Towards a Unification of Logic and Information Theory

Luis A. Lastras[1], Barry Trager[1], Jonathan Lenchner[1], Wojtek Szpankowski[2], Chai Wah Wu[1], Mark Squillante[1] and Alex Gray[1]

[1]IBM T.J. Watson Research Center
[2]Purdue University

January 24, 2023

Abstract

We examine the problem of efficient transmission of logical statements from a sender to a receiver under a diverse set of initial conditions for the sender and receiver's beliefs and on the goal for the communication. From the standpoint of our work, two different collections of logical statements are *equivalent* if there anything that can be proved from one collection can also be deduced from the other collection. Distinguishing between these two collections is thus unnecessary from the standpoint of our work and leads to communication cost efficiencies. In order to develop an example of an information theory for the transmission of logical statements, we focus on a simple logical system equivalent to propositional logic where a collection of logical statements can be alternately depicted as a collection of multivariate polynomial equations with coefficients and variables in a finite field. We then apply classical concepts from information theory, notably concepts for rate-distortion theory, to develop closed form expressions for the cost of communicating these logical statements. We additionally provide a theory of linear codes for implementing these communication systems that produces systems that are asymptotically bit-cost optimal in some settings. It is our belief that the scope for improving beyond our limited exploration is vast, including treating more sophisticated logical systems such as first order logic, studying different types of communication constraints and creating practical algorithms for attaining the Shannon limits.

1 Introduction

The starting point for our work is a general question that admits many interpretations and hence possible solutions:

What is the information content in a logical statement?

Our attempts to answer this question led to a number of threads most of which remain open and interesting. The work that we present in this article is an instance of one of these threads where we were able to obtain early results with some degree of mathematical maturity, but even within this thread it is clear that we were able to make all but a small dent on this fascinating question.

It is well known that a significant conceptual contribution of Shannon was to provide a definition for the concept of information that was independent of the semantics of the message being conveyed. This abstraction, while sometimes feels counter one's intuition of what we think of as information, is one of the most successful concepts in the computing and communication revolution, as it allowed us to build flexible machines that process and communicate information in a standardized way even as our messages and intentions behind those messages remain flexible.

In the Shannon sense, a bit of information represents a choice from a list, in this case with two elements. Communicating this bit to a receiver results in that receiver learning that choice. Implicit in this concept is the idea that the choice matters to the receiver in that it *distinguishes* among two options which have deeper meaning to it, but this deeper meaning is irrelevant of the purposes of efficient communication.

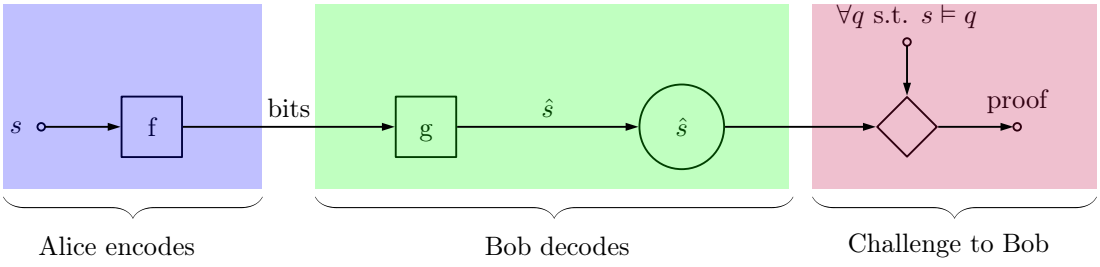


Figure 1: Prototypical communication setup in our article.

Paraphrasing, information theory concerns itself with the transmission and processing of distinguishable options, but doesn't care about what is being distinguished.

This last assertion, some might argue, isn't entirely correct. In 1959 Shannon introduced the fundamentals behind rate-distortion theory [5], which allows the designer of a compression system to specify a fidelity criterion that measures the cost of reproducing a symbol in a data source incorrectly with another symbol. Shannon's work was necessary to extend the information theory of compression to the realm of continuous valued sources which in general require infinite precision to be sent without loss, nonetheless Shannon's theory is much more general and it can be used with finite alphabet data as well and fairly arbitrary ways of measuring the cost of reproduction errors; for example already digitized audio and video is further compressed in a lossy manner because certain aspects of these signals are deemed to be less meaningful to us due to the way in which perception works in our bodies. Our reference to rate-distortion theory has a purpose: the mathematical principles behind it will be decomposed and re-assembled to create our theory.

In our work, we adopt the viewpoint that an information theory, from the classical Shannon theoretical standpoint, which captures additional elements of semantics in the quantification of information content in a message is, after all, a worthy goal. Consider the following sentences:

The light is off when it is sunny, and on when there's no sun.
The light is on when there's no sun, and off when it is sunny.
It's impossible that the light is on and its sunny; it's also impossible for both the light to be off and for there to be no sun.

Different as these sentences are in surface form, for a receiver who can recognize logic from language, all these three sentences are indistinguishable from each other from the standpoint of their implication on what we can deduce after receiving them. Yet in classical information theory, we would describe these sentences using different bit sequences, as the fact that they convey the same logical facts is irrelevant. This points to a fundamental idea that we rely on extensively on this article: if the point is to convey what appears to be the essentials in a logic statement, one can take advantage of the fact there are equivalence classes of surface forms, any of which would suffice to let a receiver acquire such essentials. Our task ahead is to translate this observation in specific mathematical forms which are amenable to treatment by the tools of information theory.

To make matters concrete, we refer the reader to Figure 1, which is the simplest scenario that we consider in this article but that otherwise it showcases the three key stages in how our agents interact. The essential goal for a communication from Alice to Bob is that Bob be able to provide proofs for certain mathematical statements presented to him after the communication from Alice to Bob took place, without any further assistance from Alice. The purpose of the communication is to equip Bob with the information necessary for him to accomplish this task. This takes place in three stages: first Alice uses mathematical assertions available to her (depicted as s) and an encoder function f to produce a sequence of bits, generally of variable length, which have some kind of dependence on s. The cost of this communication is measured by counting how many bits were sent, and it will be our goal to optimize this as much as possible averaged over multiple communications. In a second stage, these bits are transmitted to Bob, who has a decoder function g which produces some other set of mathematical statements depicted as $\hat{s}$, in general not necessarily (and almost never) matching the surface forms present in s. Finally, in a third stage Bob is challenged to produce a

proof of some mathematical assertions depicted as q, which are assumed to be entailed by s (written $s \vDash \hat{q}$). This challenge may repeat itself with other such q's. In the simple setup in this Figure 1, Bob does not posses any mathematical statements prior to the communication, but we do consider other setups which are documented in Figure 2. As it is standard in information theory, our results will consist of two components: an impossibility result (or converse) which is a lower bound on the communication cost under very general conditions for f and g, and an achievability result (or forward) in which a specific choice for f and g are shown to operate close to the limit provided by the lower bound.

The reader may be wondering: what exactly is the situation that we are attempting to model? What physical situation involves communication of logical statements in the presence of devices such as encoders and decoders? A fundamental motivation in our work is modeling of human/computer communication in a multi-turn dialog form. In this setting, a computer is attempting to communicate something to a human efficiently, possibly over multiple iterations (although here we model only one such iteration), and thus in this context it is feasible to assume the existence of the devices f and g which aid in this task. In this light, Alice is a computer, and the decoder g may be regarded as computational device in Bob's possession. To make this point of view more evident, in the figure, the large circle represents the entire mental state of the Bob after communication takes place. The diamond represents a computational capability to produce proofs of statements starting from other statements; this capability could be external to Bob (for example, a computer) or could be an algorithm Bob executes himself.

While the figure represents a very general setup, we will make an assumption on the logical system to be used which both allows us to produce sharp an interesting information theoretic results but also limits the scope of our work. Specifically, we will assume that Alice and Bob share a set of propositions and that any one mathematical statement relates these propositions using the logical operators "AND/OR/NOT" and then assigns "TRUE/FALSE" to the expression to form the mathematical statement.

For example, to describe the last line in our earlier example in this form the sender and receiver could initially agree on

$$x_0\text{: The light is on}$$
$$x_1\text{: It is sunny}$$

and then write

$$(\text{ NOT } x_0 \text{ AND NOT } x_1 \text{) is FALSE}$$
$$x_0 \text{ AND } x_1 \text{ is FALSE}$$

In order to bridge from concepts in logic to the tools of information theory, we will be leveraging the fact that mathematical assertions such as the ones exemplified earlier can be written as multivariate polynomial equations with variables and coefficients that are regarded as elements of the Galois Field with two elements or equivalently, the set of integers with standard arithmetic modulo 2. The "reasoning engine" that is then used to provide proofs of statements starting from other statements (represented as a diamond in Figures 1 and 2 can then take the form of a type multivariate polynomial division encoded in the Gröbner basis computation algorithm.

In this article we address the following problems (also illustrated in the sequence of diagrams in Figure 2) in the context of the narrower logical system described above:

1. Alice wishes to communicate to Bob enough information so that Bob can prove anything that Alice can prove, but Bob has no prior mathematical statements to help reduce the cost of this communication.

2. The same as 1), but under the assumption that Alice and Bob have shared background mathematical statements.

3. The same as 2) but with the additional assumption that Bob's knowledge is incrementally updated with additional mathematical statements to complement his existing statements.

4. Alice wishes to communicate to Bob enough information to prove a specific mathematical statement, but Bob has no prior mathematical statements to aid in this task.

5. The same as above, but where Bob does have prior mathematical statements which Alice is aware of, and potentially where some of those mathematical statements may be in contradiction with those that Alice believes to be true.

The main contribution of our article is a collection of theorems that provide sharp upper and lower bounds on the communication cost of systems that implement the patterns above. The numbering of these problems coincides with the numbering of the corresponding theorems. Theorem 1 is our simplest result, and its main contribution is to establish fundamental concepts that are used in more complex settings. Theorem 2 is a relatively simple elaboration on Theorem 1, intended to highlight the contribution of Theorem 3, which makes the stronger assumption that Bob must be taught incrementally, yet concludes that precisely the same limit as in Theorem 2 applies. As such, Theorems 2 and 3 imply a result that we nickname "the optimality of incremental communication". Theorems 4 and 5, which bring further communication efficiencies by limiting the scope of the goal for the communication, are what we call the "less is more" theorems, because of a surprising revelation that communication optimal strategies often end up describing to Bob more mathematical facts than he strictly needed to prove what the communication intended to enable him to prove. Theorems 1 through 4 can be regarded as quantifying the cost of reducing "ignorance". An interesting twist happens when you allow for Bob to believe in the initial state mathematical statements which are not entirely consistent with those believed by Alice. In that case, one may regard Bob to be in a state of misinformation (under the asymmetric assumption that Alice's knowledge is "true"). Theorem 5 then can be seen as a rudimentary model of the cost of communication for the goals of reducing ignorance and misinformation simultaneously.

The distinction between Theorems 2 and 3 is somewhat subtle, so it is worth pausing to make it crisp. As discussed earlier, a motivation for our work is to model a situation where a computer (Alice) is efficiently communicating logical statements to a human (Bob). If Bob already has prior mathematical statements in his possession, it feels natural to assume that whatever Alice sends to Bob should result in adding net new mathematical statements to Bob's mind (Theorem 3). However, upon further examination, one realizes that this is a restriction that could result in a larger communication cost; a more general setup would be to assume that g directly produces whatever ends up being in Bob's mind (Theorem 2), which includes the possibility above as a sub-case. Alas, the limits for Theorems 2 and 3 are the same, thus concluding the optimality of incremental communication, at least under our current assumptions.

As it is common in Shannon theory, the results above leave open the question of how to implement practical algorithms that approach the corresponding bounds, particularly for Theorems 4 and 5 where new ideas are necessary. Our contribution in this direction is a construction based on linear codes that is shown to be optimal for certain scenarios, leaving open the question of how to implement these systems in general.

We do not claim that the communication patterns together with the logical system assumptions have immediate practical value, but rather suggest the value of our work lies in identifying a problem formulation where fundamental concepts in information theory, algebra and logic appear to collaborate in a compelling way in problems pertaining transmission of logical information. We hope that our work may spark additional efforts which could result in more concrete implications. To assist the reader in understanding future possibilities as well as the limitations of our work, we offer the following ideas:

- Can we extend some of these results to first order logic, or other more interesting logical inference systems?
- When Bob has access to prior mathematical facts, they are assumed to also be known to Alice. What if they are not?
- We are making an asymmetric assumption that Alice is the one that teaches Bob. How would we model a more collaborative situation?
- Our linear code construction is optimal only under a narrow scenario. Can we design practical codes for the entire set of problems?
- We are currently modeling only one communication exchange. How would we model a multi-turn dialog?

We acknowledge that our discourse is one of many proposals in which words like "logic" and "information" appear in the same context. Devlin's work on logic and information [2] postulates that the concept of information can be studied through empirical observation methods as one does in physical sciences, and then proposes an extension to first order logic based on his own observations about how information appears to behave in a variety of circumstances. Devlin's work is highly conceptual (indeed, some may argue, philosophical); in comparison in our work some of the more difficult conceptual questions are delayed in favor of a crisp proposal on how to combine information theory, algebra and logic reasoning, admittedly applicable in a relatively narrow set of scenarios. Ellerman [4] examines the mathematical foundations of information and constructs a theory around the concept of "information as distinctions"; Ellerman goes on to define "logical entropy" and shows how Shannon's entropy can be seen as a special case of logical entropy. The view of information as distinctions is deeply embedded in our work as well, however our path led us towards implementing the notion of distinguishability by leveraging concepts from Shannon's rate-distortion theory, rather than coming up with a different foundational concept.

2 Mathematical preliminaries

As stated in the introduction, our work will be developed under the confines of a logical system that may be seen as zeroth-order logic. A total of m propositions are shared between Alice and Bob. We will promptly move towards representing and operating on logical statements that involve these propositions using the mathematics of algebra over finite fields, to be introduced next. Let K denote a finite field. The set of all polynomials whose terms have coefficients in K and over variables $x_1, \cdots, x_m$ is denoted $K[x_1, \cdots, x_m]$. All our results will apply under this general assumption on K, and thus in principle they could be used to treat problems in multi-valued logic. This possibility, however, is an unnecessary distraction for the exposition of the remainder of the article, so we will restrict our attention to $K = GF[2]$, the finite field of size 2, which comprises the binary alphabet $\{0, 1\}$ together with multiplication $(\cdot)$ and addition $(+)$ where the latter respectively correspond to the binary AND and XOR operations.

In order to properly associate the variables $x_1, \cdots, x_m$ to true/false status for the m propositions, we will assume throughout the article that these variables satisfy the conditions

$$x_i^{|K|} - x_i = 0. \tag{1}$$

These conditions are called the *field polynomials*, and they imply that $x_i \in K$. We associate 0 with the logical value false and 1 with true, and thus the variables $x_1, \cdots, x_m$ denote whether the corresponding propositions are false or true.

Let $\mathcal{P}$ denote the operator that returns the set of all sets of items that belong to the argument to the operator. Let $\mathcal{P}(K[x_1, \cdots, x_m])$ denote the set of all sets of polynomials in $K[x_1, \cdots, x_m]$. Any choice of element $s \in \mathcal{P}(K[x_1, \cdots, x_m])$ represents a collection of logical statements over the propositions by constructing a set of polynomial equations that equate each polynomial in s to zero. Similarly, any collection of arbitrary logical statements about these propositions can be expressed using an element of $\mathcal{P}(K[x_1, \cdots, x_m])$. To establish this, let $\text{prop}_a, \text{prop}_b$ denote propositions and a, b corresponding binary variables, and then we use the following equivalence table in order to create a mapping between logical and polynomial forms:

logical form	polynomial form
$\neg\text{prop}_a$	$a + 1$
$\text{prop}_a \vee \text{prop}_b$	$a + b + a \cdot b$
$\text{prop}_a \wedge \text{prop}_b$	$a \cdot b$
$\text{prop}_a \oplus \text{prop}_b$	$a + b$
$\text{prop}_a \vDash \text{prop}_b$	$a \cdot (1 + b) + 1$

(2)

One of the benefits of using the algebra machinery underlying the polynomial representation is the ability to borrow concepts and computational devices from algebra in order to do useful things such as proving or disproving logic statements by assuming others. A concept called *algebraic set* is particularly useful in this task, and will prove central to every single development in our exposition.

For $s \in \mathcal{P}(K[x_1, \cdots, x_m])$, denote by s_i the ith polynomial in the set s, and define $\mathbf{v}(s)$ to be the *algebraic set* of s as follows:

$$\mathbf{v}(s) := \{x_1, \cdots, x_m : s_i(x_1, \cdots, x_m) = 0, \quad \forall s_i \in s\}. \tag{3}$$

For those versed in the language of algebra, note that we defined a algebraic set directly in terms of a set of polynomials, instead of through the *ideal* of the set of polynomials. We did so out of convenience, since in this work the concept of an algebraic set is more central than that of the ideal. Paraphrasing, $\mathbf{v}(s)$ is the set of all zeros of the set of equations obtained by equating each polynomial in s to zero. Note that the algebraic set depends on the context given by the polynomial ring $K[x_1, \cdots, x_m]$; in particular, every single assignment to the variables $x_1, \cdots, x_m$ that satisfies the definition must be accounted for even if the underlying polynomial expression does not use some of the variables.

Let s, t denote the corresponding sets of polynomials. Suppose that our task is to show that s entails t. This can be done by checking the algebraic sets:

Definition 1 *We say that $s \vDash t$ if and only if $\mathbf{v}(s) \subseteq \mathbf{v}(t)$.*

There are in general many equivalent ways of expressing a set of beliefs as sets of polynomials which correspond to the same algebraic set. For example, suppose that the algebraic set comprises the set $\{(x_0 : 0, x_1 : 1), (x_0 : 1, x_1 : 0)\}$. One possible set of polynomials with this algebriac set is given by

$$\left\{ \begin{array}{c} (x_0 + 1)(x_1 + 1), \\ x_0 x_1 \end{array} \right\}. \tag{4}$$

However, alternatively, the following also has the same algebraic set:

$$\{x_0 + x_1 + 1\}. \tag{5}$$

As a result, much of the essentials in our work will be centered around algebraic sets, rather than the original sets of polynomials leading to the algebraic sets. Intuitively, a sender's knowledge is "sharper" than that of a receiver, in the sense that the sender has more facts, and thus a smaller algebraic set. Much of what we will be concerned about is sharpening the information that a receiver has is enough so as to meet some requirement set by a problem statement.

Given an arbitrary subset of K^m, it is possible to construct a corresponding set of polynomials with exactly the same algebraic set. This fact will be used multiple times in this article, and thus we phrase it as the following formal result.

Lemma 1 (Reconstruction of sets of polynomials from a proposed algebraic set) *There exists a function $\sigma : \mathcal{P}(K^m) \to \mathcal{P}(K[x_1, \cdots, x_m])$ such that for every a, the algebraic set of $\sigma(a)$ is exactly a. The set of all polynomials that vanish on a is generated by $\{\sigma(a), x_1^{|K|} - x_1, \ldots, x_m^{|K|} - x_m\}$.*

Proof. Let K be an arbitrary finite field which implies $a \subseteq K^m$ is a finite set. First we observe that, given a point $p \in K$, there exists a polynomial $I_p \in K[x]$ such that $I_p(p) = 1$ and $I_p(q) = 0$ for all $q \in K \mid q \neq p$. Recall that $x^{|K|} - x$ is a field polynomial whose roots are all the elements of K. Then $(x^{|K|} - x)/(x - p)$ is a polynomial that vanishes at all points other than p, but takes the value -1 at p, so we define $I_p := -(x^{|K|} - x)/(x - p)$. Note that when $|K| = 2$, this construction simplifies to $I_p = x + p + 1$. Now let $c = (c_1, \ldots, c_m) \in a$; the polynomial $P_c = \prod_{i=1}^{m} I_{c_i}[x_i]$ takes the value 1 at the point $c \in a$, and takes the value 0 at all other points of a. Thus the polynomial $\sigma(a) = -1 + \sum_{c \in a} P_c$ takes the value zero at each point of a and is nonzero at all other points of K^m. To restrict the zeros to lie inside K^m we will also include the field polynomials $\{x_1^{|K|} - x_1, \ldots, x_m^{|K|} - x_m\}$. Then $\{\sigma(a), x_1^{|K|} - x_1, \ldots, x_m^{|K|} - x_m\}$ is the largest ideal whose zero set is precisely a, and thus it contains all polynomials vanishing on a. □

Whenever a set in $\mathcal{P}(K[x_1, \cdots, x_m])$ is deterministic we shall use a lowercase letter (e.g., s). Whenever it is random, we shall use an uppercase letter (e.g. S). To evaluate the performance of candidate algorithms we adopt the *expected performance* metric where the average performance is computed according to a given distribution. Our choice of distribution is motivated by the earlier remarks on the prominence of algebraic sets;

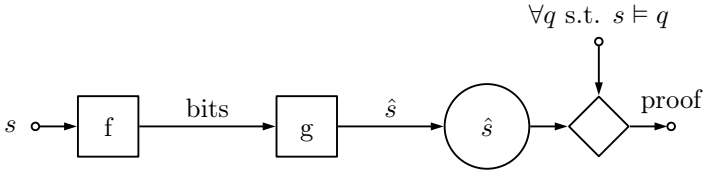
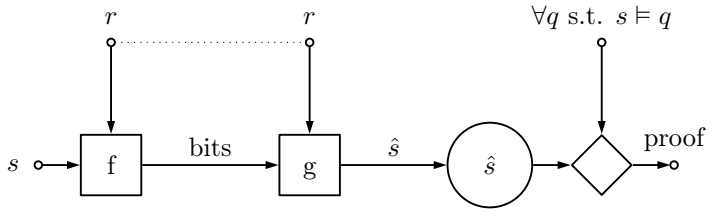
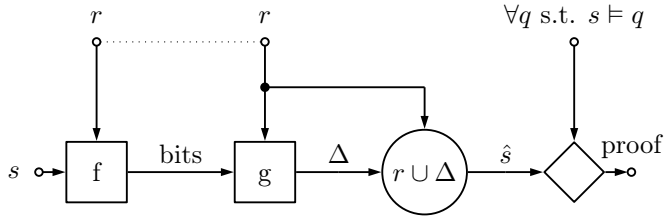
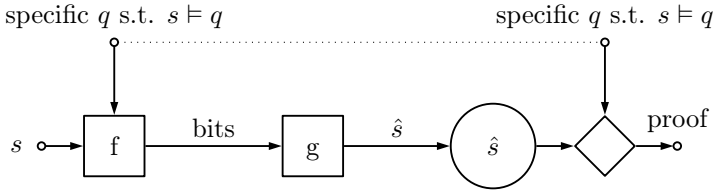
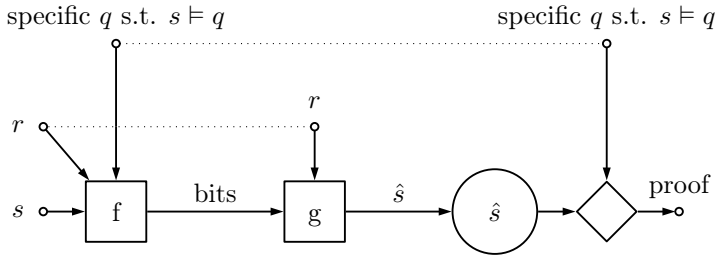
Communication diagram	Assumptions	Shannon limit
	if $s \vDash q$ then $s \vDash \hat{s} \vDash q$	$H(p_s)$
	if $s \vDash q$ then $s \vDash \hat{s} \vDash q$ $s \vDash r$	$p_r H(p_s)$
	if $s \vDash q$ then $s \vDash (r, \Delta) \vDash q$ $s \vDash r$ Δ is required to have "net new information": $\forall w$ such that $r \vDash w, w \notin \Delta$	$p_r H(p_s)$
	if $s \vDash q$ then $s \vDash \hat{s} \vDash q$	$\Lambda(p_s, 1 - p_q)$
	if $s \vDash q$ then $s \vDash \hat{s} \vDash q$	$p_r \Lambda(p_{s\|r}, 1 - p_{q_r}) +$ $(1 - p_r)\Lambda(p_{s\|\bar{r}}, 1 - p_{q\|\bar{r}})$

Figure 2: Sample communication scenarios covered in this article. The circle stands for the knowledge that the receiver has after the transmission has taken place. The diamond is a computational device that is capable of producing a proof of a query (fed from the top) using the knowledge fed on the left.

we will use the letters S, T, U as generic names for the random sets that follow. Letting $S \in \mathcal{P}(K[x_1, \cdots, x_m])$ be a random set of polynomials, we say that S follows a p_s-i.i.d. law if the elements of $\mathbf{v}(S)$ are chosen from K^m independently at random with probability p_s. Similarly, let $S, T \in \mathcal{P}(K[x_1, \cdots, x_m])$ be two random sets of polynomials such that $S \vDash T$ so that they have nested algebraic sets, i.e., $\mathbf{v}(S) \subseteq \mathbf{v}(T)$. We say that the probability law governing them is a (p_s, p_t)-i.i.d. law ($p_s \leq p_t$) if this law is constructed as follows: for every element of K^m, with probability p_s assign it to both $\mathbf{v}(S)$ and $\mathbf{v}(T)$, and with probability $p_t - p_s$ assign it to $\mathbf{v}(T)$ only – by construction this way, we ensure the nesting requirement. Note that there will be on average a fraction $1 - p_t$ of elements of K^m in neither $\mathbf{v}(S)$ nor $\mathbf{v}(T)$ at the end of this process. Finally, let $S, T, U \in \mathcal{P}(K[x_1, \cdots, x_m])$ be three random sets of polynomials with $S \vDash T$, and with U being a "conditioning" set of facts shared by two parties. We say that S, T, U follow a $(p_u, p_{s|u}, p_{t|u}, p_{s|\tilde{u}}, p_{t|\tilde{u}})$-i.i.d. law if their respective algebraic sets $\mathbf{v}(S), \mathbf{v}(T), \mathbf{v}(U)$ are governed by the following law. First, the elements of $\mathbf{v}(U)$ are chosen from K^m independently at random with probability p_u, thereby also creating the complement $\mathbf{v}(U)^c$ at the same time. Then, $\mathbf{v}(S)$ and $\mathbf{v}(T)$ are constructed by sampling from $\mathbf{v}(U)$ using an $(p_{s|u}, p_{t|u})$-i.i.d law (instead of sampling from K^m), and they each are further augmented by sampling from $\mathbf{v}(U)^c$ using an $(p_{s|\tilde{u}}, p_{t|\tilde{u}})$-i.i.d. law.

We will be using in our development a universal code for integers due to Elias called Elias delta encoding [3]. This code assigns codewords to each integer, where the length of a codeword for the integer $n \geq 1$ is given by

$$\begin{aligned} \mathbf{len}(\text{elias}_\delta(n)) &= \lfloor \log_2(n) \rfloor + 2\lfloor \log_2(1 + \lfloor \log_2(n) \rfloor) \rfloor + 1 \text{ bits} && (6) \\ &\leq \log_2(n) + 2\log_2 \log_2(n) + 3 \text{ bits.} && (7) \end{aligned}$$

In our work, the function f in any of the settings in consideration (Figure 2) is capable of producing a variable number of bits, as this is a more flexible setting than assuming a fixed number of bits. However, an additional complication is that it may not be easy to determine when these bits start and finish in an otherwise arbitrary bit sequence. To resolve this matter, we will rely on the standard concept from information theory of prefix-free codes. Let $\mathcal{C} \subseteq \{0,1\}^*$ be a set of codewords. We say that $\mathcal{C}$ is prefix free if for all distinct $c_1, c_2 \in \mathcal{C}$, c_1 is not a prefix of c_2. Let f be a function, called an encoder, whose image is a subset of $\{0,1\}^*$, the set of all finite-length sequences drawn from $\{0,1\}$, irrespective of their length. We say that *the code implied by* f is prefix free if the image of f is prefix free.

3 Fundamental information theoretic results

We now describe a series of problems of increasing complexity that will allow us to gradually introduce the essential ideas in our work. In all such problems, a fundamental expression closely related to Shannon's binary entropy appears as part of the solution: for any $a, b \geq 0$, we have

$$\Lambda(a,b) = a \log \left(\frac{a+b}{a} \right) + b \log \left(\frac{a+b}{b} \right) = (a+b) H \left(\frac{a}{a+b} \right), \tag{8}$$

where we additionally define $\Lambda(0,b) = \Lambda(a,0) = 0$. The greek letter Λ is chosen for this function in reference to its apparent emergence in problems involving Logic. This expression satisfies the following basic properties:

Lemma 2 (Elementary properties of $\Lambda(a,b)$) *The function $\Lambda(a,b)$ is concave $\cap$ over the domain $[0,+\infty) \times [0,+\infty)$. If $\Delta_a, \Delta_b \geq 0$ with at least one of them being strictly positive, then $\Lambda(a+\Delta_a, b+\Delta_b) > \Lambda(a,b)$. If $a+b<1$, then for any mixture parameter $\lambda \in [0,1]$, $\Lambda(a,b) < H(\lambda a + (1-\lambda)b)$.*

The proof of this result can be found in Section 6.

3.1 Communicating logical information

In the first problem, illustrated at the top of Figure 2, a sender wishes to send a set of logical statements to the receiver, with the property that whatever logical inferences the sender can make starting from those facts, the receiver can do the same. The receiver in this problem is a "blank slate" – i.e., it knows no statements whatsoever. Crucially, the receiver is not required to reproduce the particular way in which the

sender's logical statements are described, but rather it just needs to be able to retrieve a set of statements that are functionally equivalent; this is why we refer to this problem as *communicating logical information*. The intention of this result is mainly to introduce notation and concepts as we build up to more interesting cases.

As an example, assume that Alice and Bob have agreed on a number of binary variables ($m = 3$), and Alice has in her possession the following facts relating the binary variables x_1, x_2, x_3:

$$\begin{aligned} x_1x_2x_3 &= 0; \\ (1+x_1)(1+x_2)(1+x_3) &= 0. \end{aligned} \tag{9}$$

Alice's goal is to efficiently communicate to Bob enough information for Bob to prove any fact about x_1, x_2, x_3 that can be deduced from Alice's facts. A first significant observation is that the Alice's representation of her knowledge is by no means unique. For example, Alice's facts could be represented alternatively by the single equation

$$x_1x_2 + x_1x_3 + x_2x_3 + x_1 + x_2 + x_3 + 1 = 0, \tag{10}$$

which, as the reader can check, is true if and only if (9) is true. Bob does not really care if, after transmission, he recovers (10) instead of (9). This leads to the idea that an efficient transmission of Alice's logical information to Bob should center on the *algebraic set* of Alice's facts, which is identical for both examples above.

We refer the reader to the top of Figure 3, which illustrates two sample algebraic sets $q^{'}$, $q^{''}$ with the property that $s \vDash q^{'}$ and $s \vDash q^{''}$. Since $\mathbf{v}(\hat{s})$ must include $\mathbf{v}(s)$ and be included within $\mathbf{v}(q^{'})$ and $\mathbf{v}(q^{''})$ (and for that matter, any $\mathbf{v}(q) \supseteq \mathbf{v}(s)$), the illustration suggests that we must find a way to send $\mathbf{v}(s)$ itself.

A solution to this problem is for Alice and Bob to agree ahead of time on a method for communicating the algebraic set, and to equip Alice with the necessary mechanism to deduce her algebraic set and for Bob to reconstruct any arbitrary alternate representation of Alice's facts consistent with the transmitted algebraic set. The fundamental problem we center on here is the first one - how do we transmit this algebraic set efficiently? A general way to do this might start by first Alice sending first to Bob the *size* of the algebraic set. Once this size is transmitted, all possible algebraic sets can be enumerated by both Alice and Bob and then Alice can simply send the index of the algebraic set she has in her possession. Techniques for enumerating and efficiently sending and receiving indices from such enumeration can be found in Cover's work on enumerative source coding [1].

This overall strategy is described in the bottom of Figure 3. One can even ask the question: is this algorithm optimal under some circumstance? The following theorem answers this question in a specific setting where we assume that the algebraic set that Alice has follows a simple "i.i.d. law", described in the Mathematical Preliminaries:

Theorem 1 (Logical information) *For any $m \geq 1$, let $S_m \in \mathcal{P}(K[x_1, \cdots, x_m])$ follow a p_s-i.i.d. law. Let the encoder f_m and decoder g_m be functions*

$$f_m : \mathcal{P}(K[x_1, \cdots, x_m]) \rightarrow \{0,1\}^*, \tag{11}$$

$$g_m : \{0,1\}^* \rightarrow \mathcal{P}(K[x_1, \cdots, x_m]), \tag{12}$$

respectively. Then

$$\Lambda(p_s, 1-p_s) = H(p_s) \leq \min_{f_m, g_m} |K|^{-m} E_S\left[\mathbf{len}(f(S_m))\right] \leq H(p_s) + O\left(\frac{\log |K|^m}{|K|^m}\right), \tag{13}$$

where the minimization is over f_m, g_m such that f_m is prefix free and such that, if $s \vDash q$, then $s \vDash g_m(f_m(s)) \vDash q$.

The proof of this result can be found in Section 6. Notice that the only requirements being placed on the encoder/decoder are that they have sets of logical statements (expressed as polynomial equations with binary coefficients and variables) as input and output, respectively, and that the receiver be able to prove anything that the sender would be able to prove. While the upper bound is in essence an estimate of a formal

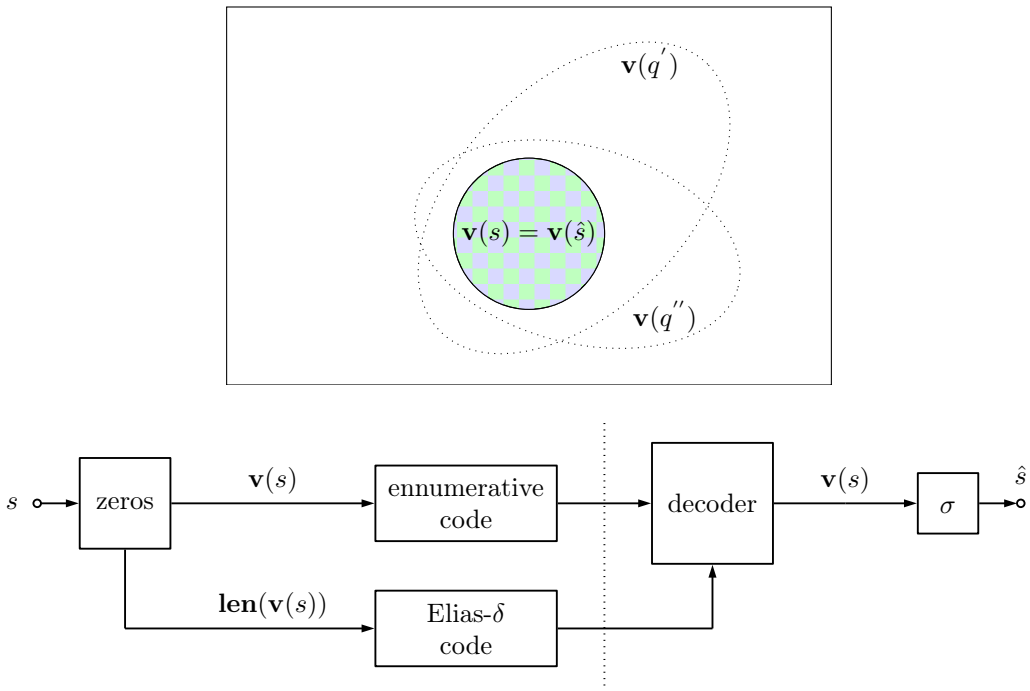


Figure 3: Proof strategy for Theorem 1.

version of the algorithm outlined above, the lower bound describes the limit that any possible algorithm that meets the requirements in the theorem will encounter. The reader may note that we introduced a seemingly unnecessary complication in the theorem statement through the equation $\Lambda(p_s, 1-p_s) = H(p_s)$ - the reason we do this is because the Λ function will play a role in more sophisticated versions of our problem.

3.2 Background mathematical statements and the optimality of incremental communications

To continue in our journey, we observe that in the earlier setup, the receiver had no prior mathematical statements whatsoever. But what if there is already pre-existing background with mathematical statements shared by the receiver and the sender? We will be using r to denote the set of polynomials serving as background and we will be making the simplifying assumption that $s \vDash r$. We refer the reader to Figure 4, which illustrates at the top the relationship between the various algebraic sets in question. In some respects, the effect that the presence of the background r has in the problem is rather elementary: we still want to send somehow $\mathbf{v}(s)$ however can now leverage the fact that both Alice and Bob know that such set must be contained within $\mathbf{v}(r)$. In the same figure we show an updated strategy: both Alice and Bob first compute $\mathbf{v}(r)$, and then Alice uses it to enumerate all possible subsets of $\mathbf{v}(r)$ of size $\mathbf{len}(\mathbf{v}(s))$ that happen to be contained within $\mathbf{v}(r)$, and sends the index of the algebraic set she has in her possession. Bob then recovers $\mathbf{v}(s)$ from the index that Alice sent Bob using that enumeration. Formally, this discussion is encapsulated in the following result:

Theorem 2 (Communicating logical information in the presence of background statements) *For any $m \geq 1$, let $S_m, R_m \in \mathcal{P}(K[x_1, \cdots, x_m])$ represent the sender's logical statements and the background logical statements, with the property that $S_m \vDash R_m$ and in particular, S_m, R_m follow a (p_s, p_r)-i.i.d. law. Let the encoder f_m and decoder g_m be functions*

$$f_m : \mathcal{P}(K[x_1, \cdots, x_m])^2 \rightarrow \{0,1\}^* \tag{14}$$

$$g_m : \{0,1\}^* \times \mathcal{P}(K[x_1, \cdots, x_m]) \rightarrow \mathcal{P}(K[x_1, \cdots, x_m]). \tag{15}$$

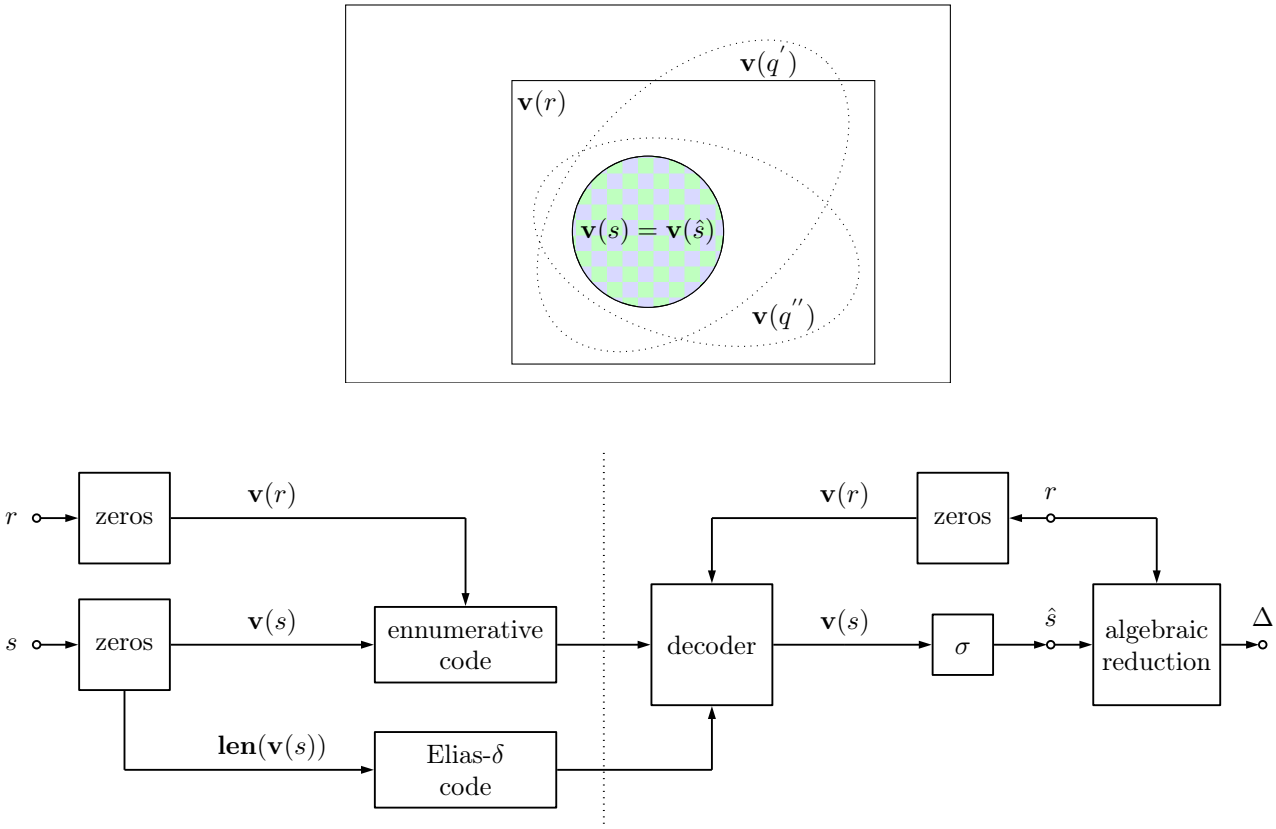


Figure 4: Proof strategy for Theorem 2 and Theorem 3.

Then

$$p_r H(p_s) \leq \min_{f_m, g_m} |K|^{-m} E_{S_m, R_m} \left[\mathbf{len}(f_m(S_m, R_m)) \right] \leq p_r H(p_s) + O\left(\frac{\log |K|^m}{|K|^m} \right), \tag{16}$$

where the minimization is over f_m, g_m such that the code implied by f_m is prefix free and such that, if $s \vDash r$ and $s \vDash q$, then $s \vDash g_m(f_m(s, r), r) \vDash q$.

The point to note from this result is that compared to that of Theorem 1 a multiplicative factor p_r reduces the total communication cost thanks to the presence of shared background logical statements.

3.2.1 Incremental communications

The reader may have noticed that in the way we set up Theorem 2, the decoder g combines both what Alice sent as well as the shared background information r to produce the *entirety* of Bob's beliefs, from which he can then prove whatever Alice could prove. However, the class of such decoders may be too large. A reasonable requirement is that whatever Alice sent to Bob should be presented to Bob as additional mathematical statements on top of the ones he already had in his mind. The difference between these two communication paradigms is illustrated in the second and third rows of Figure 2, where the output of the decoder in the third row is depicted as Δ, and where the state of Bob's mind after the communication is then $\Delta \cup r$.

We now convert the intuition above into a specific mathematical formulation. Let $\Delta, r \in \mathcal{P}(x_1, \cdots, x_m)$, then we say that Δ is an *incremental communication* over r if for any $w \in \mathcal{P}(x_1, \cdots, x_m)$ such that $r \vDash w$, then $w \notin \Delta$. Colloquially, if you are in possession of the facts r, Δ is an incremental communication over r if anything you can prove using r is not included in Δ. The result for incremental communications is then as follows:

Theorem 3 (Incremental communications) *For any $m \geq 1$, let $S_m, R_m \in \mathcal{P}(K[x_1, \cdots, x_m])$ represent the sender's logical statements and the background logical statements, with the property that $S_m \vDash R_m$ and in particular, S_m, R_m follow a (p_s, p_r)-i.i.d. law. Let the encoder f_m and decoder g_m be functions*

$$f_m : \mathcal{P}(K[x_1, \cdots, x_m])^2 \rightarrow \{0,1\}^* \tag{17}$$

$$g_m : \{0,1\}^* \times \mathcal{P}(K[x_1, \cdots, x_m]) \rightarrow \mathcal{P}(K[x_1, \cdots, x_m]). \tag{18}$$

Then

$$p_r H(p_s) \leq \min_{f_m, g_m} |K|^{-m} E_{S_m, R_m} \left[\mathbf{len}(f_m(S_m, R_m)) \right] \leq p_r H(p_s) + O\left(\frac{\log |K|^m}{|K|^m} \right), \tag{19}$$

where the minimization is over f_m, g_m such that the code implied by f_m is prefix free and such that, if $s \vDash r$ and $s \vDash q$, then $g_m(f_m(s,r), r)$ is an incremental communication over r and $s \vDash (g_m(f_m(s,r), r) \cup r) \vDash q$.

The most important observation to make about Theorems 2 and 3 is that the fundamental Shannon limit is exactly the same: $p_r H(p_s)$. This is, in our opinion, by no means evident, and thus we nick-name it as a result on the "optimality of incremental communications". The way Theorem 3 is proved is by taking the upper bound of Theorem 2 and then post-processing the output of the decoder so as to satisfy the restriction that the decoder output must innovate over r.

This is a very general pattern that goes well beyond this specific discussion, so it is worth pausing to make the tool for implementing the post-processing evident. In the lemma statement below, we use the letters u, v, w to denote generic sets of polynomials. The reader should identify u with the output of the decoder in Theorem 2 which is to be post-processed and v with the shared background information.

Lemma 3 (Post-processing to obtain incremental communications) *There exists a function*

$$\Delta : \mathcal{P}(K[x_1, \cdots, x_m])^2 \rightarrow \mathcal{P}(K[x_1, \cdots, x_m]) \tag{20}$$

such that for any $u, v, w \in \mathcal{P}(K[x_1, \cdots, x_m])$ with the property that if $u \vDash v$ and $v \vDash w$ then $w \notin \Delta(s, r)$ and $\mathbf{v}(\Delta(u, v) \cup v) = \mathbf{v}(u)$.

The proof of this Lemma can be found in Section 6.

3.3 Logical information in the context of a goal: the "less is more" theorems

We continue our journey towards making the communication patterns we are exploring more sophisticated. We refer the reader now to the last two rows of Figure 2. In here, the purpose of Alice is not to communicate to Bob enough to prove anything she can prove; rather it is to communicate to him the minimal amount of information needed to prove a *specific set of statements* q (and of course, having achieved that, any other statement that is entailed by q). The social situation one could picture in this setup is that of a teacher (Alice) that wants to equip a student (Bob) with the necessary logical information to prove anything that could be proved from q some time in the future. If one assumes that q is available to Bob at the time of communication, the problem becomes trivial, as all that Alice has to do at that point is to instruct Bob to incorporate q to its set of mathematical facts. The problem is highly non-trivial though if Bob is challenged to produce a proof for q only after the communication took place. Our work in this subsection is tightly connected with the results from Section 4, which are presented as a fully contained section involving core ideas from Shannon's rate-distortion theory.

We refer the reader to the top of Figure 5. In here we illustrate $\mathbf{v}(s)$ in green, and $\mathbf{v}(q)$ in white. One fairly obvious strategy to solve this problem would be to send to Bob enough information to reconstruct $\mathbf{v}(q)$; then Bob will be able to prove anything that can be proved starting from q. Alternately, one could send to Bob enough information to reconstruct $\mathbf{v}(s)$ itself, which would allow Bob to prove potentially even more mathematical statements. For the purposes of quantifying the cost of either of these two strategies, let's assume that S, Q follow a (p_s, p_q)-i.i.d. law (with $p_s < p_q$), then by choosing the best from either of these two strategies, we would be spending a total of

$$\min\{H(p_s), H(p_q)\} \tag{21}$$

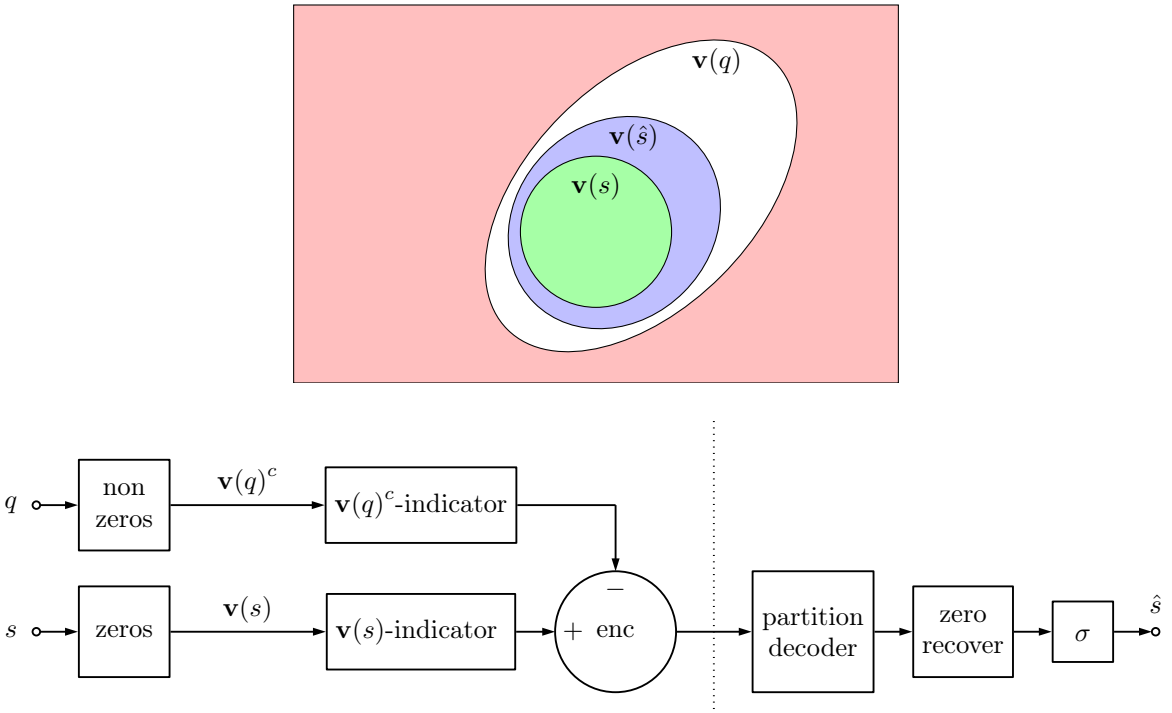


Figure 5: Proof strategy for Theorem 4

bits. Yet, as we will show soon, this strategy is in general *suboptimal*. This was a surprising revelation to us, so we want to equip the reader with the insight that we obtained when proving this result. From the top of Figure 5, it should be apparent that not only could Alice use either of the two strategies above to communicate to Bob; in fact she has the freedom to send any possible $\mathbf{v}(\hat{s})$ that satisfies $\mathbf{v}(s) \subseteq \mathbf{v}(\hat{s}) \subseteq \mathbf{v}(q)$; one such example $\mathbf{v}(\hat{s})$ is illustrated in blue in the Figure.

The existence of this freedom implies that in general, we can beat the estimate (21). This result is presented next:

Theorem 4 (Contextual logical information) *For any $m \geq 1$, let $S_m, Q_m \in \mathcal{P}(K[x_1, \cdots, x_m])$ represent the sender's logical statements and the query, with the property that $S_m \vDash Q_m$ and in particular, S_m, Q_m follow a (p_s, p_q)-i.i.d. law. Let the encoder f_m and decoder g_m be functions*

$$f_m : \mathcal{P}(K[x_1, \cdots, x_m])^2 \quad \to \quad \{0,1\}^* \tag{22}$$

$$g_m : \{0,1\}^* \quad \to \quad \mathcal{P}(K[x_1, \cdots, x_m]). \tag{23}$$

Then

$$\Lambda(p_s, 1-p_q) \leq \min_{f_m, g_m} |K|^{-m} E_{S_m, Q_m}\left[\mathbf{len}(f_m(S_m, Q_m))\right] \leq \Lambda(p_s, 1-p_q) + O\left(\frac{\log |K|^m}{|K|^m}\right), \tag{24}$$

where the minimization is over f_m, g_m such that the code implied by f_m is prefix free and such that, if $s \vDash q$, then $s \vDash g_m(f_m(s,q)) \vDash q$.

From the assumption that $p_q > p_s$, then $p_s + 1 - p_q < 1$ and thus then from Lemma 2, we obtain

$$\Lambda(p_s, 1-p_q) < \min\{H(p_s), H(1-p_q)\} = \min\{H(p_s), H(p_q)\} \tag{25}$$

A rather peculiar outcome form this result is that communication optimal strategies in general leave Bob with more knowledge than he strictly needed to prove q at the time in which he is challenged to do so.

This can be seen from the geometry implied in the algebraic sets on top of Figure 5. Notice that when $\mathbf{v}(\hat{s})$ is a strict subset of $\mathbf{v}(q)$ and of course contains $\mathbf{v}(s)$, we can prove all those mathematical facts $q^{'}$ whose algebraic set satisfies $\mathbf{v}(q^{'}) \subset \mathbf{v}(q)$ in a proper subset sense. None of those $q^{'}$ are such that $q \vDash q^{'}$, and thus it seems that Bob ended up learning more than Alice bargained to communicate.

Because of this counter-intuitive observation, we call this type of result a "less is more" theorem. The upper bound of Theorem 4 involves a novel form of encoding which is illustrated in the bottom of Figure 5 with a circle with two inputs, one of which is marked with a "+" symbol and the other one with a "-" symbol. The meaning of this device is that it produces an efficiently encoded index to a set that explicitly *includes* whatever set was passed under "+" and *excludes* whatever set was passed under "-". As it can be seen from the figure, the set passed to the positive hook is $\mathbf{v}(s)$, and the set to be excluded is $\mathbf{v}(q)^c$; these two are denoted in green and red, respectively, in the top of the Figure. This device is discussed in more length in Section 4, and it is where the most profound connection to Shannon's rate-distortion theory emerges. In particular, we would like to steer the reader's attention to our result on linear codes (subsection 4.2) for implementing systems that approach the limits specified in Theorem 6, which is the fundamental block for both of our "less is more" theorems.

3.3.1 A "less is more" theorem that accounts for ignorance and misinformation

We conclude our series of results by adding to the result from Theorem 4 the possibility that the Bob has background information that Alice is aware of, just like Theorem 2 innovated over Theorem 1. However, to keep matters interesting, we bring yet an additional level of sophistication to the problem set up. In particular, in this result we will no longer make the assumption that $s \vDash r$, this is, that the receiver side information is logically consistent with that of s.

To understand the proposed situation, we ask the reader to refer to the top of Figure 6. In the left, we illustrate the same situation that we had in the top of Figure 4, but with the addition of background information r with the property that $s \vDash r$. Now suppose that it is no longer the case that $s \vDash r$, then what we obtain is the more general setup to the right of the top of Figure 5, where as it can be appreciated, $\mathbf{v}(r)$ no longer contains $\mathbf{v}(s)$ fully and where the complement of $\mathbf{v}(r)$ is patterned with a "dotted" fill.

Our result for this scenario is presented next:

Theorem 5 (Conditional logical information in the context of a goal) *For any $m \geq 1$, let S_m, Q_m, R_m represent the sender's logical statements, the query and the receiver logical statements, respectively, with the property that $S_m \vDash Q_m$ and in particular, S_m, Q_m, R_m follow a $(p_r, p_{s|r}, p_{q|r}, p_{s|\tilde{r}}, p_{q|\tilde{r}})$-i.i.d. law. Let the encoder f_m and decoder g_m be functions*

$$f_m : \mathcal{P}(K[x_1, \cdots, x_m])^3 \quad \rightarrow \quad \{0,1\}^* \tag{26}$$

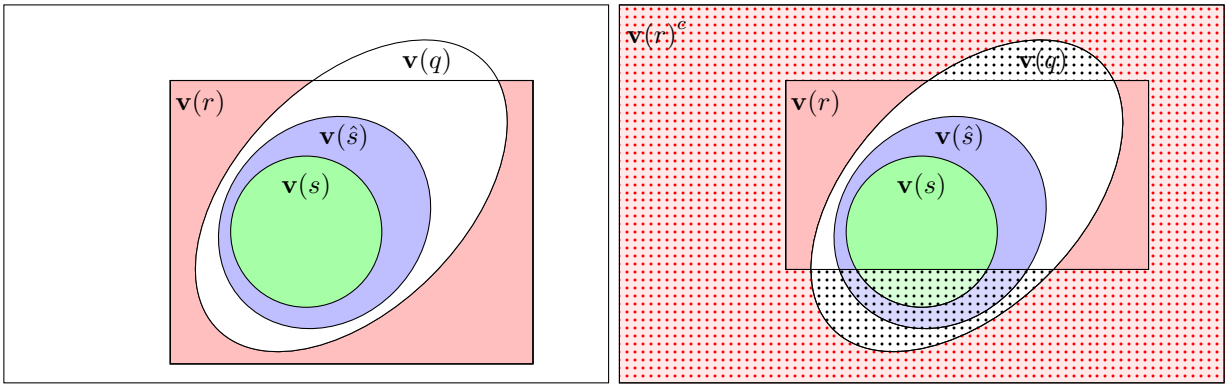
$$g_m : \{0,1\}^* \times \mathcal{P}(K[x_1, \cdots, x_m]) \quad \rightarrow \quad \mathcal{P}(K[x_1, \cdots, x_m]). \tag{27}$$

Then

$$\mathcal{L} \leq \min_{f_m, g_m} |K|^{-m} E_{S_m, Q_m, R_m} \left[\mathbf{len}(f_m(S_m, Q_m, R_m)) \right] \leq \mathcal{L} + O\left(\frac{\log |K|^m}{|K|^m} \right) \tag{28}$$

where $\mathcal{L} = p_r \Lambda(p_{s|r}, 1 - p_{q|r}) + (1 - p_r)\Lambda(p_{s|\tilde{r}}, 1 - p_{q|\tilde{r}})$, and where the minimization is over f_m, g_m such that the code implied by f_m is prefix-free, and such that if $s \vDash q$ then $s \vDash g_m(f_m(s, q, r), r) \vDash q$.

The strategy for proving the upper bound for this result is illustrated in Figure 6. While the figure appears formidable, upon further examination its elements are quickly decomposed into elements that should be familiar to the reader now. At the highest level, the problem is simply split in two: because both Alice and Bob know r, they can create a dual strategy: one to handle sending whatever piece of $\mathbf{v}(\hat{s})$ that will intersect with $\mathbf{v}(r)$, and the other one to handle the same but that intersects with $\mathbf{v}(r)^c$; this explains why there is vertical symmetry on the figure. Then, focusing on, say, only the top half of the diagram, we realize that the resulting system is in essence a combination of the strategies used in Theorems 2 and 4. As a result, the fundamental device for efficiently sending partitions, denoted by a circle with the $+$ and $-$ hooks (and fully addressed in Theorem 6) is used twice.



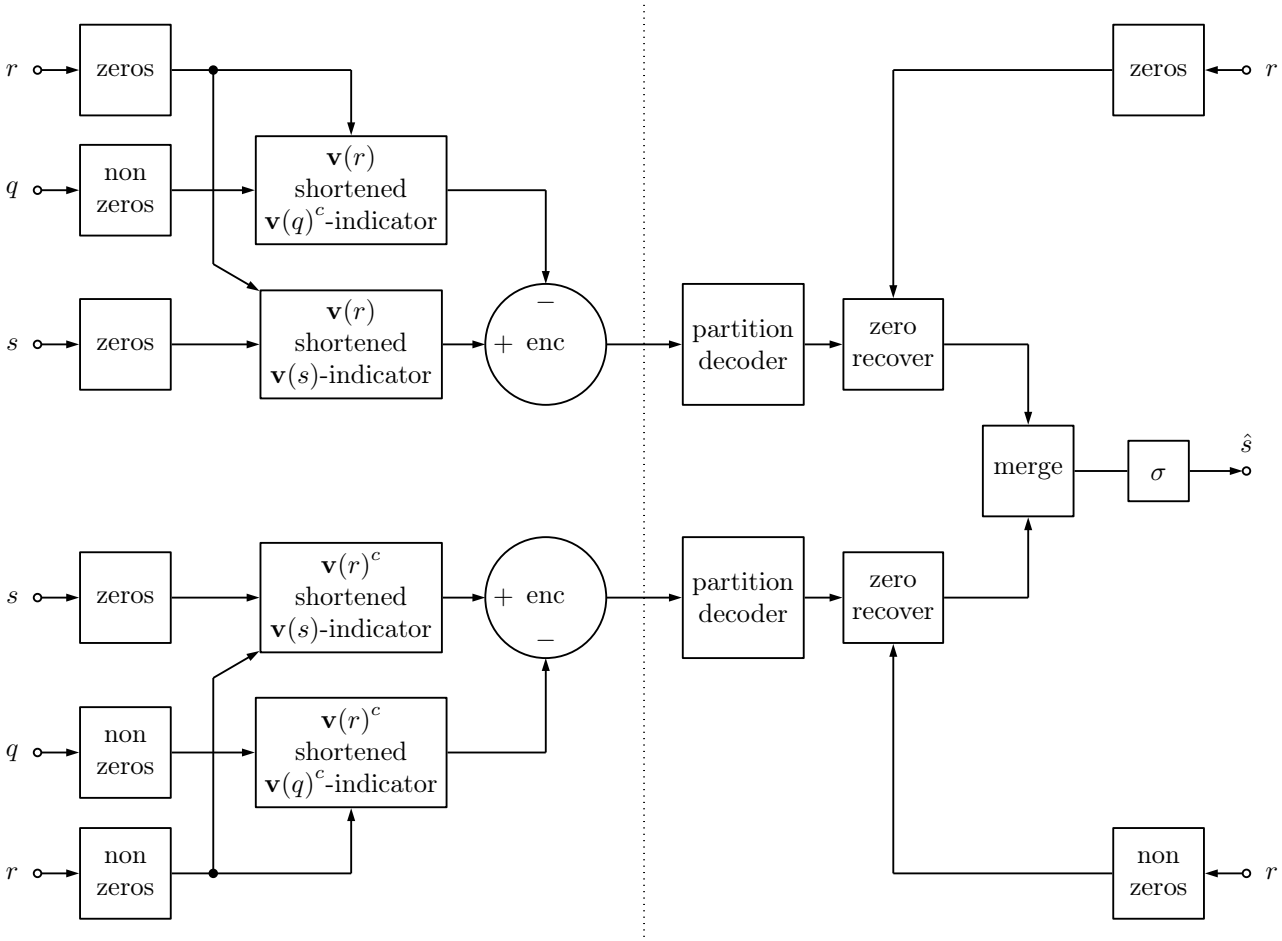


Figure 6: Proof strategy for Theorem 5

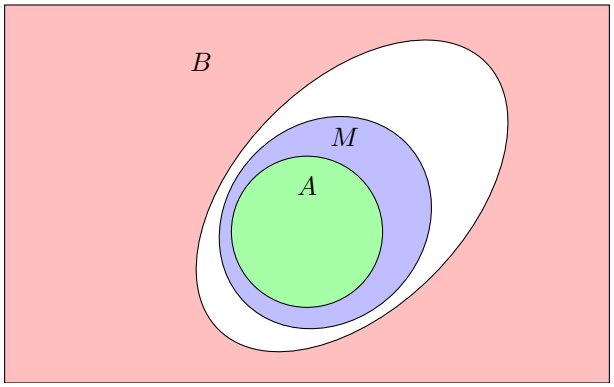


Figure 7: An example of a partition $\{M, M^c\}$ that separates A from B.

The discussion above should help explain to the reader the physical meaning of the terms in the expression below:

$$p_r \Lambda(p_{s|r}, 1 - p_{q|r}) + (1 - p_r)\Lambda(p_{s|\tilde{r}}, 1 - p_{q|\tilde{r}}) \tag{29}$$

The first term can be seen as a weighted version of the term that had already surfaced in Theorem 4; the weighting factor is simply accounting for the logical information that Bob has that happens to match Alice's logical information present in s. One may regard this term as accounting for the cost of improving a state of *ignorance* that Bob has with respect to Alice's knowledge. The second term is also very similar looking, but in this case, one may regard it as explicitly accounting for the cost of correcting a state of *misinformation* that Bob has with respect to Alice. One shall not read too much from our use of these words in terms of potential applicability to important ongoing societal problems; our models are at best very rudimentary approximations to specific communication paradigms and we make no claim of their applicability to real scenarios.

We finally note that Theorem 5 is in fact a superset of Theorems 1, 2 and 4, as those can be retrieved from it by specializing properly. First, we note that Theorem 4 corresponds to the case where $p_r = 1$, and then choosing $r = \emptyset$ when substituting in f_m, g_m. Next, Theorem 2 is obtained by setting $p_{s|r} = p_{q|r}$, and plugging s wherever q appears. Finally, 1 is obtained by combining the two strategies above.

4 The partition compression problem

In this section, we discuss a data compression problem of central relevance to the problems considered in this article. This section can be largely be read in isolation from the other parts of the paper, however its deep meaning to us is only revealed when seeing how we use this section's results in context. We will introduce both the connection to Rate-Distortion theory, as well as preliminary work towards practical techniques for implementing solutions to this problem. We refer the reader to Figure 7. Imagine one has n items as well as two non-intersecting subsets of those n items, which we call A and B, and one is interested in sending efficiently to a receiver a subset M which contains A but excludes B, or alternately stated, a partition that separates A from B. One solution is to send $M = A$ or $M = B$, whichever is cheapest to send, but it turns out there's generally a better solution. Let $X^n \in \{0, 1, \otimes\}^n$ be a vector with $X_i = 0$ if the ith element is inside of A, $X_i = 1$ if the ith element is inside of B and $x = \otimes$ if the ith element is neither in A nor in B. Assume that the entries of X^n are distributed i.i.d. with a distribution $(p_a, p_b, 1 - p_a - p_b)$. Define a distortion metric to be a function

$$\rho : \{0, 1, \otimes\} \times \{0, 1\} \to \{0, 1\} \tag{30}$$

using this matrix:

	0	1
0	0	1
1	1	0
$\otimes$	0	0

(31)

An encoder and decoder are functions $f_n : \{0,1,\otimes\}^n \to \{0,1\}^*$ and $g_n : \{0,1\}^* \to \{0,1\}^n$, respectively. The problem is to find upper and lower bounds for

$$\min_{f_n,g_n} n^{-1} E_{X^n} \left[\mathbf{len}(f_n(X^n))\right]$$

subject to the condition that for any $x_1^n \in \{0,1,\otimes\}^n$,

$$\sum_{i=0}^{n-1} \rho(x_i, g_n(f_n(x_1^n))_i) = 0 \tag{32}$$

The solution to this problem can be obtained by an application of Shannon's Rate-Distortion theory:

Theorem 6 (Shannon limit for partition compression) *For any $n \geq 1$, let $X_1^n \in \{0,1,\otimes\}^n$ be a random vector with i.i.d. entries each distributed according to a law $(p_a, p_b, 1-p_a-p_b)$. Let f_n, g_n be encoder and decoder functions as defined earlier. Then*

$$\Lambda(p_a,p_b) \leq \min_{f_n,g_n} n^{-1} E_X \left[\mathbf{len}(f_n(X_1^n))\right] \leq \Lambda(p_a,p_b) + 2\frac{\log_2\left(n\Lambda(p_a,p_b)\right)}{n} + \frac{3}{n}$$

where the minimization is over f_n, g_n, such that the code implied by f_n is prefix-free and that condition (32) is met.

From Lemma 2, if $p_a + p_b < 1$, then we can deduce that $\Lambda(p_a,p_b) < \min\{H(p_a), H(p_b)\}$, and thus this result predicts the existence of partition compression techniques which are more efficient than the "naive" solution of sending the cheapest of the sets (A or B). Fortunately, it is not very difficult to obtain examples of this phenomenon.

Consider the case $n = 6$, and where the sets A and B consist each of precisely only one of the numbers $\{0,\cdots,5\}$. One way to construct a partition that separates A from B is to simply, say, send the index of the singleton member of A. This clearly requires a total of $\log_2 6$ bits in average. Let's identify membership in A with a 0 and membership in B with a 1. Observe that any two columns of the following binary matrix contain at least one row with the pattern `"0 1"` and another row with the pattern `"1 0"`:

```
0 0 0 1 1 1
0 1 1 0 1 0
1 0 1 1 0 0
1 1 0 0 0 1
```

Therefore, for any possible assignment of different singleton items to A and B we can find at least one of four rows which partition A and B. Assuming both sender and receiver have the matrix above, then sending a row requires 2 bits, which is less than $\log_2 6$ bits. The reader may point out that this simple example makes a strong assumption in that the sizes of the A, B sets are known to both sender and receiver. This can be addressed by sending the size itself as well, which for larger problems, becomes a negligible overhead.

This small example can be in fact extended easily. Notice that the columns of the matrix are exactly the set of all binary patterns with 2 ones (in the parlance of coding theory, the columns have weight 2), and thus this poses the interesting question of what the properties are of matrices whose column weights is constant. One such property is easy to deduce, as stated in the following result:

Lemma 4 (Constant column weight codes) *Let c be a $t \times n$ binary matrix where every column has exactly the same weight w, and any two columns are different. Then the result code partitions sets a, b each comprising exactly one (but different) integer in the set $\{0,\cdots,n-1\}$.*

Proof. Let i, j be the indices of any two distinct columns of the matrix c. The problem is to demonstrate that there is a row k such that $[c_{k,i}, c_{k,j}] = [0, 1]$ and that there is another row $k^{'}$ where $[c_{k,i}, c_{k,j}] = [1, 0]$. Suppose that neither of these conditions is true, then we deduce that $[c_{k,i}, c_{k,j}] \in \{[0, 0], [1, 1]\}$ for all $0 \leq k < t$ and thus necessarily the two columns are identical, which contradicts the assumption of the Lemma. Suppose that, say, the first condition is true, but not the second one. Then it must be the case that the second column indexed by j has a strictly larger weight than the column indexed by i, which is also a contradiction of the assumptions in the Lemma. The case where the second condition is true but not the first one is dealt with similarly.

4.1 Achievability proof for Theorem 6

In this subsection we give a proof of the achievability result (upper bound) in Theorem 6. While we could have simply invoked a result from the Rate-Distortion literature, we choose to provide a proof because our particular restricted setting allows us to provide a much simpler proof including a useful second order upper bound which will play a role in our other proofs. We will not provide a proof of the lower bound of Theorem 6, as this is a special case of the classic $R(D)$ literature, and anyhow our other proofs will provide their own specialized lower bound. As we will show, the expected performance of a random matrix with the density of ones appropriately tuned will asymptotically approach the Shannon limit $\Lambda(p_a, p_b)$. This implies, as per the classic random coding argument of Shannon, the existence of a deterministic code with performance at least as good as the expected performance of the random code.

Let $X_1, \cdots, X_n \in \{0, 1, \otimes\}^n$ be an input random vector, and let $\Psi \subseteq \{0, \cdots, n-1\}$ denote the random positions where X_i is taking on 0 or 1, and thus where we want to enforce a bit pattern. For any arbitrary $z \in \{0, 1\}^n$, let $[z]_\Psi$ denote the $1 \times |\Psi|$ vector obtained by extracting from z the columns indexed by Ψ.

Let C be a random binary matrix with an infinite number of rows $\{C_1, C_2, \cdots\}$ and each with n columns. Assume its entries are chosen i.i.d. according to the distribution

$$P(C_{i,j} = 0) = \frac{p_a}{p_a + p_b} \tag{33}$$

The sender scans the matrix C from top to bottom until it finds the first row J that satisfies the following condition:

$$[C_J]_\Psi = X_\Psi \tag{34}$$

and then sends the index J of that row. The receiver then recovers the row from the index. Let $N_0 = \{i : X_i = 0\}$ and $N_1 = \{i : X_i = 1\}$. The probability of a row of C satisfying (34), conditional on Ψ, is

$$P([C_i]_\Psi = X_\Psi | \Psi) = \left(\frac{p_a}{p_a + p_b}\right)^{N_0} \left(\frac{p_b}{p_a + p_b}\right)^{N_1} \tag{35}$$

and therefore

$$E\left[J|\Psi\right] = 2^{\left(-N_0 \log_2 \frac{p_a}{p_a+p_b} - N_1 \log_2 \frac{p_b}{p_a+p_b}\right)} \tag{36}$$

To send the index J we will be using Elias' δ encoding. Note that $E_\Psi\left[\log_2\left(E\left[J|\Psi\right]\right)\right] = n\Lambda(p_a, p_b)$. We can then upper bound the performance of the code as

$$\begin{aligned}
E_J\left[\mathbf{len}(\mathrm{elias}_\delta(J))\right] &= E_\Psi\left[E_J\left[\mathbf{len}(\mathrm{elias}_\delta(J))|\Psi\right]\right] && (37)\\
&\leq E_\Psi\left[E_J\left[\log_2 J + 2\log_2(\log_2 J) + 3\ |\Psi\right]\right] && (38)\\
&\leq E_\Psi\left[\log_2 E_J\left[J|\Psi\right] + 2\log_2(\log_2 E_J\left[J|\Psi\right]) + 3\right] && (39)\\
&\leq E_\Psi\left[\log_2 E_J\left[J|\Psi\right]\right] + 2\log_2(E_\Psi\left[\log_2 E_J\left[J|\Psi\right]\right) + 3 && (40)\\
&\leq n\Lambda(p_a, p_b) + 2\log_2(n\Lambda(p_a, p_b)) + 3 && (41)
\end{aligned}$$

where in addition to the estimate of the performance of Elias' code, we used the concavity $\cap$ of the logarithm. This concludes the proof of the upper bound.

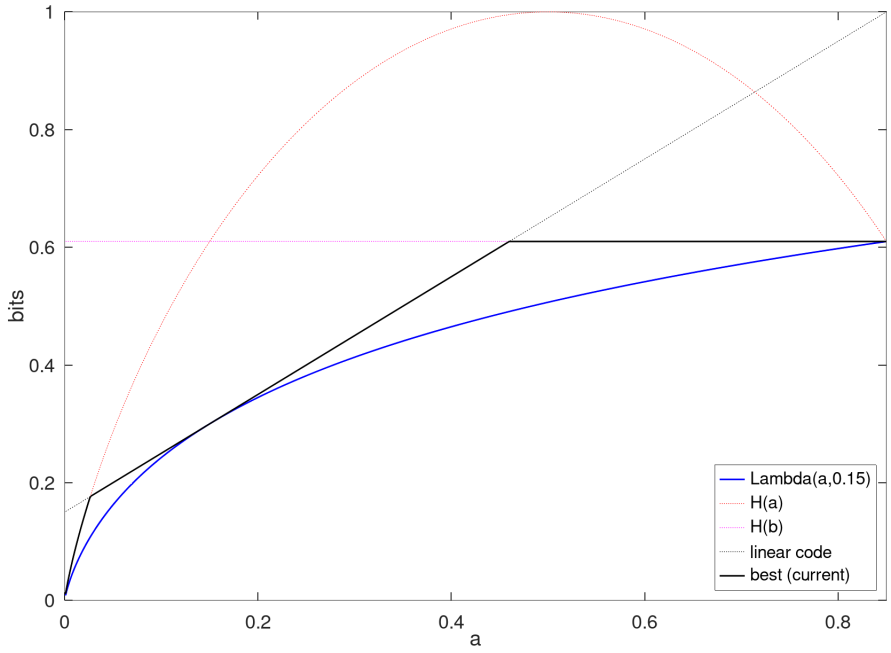


Figure 8: A comparison of the fundamental $\Lambda(a, b)$ bound to currently known practical techniques for implementing partition compression.

4.2 Linear codes for partition compression

We now present a general algorithm for partition compression based on linear codes over $GF(2)$. This algorithm can be applied even when we don't have any statistical model of the underlying sets that we are creating a partition for. In the special case where these sets are chosen according to an $(p_a, p_b, p_\otimes)$-i.i.d. law with $p_a = p_b$, we will show that this algorithm is asymptotically optimal. The attractiveness of linear codes stems from the fact they are easier to implement in practice, in contrast with the the existence proof presented in subsection 4.1 in which search over an exponentially large list was required.

Similar to Subsection 4.1, in this algorithm, we assume that encoder and decoder share a random matrix with n columns and an infinite number of rows, which we shall refer to as G; the entries of this matrix will be in $GF(2)$. However the number of rows we will be effectively be using is approximately the logarithm of the number of rows in the earlier existence result. Let G_r denote the matrix obtained by extracting the first r rows from G. The algorithm starts by finding the first row J such that the equation

$$[M \cdot G_J]_\Psi = X_\Psi \tag{42}$$

can be solved for some $M \in \{0,1\}^{1\times J}$. The sender then sends the integer J to the receiver using

$$\mathbf{len}(\text{elias}_\delta(J)) \tag{43}$$

bits, followed by the J bits in the message M. The receiver then decodes J, and then computes $M \cdot G_J$ to retrieve the partition.

We now analyze the expected performance of this algorithm:

$$\begin{aligned} n^{-1}E_J\left[\mathbf{len}(\text{elias}_\delta(J))\right] &\leq E_J\left[J + \log_2 J + 2\log_2(\log_2 J) + 3\right] && (44)\\ &= n^{-1}E_\Psi\left[E_J\left[J + \log_2 J + 2\log_2(\log_2 J) + 3|\Psi\right]\right] && (45)\\ &\leq n^{-1}E_\Psi\left[E_J\left[J|\Psi\right] + \log_2 E_J\left[J|\Psi\right] + 2\log_2(\log_2 E_J\left[J|\Psi\right]) + 3\right] && (46)\end{aligned}$$

We next upper bound $E_J\left[J|\Psi\right]$. A sufficient condition to be able to solve Equation (42) is that G_J has full

column rank. The expected number of rows (conditional on Ψ) until this happens is given by

$$E_J\left[J|\Psi\right] = \sum_{i=0}^{|\Psi|-1} \frac{1}{1-2^{i-|\Psi|}} \leq |\Psi| + 2 \tag{47}$$

Continuing from inequality (46), we obtain an upper bound on performance of

$$n^{-1} E_\Psi \left[|\Psi| + \log_2(|\Psi|+2) + 2\log_2\left(\log_2(|\Psi|+2)\right) + 5\right] \tag{48}$$

To understand how good this is, we assume an $(p_a, p_b, p_\otimes)$-i.i.d. law for the sets A, B and therefore $E\left[|\Psi|\right] = n(p_a + p_b)$, which results in the following upper bound on average performance:

$$p_a + p_b + \frac{\log_2\left(n(p_a+p_b)+2\right)}{n} + 2\frac{\log_2\left(\log_2\left(n(p_a+p_b)+2\right)\right)}{n} + \frac{5}{n} \tag{49}$$

Now assume that $p_a = p_b$. The Shannon limit for this setting is given by

$$\Lambda(p_a, p_b) = \Lambda(p_a, p_a) = 2p_a \tag{50}$$

which is the same performance as in (49) asymptotically as n grows. If $p_a \neq p_b$ then $\Lambda(p_a, p_b) < p_a + p_b$ and thus our linear code construction is not optimal.

We refer the reader to Figure 8 which demonstrates the performance of three different coding strategies: send the set A, send the set B, our use the linear code above, together with the corresponding Shannon bound.

5 Concluding remarks

We started our article by posing a motivating question: *What is the information content in a logical statement?*. We recognize this to be an ambitious question for which we have barely scratched the surface. Our main strategy here was to recognize that this question would inevitably lead us to identify what is essential in a logical statement, and therefore what makes any two statements functionally distinguishable. We recognized the next step to be one where we needed to identify a mathematical framework where we could pose the distinguishability property in a crisp manner, and we choose to focus on a simple logical system which was easy to map to the mathematics of multivariate polynomials over finite fields. From there, it was relatively straightforward to produce sharp information theoretic results for a few communication patterns, including a nice result on the optimality of incremental communication in a specific communication setting, which we believe is an example of how we might approach the general task of modeling more accurate physical communication settings. We then considered a more complex setting where the goal for the communication is narrower, and were surprised to discover the "less is more" theorems which revealed a profound connection to Shannon's rate-distortion theory, and to the discovery of a new function, which we call the Λ function, which appears to play a role in these logic communication problems. We also made some partial progress towards the identification of practical methods for implementing algorithms that approach the Shannon limits in the "less is more" theorems.

We believe that our results do offer a compelling proposal of how ideas from the fields of logic, information theory and algebra can be combined to give preliminary answers to very intriguing questions. We hope that the readers feel inspired by these initial thoughts to follow up with their unique perspective.

6 Proofs

6.1 Proof of Lemma 2

Lemma 2 (Elementary properties of $\Lambda(a,b)$) *The function $\Lambda(a,b)$ is concave $\cap$ over the domain $[0,+\infty)\times[0,+\infty)$. If $\Delta_a, \Delta_b \geq 0$ with at least one of them being strictly positive, then $\Lambda(a+\Delta_a, b+\Delta_b) > \Lambda(a,b)$. If $a+b<1$, then for any mixture parameter $\lambda \in [0,1]$, $\Lambda(a,b) < H(\lambda a + (1-\lambda)b)$.*

Proof. From its definition, it is clear that $\Lambda(a,b) \geq 0$ for $a,b \geq 0$. The gradient and Hessian of $\Lambda(a,b)$ are given by

$$\nabla\Lambda(a,b) = \begin{bmatrix} \log\frac{a+b}{a} \\ \log\frac{a+b}{b} \end{bmatrix}, \qquad \text{Hess}(\Lambda(a,b)) = \begin{bmatrix} -\frac{b}{a(1+b)} & \frac{1}{a+b} \\ \frac{1}{a+b} & -\frac{a}{b(a+b)} \end{bmatrix}. \tag{51}$$

Over $a,b \in (0,\infty)$, the gradient is strictly positive. As a consequence, the function is monotonically increasing when one of the arguments is fixed, and thus $\Lambda(a,b) \leq \Lambda(a,b+\Delta_b) \leq \Lambda(a+\Delta_a, b+\Delta_b)$ with at least one of those inequalities being strict due to the assumption. Moreover, the Hessian is positive semi-definite since its eigenvalues are 0 and $-\frac{a^2+b^2}{ab(a+b)}$, the latter always being strictly negative in the same domain. From this we conclude that the function is concave $\cap$. Because $\Lambda(a,0) = \Lambda(0,b) = 0$, the Lemma assertions can be extended to the full domain $[0,+\infty)\times[0,+\infty)$. To prove the last statement, first note that $\Lambda(a,b) = \Lambda(b,a)$. For a given $\lambda \in [0,1]$, let $a_\lambda = \lambda a + (1-\lambda)b$ and $b_\lambda = \lambda b + (1-\lambda)a$. Due to the concavity $\cap$ of the Λ function, we deduce that

$$\Lambda(a,b) = \lambda\Lambda(a,b) + (1-\lambda)\Lambda(b,a) \leq \Lambda(a_\lambda, b_\lambda). \tag{52}$$

Next note that, since $a+b<1$, then $a_\lambda + b_\lambda < 1$, and from the monotonicity property proved earlier we have

$$\Lambda(a_\lambda, b_\lambda) < \Lambda(a_\lambda, 1-a_\lambda) = H(a_\lambda). \tag{53}$$

6.2 Proof of Theorem 1

Theorem 1 (Logical information) *For any $m \geq 1$, let $S_m \in \mathcal{P}(K[x_1,\cdots,x_m])$ follow a p_s-i.i.d. law. Let the encoder f_m and decoder g_m be functions*

$$f_m : \mathcal{P}(K[x_1,\cdots,x_m]) \rightarrow \{0,1\}^*, \tag{11}$$
$$g_m : \{0,1\}^* \rightarrow \mathcal{P}(K[x_1,\cdots,x_m]), \tag{12}$$

respectively. Then

$$\Lambda(p_s,1-p_s) = H(p_s) \leq \min_{f_m,g_m} |K|^{-m} E_S\left[\mathbf{len}(f(S_m))\right] \leq H(p_s) + O\left(\frac{\log|K|^m}{|K|^m}\right), \tag{13}$$

where the minimization is over f_m, g_m such that f_m is prefix free and such that, if $s \vDash q$, then $s \vDash g_m(f_m(s)) \vDash q$.

Proof. We begin with the proof of the lower bound. Let f_m, g_m satisfy the conditions for the minimization. The starting point for this proof is the classical result from information theory proved using Kraft's inequality as follows.

Lemma 5 *Let $\{l_i\}$ be the codeword lengths of a binary code that is prefix free. Assume a distribution over these codewords, and let C be a random codeword drawn according to that distribution. Then*

$$E_C\left[\mathbf{len}(C)\right] \geq H(C). \tag{54}$$

Using this result together with the assumption that the code implied by f_m is prefix free, we can write

$$E_S\left[\mathbf{len}(f_m(S))\right] = E_{f_m(S)}\left[\mathbf{len}(f_m(S))\right] \geq H(f_m(S)) \geq H(g_m(f_m(S))) \geq H(\mathbf{v}(g_m(f_m(S)))), \tag{55}$$

where the last two inequalities follow from the fact that deterministic functions of random variables cannot increase entropy. Next, due to the assumptions for any q such that $\mathbf{v}(s) \subseteq \mathbf{v}(q)$, we have

$$\mathbf{v}(s) \subseteq \mathbf{v}(g_m(f_m(s))) \subseteq \mathbf{v}(q).$$

In particular, choosing $Q = S$ we obtain $\mathbf{v}(S) = \mathbf{v}(g_m(f_m(S)))$. Substituting this in the right hand side of (55), we conclude

$$E_S\left[\mathbf{len}(f_m(S))\right] \geq H(\mathbf{v}(S)) = |K|^m H(p_s) \tag{56}$$

due to the assumption that S follows a p_s-i.i.d. law. This completes the proof of the lower-bound result.

Turning to the proof of the upper bound, we first specify the encoder f_m. For a given size ξ of an algebraic set, there are a total of

$$\left(\begin{array}{c} |K|^m \\ \xi \end{array} \right) \leq 2^{|K|^m H\left(\frac{\xi}{|K|^m}\right)} \tag{57}$$

possible algebraic sets of this size. Let $\mathrm{enum}_\xi : \{r \in \mathcal{P}(\{0,1\}^m) : |r| = \xi\} \to \{0,1\}^*$ be a function that maps each possible algebraic set of size k to a fixed-length binary encoding of the integers $\left\{1, \cdots, \left(\begin{array}{c} |K|^m \\ \xi \end{array} \right)\right\}$, which is an integer that uniquely determines such a set. Hence, in particular,

$$\mathbf{len}(\mathrm{enum}_{|\mathbf{v}(S)|}(\mathbf{v}(S))) \leq |K|^m H\left(\frac{|\mathbf{v}(S)|}{|K|^m}\right) + 1, \tag{58}$$

where we regard in the above, the binary entropy function $H(p) = -p\log_2(p) - (1-p)\log_2(1-p)$ is evaluated on the random variable $|\mathbf{v}(S)|/|K|^m$, and thus the result of the evaluation is also a random variable.

For a given algebraic set $r \in \mathcal{P}(\{0,1\}^m)$ we define the encoder f_m as a concatenation of two separate encodings:

$$f_m(s) = \mathrm{elias}_\delta(|\mathbf{v}(s)|)\mathrm{enum}_{|\mathbf{v}(s)|}(\mathbf{v}(s)) \tag{59}$$

Since both of the codes implied by each encoding are prefix-free, the concatenation is also prefix free. Finally, we let g_m be the decoder that recovers the algebraic set $\mathbf{v}(s)$ from the output of f_m, and then returns $\sigma(s)$ (c.f. Lemma 1).

We note that by construction, if q is such that $s \vDash q$, then since $\mathbf{v}(s) = \mathbf{v}(g_m(f_m(s))$, then $g_m(f_m(s)) \vDash q$ as well, and thus we have met the conditions of the Theorem.

The estimate for the overall cost of the encoding can be done by estimating each of the terms in (59) independently:

$$|K|^m H\left(\frac{|\mathbf{v}(S)|}{|K|^m}\right) + \log_2|\mathbf{v}(S)| + 2\log_2\left(\log_2|\mathbf{v}(S)|\right) + 4 \tag{60}$$

bits. Taking the expectation with respect to S, using the concavity $\cap$ of the logarithm and entropy function, and normalizing by $|K|^m$, we obtain an upper estimate of

$$H(p_s) + \frac{\log_2\left(p_s|K|^m\right)}{|K|^m} + 2\frac{\log_2\left(\log_2\left(p_s|K|^m\right)\right)}{|K|^m} + \frac{4}{|K|^m} \tag{61}$$

This completes the proof of the Theorem.

6.3 Proofs for Theorems 2 and 3

Theorem 2 is a special case of Theorem 5, where we make the query q precisely coincide with q and where we eliminate the possibility of misinformation by setting $p_r = 1$.

Since the space of all possible encoder/decoder pairs in Theorem 3 is a subset of those in Theorem 2, the lower bound for the latter is also a lower bound for the former. Therefore, all that remains is to prove the upper bound for Theorem 3. As previewed, this is done by post-processing the output of the decoder in the upper bound of Theorem 2 using the following result, for which we now provide a proof:

Lemma 3 (Post-processing to obtain incremental communications) *There exists a function*

$$\Delta : \mathcal{P}(K[x_1, \cdots, x_m])^2 \to \mathcal{P}(K[x_1, \cdots, x_m]) \tag{20}$$

such that for any $u, v, w \in \mathcal{P}(K[x_1, \cdots, x_m])$ with the property that if $u \vDash v$ and $v \vDash w$ then $w \notin \Delta(s, r)$ and $\mathbf{v}(\Delta(u, v) \cup v) = \mathbf{v}(u)$.

Proof. We need a way to remove polynomials w from u such that $v \vDash w$. First, we observe that $v \vDash w$ if and only if w is contained in the ideal generated by $\{v, x_1^{|K|} - x_1, \ldots, x_m^{|K|} - x_m\}$. Next we need to choose a total ordering on the monomials in $\mathcal{P}(K[x_1, \cdots, x_m])$ that respects multiplication, i.e., if $m_1 < m_2$ then $mm_1 < mm_2$ and $1 \leq m$ for all monomials m. This ordering provides a leading term for each polynomial. Given polynomials p and q, if some monomial of q is divisible by the leading term of p, we can remove that monomial by subtracting a multiple of p from q. We can continue this process until no monomials of q are divisible by the leading term of p, at which point we say that q is reduced with respect to p. Similarly, given a finite collection of polynomials $\{p_1, \ldots, p_k\}$, we will call a polynomial q reduced with respect to the p_i's if no monomial of q is divisible by a leading monomial of some p_i. Among all sets of generators G_I of an ideal I, those which have the property that $p \in I$ if and only if p can be reduced to zero by G_I are called a Gröbner basis for the ideal I. Let G_v be a Gröbner basis for the ideal generated by $\{v, x_1^{|K|} - x_1, \ldots, x_m^{|K|} - x_m\}$. We define $\Delta(u, v)$ to be the reduction of the generators of u with respect to G_v. The reduction process guarantees that all polynomials w such that $v \vDash w$ will reduce to zero, and thus they are no longer contained in $\Delta(u, v)$. Since $\mathbf{v}(u) \subseteq \mathbf{v}(v)$, we have $\mathbf{v}(u \cup v) = \mathbf{v}(u)$. Since all solutions of v lie in K^m, we obtain $\mathbf{v}(v) = \mathbf{v}(v, x_1^{|K|} - x_1, \ldots, x_m^{|K|} - x_m)$, so $\mathbf{v}(u \cup \{v, x_1^{|K|} - x_1, \ldots, x_m^{|K|} - x_m\}) = \mathbf{v}(u)$. But $\Delta(u, v)$ and u are the same modulo $\{v, x_1^{|K|} - x_1, \ldots, x_m^{|K|} - x_m\}$, and thus $\mathbf{v}(\Delta(u, v) \cup v) = \mathbf{v}(\Delta(u, v) \cup \{v, x_1^{|K|} - x_1, \ldots, x_m^{|K|} - x_m\}) = \mathbf{v}(u \cup \{v, x_1^{|K|} - x_1, \ldots, x_m^{|K|} - x_m\}) = \mathbf{v}(u \cup v) = \mathbf{v}(u)$.

6.4 Proof of Theorem 4

Theorem 4 (Contextual logical information) *For any $m \geq 1$, let $S_m, Q_m \in \mathcal{P}(K[x_1, \cdots, x_m])$ represent the sender's logical statements and the query, with the property that $S_m \vDash Q_m$ and in particular, S_m, Q_m follow a (p_s, p_q)-i.i.d. law. Let the encoder f_m and decoder g_m be functions*

$$f_m : \mathcal{P}(K[x_1, \cdots, x_m])^2 \rightarrow \{0,1\}^* \tag{22}$$

$$g_m : \{0,1\}^* \rightarrow \mathcal{P}(K[x_1, \cdots, x_m]). \tag{23}$$

Then

$$\Lambda(p_s, 1-p_q) \leq \min_{f_m, g_m} |K|^{-m} E_{S_m, Q_m}\left[\mathbf{len}(f_m(S_m, Q_m))\right] \leq \Lambda(p_s, 1-p_q) + O\left(\frac{\log |K|^m}{|K|^m}\right), \tag{24}$$

where the minimization is over f_m, g_m such that the code implied by f_m is prefix free and such that, if $s \vDash q$, then $s \vDash g_m(f_m(s, q)) \vDash q$.

Proof. As with Theorem 1, we start by invoking Kraft's inequality through Lemma 5 and write

$$E_{S,Q}\left[\mathbf{len}(f_m(S, Q))\right] \geq H(f_m(S, Q)) \geq H(g_m(f_m(S, Q))) \tag{62}$$

The direction of the proof now diverges with respect to that of Theorem 1:

$$H(g_m(f_m(S, Q))) \stackrel{(a)}{=} H(g_m(f_m(S, Q))) - H(g_m(f_m(S, Q))|S, Q) \tag{63}$$

$$\stackrel{(b)}{=} I(S, Q; g_m(f_m(S, Q))) \tag{64}$$

$$\stackrel{(c)}{\geq} I(\mathbf{v}(S), \mathbf{v}(Q); \mathbf{v}(g_m(f_m(S, Q)))) \tag{65}$$

where (a) follows from the fact that discrete entropy is zero when conditioning on all randomness, (b) is the definition of mutual information and (c) follows from the data processing inequality. Next, assume an arbitrary enumeration of K^m, denoting the ith item as $[K^m]_i$. Define, for $1 \leq i \leq |K|^n$,

$$X_i = \begin{cases} 0 & [K^m]_i \in \mathbf{v}(S) \\ 1 & [K^m]_i \in \mathbf{v}(Q)^c \\ \otimes & [K^m]_i \notin \mathbf{v}(S) \cup \mathbf{v}(Q)^c \end{cases} \tag{66}$$

$$Y_i = \begin{cases} 0 & [K^m]_i \in \mathbf{v}(g_m(f_m(S, Q))) \\ 1 & [K^m]_i \notin \mathbf{v}(g_m(f_m(S, Q))) \end{cases} \tag{67}$$

We note that since $S \vDash Q$, we have that $\mathbf{v}(S) \cap \mathbf{v}(Q)^c = \emptyset$, and therefore X_i is well defined. The relationship between X_1^n and Y_1^n is in general very complex, as we have few assumptions on f_m and g_m. However, some key assertions can be made. First, recall that S, Q follow a (p_s, p_q)-i.i.d. law. By construction, then the $\{X_i\}$ are distributed i.i.d. according to

$$\begin{aligned} P(X_i = 0) &= p_s \\ P(X_i = 1) &= 1 - p_q \\ P(X_i = \otimes) &= p_q - p_s \end{aligned} \tag{68}$$

Furthermore, the assumption that $s \vDash g_m(f_m(s,q)) \vDash q$ can be used to establish a useful relation between X_i and Y_i. Recall the definition of the distortion metric ρ from Equations (30) and (31). Then, the assumption implies that for all $1 \leq i \leq |K|^n$,

$$\rho(X_i, Y_i) = 0 \tag{69}$$

This is an important fact that will be used shortly. We now apply the data processing inequality once more, taking advantage of the definitions for X_i and Y_i, and continue the proof with a pattern commonly found in Rate-Distortion theory:

$$\begin{aligned} I(\mathbf{v}(S), \mathbf{v}(Q); \mathbf{v}(g_m(f_m(S,Q)))) &\overset{(d)}{\geq} I(X_1 \cdots, X_{|K|^n}; Y_1 \cdots, Y_{|K|^n}) && (70)\\ &= H(X_1 \cdots, X_{|K|^n}) - H(X_1 \cdots, X_{|K|^n} | Y_1 \cdots, Y_{|K|^n}) && (71)\\ &\overset{(e)}{=} \sum_{i=1}^{|K|^n} H(X_i) - H(X_i | Y_1 \cdots, Y_{|K|^n}, X_1 \cdots X_{i-1}) && (72)\\ &\overset{(f)}{\geq} \sum_{i=1}^{|K|^n} H(X_i) - H(X_i | Y_i) && (73)\\ &= \sum_{i=1}^{|K|^n} I(X_i; Y_i) && (74) \end{aligned}$$

where (d) follows from the data processing inequality, (e) follows from the fact that the $\{X_i\}$ are independent and from the chain rule for entropy, (f) follows from the fact that conditioning cannot increase entropy.

To complete the lower bound, we note that the expression (74) is an averaging of mutual informations where the marginal for X_i is identical for all i, but the conditionals $Q(Y_i|X_i)$ are in general different. Define a conditional distribution by averaging all those conditionals:

$$\frac{1}{|K|^m} \sum_{i=1}^{|K|^m} Q_{Y_i|X_i}(y|x) \tag{75}$$

and let $\hat{X}, \hat{Y}$ be distributed according to the marginal for X_1 and the conditional distribution above. Because of (69), it is the case that

$$E_{\hat{X},\hat{Y}} \rho(\hat{X}, \hat{Y}) = 0 \tag{76}$$

It is known that mutual information is convex $\cup$ on the conditional $Q(Y_i|X_i)$ when the marginal of X is fixed, and therefore the following bound holds:

$$\sum_{i=1}^{|K|^m} I(X_i; Y_i) \geq |K|^m I(\hat{X}; \hat{Y}) \geq |K|^m \min_{P(X,Y) \in \mathcal{D}} I(X; Y) \tag{77}$$

where the domain $\mathcal{D}$ for the minimization is defined by joint distributions for X, Y with $X \sim (p_s, 1-p_q, p_q - p_s)$ and $E[\rho(X,Y)] = 0$. The fact that such a minimization results in $\Lambda(p_s, 1-p_q)$ can be checked using standard variational methods. This concludes the proof of the lower bound.

To prove the upper bound, we construct a code as follows. Let $\psi(s,q)$ be the function defined by composition of the act of computing the respective algebraic sets $\mathbf{v}(s), \mathbf{v}(q)$ and then, from those sets, constructing the indicator functions in Equation (66), so that

$$\mathbf{X} = [X_1, \cdots, X_{|K|^m}] = \psi(S, Q) \tag{78}$$

As before, we note that the $\{X_i\}$ are i.i.d. according to Equations (68). We next invoke the upper bound half of Theorem 6, which guarantees for $m \geq 1$ the existence of $\hat{f}_m$, $\hat{g}_m$ such that

$$|K|^{-m} E_{\mathbf{X}}\left[\mathbf{len}(\hat{f}_m(\mathbf{X}))\right] \leq \Lambda(p_s, 1-p_q) + 2\frac{\log_2\left(|K|^m \Lambda(p_s, 1-p_q)\right)}{|K|^m} + \frac{3}{|K|^m} \tag{79}$$

$$\sum_{i=1}^{|K|^m} \rho(X_i, \hat{g}_m(\hat{f}_m(\mathbf{X}))_i) = 0. \tag{80}$$

We define our encoder then as

$$f_m(s,q) = \hat{f}_m(\psi(s,q)) \tag{81}$$

Finally, we define the decoder g_m as follows:

$$g_m(\text{codeword}) = \sigma\left(\{[K^m]_i : \hat{g}_m(\text{codeword})_i = 0\}\right) \tag{82}$$

where σ is the polynomial set recovery function from Lemma 1 and where codeword $\in \{0,1\}^*$ is meant to be precisely $f_m(s,q)$ when the encoder and decoder are being used simultaneously. We conclude the proof by noting that by construction, if $s \vDash q$ we have $s \vDash g_m(f_m(s,q)) \vDash q$, and furthermore

$$|K|^{-m} E_{S,Q}\left[\mathbf{len}(f_m(S,Q))\right] \leq \Lambda(p_s, 1-p_q) + 2\frac{\log_2\left(|K|^m \Lambda(p_s, 1-p_q)\right)}{|K|^m} + \frac{3}{|K|^m}. \tag{83}$$

6.5 Proof of Theorem 5

Theorem 5 (Conditional logical information in the context of a goal) *For any $m \geq 1$, let S_m, Q_m, R_m represent the sender's logical statements, the query and the receiver logical statements, respectively, with the property that $S_m \vDash Q_m$ and in particular, S_m, Q_m, R_m follow a $(p_r, p_{s|r}, p_{q|r}, p_{s|\tilde{r}}, p_{q|\tilde{r}})$-i.i.d. law. Let the encoder f_m and decoder g_m be functions*

$$f_m : \mathcal{P}(K[x_1, \cdots, x_m])^3 \to \{0,1\}^* \tag{26}$$

$$g_m : \{0,1\}^* \times \mathcal{P}(K[x_1, \cdots, x_m]) \to \mathcal{P}(K[x_1, \cdots, x_m]). \tag{27}$$

Then

$$\mathcal{L} \leq \min_{f_m, g_m} |K|^{-m} E_{S_m, Q_m, R_m}\left[\mathbf{len}(f_m(S_m, Q_m, R_m))\right] \leq \mathcal{L} + O\left(\frac{\log |K|^m}{|K|^m}\right) \tag{28}$$

where $\mathcal{L} = p_r \Lambda(p_{s|r}, 1-p_{q|r}) + (1-p_r)\Lambda(p_{s|\tilde{r}}, 1-p_{q|\tilde{r}})$, and where the minimization is over f_m, g_m such that the code implied by f_m is prefix-free, and such that if $s \vDash q$ then $s \vDash g_m(f_m(s,q,r), r) \vDash q$.

Proof. The proof of this result builds upon the ideas in the proof of Theorem 2 and thus we will reference it as needed. We start using the law of total expectation to write

$$E_{S,Q,R}\left[\mathbf{len}(f_m(S,Q,R))\right] = E_R\left[E_{S,Q}\left[\mathbf{len}(f_m(S,Q,R))|R\right]\right] \tag{84}$$

We invoke again Kraft's inequality through Lemma 5 and write

$$E_{S,Q}\left[\mathbf{len}(f_m(S,Q,r))|R=r\right] = E_{f_m(S,Q,r)}\left[\mathbf{len}(f_m(S,Q,r))|R=r\right] \geq H(f_m(S,Q,r)|R=r) \tag{85}$$

From here, we apply the same ideas as in Theorem 2 to obtain

$$E_{S,Q}\left[\mathbf{len}(f_m(S,Q,r))|R=r\right] \geq I(\mathbf{v}(S), \mathbf{v}(Q); \mathbf{v}(g_m(f_m(S,Q,r),r))|R=r) \tag{86}$$

As before, define, for $1 \leq i \leq |K|^n$,

$$\begin{aligned} X_i &= \begin{cases} 0 & [K^m]_i \in \mathbf{v}(S) \\ 1 & [K^m]_i \in \mathbf{v}(Q)^c \\ \otimes & [K^m]_i \notin \mathbf{v}(S) \cup \mathbf{v}(Q)^c \end{cases} \\ Y_i &= \begin{cases} 0 & [K^m]_i \in \mathbf{v}(g_m(f_m(S,Q,r),r)) \\ 1 & [K^m]_i \notin \mathbf{v}(g_m(f_m(S,Q,r),r)) \end{cases} \end{aligned}$$

Similarly, as before note that the assumption that $s \vDash g_m(f_m(s,q;r),r) \vDash q$ implies that for all $1 \leq i \leq |K|^n$,

$$\rho(X_i, Y_i) = 0 \tag{87}$$

We continue applying the ideas in the proof of Theorem 2 and obtain

$$\begin{aligned} I(\mathbf{v}(S), \mathbf{v}(Q); \mathbf{v}(g_m(f_m(S,Q,r),r))|R=r) &\overset{(e)}{\geq} I(X_1 \cdots, X_{|K|^n}; Y_1 \cdots, Y_{|K|^n}|R=r) && (88) \\ &= I(X_1 \cdots, X_{|K|^n}; Y_1 \cdots, Y_{|K|^n}|\mathbf{v}(R)=\mathbf{v}(r)) && (89) \end{aligned}$$

where the last step follows from the fact that the following Markov chain holds:

$$R \to \mathbf{v}(R) \to (\mathbf{v}(S), \mathbf{v}(Q))$$

We will next argue that conditioned on $\mathbf{v}(R) = \mathbf{v}(r)$, the $\{X_i\}$ are independent, and distributed according to at most two distributions. This follows from the assumption that S, R, Q follow a $(p_r, p_{s|r}, p_{q|r}, p_{s|\tilde{r}}, p_{q|\tilde{r}})$-i.i.d. law. Then by construction, conditioned on $\mathbf{v}(R) = \mathbf{v}(r)$ the collection of random variables

$$\{X_i : [K^m]_i \in \mathbf{v}(r)\}$$

are i.i.d. according to the following law:

$$\begin{aligned} P(X_i = 0) &= p_{s|r} && (90) \\ P(X_i = 1) &= 1 - p_{q|r} && (91) \\ P(X_i = \otimes) &= p_{q|r} - p_{s|r} && (92) \\ & && (93) \end{aligned}$$

Similarly, the collection of random variables

$$\{X_i : [K^m]_i \notin \mathbf{v}(r)\}$$

are i.i.d. according to the following law:

$$\begin{aligned} P(X_i = 0) &= p_{s|\tilde{r}} && (94) \\ P(X_i = 1) &= 1 - p_{q|\tilde{r}} && (95) \\ P(X_i = \otimes) &= p_{q|\tilde{r}} - p_{s|\tilde{r}} && (96) \\ & && (97) \end{aligned}$$

Finally, by construction, the two kinds of X_is are mutually independent. Using the fact that the $\{X_i\}$ are independent, we can now write

$$\begin{aligned} I(\mathbf{v}(S), \mathbf{v}(Q); \mathbf{v}(g(f(S,Q,r),r))|R=r) &\overset{(e)}{\geq} \sum_{i=1}^{|K|^n} I(X_i; Y_i|\mathbf{v}(R)=\mathbf{v}(r)) && (98) \\ &= \sum_{i \in \mathbf{v}(r)} I(X_i; Y_i|\mathbf{v}(R)=\mathbf{v}(r)) + \sum_{i \notin \mathbf{v}(r)} I(X_i; Y_i|\mathbf{v}(R)=\mathbf{v}(r)) && (99) \\ &\overset{(g)}{\geq} |\mathbf{v}(r)| \min_{X \sim (p_{s|r}, 1-p_{q|r}, p_{q|r}-p_{s|r}), Q_{Y|X}: E[\rho(X,Y)]=0} I(X;Y) && (100) \\ &\quad + |\mathbf{v}(r)^c| \min_{X \sim (p_{s|\tilde{r}}, 1-p_{q|\tilde{r}}, p_{q|\tilde{r}}-p_{s|\tilde{r}}), Q_{Y|X}: E[\rho(X,Y)]=0} I(X;Y) && (101) \\ &\geq |\mathbf{v}(r)| \Lambda(p_{s|r}, 1-p_{q|r}) + |\mathbf{v}(r)^c| \Lambda(p_{s|\tilde{r}}, 1-p_{q|\tilde{r}}) && (102) \end{aligned}$$

Finally note that $E_R\left[|\mathbf{v}(R)|\right] = |K^m| p_r$ and thus putting everything together

$$E_{S,Q,R}\left[\mathbf{len}(f_m(S,Q,R))\right] \geq |K^m| p_r \Lambda(p_{s|r}, 1-p_{q|r}) + |K^m|(1-p_r)\Lambda(p_{s|\tilde{r}}, 1-p_{q|\tilde{r}}) \tag{103}$$

We now prove the upper bound result. Let f_m and g_m be the encoder and decoder functions that we will be constructing.

Let $\{U_i\}_{i=1}^{\infty}$ be a collection of i.i.d. random variables each over $\{0,1,\otimes\}$ distributed according to $\{p_{s|r}, 1-p_{q|r}, p_{q|r}-p_{s|r}\}$; similarly let $\{V_i\}_{i=1}^{\infty}$ be a collection of i.i.d. random variables each over $\{0,1,\otimes\}$ distributed according to $\{p_{s|\tilde{r}}, 1-p_{q|\tilde{r}}, p_{q|\tilde{r}}-p_{s|\tilde{r}}\}$. Using Theorem 6 twice we can guarantee for any $j,k \geq 1$ of the existence of $f_j^{(U)}$, $g_j^{(U)}$, $f_k^{(V)}$, $g_k^{(V)}$ such that

$$\begin{aligned}
E_{\mathbf{U}^j}\left[\mathbf{len}(f_j^{(U)}(\mathbf{U}^j))\right] &\leq j\Lambda(p_{s|r}, 1-p_{q|r}) + 2\log_2\left(j\Lambda(p_{s|r}, 1-p_{q|r})\right) + 3 &(104)\\
\sum_{i=1}^{j} \rho(U_i, g_j^{(U)}(f_j^{(U)}(\mathbf{U}^j))_i) &= 0 &(105)\\
E_{\mathbf{V}^k}\left[\mathbf{len}(f_k^{(V)}(\mathbf{V}^k))\right] &\leq k\Lambda(p_{s|\tilde{r}}, 1-p_{q|\tilde{r}}) + 2\log_2\left(k\Lambda(p_{s|\tilde{r}}, 1-p_{q|\tilde{r}})\right) + 3 &(106)\\
\sum_{i=1}^{k} \rho(V_i, g_k^{(V)}(f_k^{(V)}(\mathbf{V}^{\mathbf{k}}))_i) &= 0 &(107)
\end{aligned}$$

Next we assemble an encoder and decoder for our setup. Given two binary strings $c_1, c_2 \in \{0,1\}^*$, let $c_1 c_2$ denote the string resulting from concatenating the two individual strings. Recall the definition of ψ in the discussion preceding (78). We define our encoder as

$$f_m(s,q,r) = f_{|\mathbf{v}(r)|}^{(U)}\left([\psi(s,q)]_{\mathbf{v}(r)}\right) f_{|\mathbf{v}(r)^c|}^{(V)}\left([\psi(s,q)]_{\mathbf{v}(r)^c}\right) \tag{108}$$

We pause to observe that conditioned on $\mathbf{v}(R) = \mathbf{v}(r)$, $[\psi(s,q)]_{\mathbf{v}(r)}$ is an i.i.d. vector with entries over $\{0,1,\otimes\}$ and distributed according to $\{p_{s|r}, 1-p_{q|r}, p_{q|r}-p_{s|r}\}$. Similarly, $[\psi(s,q)]_{\mathbf{v}(r)^c}$ is an i.i.d. vector with entries over $\{0,1,\otimes\}$ and distributed according to $\{p_{s|\tilde{r}}, 1-p_{q|\tilde{r}}, p_{q|\tilde{r}}-p_{s|\tilde{r}}\}$ under the same conditioning. As a consequence

$$\begin{aligned}
E_{S,Q,R}\left[\mathbf{len}(f_m(S,Q,R)|\mathbf{v}(R)\right] &\leq |\mathbf{v}(R)|\Lambda(p_{s|r}, 1-p_{q|r}) + |\mathbf{v}(R)^c|\Lambda(p_{s|\tilde{r}}, 1-p_{q|\tilde{r}}) + 6 &(109)\\
&\quad +2\log|\mathbf{v}(R)|\Lambda(p_{s|r}, 1-p_{q|r}) + 2\log|\mathbf{v}(R)^c|\Lambda(p_{s|\tilde{r}}, 1-p_{q|\tilde{r}}) &(110)
\end{aligned}$$

and using the law of total expectations, and the concavity $\cap$ of the logarithm, we obtain

$$\begin{aligned}
|K|^{-m} E_{S,Q,R}\left[\mathbf{len}(f_m(S,Q,R))\right] &\leq p_r\Lambda(p_{s|r}, 1-p_{q|r}) + (1-p_r)\Lambda(p_{s|\tilde{r}}, 1-p_{q|\tilde{r}}) + \frac{6}{|K|^m} &(111)\\
&\quad +2\frac{\log|K|^m p_r \Lambda(p_{s|r}, 1-p_{q|r})}{|K|^m} &(112)\\
&\quad +2\frac{\log|K|^m (1-p_r)\Lambda(p_{s|\tilde{r}}, 1-p_{q|\tilde{r}})}{|K|^m} &(113)
\end{aligned}$$

Now we construct the decoder. First, the code that Theorem 6 guarantees existence of is prefix free, and thus the output of $f_m(s,q,r)$ can be decoded sequentially. The corresponding decodings are

$$\begin{aligned}
\text{roots}(\mathbf{v}(r)) &= g_{|\mathbf{v}(r)|}^{(U)}\left(f_{|\mathbf{v}(r)|}^{(U)}\left([\psi(s,q)]_{\mathbf{v}(r)}\right)\right) &(114)\\
\text{roots}(\mathbf{v}(r)^c) &= g_{|\mathbf{v}(r)^c|}^{(V)}\left(f_{|\mathbf{v}(r)^c|}^{(V)}\left([\psi(s,q)]_{\mathbf{v}(r)^c}\right)\right) &(115)
\end{aligned}$$

Now, assume that q is such that $s \vDash q$. Then by we know that

$$\begin{aligned}
\mathbf{v}(s) \cap \mathbf{v}(r) \subseteq \text{roots}(\mathbf{v}(r)) \subseteq \mathbf{v}(q) \cap \mathbf{v}(r) &(116)\\
\mathbf{v}(s) \cap \mathbf{v}(r)^c \subseteq \text{roots}(\mathbf{v}(r)^c) \subseteq \mathbf{v}(q) \cap \mathbf{v}(r)^c &(117)
\end{aligned}$$

and as a consequence

$$\mathbf{v}(s) \subseteq (\text{roots}(\mathbf{v}(r)) \cup \text{roots}(\mathbf{v}(r)^c)) \subseteq \mathbf{v}(q) \tag{118}$$

The output of the decoder g_m would then be

$$\sigma\,(\text{roots}(\mathbf{v}(r)) \cup \text{roots}(\mathbf{v}(r)^c)) \tag{119}$$

This ensures that the requirement $g_m(f_m(s, q, r)) \vDash q$ is satisfied.

References

[1] T. Cover. Enumerative source encoding. *IEEE Transactions on Information Theory*, 19(1):73–77, 1973.

[2] K. Devlin. *Logic and Information.* Cambridge University Press, 1991.

[3] P. Elias. Universal codeword sets and representations of the integers. *IEEE Transactions on Information Theory*, 21(2):194–203, 1975.

[4] D. Ellerman. *New Foundations for Information Theory: Logical Entropy and Shannon Entropy.* 2021.

[5] C. E. Shannon. *Coding Theorems for a Discrete Source With a Fidelity CriterionInstitute of Radio Engineers, International Convention Record, vol. 7, 1959.*, pages 325–350. 1959.