entropy

MDPI

*Article*

# Mutual Information and Multi-Agent Systems

**Ira S. Moskowitz [1,*], Pi Rogers [2] and Stephen Russell [3]**

[1] Naval Research Laboratory, Code 5580, Washington, DC 20375, USA
[2] 2022 SEAP Summer Intern at the Naval Research Laboratory, Washington, DC 20375, USA
[3] Jackson Health System, 1500 NW, 12th Ave, Miami, FL 33136, USA
* Correspondence: ira.moskowitz@nrl.navy.mil

**Abstract:** We consider the use of Shannon information theory, and its various entropic terms to aid in reaching optimal decisions that should be made in a multi-agent/Team scenario. The methods that we use are to model how various agents interact, including power allocation. Our metric for agents passing information are classical Shannon channel capacity. Our results are the mathematical theorems showing how combining agents influences the channel capacity.

**Keywords:** multi-agent system; mutual information; channel capacity; information geometry



**Citation:** Moskowitz, I.S.; Rogers, P.; Russell, S. Mutual Information and Multi-Agent Systems. *Entropy* **2022**, *24*, 1719. https://doi.org/10.3390/e24121719

Academic Editor: Éloi Bossé

Received: 21 October 2022
Accepted: 19 November 2022
Published: 24 November 2022



## 1. Introduction

Advances in machine intelligence have led to an increase in human-agent teaming. In this context, one or more machines act as semi-autonomous or autonomous agents interacting with other machine teammates and/or their human proxies. This phenomenon has led to cooperative work models where the role of an agent can be, interchangeably, a human, or machine, support system. Human counterparts that interact with automation become less like operators, supervisors, or monitors, and more like equal-authority peers.

Critical to the success of any team is efficient and effective communication. Multi-agent systems are no different. Information sharing is a key element in building collective cognition, and it enables agents to cooperate and ultimately achieve shared goals successfully. Information sharing, or communication, provides the foundation for a team's success. In complex multi-agent engagements, information is not always universally available to all agents. Such engagements are often characterized by distributed entities with limited communication channels among them, where no agent has a complete view of the solution space, and information relevant to team goals only becomes available to team members in spontaneous, unpredictable and even unanticipated ways. Moreover, there is always a resource cost to inter-agent communication. Finding highly efficient and effective communication patterns is a recurring problem in any multi-agent system, particularly if the system agents are distributed.

We are concerned with how a Multi-agent System (MAS) [1], or Team, sends information between agents or teammates. By "how" we mean "how" in an information theoretic [2] sense—in particular, we do not concentrate on the mechanics or physics of the transmission other than how it impacts information theory. We are concerned with what strategy an agent can to use to maximize its information flow to another agent. From an information geometric standpoint, we only use a simple metric in this article, but lay the ground work for more complex Riemannian metrics. We are concerned with a transmitting agent sending a small amount of distinct symbols in a fixed time. In fact, we restrict ourselves to two symbols to develop our theory (A list of notation is at the end of the article.). We are using a mathematical approach to model the communication between two agents. The equations we present are based on a series of assumptions that we will explain.

We assume that an agent sends two symbols to another agent. We refer to the symbols as "0" or "1". We are concerned with the fidelity of how the symbols are passed. All

symbols take the same time to pass. We will be looking at the (Shannon) capacity as one agent attempts to send a symbol to another agent.

Our scenario is illustrated in Figures 1 and 2. The first agent $A_X$ sends a 0 or 1 to the second agent $A_Y$. We have a clock and the unit of time is $t$. Every $t$, $A_X$ transmits the symbol to $A_Y$. We assume that the symbol is received within the same time unit (i.e., we assume instantaneous transmission speeds during each interval $t$). There is no feedback (which, for the channels we analyze, would not change the capacity anyway (p. 520 [3])) from $A_Y$ to $A_X$, and the transmission is considered to be memoryless (quoting [4], "…channel is memoryless if the probability distribution of the output depends only on the input at that time and is conditionally independent of previous channel inputs or outputs"). Furthermore, it is implicit that the channel statistics never change (sometime the literature refers to this as a "stationary" condition).

To summarize the above, we have a Discrete Memoryless Channel (DMC) between $A_X$ and $A_Y$. This channel measures information flow in terms of bits per symbol (since $t$ does not vary). We let $X$ represent the input distribution to this DMC, and we let $Y$ denote the output random variable.

The probability for the random variable $X$ is given by $P(X = i), i = 0, 1$; it is the probability that $A_X$ inputs symbol $i$, and $P(Y = j), j = 0, 1$ is the probability that $A_Y$ received symbol $j$. The input distribution $X$ is determined by the transmission fidelity of $A_X$. In particular,

$$x = P(X = 0) = x, \ \bar{x} := P(X = 1) = 1 - x \ . \tag{1}$$

Whereas the output distribution $Y$ is determined by the (assumed to be well-defined) conditional distribution between $X$ and $Y$, and the input distribution. Thus,

$$P(Y = j) = \sum_i P(Y = j | X = i) \cdot P(X = i) \ . \tag{2}$$

The approach presented in this paper follows from [2,5–7].

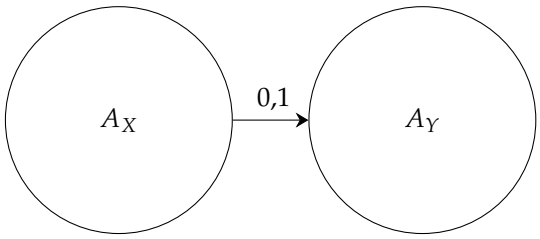


**Figure 1.** Heuristic figure of $A_X$ transmitting a bit to $A_Y$.

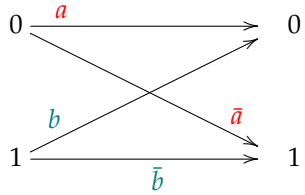


**Figure 2.** The noisy channel diagram corresponding to the first figure.

The conditional probabilities of the DMC is given by a $2 \times 2$ matrix $M_1$, where (Please keep in mind the swapping of the indices, and, as we had above for $\bar{x}$, that notationally $\bar{*} := 1 - *$ . Furthermore, the convention is that a conditional probability is fixed for all $P(X = i)$, even if that probability is 0. In the next footnote, we address the impact of this with respect to (w.r.t.) information theory).

$$m_{i,j} := P(Y = j | X = i) \text{ and} \tag{3}$$

$$M_1 = \begin{pmatrix} m_{0,0} & m_{0,1} \\ m_{1,0} & m_{1,1} \end{pmatrix} = \begin{pmatrix} P(Y=0|X=0) & P(Y=1|X=0) \\ P(Y=0|X=1) & P(Y=1|X=1) \end{pmatrix} =: \begin{pmatrix} a & \bar{a} \\ b & \bar{b} \end{pmatrix} . \tag{4}$$

Note that $(a, b) \in [0,1] \times [0,1]$.

Before we continue with the mathematics let us put this research into some more perspective. Von Neumann's [8] seminal work had no concept of "Teamwork", which is at the core of what we are discussing. Sliwa's [9] review suggests that minimum communication channels are more important when context is understood during teamwork, a suggestion opposite to our work in this article which we hope to test in the future. Lawless [10] suggests that maximized channels become more important when Teams confront uncertainty in their environment. Schölkopf et al. [11] suggest that i.i.d. data are insufficient to reconstruct whatever social event is being captured, that something is missing and a new approach must be innovated, our goal in this article. Our results will be discussed in situ for maximum effect.

### *1.1. Entropy and Mutual Information*

We extend our random variables to allow more than two possible outcomes, and give the following definitions with the most generality possible. We now have $I+1$ possible inputs, and $J+1$ possible outcomes.

Given a discrete random variable $V$, we define the entropy of $V$ as (By convention log is the base 2 logarithm, and ln is the natural logarithm. Furthermore, we are able to extend the definitions (p. 19 [4]), as is standard, so that $0 \log(0) = 0 \log(0/0) = 0$. These conventions allows the most general derivation of (8) from (7)).

$$H(V) := -\sum_j P(V = v_j) \log P(V = v_j) .$$

If $z \in [0,1]$, then we define the binary entropy function of $z$ as

$$h(z) := -z \log(z) - (1-z) \log(1-z) .$$

Note that if $B$ is a binary random variable taking the values 0 or 1, then $H(B) = h(P(B=0))$. In fact, we simplify the notation and express the probability of the event $\{V = v_k\}$ as

$$p_v(v_k) = P(V = v_k) .$$

Furthermore, when it is clear which distribution we are using, we further simplify the notation and just write $p(v_k)$. Thus,

$$H(B) = h(p(0)) .$$

Given two discrete random variables $V, W$, we define [2] the conditional entropy of $V$ given $W$ as

$$H(V|W) := -\sum_i p_w(w_i) \sum_j p_{v|w}(v_j|w_i) \log p_{v|w}(v_j|w_i) , \tag{5}$$

where, as in the $2 \times 2$ case

$$P(v_j|w_i) := m_{i,j}, i = 0, 1, \ldots, I; j = 0, 1, \ldots, J,$$

forming the channel matrix (Of course, as in the $2 \times 2$ case, conditional probability is only defined when $p(w_i) \neq 0$. However, as we note below, such terms are dealt with by using the limiting value of the constant conditional probability term which makes our mutual information calculations consistent, keeping in mind that $0 \log *$ is always taken to be 0. Furthermore, keep in mind that a distribution that achieves capacity for a 2-input channel (the subject of this paper) never has either probability value as zero of course (Ref. [12]

gives better bounds). There are, however, $3 \times 2$ channels for which this does not hold, for example $\begin{pmatrix} 1 & 0 \\ 0.8 & 0.2 \\ 0 & 1 \end{pmatrix}$ which has an optimizing input distribution of $(0.5, 0, 0.5)$.)

$$M = \begin{pmatrix} p(v_0|w_0) & p(v_1|w_0) & \dots & p(v_J|w_0) \\ p(v_0|w_1) & p(v_1|w_1) & \dots & p(v_J|w_1) \\ \vdots & \vdots & \ddots & \vdots \\ p(v_0|w_I) & p(v_1|w_I) & \dots & p(v_J|w_I) \end{pmatrix} . \tag{6}$$

We define the mutual information between $V$ and $W$ by [2]

$$I(V,W) := H(V) - H(V|W) = H(W) - H(W|V) =: I(W,V) . \tag{7}$$

Using (5) and (7), and some substitutions [4] (again, division by 0 is taken care of in the usual way by using limiting values ([Section 2.3] [4]), we find that

$$I(V,W) = \sum_{j,i} p(v_j, w_i) \log \left( \frac{p(v_j, w_i)}{p(v_j)p(w_i)} \right) . \tag{8}$$

We now give Shannon's definition [2] of (channel) capacity. It has been well-studied since its inception. We will not delve into the Noisy Coding Theorem, or any of the other results which showcase its importance. Rather, we will assume in this paper that capacity is a standard measure of how much information a channel can transmit in an essentially noise-free manner [2,4]. The traditional units of capacity and mutual information are accepted in this article; they are bits per channel usage, which in our scenario is equivalent to bits per $t$.

**Definition 1.** *We consider $W$ to be the input random variable to a DMC. The capacity $C$ of the DMC is*

$$C := \sup_{\{p(w_i)\}} I(V,W) . \tag{9}$$

The optimization is taken over all possible distributions of $W$ with its fixed values $w_i$. The supremum is actually achieved and can be taken as a maximum [2,4]. Note that when trying to compare the magnitude of the channel capacity (with the same number of inputs), it suffices to compare the mutual information for all $x$ values. Of course the two channels may have different optimizing distributions. Note the principle (and similar principles) that if $\forall x, I(CH_1, x) \leq I(CH_2, x)$ and if $CH_1$ achieves capacity at $x'$, then $C(CH_1) = I(CH_1, x') \leq I(CH_2, x') \leq C(CH_2)$.

Of course swapping rows, or swapping columns from the channel matrix (6) is just notational and leaves capacity unchanged. However, we end this subsection with some interesting results in information theory—some obvious, some not so obvious.

**Property 1.** *Removing a row from the channel matrix (6) never increases the capacity.*

**Proof.** Not using a channel input cannot increase mutual information. This is equivalent to using input probability distributions which are always zero for a particular index; therefore, the capacity can never be greater since capacity is the maximum over all input distributions. □

**Property 2.**

*A—For any input probability, combining (by adding two columns to form one column hence reducing the channel matrix from $n \times m$ to $n \times m - 1$ as illustrated below with $Q, Q'$) two columns of a channel matrix will never increase mutual information.*

*B—For input probabilities with all terms non-zero, the mutual information will stay the same iff one of the combined columns is a multiple of the other. Otherwise, the uncombined channel has a larger mutual information and hence a larger capacity. (Note, that for a 2-input channel [12] has shown that the capacity achieving distribution has both probabilities in the interval* $[\frac{1}{e}, 1 - \frac{1}{e}]$ *so we can apply this property to the capacity directly.)*

**Proof.** A:

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a+b & c \\ d+e & f \\ g+h & i \end{pmatrix}$$

The Data-Processing Inequality (Cascade of Channels) [3] shows that the capacity of the third channel above cannot be greater than that of the first channel. That is, processing one channel into another can never increase the information sent. The actual statement of the inequality is for mutual information. However, we use the probability that maximizes the mutual information of the first channel (which is its capacity), and therefore, it is less than or equal to the mutual information of the third channel which is less than or equal to the third channel's capacity. This argument holds for any initial channel matrix (with adjustments to the second matrix), not just the $3 \times 3$ matrix, or the columns we chose, for simplicity above.

B: Without loss of generality (WLOG), combine the first two columns of $n$ by $m$ channel matrix (note how the indices are reversed as compared to (6))

$$Q = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1m} \\ q_{21} & q_{22} & \cdots & q_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ q_{n1} & q_{n2} & \cdots & q_{nm} \end{pmatrix} \text{ (uncombined)}$$

to make

$$Q' = \begin{pmatrix} q_{11} + q_{12} & \cdots & q_{1m} \\ q_{21} + q_{22} & \cdots & q_{2m} \\ \vdots & \ddots & \vdots \\ q_{n1} + q_{n2} & \cdots & q_{nm} \end{pmatrix} \text{ (combined) .}$$

For $Q$, the output symbols are $y_j$, where $j$ goes from 1 to $m$. For $Q'$, they are the same, but with $y_1$ and $y_2$ replaced by $y_1 \cup y_2$. For both channels, the input symbols are $x_i$, with input probability vector $p$ defined as

$$p_i := p(x_i).$$

Therefore,

$$p(y_1) = \sum_{i=1}^{n} p_i q_{i1}$$

and

$$p(y_2) = \sum_{i=1}^{n} p_i q_{i2} \, .$$

If either of these last two relations are 0, WLOG we assume $p(y_1) = 0$. This assumption means column 1 of $Q$ must be a 0 column (since the input probabilities are positive), so it contributes 0 to the mutual information. Therefore, the mutual informations are equal, and one column is a constant multiple of the other. Now that we have dealt with this case, we can assume $y_1$ and $y_2$ are positive for the remainder of this proof. For fixed $p$, the mutual information of an $n$ by $m$ channel is

$$I = \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i) p(y_j|x_i) \log \frac{p(y_j|x_i)}{p(y_j)} \, .$$

Columns 3 through $m$ of $Q$ and 2 through $m-1$ of $Q'$ are the same, so their contributions to mutual information are the same. Therefore, we only need to consider columns 1 and 2 of $Q$ and column 1 of $Q'$. Let $\bar{I}$ be these columns' mutual information, that is,

$$\bar{I}(Q) = \sum_{i=1}^{n} p_i q_{i1} \log \frac{q_{i1}}{p(y_1)} + \sum_{i=1}^{n} p_i q_{i2} \log \frac{q_{i2}}{p(y_2)} \text{ , and}$$

$$\bar{I}(Q') = \sum_{i=1}^{n} p_i \left[ (q_{i1} + q_{i2}) \log \frac{q_{i1} + q_{i2}}{p(y_1) + p(y_2)} \right] \text{ , since } p(y_1 \cup y_2) = p(y_1) + p(y_2), \text{etc.}$$

Note that $\bar{I}(Q)$ can also be written as

$$\bar{I}(Q) = \sum_{i=1}^{n} p_i \left[ q_{i1} \log \frac{q_{i1}}{p(y_1)} + q_{i2} \log \frac{q_{i2}}{p(y_2)} \right].$$

The log sum inequality [4] states that, for a series of non-negative numbers $a_k$ and $b_k$ with sums $a$ and $b$, respectively, where $k$ goes from 1 to $K$, then

$$\sum_{i=1}^{K} a_i \log \frac{a_k}{b_k} \leq a \log \frac{a}{b},$$

with equality iff $\frac{a_k}{b_k}$ are equal for all $i$. By applying this inequality to the above terms in square braces, we have that $\bar{I}(Q') \leq \bar{I}(Q)$, with equality iff $\frac{q_{1i}}{p(y_1)} = \frac{q_{2i}}{p(y_2)}$ for all $i$. Since $p(y_1)$ and $p(y_2)$ are nonzero and independent of $i$, this is true iff column 1 of $Q$ is a constant multiple $p(y_1)/p(y_2)$ of column 2. In fact, this also shows that $p(y_1)$ is a constant multiple of $p(y_2)$, regardless of the all of the positive input probabilities. □

### *1.2. Back to Our Binary-Input Binary-Output DMC, the (2,2) Channel*

Restating (1) and following the approach of [13] :

$$\begin{aligned} &x = P(X=0),\ \bar{x} = P(X=1) \text{ and we define} \\ &y := P(Y=0),\ \text{thus } \bar{y} = P(Y=1)\ . \end{aligned} \tag{10}$$

The above expressions simplify for our DMC under investigation. Using (1) and (2), we have that the distribution of $Y$ is

$$(y, \bar{y}) = (x, \bar{x}) \begin{pmatrix} a & \bar{a} \\ b & \bar{b} \end{pmatrix} = \Big( (a-b)x + b,\ 1 - \big[ (a-b)x + b \big] \Big)$$

We now define a differentiable function $f(x), x \in [0,1]$ by

$$f(x) := (a-b)x + b = ax + b\bar{x}\ , \tag{11}$$

which gives us

$$(y, \bar{y}) = (f(x), \overline{f(x)})\ . \tag{12}$$

Thus,

$$H(Y) = h(y) = h(f(x))\ . \tag{13}$$

From (5), we have that

$$\begin{aligned} H(Y|X) &= -\Big\{ x \big[ a \log a + \bar{a} \log \bar{a} \big] + \bar{x} \big[ b \log b + \bar{b} \log \bar{b} \big] \Big\} \\ &= x \cdot h(a) + \bar{x} \cdot h(b)\ . \end{aligned} \tag{14}$$

Putting the above together gives us

$$I(Y,X) = h(f(x)) - x \cdot h(a) - \bar{x} \cdot h(b) \ . \tag{15}$$

Using (9), we have that the capacity of the *(2,2) channel* is

$$C_{2,2} = \max_x I(Y,X) = \max_x \left[ h(f(x)) - x \cdot h(a) - \bar{x} \cdot h(b) \right] \ . \tag{16}$$

So, for the *(2,2) channel*, the capacity calculation boils down to a (not so simple) calculus problem. Silverman [14] was the first to express the closed form result (see also [5,13] and ([Equation (5)] [7]) for derivations and alternate expressions).

$$C_{2,2}(a,b) = \log\left( 2^{\frac{\bar{a}\cdot h(b) - \bar{b}\cdot h(a)}{a-b}} + 2^{\frac{b\cdot h(a) - a\cdot h(b)}{a-b}} \right) \ , \text{ where } C(a,a) := 0 \ , \tag{17}$$

which is a continuous function on the unit square $[0,1] \times [0,1]$. It is trivial to show that capacity is continuous on the unit square without the main diagonal $a = b$. However, to prove continuity on the entire unit square requires some work and uses the fact that (15) is continuous in $a$, $b$, and $x$ see ([Section 2.4] [15]).

One can easily show that (see Figure 3)

$$C_{2,2}(a,b) = C_{2,2}(b,a) = C_{2,2}(\bar{a},\bar{b}) \ , \tag{18}$$

by simple algebraic substitution. Additionally, this tells us that $C_{2,2}(a,b) = C_{2,2}(\bar{b},\bar{a})$ also.

$C_{2,2}(a,b) = C_{2,2}(b,a)$ is equivalent to capacity being symmetric across the line $b = a$, and $C_{2,2}(a,b) = C_{2,2}(\bar{b},\bar{a})$ is equivalent to capacity being symmetric when across the line $b = -a + 1$ (simple geometry proves this). This result is illustrated in Figure 3. Thus, capacity has a quadrant of the unit square as its principal domain (see ([Figure 1] [14])).

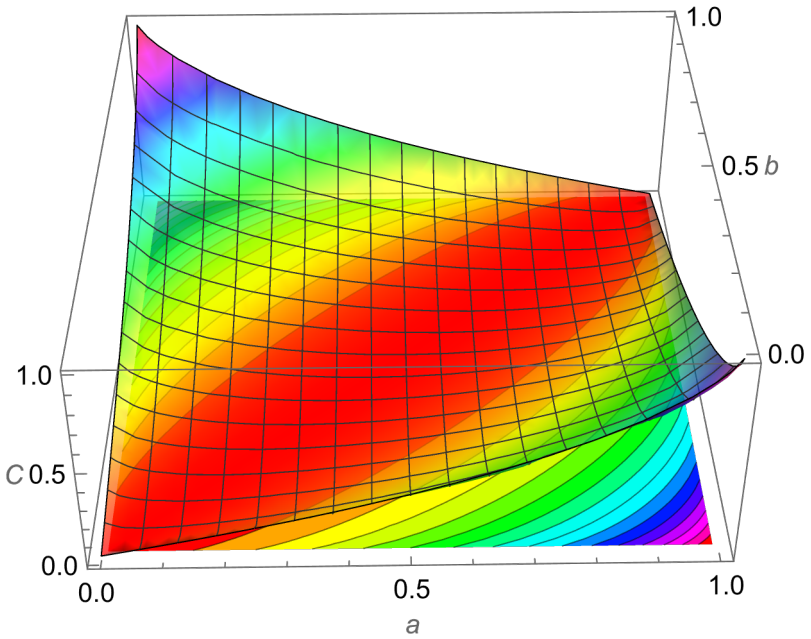


**Figure 3.** Plot of $C_{2,2}(a,b)$ along with its level set contours. This figure shows the symmetries (18) about the lines $y = x$ and $y = -x + 1$ as seen by how the countours can be folded onto each other across the two lines. $C$ is the capacity.

### *1.3. Power/Fidelity Constraints of $C_{2,2}$*

We consider the situation where we attempt to increase the capacity by adjusting the terms $a$ and $b$. Ideas like this for a Team's interdependence, with a different measurement

and no mention of information theory, were discussed in [1]. However, the values of $a, b$ are a function of the transmitting environment from $A_X$ to $A_Y$. If the agents were all-powerful, that could simply adjust $a$ to be 1, and $b$ to be 0 (or visa versa) to achieve a channel of maximal capacity $C_{2,2} = 1$.

1.3.1. Positive Channels

Let us start by considering positive channels [6], that is $a > b$. Note if $a < b$, we have a negative channel, and if $a = b$, we have a 0-capacity channel. Of course, no matter what $C(a, b), \geq 0$. However, if we are at a point (a,b), is it better to increase $a$, decrease $b$, or some combination thereof? Implicit in this question is that we stay in the domain of positive channels (under the line $b = a$).

**Definition 2.** *We say that we have a power constraint $P$ when we are at the channel given by (a,b) and the most we can adjust the channel is to $(a', b')$ where the standard Euclidean distance (its $l^2$ norm) between $(a, b)$ and $(a', b')$ is no more that $P$.*

In terms of Information Geometry [1], our distance is obtained from the Riemannian metric

$$ds^2 = da^2 + db^2 \,. \tag{19}$$

Of course we can generalize this to a more general metric of the form

$$ds^2 = E\, da^2 + F\, da \cdot db + G\, db^2 \,, \tag{20}$$

which would put us in a non-Euclidean situation. This non-trivial situation may be necessary if $a$ and $b$ relate differently to various transmission characteristics.

It is shown in ([Theorem 4.9] [6]) that if we restrict ourselves to positive channels, that the capacity increases as $a$ increases, and decreases as $b$ increases. This result makes physical sense in terms of adding or decreasing noise. Now consider the (closed) disk of radius $r$ about the point $(a, b)$, denoted as $D_r(a, b)$. We assume that $r$ is small enough so that $D_r(a, b)$ is composed only of positive channels.

**Example 1.** *We illustrate this situation in Figure 4 by the channels that are in the disk or radius 0.15 about the point (0.6,0.2).*

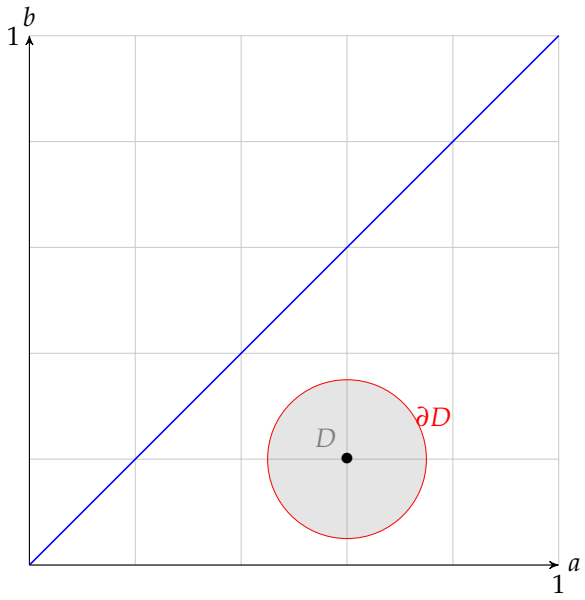


**Figure 4.** Closed disk $D$ of radius 0.15, about the point (0.6,0.2), that consists only of positive channels. The boundary of the disk is the circle $\partial D$.

**Theorem 1.** *Given a closed disk $D_r(a,b)$ consisting of positive channels, the maximum capacity is achieved and occurs on the boundary circle $\partial D_r(a,b)$.*

**Proof.** Since $C_{2,2}(a,b)$ is a continuous function on the compact set $D_r(a,b)$ , it is has a maximum denoted as $C_M$. Assume that the maximum is achieved at an interior point $(a',b') \in D_r(a,b)$. By ([Theorem 4.9] [6]) we know that increasing $a'$ increases capacity, which contradicts $C_M$ being achieved at the interior point $(a',b')$. □

We note that the above theorem still holds for non-positive channels by a simple adjustment of the proof.

Example 1 is illustrated in Figure 4 and is examined again in Figures 5 and 6, where we can see the level sets of $C_{2,2}$ and the surface plot of capacity. Furthermore, numerical calculations show that the maximum of capacity for the closed disk is obtained at the boundary points $(0.68, 0.07)$ and has a value of 0.32 .

Of course, as the center of the disk and the radius vary, so does the relative position of the point on the circle that capacity is achieved at. What is interesting is that it is not obvious where this point should be. We will explain this further. For a positive channel, increasing $a$ brings increased capacity, whereas decreasing $b$ results in increased capacity. So, considering our example of the disk centered at $(0.6, 0.2)$ with radius 0.15, one might think that this critical point is when $b$ is decreased by the amount that $a$ is increased—this being the point on the boundary circle at $2\pi - \frac{\pi}{4} = 5.50$ radians, which only gives us a capacity of 0.31 . However, numerical methods tell us that the actual maximum occurs are 5.25 radians with a value, as noted, of 0.32. Of course, for this example, the difference is not much, but this result is relative to the size of the disk. What is important is that the actual critical point depends on the disk's position to the two lines $b = a$ and $b = 1 - a$. We do note that when the disk is centered on the line $b = 1 - a$, that $2\pi - \frac{\pi}{4}$ radians is the correct position for the critical point. One can also see this by examining the capacity level sets in Figure 5.

Of course, we are using an $l^2$ metric which has a metric ball of a disk. If, for example, we used an $l^1$ metric, the ball would be a square rotated by 45 degrees.

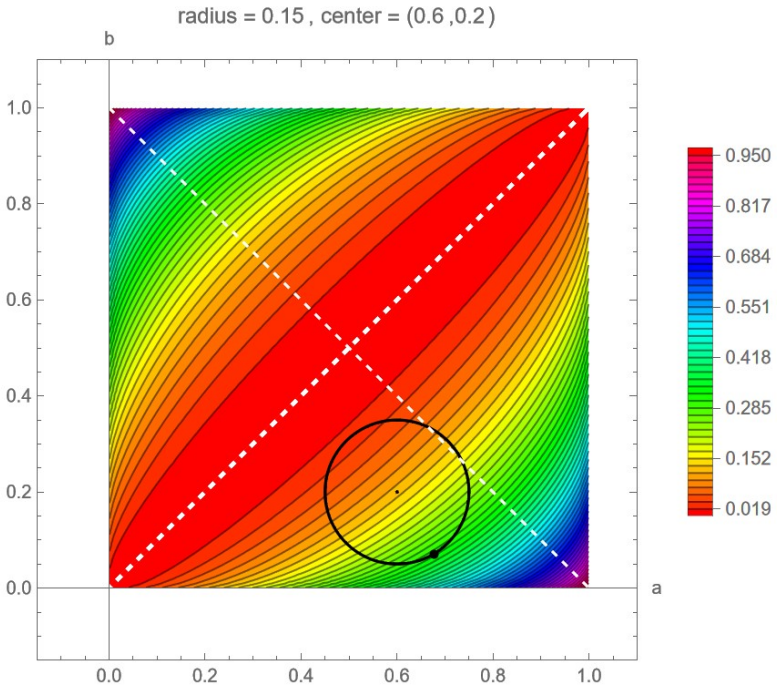


**Figure 5.** Example 1 illustrated with level sets of capacity with more detail than Figure 4.

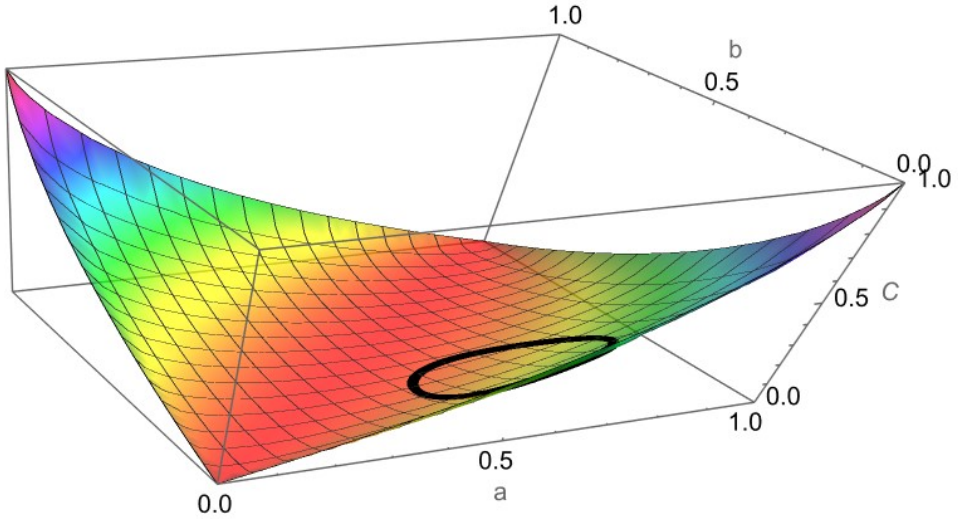


**Figure 6.** Same as Figure 5, but with a 3D perspective.

#### 1.3.2. Power

We assume that the transmitting agent $A_X$ has adjustable power $P$. This power allows the transmission capabilities of $A_X$ to vary. By way of example, say that $A_X$ transmits with fidelity $a = 0.6, b = 0.2$. Now, $A_X$ is given an increase in its transmitting power that allows it to change $(a, b)$ to $(a', b')$ such that the "distance" between the two points is less than $P$. Consider that we use the $L_2$ Euclidean norm and set $P = 0.15$. This tells us that all such points $(a', b')$ are in the disk of radius 0.15 about the center $(0.6, 0.2)$. We note that this is a rudimentary concept of power. Power helps a transmission when we are restricted to the bottom quarter of this disk and where $a$ is increasing (giving more transmission fidelity) and $b$ is decreasing (more transmission fidelity—recall that $b$ gives us the probability of a 1 going to the opposite symbol 0). However, the conclusion is still the same point that we made and illustrated above.

#### 1.3.3. Results and Discussion

We end this section with a brief summary. We have discussed how one agent can pass Shannon information to another and how changing the transmission characteristics can increase or decrease this information transfer. We have used capacity as our metric for information transfer. Let us now progress to multiple agents. We have also proven some information theoretic properties for the reader (Properties 1 & 2).

In the situation that we discussed in this section where there are two transmitting agents and one receiving agent, we denote the channel as $M_1$, which is given by the channel matrix (In this article, we freely identify a channel with its matrix. Furthermore, for a $2 \times 2$ channel, we identify the channel as the ordered 2-tuple $(a, b)$ also.) $M_1$ described earlier (4). We denote that channel capacity as $C(M_1)$ which we have analyzed as $C_{2,2}$ in this section.

### 2. Two Transmitting Agents

Say we have two transmitting agents, $A_{X_1}$ and $A_{X_2}$ acting independently with respect to each other. Assume they have the same transmitting characteristics; that is, the channel matrices are the same. The receiving agent $A_Y$ gets symbols from both transmitting agents. How does this impact the information flow to $A_Y$?

In our scenario, $A_{X_1}$ and $A_{X_2}$ both sense the same environment. That is, they both wish to send a 0 or they both wish to send a 1. So, as before, the possible inputs are 0 or 1, but the outputs are of the form

$$(0,0), (0,1), (1,0), (1,1) \tag{21}$$

since we are assuming that the noise affects each transmitting agent independently. Keep in mind that both $A_{X_1}$ and $A_{X_2}$ are both attempting to transmit the same symbol.

The output that $A_Y$ uses is given by the random variable $Y$.

$$\begin{aligned}
&(0,0) \text{ is taken to be the symbol } Y = O_{0,0}\\
&(0,1) \text{ is taken to be the symbol } Y = O_{0,1}\\
&(1,0) \text{ is taken to be the symbol } Y = O_{1,0}\\
&(1,1) \text{ is taken to be the symbol } Y = O_{1,1}\ .
\end{aligned}$$

We denote $P(Y = O_{i,j}) =: y_{i,j}$. Our channel matrix is $2 \times 4$ and is

$$\begin{aligned}
M_2 &= \begin{pmatrix} P(Y = O_{0,0}|X = 0) & P(Y = O_{0,1}|X = 0) & P(Y = O_{1,0}|X = 0) & P(Y = O_{1,1}|X = 0)\\ P(Y = O_{0,0}|X = 1) & P(Y = O_{0,1}|X = 1) & P(Y = O_{1,0}|X = 1) & P(Y = O_{1,1}|X = 1)\end{pmatrix}\\
&= \begin{pmatrix} a^2 & a\bar{a} & \bar{a}a & \bar{a}^2\\ b^2 & b\bar{b} & \bar{b}b & \bar{b}^2\end{pmatrix}.
\end{aligned}$$

We note that the second and third columns of the above channel matrix are identical. This has implications for the mutual information and, of course, the capacity of the channel.

Let us look at this in more generality. Say we have two channel matrices

$$M^3 = \begin{pmatrix} \alpha & 2\epsilon & \delta\\ \beta & 2\gamma & \phi\end{pmatrix} \quad \text{and} \quad M^4 = \begin{pmatrix} \alpha & \epsilon & \epsilon & \delta\\ \beta & \gamma & \gamma & \phi\end{pmatrix}.$$

Both channels have the same input random variable $X$ as above. The output random variables are $Y^3$ and $Y^4$, respectively.

Let us consider the $M^3$ channel first. $Y^3$ has probability values $y_i := P(Y^3 = i)$ as follows

$$\begin{aligned}
(y_1, y_2, y_3) &= (\alpha x + \beta\bar{x}, 2\epsilon x + 2\gamma\bar{x}, \delta x + \phi\bar{x})\ .\text{So,}\quad H(Y^3)\\
&= -[(\alpha x + \beta\bar{x})\log(\alpha x + \beta\bar{x}) + (2\epsilon x + 2\gamma\bar{x})\log(2\epsilon x + 2\gamma\bar{x}) + (\delta x + \phi\bar{x})\log(\delta x + \phi\bar{x})]\ ,\\
H(Y^3|X) &= -x\Big[\alpha\log(\alpha) + 2\epsilon\log(2\epsilon) + \delta\log(\delta)\Big]\\
&\quad - \bar{x}\Big[\beta\log(\beta) + 2\gamma\log(2\gamma) + \phi\log(\phi)\Big]\ . \qquad (22)
\end{aligned}$$

The mutual information is $I(Y, X) = H(Y) - H(Y|X)$. We expand the mutual information into the sum of two functions. The first function is from the first and last columns, and the second function is from the middle column. That is

$$I(Y^3, X) = F_1^3(\alpha, \beta, \delta, \phi, x) + F_2^3(\epsilon, \gamma, x)\ ,\ \text{where}$$

$$\begin{aligned}
F_2^3 &= -2\epsilon x\log(2\epsilon x + 2\gamma\bar{x}) - 2\gamma\bar{x}\log(2\epsilon x + 2\gamma\bar{x}) + 2\epsilon x\log(2\epsilon) + 2\gamma\bar{x}\log(2\gamma)\\
&= 2\epsilon x\log\left(\frac{2\epsilon}{2\epsilon x + 2\gamma\bar{x}}\right) + 2\gamma\bar{x}\log\left(\frac{2\gamma}{2\epsilon x + 2\gamma\bar{x}}\right)\\
&= 2\epsilon x\log\left(\frac{\epsilon}{\epsilon x + \gamma\bar{x}}\right) + 2\gamma\bar{x}\log\left(\frac{\gamma}{\epsilon x + \gamma\bar{x}}\right).
\end{aligned}$$

Now let us consider the $M^4$ channel. As above

$$(y_1, y_2, y_3, y_4) = (\alpha x + \beta\bar{x}, \epsilon x + \gamma\bar{x}, \epsilon x + \gamma\bar{x}, \delta x + \phi\bar{x})\ .$$

$$H(Y^4) = -\Big[(\alpha x + \beta\bar{x})\log(\alpha x + \beta\bar{x}) + (\epsilon x + \gamma\bar{x})\log(\epsilon x + \gamma\bar{x}) + (\epsilon x + \gamma\bar{x})\log(\epsilon x + \gamma\bar{x}) + (\delta x + \phi\bar{x})\log(\delta x + \phi\bar{x})\Big] \quad (23)$$

$$= -\Big[(\alpha x + \beta\bar{x})\log(\alpha x + \beta\bar{x}) + 2(\epsilon x + \gamma\bar{x})\log(\epsilon x + \gamma\bar{x}) + (\delta x + \phi\bar{x})\log(\delta x + \phi\bar{x})\Big] . \quad (24)$$

$$H(Y^4|X) = -x\Big[\alpha\log(\alpha) + \epsilon\log(\epsilon) + \epsilon\log(\epsilon) + \delta\log(\delta)\Big] - \bar{x}\Big[\beta\log(\beta) + \gamma\log(\gamma) + \gamma\log(\gamma) + \phi\log(\phi)\Big] \quad (25)$$

$$= -x\Big[\alpha\log(\alpha) + 2\epsilon\log(\epsilon) + \delta\log(\delta)\Big] - \bar{x}\Big[\beta\log(\beta) + 2\gamma\log(\gamma) + \phi\log(\phi)\Big] . \quad (26)$$

As above we expressthe mutual information as

$$I(Y^3, X) = F_1^3(\alpha, \beta, \delta, \phi, x) + F_2^3(\epsilon, \gamma, x)$$

and we have that

$$\begin{aligned} F_2^4 &= -2\epsilon x\log(\epsilon x + \gamma\bar{x}) - 2\gamma\bar{x}\log(\epsilon x + \gamma\bar{x}) + 2\epsilon x\log(\epsilon) + 2\gamma\bar{x}\log(\gamma) \\ &= 2\epsilon x\log\left(\frac{\epsilon}{\epsilon x + \gamma\bar{x}}\right) + 2\gamma\bar{x}\log\left(\frac{\gamma}{\epsilon x + \gamma\bar{x}}\right) = F_2^3 . \end{aligned}$$

A quick inspection tells us that $F_1^4 = F_1^3$; thus, the mutual information of both channels is the same. This result is not surprising because if we combine output symbols where the channel matrix has identical rows, we lose nothing as far as the output information is concerned—there is no extra value in looking at the output symbols separately. This makes sense, and is also what our mathematics have shown.

Let us keep in mind that we wish to find $C(M_2)$, the capacity of the Shannon channel when there are two transmitting agents. (To keep our notation consistent, $C(a, b)$ is the capacity given by the corresponding $2 \times 2$ channel matrix as in (4), whereas $C(*)$ is the capacity of the channel given by $*$).

**Theorem 2.** $C(M_2) \geq C(M_1)$.

**Proof.** $M_2$ has four output symbols which are in essence 2-vectors. We ignore the second component of the vector. Therefore, we collapse the first and third symbol to $a$, and the second and fourth to $\bar{a}$. This results in $M_1$, and since using more output symbols never lowers capacity, by Property 2 (also, a code that works for $M_1$ works for $M_2$ as well by collapsing the symbols), we are done. (Later in the paper we do better than this result with Corollary 1 to Theorem 6.) □

We now form another channel related to what we discussed above. Say now that the receiving agent receives the symbols without any order. Therefore, instead of a 2-vector, the output is one of the three multisets $[0, 0], [1, 0], [1, 1]$ with

$$P(Y = [0, 0]) = a^2, P(Y = [1, 0]) = 2\bar{a}a, P(Y = [1, 1]) = \bar{a}^2 .$$

We call this channel $M_{2-}$, and its channel matrix is

$$M_{2-} = \begin{pmatrix} a^2 & 2\bar{a}a & \bar{a}^2 \\ b^2 & 2\bar{b}b & \bar{b}^2 \end{pmatrix} .$$

From what we discussed above with $M^4$ and $M^3$, we see that

**Theorem 3.**

$$C(M_{2-}) = C(M_2)\ .$$

Let us examine the bounds in Theorem 1 above. We will see that, not surprisingly except for special cases, $C(M_2) > C(M_1)$. Figure 7 is a plot of $C(M_2) - C(M_1)$ as a function of $(a, b)$.

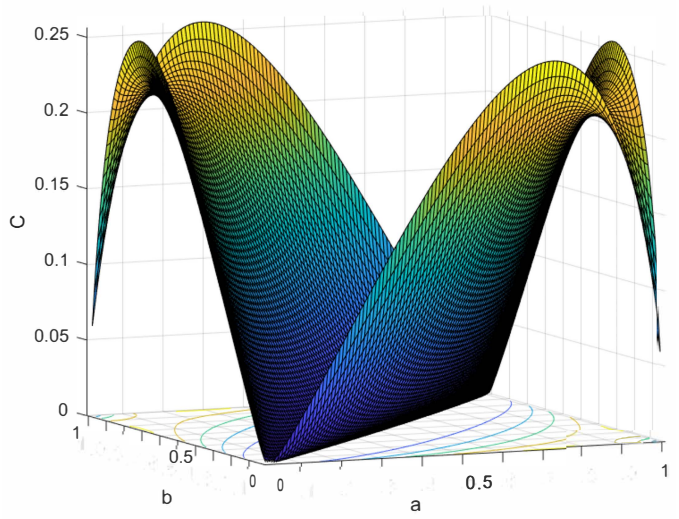


**Figure 7.** The plot $C(M_2) - C(M_1)$, of course the $C$ axis is now measuring the difference in the capacities (in units of bits per $t$).

From Figure 7, we see that except for the line $b = a$ (where both channels $M_1$ and $M_2$ have 0 capacity), and at $(a, b) = (1, 0)$ or $(a, b) = (0, 1)$ (where both channels have capacity 1), that $C(M_2) > C(M_1)$. We note that for $M_2$ and the other higher dimensional channels that we will discuss, there is to our knowledge no closed form as there is for $M_1$. Therefore, for our calculations of capacity, we rely upon numerical results from the Blahut-Arimoto capacity algorithm [16,17].

*Results and Discussion*

In this section, we have laid the groundwork for $n$ transmitting agents. We derived some capacity results. We concentrated on the effects of going from 1 to 2 transmitting agents. What happens as we go to three or more transmitting agents?

## 3. Multiple Transmitting Agents

We have the canonical representation for the channel of $n$ transmitting agents, and we denote this canonical channel matrix as $M_{\underline{n}}$, which is formed by taking the output of channel $M_{\underline{n-1}}$ (Note, due to the simplicity of the construction for "small" channels, we have that $M_{\underline{1}} = M_1, M_{\underline{2}} = M_2$.) and adding a 0 or a 1 to it. For $M_{\underline{3}}$ this results in

$$M_{\underline{3}} = \begin{pmatrix} a^3 & a^2\bar{a} & a^2\bar{a} & a\bar{a}^2 & a^2\bar{a} & a\bar{a}^2 & a\bar{a}^2 & \bar{a}^3 \\ b^3 & b^2\bar{b} & b^2\bar{b} & b\bar{b}^2 & b^2\bar{b} & b\bar{b}^2 & b\bar{b}^2 & \bar{b}^3 \end{pmatrix}.$$

This comes from taking the output for two agents as given in canonical form by (21) and extending it to

$$(0,0,0), (0,0,1), (0,1,0), (0,1,1), (1,0,0), (1,0,1), (1,1,0), (1,1,1)\ .$$

**Theorem 4.** *Rearranging outputs/columns of a channel matrix does not affect capacity.*

**Proof.** By looking at the expression for mutual information, we see that changing the order of arithmetic operations leaves it unchanged. This result follows, since capacity is the maximum of mutual information. □

Therefore, we can permute the columns of $M_{\underline{n}}$ and obtain a new matrix $M_n$, which has the same capacity, that is $C(M_n) = C(M_{\underline{n}})$, and is given below.

$$M_n = \begin{pmatrix} a^n & a^{n-1}\bar{a} & \dots & a^{n-1}\bar{a} & a^{n-2}\bar{a}^2 & \dots & a^{n-1}\bar{a} & \dots & \bar{a}^n \\ b^n & b^{n-1}\bar{b} & \dots & b^{n-1}\bar{b} & b^{n-2}\bar{b}^2 & \dots & b^{n-1}\bar{b} & \dots & \bar{b}^n \end{pmatrix}. \tag{27}$$

Look at the above theorem in terms of the columns of $M_n$. Let us use $M_3$ as an example.

$$M_3 = \begin{pmatrix} a^3 & a^2\bar{a} & a^2\bar{a} & a^2\bar{a} & a\bar{a}^2 & a\bar{a}^2 & a\bar{a}^2 & a^3 \\ b^3 & b^2\bar{b} & b^2\bar{a} & b^2\bar{b} & b\bar{b}^2 & b\bar{b}^2 & b\bar{b}^2 & b^3 \end{pmatrix}. \tag{28}$$

Collapsing the output in this situation is equivalent to interchanging the 4th and 5th columns (which does not change capacity) and forming the matrix $M_{3c}$.

$$M_{3c} = \begin{pmatrix} a^3 & a^2\bar{a} & a^2\bar{a} & a\bar{a}^2 & a^2\bar{a} & a\bar{a}^2 & a\bar{a}^2 & a^3 \\ b^3 & b^2\bar{b} & b^2\bar{a} & b\bar{b}^2 & b^2\bar{b} & b\bar{b}^2 & b\bar{b}^2 & b^3 \end{pmatrix}. \tag{29}$$

As above when we looked at $M^3$ and $M_4$, we see that we may form the channel where we identify output symbols with the same conditional probabilities for both inputs. This give us the channel $M_n-$ , where

$$M_{n-} = \begin{pmatrix} a^n & na^{n-1}\bar{a} & \binom{n}{2}a^{n-2}\bar{a}^2 & \dots & na\bar{a}^{n-1} & \bar{a}^n \\ b^n & nb^{n-1}\bar{b} & \binom{n}{2}b^{n-2}\bar{b}^2 & \dots & nb\bar{b}^{n-1} & \bar{b}^n \end{pmatrix}. \tag{30}$$

**Theorem 5.** $C(M_n) = C(M_{n-})$

**Proof.** As above for $M_2$ in Theorem 3, or we can just use Property 2 repeatedly. □

The reason we introduce $M_{n-}$ is that it is a cleaner way to express the channel, and the calculations are simpler than that of $M_n$. For example, $M_8$ is a $2 \times 256$ matrix, whereas $M_{8-}$ is a $2 \times 9$ matrix. This obviously makes the coding issues easier. Now we examine Figure 8, which is the difference between $C(M_8)$ and $C(M_1)$.

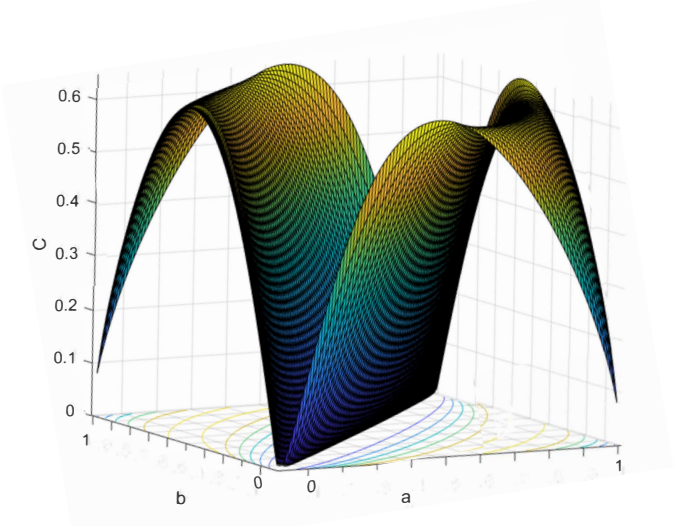


**Figure 8.** $C(M_8) - C(M_1)$.

When we compare Figure 8 to Figure 7, we easily see that $C(M_n)$ grows, except for the endpoints and the line $b = a$ (which stay at 0) as n grows.

*Nota Bene* We now look at the prior illustrative results in terms of a more general encompassing theory. We included much of Section 2 so that the reader who is not familiar with some of the "tricks" will have a feel for why the more general results hold.

**Theorem 6.** *$C(M_{n+1}) \geq C(M_n)$ for any positive integer $n$.*

**Proof.** (The proof is the same as for the above when $n = 1$.) $M_n$ can be obtained from $M_{n+1}$ by combining certain columns together; the result follows from Property 2. □

**Corollary 1.** *$C(M_{n+1}) > C(M_n)$, except for $(1,0)$ and $(0,1)$ where they both have capacity 1, and the line $b = a$ where they both have capacity 0.*

**Proof.** We show the proof in three steps.

1. If $a = b$, $C(M_n) = C(M_{n+1}) = 0$ since the rows are identical. In this case, it is trivial to show that $H(Y) = H(Y|X)$ (the output has no idea what the channel input was). One can see this by the fact that $x \cdot a^q \bar{a}^{n-q} + \bar{x} \cdot a^q \bar{a}^{n-q} = a^q \bar{a}^{n-q}$. In short, the capacities are equal.
2. If $(a,b) = (1,0)$ or $(a,b) = (0,1)$, both $M_n$ and $M_{n+1}$ are both the $2 \times 2$ identity matrix with zero columns added in; hence, $C(M_n) = C(M_{n-1}) = 1$. In short, the channel capacities are equal.
3. Now, excluding the special cases where $a = b$, $(a,b) = (1,0)$, or $(a,b) = (0,1)$, by Property 2, we only have to show that here are two combined columns that are not multiples of each other.

By excluding the special cases, we cannot use the endpoints of the unit square; therefore, $a$ or $b$ must be in $(0,1)$. WLOG, we *assume* that $0 < a < 1$.

Consider a generic column of $M_n$; it is of the form $c = \begin{pmatrix} a^e \bar{a}^{n-e} \\ b^e \bar{b}^{n-e} \end{pmatrix}, e \in \{0, \ldots, n\}$. By construction, $M_{n+1}$ has two columns, $c_1 = \begin{pmatrix} a \cdot a^e \bar{a}^{n-e} \\ b \cdot b^e \bar{b}^{n-e} \end{pmatrix}$ and $c_2 = \begin{pmatrix} \bar{a} \cdot a^e \bar{a}^{n-e} \\ \bar{b} \cdot b^e \bar{b}^{n-e} \end{pmatrix}$, that when combined result in column $c$. If $c_1$ is not a constant multiple of $c_2$, we will have shown that $C(M_{n+a}) > C(M_n)$. Assume the opposite—that is, $c_1 = k \cdot c_2$; since neither $a$ or $\bar{a}$ is 0 we have that $a = k\bar{a}$. Then $a = k\bar{a}$ is equivalent to $a = \frac{k}{k+1}, k \neq 0$. We now have three cases for b.

- $b = 0$. In this case, $\bar{b} = 1$ and we only look at the last column of $M_n$, so we let $c = \begin{pmatrix} \bar{a}^n \\ \bar{b}^n \end{pmatrix} = \begin{pmatrix} \bar{a}^n \\ 1 \end{pmatrix}$. Since we are assuming that $c_1 = k \cdot c_2$, we have that $0 = 0 \cdot 1 = b \cdot 1 = k \cdot \bar{b} \cdot 1 = k$, which is impossible.
- $b = 1$. Using the same argument as above, just replace the last column of $M_n$ with the first. So again, it is impossible that the columns are multiples.
- $0 < b < 1$. As above for $a$, we also have that $b = \frac{k}{k+1}$. This tells us that $a = b$ which has been ruled out.

Thus, we have shown the existence of two columns of $M_{n+1}$ that are not multiples of each other and combine them into a column of $M_n$. □

**Theorem 7.** *$\lim_{n\to\infty} C(M_n) = 1$, except for when $b = a$, and in that case, the channel capacity is 0.*

**Proof.** WLOG, we assume $a > b$. We can do this because of the constraint $a \neq b$ and the fact that the rows of a channel matrix can be interchanged without affecting its capacity. Take a positive $\varepsilon << \frac{a-b}{2}$ be fixed. For a large enough $N$, we can always find a rational number $m(n)$ for any $n > N$ such that $\bar{a} + \varepsilon < m < \bar{b} - \varepsilon < 1$ and $nm \in \mathbb{Z}$. (The $\varepsilon$ padding prevents $m$ from converging to $\bar{a}$ or $\bar{b}$). This result is guaranteed to exist for sufficiently large $N$.

Given $0 \leq b < a \leq 1$, let $x = \bar{a} + \varepsilon, y = \bar{b} - \varepsilon$, giving us $0 \leq x < y \leq 1$. Certainly there exists a positive integer $N$ such that $1/N < y - x$. Therefore, for any integer $n \geq N$, we have that $1/n < y - x$. Consider $(x, y)$ as a sub-interval of $[0, 1]$. For any $n \geq N$, consider the largest integer $W$ such that $W(1/n) \leq x$. Look at $(W + 1)(1/n)$; by the definition of $W$, this must be greater than $x$. However, since $1/n < y - x$, we have that $(W + 1)(1/n) < y$. We let $m = (W + 1)(1/n)$. Keep in mind two characteristics of $m$ as a function of $n$:

1. Since $W$ is an integer, $mn \in \mathbb{Z}$, and,
2. $mn < n$, since $m < 1$.

Let $M'_n$ be the channel matrix $M_{n-}$, but modified as follows: all outputs $y_k$ for $k \leq mn$ are combined into $y'_0$, and all of the other outputs are combined into $y'_1$. The channel matrix then looks like this:

$$M'_n = \begin{pmatrix} P(y'_0|x_0) & P(y'_1|x_0) \\ P(y'_0|x_1) & P(y'_1|x_1) \end{pmatrix},$$

where

$$(Y = y'_0) = (Y = y_0) \cup (Y = y_1) \cup \ldots \cup (Y = y_{mn}) \subsetneq (Y = y_o) \cup \ldots \cup (Y = y_n) \text{ and}$$

$$P(y'_0|x_0) = \sum_{i=0}^{mn} P(y_i|x_0) \text{ , with } P(y_i|x_0) = \binom{n}{i} a^{n-i} \bar{a}^i \text{ .}$$

(Keep in mind that we are dealing with the binomial random variable $S_n$, where $i$ is the number of successes in $n$ Bernoulli trials, with the probability of success $\bar{a}$, $P(S_n = i) = \binom{n}{i} a^{n-i} \bar{a}^i$).

$$\therefore \; P(y'_0|x_0) = \sum_{i=0}^{mn} \binom{n}{i} a^{n-i} \bar{a}^i \text{ .}$$

If we let $\Phi(x)$ be the cumulative standard normal distribution function, the De-Moivre Laplace limit theorem [18] states that (when we take $c, d$ as integers)

$$P\left(c < \frac{S_n - n\bar{a}}{\sqrt{na\bar{a}}} < d\right) \to \Phi(d) - \Phi(c) \text{ as } n \to \infty; \text{ thus,}$$

$$P\left(\frac{c - \bar{a}}{\sqrt{na\bar{a}}} < \frac{S_n - n\bar{a}}{\sqrt{na\bar{a}}} < \frac{d - \bar{a}}{\sqrt{na\bar{a}}}\right) \to \Phi\left(\frac{d - \bar{a}}{\sqrt{na\bar{a}}}\right) - \Phi\left(\frac{c - \bar{a}}{\sqrt{na\bar{a}}}\right) \text{ as } n \to \infty, \text{ and}$$

$$P(c \leq S_n \leq d) \to \Phi\left(\frac{d - \bar{a}}{\sqrt{na\bar{a}}}\right) - \Phi\left(\frac{c - \bar{a}}{\sqrt{na\bar{a}}}\right) \text{ as } n \to \infty \text{ .}$$

This step leaves us with

$$\sum_{i=c}^{d} \binom{n}{i} a^{n-i} \bar{a}^i \to \Phi\left(\frac{d - n\bar{a}}{\sqrt{na\bar{a}}}\right) - \Phi\left(\frac{c - n\bar{a}}{\sqrt{na\bar{a}}}\right) \text{ as } n \to \infty \text{ .} \tag{31}$$

Thus, the De-Moivre Laplace limit theorem gives us (with $c = 0, d = mn$):

$$\begin{aligned} \lim_{n\to\infty} P(y'_0|x_0) &= \lim_{n\to\infty} \left[\Phi\left(\frac{mn - n\bar{a}}{\sqrt{na\bar{a}}}\right) - \Phi\left(\frac{-n\bar{a}}{\sqrt{na\bar{a}}}\right)\right] \\ &= \lim_{n\to\infty} \Phi\left(\sqrt{n}\frac{m - \bar{a}}{\sqrt{a\bar{a}}}\right) - \lim_{n\to\infty} \Phi\left(\sqrt{n}\frac{-\bar{a}}{\sqrt{a\bar{a}}}\right) \text{ .} \end{aligned}$$

Since $a$ and $\bar{a}$ are positive, then $-\frac{\bar{a}}{\sqrt{a\bar{a}}}$ is negative, giving

$$\lim_{n\to\infty} \sqrt{n}\frac{-\bar{a}}{\sqrt{a\bar{a}}} = -\infty, \text{ and}$$

$$\lim_{n\to\infty} \Phi\left(\sqrt{n}\frac{-\bar{a}}{\sqrt{a\bar{a}}}\right) = 0 \text{ .}$$

If $m < \bar{a}$, then $\frac{m-\bar{a}}{\sqrt{a\bar{a}}}$ is negative. However, if $m > \bar{a}$, it is positive, giving (Even though $m$ changes as $n$ changes, the value of $\sqrt{n}\frac{m-\bar{a}}{\sqrt{a\bar{a}}}$ remains greater than or equal to $\sqrt{n}\frac{\varepsilon}{\sqrt{a\bar{a}}}$ for $m > \bar{a} + \varepsilon$. Since $\sqrt{n}\frac{\varepsilon}{\sqrt{a\bar{a}}}$ approaches $\infty$, so does $\sqrt{n}\frac{m-\bar{a}}{\sqrt{a\bar{a}}}$. The same logic can also be used for the $m < \bar{a} - \varepsilon$ case.)

$$\lim_{n\to\infty} \sqrt{n}\frac{m-\bar{a}}{\sqrt{a\bar{a}}} = \begin{cases} -\infty & \text{if } m < \bar{a} - \varepsilon \\ \infty & \text{if } m > \bar{a} + \varepsilon \text{ ; and} \end{cases}$$

$$\lim_{n\to\infty} \Phi\left(\sqrt{n}\frac{m-\bar{a}}{\sqrt{a\bar{a}}}\right) = \begin{cases} 0 & \text{if } m < \bar{a} - \varepsilon \\ 1 & \text{if } m > \bar{a} + \varepsilon \end{cases}$$

$$\therefore \lim_{n\to\infty} P(y_0'|x_0) = 1 - 0 = 1 \,.$$

Thus, we have that

$$\lim_{n\to\infty} P(y_1'|x_0) = 0 \,.$$

$P(y_0'|x_1)$ beahves the same, but with $a$ replaced by $b$. Since $\bar{a} + \varepsilon < m < \bar{b} - \varepsilon$, then the $\lim_{n\to\infty} P(y_0'|x_0) = 1$ and $\lim_{n\to\infty} P(y_0'|x_1) = 0$; thus,

$$\lim_{n\to\infty} M_n' = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

which has a channel capacity of 1. Since $M_n'$ was formed by combining the outputs of $M_n$, then $C(M_n') \leq C(M_n) \leq 1$. Therefore, by the squeeze theorem, $\lim_{n\to\infty} C(M_n) = 1$. □

*Results and Discussion*

The theorems presented in this section shows what happens as the number of transmitters grows. The ultimate result of this section was Theorem 7, which used a rather non-trivial application of the Central Limit Theorem. At this point, the seemingly obvious but difficult result that we proved, i.e., that as the number of transmitting agents grows, so does the reliability of the channel in terms of its capacity. This result, of course, is in line with the similar result that if we have a code that consisted of repeating a symbol many times the error rate is small (the transmission rate may be low, but this does not apply to our agent examples).

## 4. Non-Identical Transmitting Agents

In a shift, say we start with only two transmitting agents, but their noise characteristics are different. Of course, keep in mind that in this situation, we have assumed that there is a master transmitter using the $X$ agent to communicate with $Y$. The master transmitter picks the input symbols and the transmitting agents do their best to communicate by forming one encompassing Shannon channel. We have shown above that, if all of the agents share the same assumption for $(a, b)$, the channel capacity increases as the number of agents increase. However, what happens if the $(a, b)$ are different for the various agents? Are we better off only using a subset of agents, or is it still best to use as many agents as possible? We partially answer those questions below.

Let $M_1^1$ be the channel matrix for agent 1, and $M_1^2$ be the channel matrix for agent 2.

$$M_1^1 = \begin{pmatrix} a & \bar{a} \\ b & \bar{b} \end{pmatrix},$$

$$M_1^2 = \begin{pmatrix} c & \bar{c} \\ d & \bar{d} \end{pmatrix}.$$

The output is such that the receiving agent uses the ordering of agent 1 first, then agent 2. If the agents wish to send a signal of 0, the possible outputs, expressed via their probabilities, are

$$\begin{aligned} P(0,0) &= ac \\ P(0,1) &= a\bar{c} \\ P(1,0) &= \bar{a}c \\ P(1,1) &= \bar{a}\bar{c} \end{aligned}$$

If the agents wish to send a signal of 1 instead, we have

$$\begin{aligned} P(0,0) &= bd \\ P(0,1) &= b\bar{d} \\ P(1,0) &= \bar{b}d \\ P(1,1) &= \bar{b}\bar{d}\,. \end{aligned}$$

This gives us a combined channel matrix for both agents who are transmitting as $M_2^{1,2}$, where

$$M_2^{1,2} = \begin{pmatrix} ac & a\bar{c} & \bar{a}c & \bar{a}\bar{c} \\ bd & b\bar{d} & \bar{b}d & \bar{b}\bar{d} \end{pmatrix}. \tag{32}$$

We use our own notation to express the above channel as the tensor product ,

$$(a,b) \otimes (c,d)\,.$$

We know, by Property 2, that collapsing output symbols does not increase capacity. However, if we collapse $y_1$ and $y_2$ into $y_{1'}$ and $y_3$ and $y_4$ into $y_{2'}$, we have a channel matrix of $M_{2'}^{1,2}$:

$$M_{2'}^{1,2} = \begin{pmatrix} ac + a\bar{c} & \bar{a}c + \bar{a}\bar{c} \\ bd + b\bar{d} & \bar{b}d + \bar{b}\bar{d} \end{pmatrix} = \begin{pmatrix} a & \bar{a} \\ b & \bar{b} \end{pmatrix}.$$

Thus, $C(M_2^{1,2}) \geq C(M_{2'}^{1,2}) = C(M_1^1)$.

Now let us combine the first and third outputs of $M_2^{1,2}$ into $y_{1''}$ and the second and fourth outputs into $y_{2''}$. This gives us a channel matrix $M_{2''}^{1,2}$.

$$M_{2''}^{1,2} = \begin{pmatrix} ac + \bar{a}c & a\bar{c} + \bar{a}\bar{c} \\ bd + \bar{b}d & b\bar{d} + \bar{b}\bar{d} \end{pmatrix} = \begin{pmatrix} c & \bar{c} \\ d & \bar{d} \end{pmatrix}.$$

Thus, $C(M_2^{1,2}) \geq C(M_{2''}^{1,2}) = C(M_1^2)$. This result leads us to the next theorem:

**Theorem 8.** *As the number of agents increase, no matter if they have different channel noises, the total channel capacity is non-decreasing.*

**Proof.** In the above discussion we have show that

$$\begin{aligned} C(M_2^{1,2}) &\geq C(M_{2'}^{1,2}) = C(M_1^1) \\ C(M_2^{1,2}) &\geq C(M_{2''}^{1,2}) = C(M_1^2). \end{aligned}$$

Therefore, by repeating the same argument we see that as we add extra agents the capacity can never decrease. □

In fact, as before when the agents had identical characteristics, the channel capacity, except for special cases (dependent columns, a capacity 0 or 1, etc.), is greater than that for separate agents. One can see this by examining the channel matrix—if you unpack the

outputs and find that the statistics are different, extra information is learned. Let us now look at the special case of combining a channel with a 0-channel.

**Theorem 9.** *For any zero channel given by* $(e,e), e \in [0,1]$*, we find that*

$$C((a,b) \otimes (e,e)) = C(a,b) \ .$$

**Proof.** If we can show that the mutual information of $(a,b) \otimes (e,e)$ is given by (15), we are done. The channel matrix for this situation is

$$\begin{pmatrix} ae & a\bar{e} & \bar{a}e & \bar{a}\bar{e} \\ be & b\bar{e} & \bar{b}e & \bar{b}\bar{e} \end{pmatrix} .$$

Let $u := ax + b\bar{x}$, and we find that $\bar{u} = \bar{a}x + \bar{b}\bar{x}$. Further,

$$\begin{aligned} Y = (y_1, y_2, y_3, y_4) &= (aex + be\bar{x}, a\bar{e}x + b\bar{e}\bar{x}, \bar{a}ex + \bar{b}e\bar{x}, \bar{a}\bar{e}x + \bar{b}\bar{e}\bar{x}) \\ &= (ue, u\bar{e}, \bar{u}e, \bar{u}\bar{e}) \ , \end{aligned}$$

$$\begin{aligned} H(Y) &= -[ue\log(ue) + u\bar{e}\log(u\bar{e}) + \bar{u}e\log(\bar{u}e) + \bar{u}\bar{e}\log(\bar{u}\bar{e})] \\ &= -[ue(\log(u) + \log(e)) + u\bar{e}(\log(u) + \log(\bar{e})) \\ &\quad + \bar{u}e(\log(\bar{u}) + \log(e)) + \bar{u}\bar{e}(\log(\bar{u}) + \log(\bar{e}))] \\ &= -[u\log(u) + \bar{u}\log\bar{u} + e\log(e) + \bar{e}\log(\bar{e})] \\ &= h(u) + h(e) \ , \text{ and} \end{aligned}$$

$$\begin{aligned} H(Y|X) &= -x[ae\log(ae) + a\bar{e}\log(a\bar{e}) + \bar{a}e\log(\bar{a}e) + \bar{a}\bar{e}\log(\bar{a}\bar{e})] \\ &\quad -\bar{x}\left[be\log(be) + b\bar{e}\log(b\bar{e}) + \bar{b}e\log(\bar{b}e) + \bar{b}\bar{e}\log(\bar{b}\bar{e})\right] . \end{aligned}$$

Now again using the log of a product as the sum of the logs, then grouping like log terms, this results in

$$H(Y|X) = x[h(a) + h(e)] + \bar{x}[h(b) + h(e)] = x \cdot h(a) + \bar{x} \cdot (b) + h(e) \ ,$$

and we see that $H(Y) - H(Y|X) = h(ax + b\bar{x}) - x \cdot h(a) - \bar{x} \cdot h(b)$ . □

*Results and Discussion*

In this section, we showed what happens when two transmitting agents with different noise characteristics are used. Our important result was that as the number of agents increase, no matter if they have different channel noises, the total channel capacity is non-decreasing. As with many of our results it relied upon the algebra of mutual information giving common sense answers. However, without proofs we just have intuition to rely upon.

## 5. Resource Allocation

We now concern ourselves with the physical limitations of the receiving agent. We assume that the receiving agent has a limited resource $\mathcal{R}$ that it can use to receive messages. To the extent possible, the receiving resource, $\mathcal{R}$, may be measured in terms of various antennas or various allocations of frequencies, etc. It is not our goal in this article to discuss the engineering of the receiving agent in general. Rather, we accept it as a given.

Upon completion of the mathematics in this section, the results do not seem surprising. That is good! It shows that our intuition is correct and it lays a foundation for dealing with many agents and non-linear allocation schemes (where we lose elements of intuition). Furthermore, aside from linearity, we based our allocation scheme on a Euclidean metric; it is not at all clear if an information geometric-style Riemannian metric be used instead. That is beyond the scope of the article.

Let us take the simplest case where there are two transmitting agent $A_{X_1}$ and $A_{X_2}$. As before, $A_{X_i}$ has channel matrix $M_i$. We model noise affecting each channel in a linear manner. Suppose that an agent $A_X$ is given, as before, by its channel matrix

$$M_1 = \begin{pmatrix} a & \bar{a} \\ b & \bar{b} \end{pmatrix}.$$

How does noise, which results from the receiving agent not allocating enough of its resources to $A_X$, change this channel matrix? The channel $(a, b)$ is a point in $[0, 1] \times [0, 1]$. Consider the shortest path from $(a, b)$ to the main diagonal (which consists of zero-capacity channels). View $[0, 1] \times [0, 1]$ as sitting $\mathbb{R}^2$ and consider the straight line $y = -x + (a + b)$. This line is orthogonal to the straight diagonal line of zero-capacity channels, goes through the point $(a, b)$, and intersects the line for the zero-capacity channels at $(\frac{a+b}{2}, \frac{a+b}{2})$. The line segment of interest is given parametrically for $t \in [0, 1]$ as

$$(1 - t)\Big(a, b\Big) + t\Big(\frac{a + b}{2}, \frac{a + b}{2}\Big).$$

We model noise as moving on this new line segment from the point $(a, b)$ to the point $(\frac{a+b}{2}, \frac{a+b}{2})$. No noise corresponds to $t = 0$, total noise to $t = 1$; that is, we use $t$ as a measure of the noise normalized in a linear manner between 0 and 1.

EXAMPLE : Let $(a, b) = (0.8, 0.4)$. If $t = 0$, the channel is given as $(0.8, 0.4)$ and the capacity is 0.12 . If $t = 1$, the channel is given as $(0.6, 0.6)$ and the capacity is 0 . Let $t = 0.9$, then the channel is given by $0.1(0.8, 0.4) + 0.9(0.6, 0.6) = (0.08, 0.04) + (0.54, 0.54) = (0.62, 0.58)$, which has a capacity of 0.001 .

Now, let $t = 0.1$, then the channel is given by $0.9(0.8, 0.4) + 0.1(0.6, 0.6) = (0.72, 0.36) + (0.06, 0.06) = (0.78, 0.42)$, which has a capacity of .10 . Note that, unsurprisingly, the cleaner channel has $C(0.8, 0.4) = 0.1246 > C(0.78, 0.42)$.

What we have been discussing motivates the following our modeling definition.

**Definition 3.** *An agent $A_X$ with channel matrix $(a, b)$ requires the receiving resource $\mathcal{R}$ for its channel matrix to be unchanged. If the receiving agent only allocates $\mathcal{A}, 0 \leq \mathcal{A} \leq R$ to $A_x$, the channel matrix is modified from $(a, b)$ in the following manner,*

$$(a^{\mathcal{A}}, b^{\mathcal{A}}) = \frac{\mathcal{A}}{\mathcal{R}}(a, b) + \left(1 - \frac{\mathcal{A}}{\mathcal{R}}\right)\left(\frac{a + b}{2}, \frac{a + b}{2}\right). \tag{33}$$

Thus, $\mathcal{A} = R$ corresponds to $t = 0$ above, and $\mathcal{A} = 0$ corresponds to $t = 1$ above. As $\mathcal{A}$ decreases, the capacity "travels" the shortest path in the Euclidean metric to the line of the 0-capacity channels. This is the essence of our modeling assumption.

Note that a channel is a 0-capacity channel iff $a = b$. However, if we let $b = a$, then $\forall \mathcal{A}, (a^{\mathcal{A}}, a^{\mathcal{A}}) = (a, a)$.

**Theorem 10.** *For a non-zero channel $(a, b)$, that is, $a \neq b$, $C(a^{\mathcal{A}}, b^{\mathcal{A}})$ decreases as $\mathcal{A}$ decreases from $\mathcal{R}$ to 0.*

**Proof.** If $(a, b)$ is a positive channel, that is, if $a > b$, we have that $a^{\mathcal{A}}$ decreases and $b^{\mathcal{A}}$ increases as $\mathcal{A}$ goes from $\mathcal{R}$ to 0. This result is easily shown with algebra, but even more simply by observation of the line segment. From ([Theorem 4.9] [6]), if $(a, b)$ is a negative channel, then by symmetry of capacity about the line $b = a$, that completes the proof. □

**Corollary 2.** *If we have a 0-capacity channel $(a, b) = (e, e)$, then the $C(e^{\mathcal{A}}, e^{\mathcal{A}})$ is constant at 0 as $\mathcal{A}$ decreases.*

**Proof.** Trivial, since the line segment reduces to the point $(e, e)$ is this situation. □

*5.1. Resource Allocation Amongst Different Transmitters*

Assume that there are two transmitting agents $A_{X_1}$ with matrix $(a, b)$, and $A_{X_2}$ with matrix $(c, d)$. The difference from before is that the receiver can only allocate total resource $\mathcal{R}$ to the reception by the agents and, further, each agent requires resource $\mathcal{R}$ to prevent degradation to its channel matrix.

If $A_Y$ allocates $\mathcal{A}$ to $A_{X_1}$, we have the resulting channel matrix Equation (33) as given above. Then it allocates the remainder $\mathcal{R} - \mathcal{A}$ to $A_{X_2}$, resulting in this channel matrix

$$(c^{\mathcal{R}-\mathcal{A}}, d^{\mathcal{R}-\mathcal{A}}) = \left(1 - \frac{\mathcal{A}}{\mathcal{R}}\right)(c, d) + \frac{\mathcal{A}}{\mathcal{R}}\left(\frac{c+d}{2}, \frac{c+d}{2}\right). \tag{34}$$

Note that

$$\begin{aligned} (a^{\mathcal{R}}, b^{\mathcal{R}}) &= (a, b), \text{ with } C(a^{\mathcal{R}}, b^{\mathcal{R}}) = C(a, b), \text{ and} \\ (a^0, b^0) &= \left(\frac{a+b}{2}, \frac{a+b}{2}\right), \text{ with } C(a^0, b^0) = 0\,. \end{aligned}$$

As we have shown in the previous section, we arrive at:

$$M_2^{1,2}\,|_{\mathcal{A}} = \begin{pmatrix} a^{\mathcal{A}} \cdot c^{\mathcal{R}-\mathcal{A}} & a^{\mathcal{A}} \cdot \overline{c^{\mathcal{R}-\mathcal{A}}} & \overline{a^{\mathcal{A}}} \cdot c^{\mathcal{R}-\mathcal{A}} & \overline{a^{\mathcal{A}}} \cdot \overline{c^{\mathcal{R}-\mathcal{A}}} \\ b^{\mathcal{A}} \cdot d^{\mathcal{R}-\mathcal{A}} & b^{\mathcal{A}} \cdot \overline{d^{\mathcal{R}-\mathcal{A}}} & \overline{b^{\mathcal{A}}} \cdot d^{\mathcal{R}-\mathcal{A}} & \overline{b^{\mathcal{A}}} \cdot \overline{d^{\mathcal{R}-\mathcal{A}}} \end{pmatrix}. \tag{35}$$

Consider the situation when all of the resource is allocated to one channel; then, without the loss of generality, we let $\mathcal{A} = \mathcal{R}$, giving

$$M_2^{1,2}\,|_{\mathcal{A}=\mathcal{R}} = \begin{pmatrix} a\left(\frac{c+d}{2}\right) & a\left(1-\frac{c+d}{2}\right) & \bar{a}\left(\frac{c+d}{2}\right) & \bar{a}\left(1-\frac{c+d}{2}\right) \\ b\left(\frac{c+d}{2}\right) & b\left(1-\frac{c+d}{2}\right) & \bar{b}\left(\frac{c+d}{2}\right) & \bar{b}\left(1-\frac{c+d}{2}\right) \end{pmatrix}. \tag{36}$$

Keep in mind that the above result is the channel matrix when we combine a 0-capacity channel with $(a, b)$. Intuitively, this should not change the capacity from that of $C(a, b)$. Looking at the channel matrix and thinking in terms of coding, we see that we are affecting the first and second outputs; as much as the third and fourth. Below, we present the mathematical details.

**Theorem 11.** $C\left(M_2^{1,2}\,|_{\mathcal{A}=\mathcal{R}}\right) = C(a, b)$.

**Proof.** Let us calculate $C\left(M_2^{1,2}\,|_{\mathcal{A}}\right)$. We let $\frac{c+d}{2} := \gamma$ and $q := (ax + b\bar{x})$. Thus,

$$(y_1, y_2, y_3, y_4) = \left(\gamma(ax + b\bar{x}), \bar{\gamma}(ax + b\bar{x}), \gamma(\bar{a}x + \bar{b}\bar{x}), \bar{\gamma}(\bar{a}x + \bar{b}\bar{x})\right) \text{ . Then if}$$

$$(y_1, y_2, y_3, y_4) = (\gamma q, \bar{\gamma} q, \gamma \bar{q}, \bar{\gamma}\bar{q}) \text{ , we find that}$$

$$H(Y) = h(\gamma) + h(q)\,.$$

Next we examine the conditional entropy:

$$H(Y|X) = -x\left(a\gamma \log(a\gamma) + a\bar{\gamma}\log(a\bar{\gamma}) + \bar{a}\gamma\log(\bar{a}\gamma) + \bar{a}\bar{\gamma}\log(\bar{a}\bar{\gamma})\right).$$

Again use the rule that the log of a product is the sum of the logs to arrive at:

$$H(Y|X) = H(Y) - H(Y|X) = h(ax + b\bar{x}) - xh(a) - \bar{x}h(b)\,.$$

This result is the same as the mutual information of $(a, b)$. Thus, the maximum of the mutual information for both cases remains the same. □

**Corollary 3.** $C\left(M_2^{1,2}\mid_{\mathcal{A}=0}\right) = C(c,d)$.

**Proof.** If we swap the two transmitting agents we establish the proof (details are left to the reader). □

Note that any 0-capacity channel is some $(a,b)$ channel witha 0 resource allocation. Thus,

**Corollary 4.** *Combining $(a,b)$ with a 0-capacity channel results in a channel with the same capacity as $(a,b)$.*

We arrive at the question at hand—what happens with a partial allocation to each channel? That is, in general, how does $C\left(M_2^{1,2}\mid_{\mathcal{A}}\right)$ compare to $C(a,b)$ and $C(c,d)$? Our answer follows.

Allocate Resources to $(a,b)$ and a 0-Capacity Channel

In this situation, we know that $C\left(M_2^{1,2}\mid_{\mathcal{A}=\mathcal{R}}\right) = C(a,b)$ and that $C\left(M_2^{1,2}\mid_{\mathcal{A}=0}\right) = C(c,d)$. What happens for $0 < \mathcal{A} < R$? Not surprisingly, we get the following theorem:

**Theorem 12.** *Through allocation if we combine $(a,b)$, the first channel, with $(e,e)$, the second channel, we find that $C\left(M_2^{1,2}\mid_{\mathcal{A}}\right) = C(a^{\mathcal{A}}, b^{\mathcal{A}})$.*

**Proof.** Trivial from Theorem 9. □

*5.2. More Examples*

We will find the capacity of $C\left(M_2^{1,2}\mid_{\mathcal{A}}\right)$ by using (35) for various $\mathcal{A}$ and agent matrices.

$$
\begin{aligned}
EXAMPLE \qquad & \text{Given a } 90/10 \text{ allocation} \\
\text{The first agent } M_1^1 = (0.8, 0.4), \qquad & \text{the second agent } M_1^2 = (0.7, 0.3), \quad \mathcal{A} = 0.9 \\
C(M_1^1) &= 0.1246, \quad C(M_1^2) = 0.1187 \\
(a^{\mathcal{A}}, b^{\mathcal{A}}) &= (0.78, 0.42) \\
(c^{\mathcal{R}-\mathcal{A}}, c^{\mathcal{R}-\mathcal{A}}) &= (0.52, 0.48) \\
C\left(M_2^{1,2}\mid_{\mathcal{A}}\right) &= 0.1012 \\
C\left(M_2^{1,2}\mid_{\mathcal{A}}\right) < C(M_1^1) \qquad & C\left(M_2^{1,2}\mid_{\mathcal{A}}\right) < C(M_1^2)
\end{aligned}
$$

$$
\begin{aligned}
EXAMPLE \qquad & \text{Given a } 10/90 \text{ allocation, with the same agents as above} \\
\text{The first agent } M_1^1 = (0.8, 0.4), \qquad & \text{the second agent } M_1^2 = (0.7, 0.3), \quad \mathcal{A} = 0.1 \\
C(M_1^1) &= 0.1246, \quad C(M_1^2) = 0.1187 \\
(a^{\mathcal{A}}, b^{\mathcal{A}}) &= (0.62, 0.58) \\
(c^{\mathcal{R}-\mathcal{A}}, c^{\mathcal{R}-\mathcal{A}}) &= (0.68, 0.32) \\
C\left(M_2^{1,2}\mid_{\mathcal{A}}\right) &= 0.0967 \\
C\left(M_2^{1,2}\mid_{\mathcal{A}}\right) < C(M_1^1) \qquad & C\left(M_2^{1,2}\mid_{\mathcal{A}}\right) < C(M_1^2)
\end{aligned}
$$

$$
\begin{array}{rcl}
EXAMPLE & & \text{Given a 90/10 allocation, second agent has little noise} \\
\text{The first agent } M_1^1 = (0.7, 0.3), & & \text{the second agent } M_1^2 = (0.99, 0.01), \;\; \mathcal{A} = 0.9 \\
C(M_1^1) & = & 0.1287, \;\; C(M_1^2) = 0.9192 \\
(a^{\mathcal{A}}, b^{\mathcal{A}}) & = & (0.6, 0.4) \\
(c^{\mathcal{R}-\mathcal{A}}, c^{\mathcal{R}-\mathcal{A}}) & = & (0.745, 0.255) \\
C\left(M_2^{1,2} \mid_{\mathcal{A}}\right) & = & 0.2030 \\
C\left(M_2^{1,2} \mid_{\mathcal{A}}\right) > C(M_1^1) & & C\left(M_2^{1,2} \mid_{\mathcal{A}}\right) < C(M_1^2)
\end{array}
$$

From these results, we see that both

$$
\begin{aligned}
C\left(M_2^{1,2} \mid_{\mathcal{A}}\right) &< \min\left(C(M_1^1), C(M_1^2)\right), \text{ and} \\
\min\left(C(M_1^1), C(M_1^2)\right) < C\left(M_2^{1,2} \mid_{\mathcal{A}}\right) &< \max\left(C(M_1^1), C(M_1^2)\right)
\end{aligned}
$$

are possible. In fact, equalities are also possible by using the special cases examined at the beginning of this section. However, $\max\left(C(M_1^1), C(M_1^2)\right) < C\left(M_2^{1,2} \mid_{\mathcal{A}}\right)$ is not possible. (We show this by a re-wording and then proving that $\left(M_2^{1,2} \mid_{\mathcal{A}}\right)$ cannot be larger than both $C(M_1^1)$ and $C(M_1^2)$.) Thus, we need a lemma.

**Lemma 1.** *For channels $(a, b)$ and $(c, d)$, we find that*

$$
C((a, b) \otimes (c, d)) \leq C(a, b) + C(c, d) \,, \tag{37}
$$

*with equality if $a = b$ or $c = d$.*

**Proof.** The product channel $(a, b) \times (c, d)$ is given by channel matrix

$$
\begin{pmatrix}
ac & a\bar{c} & \bar{a}c & \bar{a}\bar{c} \\
ad & a\bar{d} & \bar{a}d & \bar{a}\bar{d} \\
bc & b\bar{c} & \bar{b}c & \bar{b}\bar{c} \\
bd & b\bar{d} & \bar{b}d & \bar{b}\bar{d}
\end{pmatrix}.
$$

The capacity of this product channel equals the sum of the capacities of its component channels $(a, b)$ and $(c, d)$ (p. 85 [5]). Removing the middle two rows gives us $(a, b) \otimes (c, d)$, and, since removing a row never increases capacity, we find that

$$
C((a, b) \otimes (c, d)) \leq C((a, b) \times (c, d)) = C(a, b) + C(c, d).
$$

□

**Theorem 13.** *If we combine through an allocation $(a, b)$, the first channel, with $(c, d)$, the second channel, then $C(M_2^{1,2}|_{\mathcal{A}})$ cannot be greater than both of the individual channel's component capacities.*

**Proof.** Let

$$
M_1^1|_{\mathcal{A}} = \begin{pmatrix} a^{\mathcal{A}} & \overline{a^{\mathcal{A}}} \\ b^{\mathcal{A}} & \overline{b^{\mathcal{A}}} \end{pmatrix},
$$

$$
M_1^2|_{\mathcal{R}-\mathcal{A}} = \begin{pmatrix} c^{\mathcal{R}-\mathcal{A}} & \overline{c^{\mathcal{R}-\mathcal{A}}} \\ d^{\mathcal{R}-\mathcal{A}} & \overline{d^{\mathcal{R}-\mathcal{A}}} \end{pmatrix},
$$

so that $M_2^{1,2}|_{\mathcal{A}} = M_1^1|_{\mathcal{A}} \otimes M_1^2|_{\mathcal{R}-\mathcal{A}}$. For any input probability distribution held constant, the mutual information is convex with respect to the elements of the channel matrix ([Theorem 2.7.4] [4]). That is, for any given input probability distribution $x$, for all $a_1, a_2, b_1, b_2, t \in [0,1]$,

$$I(ta_1 + \bar{t}a_2, tb_1 + \bar{t}b_2, x) \leq t \cdot I(a_1, b_1, x) + \bar{t} \cdot I(a_2, b_2, x),$$

where $I(\alpha, \beta, x)$ is the mutual information of channel $(\alpha, \beta)$ with input distribution $x$; thus,

$$C(\alpha, \beta) = \max_x I(\alpha, \beta, x), \text{ and}$$

$$\therefore \ \forall x, C(\alpha, \beta) \geq I(\alpha, \beta, x) \ .$$

If we let $a_1 = a,\ b_1 = b,\ a_2 = b_2 = \frac{a+b}{2}, t = \frac{\mathcal{A}}{\mathcal{R}}$, we have from convexity that

$$\begin{aligned} I(a^{\mathcal{A}}, b^{\mathcal{A}}, x) &= I\left(\frac{\mathcal{A}}{\mathcal{R}}a + (1 - \frac{\mathcal{A}}{\mathcal{R}})(\frac{a+b}{2}), \frac{\mathcal{A}}{\mathcal{R}}b + (1 - \frac{\mathcal{A}}{\mathcal{R}})(\frac{a+b}{2}), x\right) \\ &\leq \frac{\mathcal{A}}{\mathcal{R}} I(a, b, x) + (1 - \frac{\mathcal{A}}{\mathcal{R}}) \cancel{I\left(\frac{a+b}{2}, \frac{a+b}{2}, x\right)} \text{ (this last term is 0)} \end{aligned}$$

for any input probability distribution $x$, because $I(e, e, x)$ always equals 0. Now, we let $\chi$ be a capacity achieving input probability (unique except for 0-channels) distribution for $(a^{\mathcal{A}}, b^{\mathcal{A}})$, giving

$$C(a^{\mathcal{A}}, b^{\mathcal{A}}) = I(a^{\mathcal{A}}, b^{\mathcal{A}}, \chi) \leq \frac{\mathcal{A}}{\mathcal{R}} I(a, b, \chi) \leq \frac{\mathcal{A}}{\mathcal{R}} C(a, b).$$

Therefore,

$$C(M_1^1|_{\mathcal{A}}) \leq \frac{\mathcal{A}}{\mathcal{R}} C(M_1^1),$$

and by replacing $\frac{\mathcal{A}}{\mathcal{R}}$ with $1 - \frac{\mathcal{A}}{\mathcal{R}}$ and repeating the above convexity argument, we find that

$$C(M_1^2|_{\mathcal{R}-\mathcal{A}}) \leq \frac{\mathcal{R}-\mathcal{A}}{\mathcal{R}} C(M_1^2).$$

By Lemma 1,

$$C(M_2^{1,2}|_{\mathcal{A}}) = C(M_1^1|_{\mathcal{A}} \otimes M_1^2|_{\mathcal{R}-\mathcal{A}}) \leq C(M_1^1|_{\mathcal{A}}) + C(M_1^2|_{\mathcal{R}-\mathcal{A}}) \text{ . Thus,}$$

$$C(M_2^{1,2}|_{\mathcal{A}}) \leq \frac{\mathcal{A}}{\mathcal{R}} C(M_1^1) + \frac{\mathcal{R}-\mathcal{A}}{\mathcal{R}} C(M_1^2) \leq \left(\frac{\mathcal{A}}{\mathcal{R}} + \frac{\mathcal{R}-\mathcal{A}}{\mathcal{R}}\right) \max(C(M_1^1), C(M_1^2)) \ .$$

Resulting in, $C(M_2^{1,2}|_{\mathcal{A}}) \leq \max(C(M_1^1), C(M_1^2))$ .

□

Thus, we have shown that $C\left(M_2^{1,2}\ |_{\mathcal{A}}\right) \leq \max\left(C(M_1^1), C(M_1^2)\right)$ and, by using Theorem 11 and Corollary 3, equality can be obtained by letting $\mathcal{A} = \mathcal{R}$ or 0, the choice depending on the underlying original channels.

Results and Discussion

In this section, we showed what happens when we have limited transmission power and want to distribute it among two transmitting agents. The theorems of this section capture the physical properties of the power allocation and happily agree with intuition.

## 6. Conclusions

We considered the use of Shannon information theory, and its various entropic terms to aid in reaching optimal decisions that should be made in a multi-agent/Team scenario. Our metric for agents passing information are classical Shannon channel capacity. Our results are the mathematical theorems in this article showing how combining agents influences the channel capacity.

We have put the idea forward of multi-agent communication on a firm information theoretic foundation. We examined simple scenarios in this paper to lay that strong foundation. We obtained results that may seem obvious, but are quite difficult to prove. We ask the reader to keep in mind that there is a big difference between "it is obvious" and "it has been shown".

From our perspective we have shown that, except for certain boundary cases, one can achieve near perfect transmission of Shannon information, provided one has a large enough number of agents.

We have used most information versus resource (power) allocation as an optimizing criterion. With regard to resource allocation, our results tell us that the best thing to do is to just use the strongest channel. This result is not surprising. However, without the mathematics to prove it, we would be relying on intuition. Furthermore, note that we only used a simple linear allocation scheme in this section, and we only combined two agents. Future work will consider non-linear allocation schemes and multiple agents to continue what we have started in this paper. Going forward, this path is especially meaningful if we adjust the Riemannian metric to influence the power allocated to each channel. For example, a geometric region with high noise levels can be reflected in the Riemannian metric by acknowledging that the $E, F, G$ terms of the metric are functions of $a$ and $b$. We will explore this direction in future work.

In addition, in future work, we will also consider more than two agents competing for the available resources, non-Euclidean Riemannian metrics, and more complicated signaling alphabets and schemes. We are also interested in information flow in the Vicsek [19] bird flocking model.

## 7. Notation

We include some of the notation that is used repeatedly throughout the article. The other notation is variants of what we give here with changes to the indices and is made clear in its first usage.

| | |
|---|---|
| MAS | Multi-agent System |
| $A_x$ | Agent $X$ |
| $M$ | A channel matrix, that is every row contains non-negative numbers that sum to 1 |
| $M_n$ | $2 \times 2n$ channel matrix, representing $n$ (transmitting) Agents |
| $H(V)$ | Entropy of the (discrete) random variable $V$ |
| $H(V\|W)$ | Conditional Entropy of the random variable $V$ conditioned on $W$ |
| $I(V,W)$ | Mutual information between the random variables $V$ and $W$ |
| $C$ | Capacity of a generic channel |
| $C_{2,2}$ | Specifically the capacity of a 1 (transmitting) agent channel |
| $M_1^1$ | A specific 1-agent channel $\begin{pmatrix} a & \bar{a} \\ b & \bar{b} \end{pmatrix}$. Note: C(a,b):=C($M_1^1$) |
| $M_1^2$ | Another 1-agent channel $\begin{pmatrix} c & \bar{c} \\ c & \bar{d} \end{pmatrix}$ |
| $M_2^{1,2}$ | The combined channel $(a,b) \otimes (c,d)$ with channel matrix $\begin{pmatrix} \text{ac} & a\bar{c} & \bar{a}c & \bar{a}\bar{c} \\ \text{bd} & b\bar{d} & \bar{b}d & \bar{b}\bar{d} \end{pmatrix}$ |
| $M_2^{1,2}\|_{\mathcal{A}}$ | Combined power allocated channel with channel matrix $= \begin{pmatrix} a^{\mathcal{A}} \cdot c^{\mathcal{R}-\mathcal{A}} & a^{\mathcal{A}} \cdot \overline{c^{\mathcal{R}-\mathcal{A}}} & \overline{a^{\mathcal{A}}} \cdot c^{\mathcal{R}-\mathcal{A}} & \overline{a^{\mathcal{A}}} \cdot \overline{c^{\mathcal{R}-\mathcal{A}}} \\ b^{\mathcal{A}} \cdot d^{\mathcal{R}-\mathcal{A}} & b^{\mathcal{A}} \cdot \overline{d^{\mathcal{R}-\mathcal{A}}} & \overline{b^{\mathcal{A}}} \cdot d^{\mathcal{R}-\mathcal{A}} & \overline{b^{\mathcal{A}}} \cdot \overline{d^{\mathcal{R}-\mathcal{A}}} \end{pmatrix}$ |
| $M_{2-}$ | $= \begin{pmatrix} a^2 & 2\bar{a}a & \bar{a}^2 \\ b^2 & 2\bar{b}b & \bar{b}^2 \end{pmatrix}$, formed from the $(a,b)$ channel |

**Author Contributions:** Conceptualization, I.S.M.; Methodology, I.S.M. and S.R.; Software, I.S.M. and P.R.; Investigation, I.S.M., P.R. and S.R.; Writing, I.S.M., P.R. and S.R. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Acknowledgments:** We thank Hans Haucke for his assistance. We are especially grateful to Ruth Irene for her helpful comments on the draft versions of this paper. A special thanks to the reviewers who encouraged us to expand the background literature citations and pointed out what was lacking in some of our explanations and discussions. We also thank them for catching typos and points that needed clarification. We thank Katarina Doctor for her discussions on domain focused interpretable machine learning. A very special thanks to the special issue editor William Lawless for his assistance.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Moskowitz, I.S. A Cost Metric for Team Efficiency. *Front. Phys. Interdiscip. Phys.* **2022**, *212*, 861633. [CrossRef]
2. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656. [CrossRef]
3. Gallager, R.G. *Information Theory and Reliable Communication*; Wiley: New York, NY, USA, 1968.
4. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley: New York, NY, USA, 2006.
5. Ash, R.B. *Information Theory*; Dover Publications: New York, NY, USA, 1965.
6. Martin, K.; Moskowitz, I.S.; Allwein, G. Algebraic Information Theory For Binary Channels. *Electron. Notes Theor. Comput. Sci.* **2006**, *158*, 289–306. [CrossRef]
7. Moskowitz, I.S.; Cotae, P.; Safier, P.N. Algebraic Information Theory and Stochastic Resonance for Binary-Input Binary-Output Channels. In Proceedings of the 46th Annual Conference on Information Science and Systems (CISS), Princeton, NJ, USA, 21–23 March 2012.
8. Neumann, J.V. *Theory of Self-Reproducing Automata*; Burks, A.W., Ed.; University of Illinois Press: Urbana, IL, USA, 1966.
9. Sliwa, J.Toward Collective Animal Neuroscience. *Science* **2021**, *374*, 397–398. [CrossRef] [PubMed]
10. Lawless, W.F. Risk Determination versus Risk Perception: A New Model of reality for Human–Machine Autonomy. *Informatics* **2022**, *9*, 30. [CrossRef]
11. Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N.R.; Kalchbrenner, N.; Goyal, A.; Bengio, Y. Toward Causal Representation Learning. *Proc. IEEE* **2021**, *109*, 612–634. [CrossRef]
12. Majani, E.E.; Rumsey, H. Two Results on Binary-Input Discrete Memoryless Channels. In Proceedings of the 1991 IEEE International Symposium on Information Theory, Budapest, Hungary, 24–28 June 1991.
13. Martin, K.; Moskowitz, I.S. Noisy Timing Channels with Binary Outputs. In *International Workshop on Information Hiding 2006*; LNCS 4437; Springer: Berlin/Heidelberg, Germany, 2007; pp. 124–144.
14. Silverman, R.A. On Binary Channels and their Cascades. *Ire Trans. Inf. Theory* **1955**, *1*, 19–27. [CrossRef]
15. Moskowitz, I.S.; Newman, R.E.; Crepeau, D.P.; Miller, A. *A Detailed Mathematical Analysis of a Class of Covert Channels Arising in Certain Anonymizing Networks*; Naval Research Laboratory Memorandum Report, NR/MR/5540–03-8691; Naval Research Laboratory: Washington, DC, USA, 2003.
16. Arimoto, S. An Algorithm for Computing the Capacity of Arbitrary Discrete Memoryless Channels. *IEEE Trans. Inf. Theory* **1972**, *18*, 14–20. [CrossRef]
17. Blahut, R. Computation of Channel Capacity and Rate-Distortion Functions. *IEEE Trans. Inf. Theory* **1972**, *18*, 460–473. [CrossRef]
18. Ross, S. *A First Course in Probability*; Macmillan: New York, NY, USA, 1976.
19. Vicsek, T.; Czirok, A.; Ben-Jacob, E.; Cohen, I.; Shochet, O. Novel type of Phase Transition in a System of Self-Driven Particles . *Phys. Rev. Lett.* **1995**, *75*, 1226–1229. [CrossRef] [PubMed]