=} \{x^n : ||x^n||^2 = nP\} \subset \mathbb{C}^n\,,$$

and $\mathbb{C}^n$ via

$$Y^n = \mathrm{diag}(H_1, \ldots, H_n)X^n + Z^n\,,$$

where H_j are iid[1] and $Z^n \sim \mathcal{N}_c(0, I_n)$.

To compute the automorphism group we first address $n = 1$ case. Here we have automorphism group given by $U(1) \times U(1)$. An element $(e^{i\theta_1}, e^{i\theta_2})$ acts as follows:

$$X \to Xe^{i\theta_1} \tag{4.8}$$
$$Y \to Ye^{i\theta_2}\,. \tag{4.9}$$

Thus, we can cancel redundant degrees of freedom by first taking the quotient of the input space by the subgroup $U(1) \times \{1\}$, which reduces the input space from $\mathbb{C}$ to $\mathbb{R}_+$ and corresponds to a map

$$x \to |x|\,.$$

Similarly, taking the quotient of the output space by the subgroup $\{1\} \times U(1)$ we obtain the equivalent model for the (single-letter) channel:

$$Y = |HX + Z|\,,$$

where now $X, Y \in \mathbb{R}^+$. Moreover, this is equivalent to the following channel

$$Y = |\sqrt{G}X + Z|\,, \tag{4.10}$$

where now G has χ^2 distribution (with two degrees of freedom).

Applying this operation to the blocklength n channel we obtain an equivalent channel which acts between

$$\mathcal{F}'_n \stackrel{\triangle}{=} \{x^n : \sum_{j=1}^n x_j^2 = nP\} \subset \mathbb{R}_+^n\,,$$

[1] This is the meaning of "fast" in the title of the section.

and $\mathbb{R}_+^n$ independently on each of n components and according to (4.10). This channel clearly has permutation symmetries and it can be shown (**TODO**) that these are the only symmetries, thus for the converted channel we have

$$\mathrm{Aut} = S_n\,,$$

whereas for the original we have $(U(1) \times U(1))^n \rtimes S_n$.

Important conclusion: For this channel the only source of symmetry is permutation and thus we (knowing that capacity achieving distribution is unique and discrete) propose to analyse the following codes:

- If for SNR P capacity achieving distribution is given by

$$P_X^* = \sum_{i=1}^{L} w_i \delta_{a_i}\,,$$

  where $a_i > 0$. Then construct (for large n) the vector

$$\mathbf{x}_0 = [\underbrace{a_1, \ldots, a_1}_{\approx n w_1}, \ldots, \underbrace{a_L, \ldots, a_L}_{\approx n w_L}]\,.$$

- Take a subgroup $G \leq S_n$, then the codebook is

$$\{g\mathbf{x}_0\,, g \in G\,.\}$$

References

[1] Y. Polyanskiy, "Channel coding: non-asymptotic fundamental limits," Ph.D. dissertation, Princeton University, Princeton, NJ, USA, 2010.

[2] G. D. Forney , Jr, "Geometrically uniform codes," *IEEE Trans. Inf. Theory*, vol. 37, no. 5, pp. 1241-1260, Sep. 1991.

[3] D. Slepian, "Group codes for the Gaussian channel", BSTJ, vol. 47, pp. 575-602, Apr. 1968.

[4] I. Ingemarsson, "Commutative group codes for the Gaussian channel", *IEEE Trans. Inform. Theory*, vol. 19, pp. 215-219, Mar. 1973.

[5] V. M. Sidelnikov, "Spherical 7-design in the 2^n-dimensional Euclidean space", ISIT-98, Cambridge, MA, USA, 1998.

[6] A. YA. Dorofeev, L.S. Kazarin, V.M. Sidelnikov and M.E. Tuzhilin, "Matrix Groups Related to the Quaternion Group and Spherical Orbit Codes"

[7] R. L. Dobrushin, "Asymptotic optimality of group and systematic codes for some channels," *Theor. Probab. Appl.*, vol. 8, pp. 47-59, 1963.

[8] G. Como, F. Fagnani, "The capacity of finite abelian group codes over symmetric memoryless channels" *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2037-2054, May 2009

[9] M. A. Karmasin, "Solution of a problem of Shannon," *Probl. Kibern.*, vol. 11, pp. 263-266, 1964 (Section III).

[10] E. C. Johnsen, "Stochastic matrices. I. Similarity via stochastic transforms," *Lin. Alg. Appl.*, vol. 29, pp.185-193, 1980.

[11] M.-K. Siu, "Which Latin Squares are Cayley Tables?" *American Math. Monthly*, vol. 98, no. 7, pp. 625-627, Aug-Sep 1991.

[12] Y. Polyanskiy and S. Verdú, "Arimoto channel coding converse and Rényi divergence," *48th Allerton Conference 2010,* Allerton Retreat Center, Monticello, IL, USA, Sep. 2010.