Electronic Colloquium on Computational Complexity, Report No. 171 (2012)

From Information to Exact Communication

Mark Braverman * Ankit Garg † Denis Pankratov ‡ Omri Weinstein §

December 2, 2012

Abstract

We develop a new local characterization of the zero-error information complexity function for two party communication problems, and use it to compute the exact internal and external information complexity of the 2−bit AND function: $\mathrm{IC}(\mathrm{AND}, 0) = C_\wedge \approx 1.4923$ bits, and $\mathrm{IC}^{\mathrm{ext}}(\mathrm{AND}, 0) = \log_2 3 \approx 1.5839$ bits. This leads to a tight (upper and lower bound) characterization of the communication complexity of the set intersection problem on subsets of $\{1, \ldots, n\}$, whose randomized communication complexity tends to $C_\wedge \cdot n \pm o(n)$ as the error tends to zero.

The information-optimal protocol we present has an infinite number of rounds. We show this is necessary by proving that the rate of convergence of the $r-$round information cost of AND to $\mathrm{IC}(\mathrm{AND}, 0) = C_\wedge$ behaves like $\Theta(1/r^2)$, i.e. that the r-round information complexity of AND is $C_\wedge + \Theta(1/r^2)$.

We leverage the tight analysis obtained for the information complexity of AND to calculate and prove the exact communication complexity of the *set disjointness* function $Disj_n(X, Y) = \neg \vee_{i=1}^n AND(x_i, y_i)$ with error tending to 0, which turns out to be $= C_{DISJ} \cdot n \pm o(n)$, where $C_{DISJ} \approx 0.4827$. Our rate of convergence results imply that an optimal protocol for set disjointness will have to use $\omega(1)$ rounds of communication, since every r-round protocol will be sub-optimal by at least $\Omega(n/r^2)$ bits of communication.

We also obtain the tight bound of $\frac{2}{\ln 2}k \pm o(k)$ on the communication complexity of disjointness of sets of size $\leq k$. An asymptotic bound of $\Theta(k)$ was previously shown by Håstad and Wigderson.

1 Introduction

Information theory as the primary mathematical tool for analyzing communication was first discovered by Shannon in the late 1940's [37]. In particular, Shannon introduced his entropy function $H(X)$ to measure the amount of information contained in a random variable X. Shannon's source coding theorem – also known as the noiseless coding theorem – postulates that in the limit the per-message cost of transmitting a stream of messages $x_1, x_2, \ldots$ independently distributed according to X is exactly $H(X)$. In the 65 years since its introduction, information theory has been developed in many different directions. An early milestone was the "one-copy" version of Shannon's theorem, attained by Huffman coding [21] – showing that on average even a single copy of $x \sim X$ can be encoded using $< H(X) + 1$ bits. Other achievements include the Slepian-Wolf theorem [41], which essentially says that an analogue of Shannon's theorem holds even when the receiver has some information about the input that is unknown to the sender. Overall, while some notable open problems remain, it is fair to say that at least in the two terminal case the data transmission problem is very well understood, with information theory being the primary tool in providing this understanding. While many of the same results could, in principle, have been obtained using direct combinatorial techniques, information theoretic formalism makes the proofs both much simpler and more illuminating.

*Department of Computer Science, Princeton University. Research supported in part by an Alfred P. Sloan Fellowship, an NSF CAREER award, and a Turing Centenary Fellowship.

†Department of Computer Science, Princeton University

‡Department of Computer Science, University of Chicago

§Department of Computer Science, Princeton University

ISSN 1433-8092

Communication complexity [43] can be viewed as the generalization of transmission problems to general tasks performed by two (or more) parties over a communication channel. Communication complexity is much more general than one-way transmission, but unlike circuit complexity, it is still amenable to lower bounds proofs by a broad range of techniques [27]. Furthermore, communication complexity lower bounds have found many applications, for example in obtaining tight bounds on streaming algorithms and data structures. In addition, some of the most promising approaches for strong circuit lower bounds that appear viable, such as Karchmer-Wigderson games and ACC lower bounds [26, 5] involve communication complexity lower bounds. Thus, at the moment, developing tools in communication complexity is one of the most promising approaches for making progress within computational complexity.

The earliest communication complexity lower bound techniques to be developed were combinatorial in nature. By representing the two-party function f using a 0/1-matrix M_f, and studying its combinatorial and analytic properties one obtains lower bounds on f's communication complexity in a variety of different models. Most existing state-of-the-art lower bounds on the communication complexity of specific problems, including recently obtained ones such as the recent lower bounds for Gap Hamming Distance [13, 39] fall into this broad category.

Despite information theory being so successful in reasoning about one-way communication, it took a while until information theory has been adopted into the communication complexity toolbox. Indeed, the first applications of information in communication complexity [1, 14] were in the context of one-way and simultaneous message communication complexity, which is most directly related to the classical transmission setting. It was not until the work of Bar-Yossef et al. [3] that these techniques were extended to the two-way setting. Further developments [4, 10, 8] showed that information-theoretic notions generalize nicely, at least to two-party communication complexity. One can define the information complexity of a task as the two-party analogue of Shannon's entropy. Shannon's entropy of a random variable X captures the amount of information contained in one sample – the least amount of information that needs to be conveyed to transmit an $x \sim X$. The *information complexity* of an interactive task T is the least amount of information about their inputs that Alice and Bob need to disclose to each other in order to perform T. Information complexity is similar to Shannon's entropy in that it captures exactly the amortized communication complexity of computing n independent copies of T over a noiseless binary channel as $n \to \infty$. Also, like Shannon's entropy, information complexity satisfies the direct sum property, i.e. it is additive: the information complexity of performing two independent tasks (T_1, T_2) is equal to the sum of the information complexities of T_1 and T_2 [10, 8].

Shannon's information theory has two broad benefits in addressing communication problems. Firstly, it gives us a set of simple yet powerful tools for reasoning about transmission problems and more broadly about relationships between interdependent random variables. Tools that include mutual information, the chain rule, and the data processing inequality [16]. It is this benefit that has been primarily used in prior works involving information-theoretic tools in communication complexity [1, 14, 12, 3, 23, 4]. Secondly, in the context of transmission problems – starting with Shannon's noiseless coding theorem – information theory is known to give tight precise bounds on rates and capacities. In fact, unlike computational complexity, where we often ignore constant, and sometimes even polylogarithmic, factors, a large fraction of results in information theory provide us with precise answers up to additive lower-order terms. For example, we know that sending a sequence of random digits would take exactly $\log_2 10 \approx 3.322$ bits per digit, and that the capacity of a binary symmetric channel with substitution probability 0.2 is exactly $1 - H(0.2) \approx 0.278$ bits per symbol. Generally speaking, prior to this work, this benefit has not been fully realized in an interactive communication complexity scenario. In this work, we explore and develop analytic machinery needed to bring tight bounds into the realm of communication complexity. In particular, we use these tools to calculate the tight communication complexity of the set disjointness function.

The set disjointness problem is one of the oldest and most studied problems in communication complexity [27]. In the two party setting, Alice and Bob are given subsets $X, Y \subset [n]$, respectively, and need to output 1 if $X \cap Y = \emptyset$, and 0 otherwise. Thus the disjointness function $Disj_n$ can be written as $Disj_n(X, Y) = \bigwedge_{i=1}^{n} (\neg X_i \vee \neg Y_i)$. In the deterministic communication complexity model, it is easy to show that $Disj_n$ has communication complexity $n + 1$. In the randomized communication complexity model – which is the focus

of this paper – an $\Omega(n)$ lower bound was first proved by Kalyanasundaram and Schnitger [25]. The proof was combinatorial in its nature. A much simpler combinatorial proof was given by Razborov a few years later [35]. In terms of upper bounds on the communication complexity of disjointness, an $n+1$ bound is trivial. No better bound was known prior to this work, although by examining the problem, one can directly convince oneself that there is a protocol for $Disj_n$ that uses only $(1-\varepsilon)n$ communication for some small $\varepsilon > 0$ – so that the deterministic algorithm is suboptimal. Another set of techniques which were successfully applied to versions of disjointness, especially in the quantum and multiparty settings [36, 15, 40] are analytic techniques. Analytic techniques such as the pattern matrix method [38], allow one to further extend the reach of combinatorial techniques.

The first information-theoretic proof of this bound was given by Bar-Yossef et al. [3]. While not materially improving the lower bound, the information-theoretic approach was extended to the multi-party number-in-hand setting [12, 23] with applications to tight lower bounds on streaming algorithms. At the core of the proof is a direct-sum reduction of proving an $\Omega(n)$ bound on $Disj_n$ to proving an $\Omega(1)$ bound on the information complexity of AND. The direct sum in this and other proofs follows from an application of the chain rule for mutual information – one of the primary information-theoretic tools. More recently, an information complexity view of disjointness lead to tight bounds on the ability of extended formulations by linear programs to approximate the CLIQUE problem [9]. This suggests that information complexity and a better understanding of the disjointness problem may have other interesting implications within computational complexity.

A problem related to disjointness is *Set Intersection* Int_n: now Alice and Bob do not want to just determine whether X and Y intersect, but both want to learn the intersection set $X \cap Y$. For this problem, even in the randomized setting, a lower bound of n bits on the communication is trivial: by fixing $X = [n]$ we see that in this special case the problem will amount to Bob sending his input to Alice (since $[n] \cap Y = Y$) – which clearly requires $\geq n$ bits. Thus the randomized communication complexity of this problems lies somewhere between n and $2n$ – the trivial upper and lower bounds. Note that the intersection problem is nothing but n copies of the two-bit AND function. Therefore, determining the communication complexity of Int_n is equivalent to determining the information complexity of the two-bit AND function by the information = amortized communication connection [10].

Essentially independently of the communication complexity line of work described above, a study of the AND/intersection problem has recently originated in the information theory community. A series of papers by Ma and Ishwar [29, 31] develops techniques and characterizations which allow one to rigorously calculate tight bounds on the communication complexity of Int_n and other amortized functions on the condition that one only considers protocols restricted to r rounds of communication. These techniques allow one to numerically (and sometimes analytically) compute the information complexity of the two-bit AND function – although the numerical computation is not provably correct for the most general unbounded-round case since the rate of convergence of r-round information complexity down to the true information complexity is unknown. Furthermore, their results about the AND function are non-constructive in the sense that they do not exhibit a protocol achieving their bounds. Nonetheless, numerical calculations produced by Ma and Ishwar do point at convergence to 1.4923 bits for the AND function [22]. As discussed below, our tight upper and lower bounds are consistent with this evidence.

The main result of this paper is giving tight bounds on the information and communication complexity of the AND, Int_n, and $Disj_n$ functions. As noted above, being able to obtain tight bounds is the second benefit information theory provides – one that has been largely untapped by the communication complexity community. In this work we begin to realize this benefit by precisely "solving" the randomized communication complexity of disjointness. We give a (provably) information-theoretic optimal protocol for the two-bit AND function. Combined with prior results – and new additional technical work – this optimality immediately gives a tight optimal randomized protocol for Int_n that uses $C_\wedge \cdot n \pm o(n)$ bits of communication and fails with a vanishing probability. Here $C_\wedge \approx 1.4923$ is an explicit constant given as a maximum of a concave analytic function. We then apply the same optimal result to obtain the optimal protocol for set disjointness, showing that the best vanishing error randomized protocol for $Disj_n$ will take $C_{DISJ} \cdot n \pm o(n)$ bits of communication, where $C_{DISJ} \approx 0.4827$ is another explicit constant (which we found to be surprisingly low).

The fact that we need the bounds to be exact throughout requires us to develop some new technical tools for dealing with information complexity in this context. For example, we show that unlike communication complexity, the randomized ε-error information complexity converges to the 0-error information complexity as $\varepsilon \to 0$.

Applying what we've learned about the AND function to the *sparse sets* regime, we are able to determine the precise communication complexity of disjointness $Disj_n^k$ where the sets are restricted to be of size at most k. Håstad and Wigderson [20] showed that the randomized communication complexity of this problem is $\Theta(k)$. We sharpen this result by showing that for vanishing error the communication complexity of $Disj_n^k$ is $\frac{2}{\ln 2}k \pm o(k) \approx 2.885k \pm o(k)$.

Interestingly the optimal protocol we obtain for AND is not an actual protocol in the strict sense of communication protocols definitions. One way to visualize it is as a game show where Alice and Bob both have access to a "buzzer" and the game stops when one of them "buzzes in". The exact time of the "buzz in" matters. If we wanted to simulate this process with a conventional protocol, we'd need the time to be infinitely quantized, with Alice and Bob exchanging messages of the form "no buzz in yet", until the buzz in finally happens. Thus the optimal information complexity of AND is obtained by an infimum of a sequence of conventional protocols rather than by a single protocol.

It turns out that the unlimited number of rounds is necessary, both for the AND function and for $DISJ_n$. Our understanding of information complexity in the context of the AND function allows us to lower bound the amount of communication needed for $DISJ_n$ if we restrict the number of rounds of interaction between players to r. $R(Disj_n, 0^+, r) \geq (C_{DISJ} + \Omega(1/r^2)) \cdot n$. In particular, any constant bound on the number of rounds means a linear loss in communication complexity. There are well-known examples in communication complexity where adding even a single round causes an exponential reduction in the amount of communication needed [32]. There are also examples of very simple transmission problems where it can be shown that two rounds are much better than one, and more than two are better yet [33, 34]. However, to our knowledge, together with a very recent independently obtained result on rounds in the communication complexity of small set intersection [11], this is the first example of a "natural" function where an arbitrary number of additional rounds is provably helpful.

Open problems

This paper shows that the information-theoretic tools and precise bounds can be extended into communication complexity where only asymptotic bounds can usually be proved. This opens up a set of problems involving extending such bounds further into communication complexity and related models using information theoretic analysis. We list problems here in an increasing order of generality. Additional open problems within the more general context of interactive coding theory can be found in [7].

Extensions to the exact communication complexity of read-once formulas. The first set of problems consists of extending the exact bounds to more general read-once formulas. Non-exact linear bounds are known, for example, for bounded-depth AND-OR trees [24, 28]. However, it is already not clear how to extend our exact results to a depth-two tree – i.e. to an OR of $\sqrt{n}$ copies of $Disj_{\sqrt{n}}$:

$$F(x_1, \ldots, x_n, y_1, \ldots, y_n) := \bigvee_{i=0}^{\sqrt{n}-1} \bigwedge_{j=1}^{\sqrt{n}} \left(\neg x_{i\sqrt{n}+j} \vee \neg y_{i\sqrt{n}+j}\right).$$

It is not hard to see that the vanishing-error randomized communication complexity of the depth-two problem is bounded from above by $\frac{C_{DISJ}}{2}n$. However, this bound is not tight, and the actual constant in front of n should be slightly lower. It would be interesting to compute this constant, or at least figure out an algorithm for computing it.

Communication and information complexity with non-negligible error. Our analysis for the information complexity of AND is carried out for the setting where error is not allowed. Consequently our communication complexity bounds apply to the setting where the error goes to 0 as n grows. It would be

interesting to understand the behavior of $\mathrm{IC}(AND_\varepsilon)$, where AND_ε is the task of computing the AND of two bits with error at most ε. By the continuity of information complexity which we prove, we know that this quantity converges to the (zero-error) information complexity of AND. Of particular interest is the rate of convergence of $\mathrm{IC}(AND_\varepsilon)$ to $\mathrm{IC}(AND)$:

Open Problem 1.1. *What is the asymptotic behavior of* $\mathrm{IC}(AND) - \mathrm{IC}(AND_\varepsilon)$ *as* $\varepsilon \to 0$*?*

One plausible conjecture is that it behaves as $\Theta(H(\varepsilon))$.

A closely related problem is computing the communication complexity of $DISJ_n$ with non-negligible error $\varepsilon > 0$. It is reasonable to assume (although would be interesting to prove) that for each ε it behaves as $C_{DISJ_\varepsilon} \cdot n \pm o(n)$ for a constant $C_{DISJ_\varepsilon} < C_{DISJ}$. We do not know how to find this constant given ε. Once again, it would also be interesting to understand the asymptotic behavior of $C_{DISJ} - C_{DISJ_\varepsilon}$.

The computability of information complexity and its rate of convergence in the number of rounds. More generally, we do not have an algorithm that given the truth table of a function $F(X,Y)$ calculates the (zero-error) information complexity of this function. Note we *can* compute the communication complexity of F^n for any n, and we have $\frac{\mathrm{CC}(F^n)}{n} \searrow \mathrm{IC}(F)$, which gives us a sequence which decreases down to $\mathrm{IC}(F)$, but we do not have a similar sequence of lower bound. We can construct similar lower bounds, following the methodology of Ma and Ishwar [31] if we fix the number of rounds r the computation is allowed to use. Thus we can compute $\mathrm{IC}_r(F)$ – the information cost of the best r-round protocol computing F. Unfortunately, once again we only know that $\mathrm{IC}_r(F) \searrow \mathrm{IC}(F)$ with no effective rate of convergence. Thus figuring this rate of convergence, or at least an upper bound on it, is sufficient for the computability of $\mathrm{IC}(F)$.

The rate of convergence of $\mathrm{IC}_r(F) \searrow \mathrm{IC}(F)$ is a very interesting question in its own right. The question is about the usefulness of additional rounds in giving an information-theoretically efficient protocol for F, and equivalently whether extra rounds of communication are useful for computing n copies of F for large n. We showed that in the case of $F = AND$, the rate of convergence is $1/r^2$. We conjecture that this is always the right rate:

Conjecture 1.2. *For all* $F(X,Y)$, $\mathrm{IC}_r(F) - \mathrm{IC}(F) = O_F(1/r^2)$.

As we've just mentioned the AND function shows this is the best we can hope for in general. An example where this bound is tight is the single-bit transmission function $F(X,Y) = X$. Its information cost is $\mathrm{IC}_r(F) = \mathrm{IC}(F) = 1$ for all $r \geq 1$. It is also an interesting open problem what asymptotic behaviors can $\mathrm{IC}_r(F) - \mathrm{IC}(F)$ exhibit as $r \to \infty$. A recent paper by Brody et. al. [11] shows a very interesting tradeoff between rounds and information for $F = EQ_m$ the equality function on m bits, but in the regime where $r \ll m$. Note that here we are interested in the rate of convergence where r is arbitrarily large compared to the number of variables in F.

The zero-error communication complexity. Another interesting question is extending the information = amortized communication results to zero-error randomized communication protocols. Currently we only know that the internal (zero-error) information complexity of a function F is equal to its vanishing-error amortized communication complexity. This connection fails when one considers zero-error amortized communication complexity. An extreme example of this is the equality function EQ_n on n-bit strings. The information complexity of this function is $O(1)$ [8], and its amortized *vanishing error* communication complexity in indeed constant [18]. On the other hand, it can be shown that the *zero-error* amortized communication complexity of EQ_n is $\Omega(n)$. We conjecture that the (average case) amortized zero-error communication complexity of functions is exactly captured by the *external* information complexity – the amount of information the parties need to reveal to an external observer to compute F:

Conjecture 1.3. *For all* F *we have* $\lim_{n\to\infty} \frac{\mathrm{CC}(F^n,0)}{n} = \mathrm{IC}^{ext}(F)$.

We note that using techniques similar to Harsha et al. [19], we can show the $\leq$ direction of this conjecture. The external information complexity of AND, which we show to be equal to $\log_2 3 \approx 1.585$, exactly matches the bound on the amortized communication complexity of AND previously established by Ahlswede and Cai

[2]. This result provides the strongest piece of evidence to date to support this conjecture. In addition, the conjecture is true for product prior distributions $\mu = \mu_x \times \mu_y$, since over product distributions the external and internal information costs coincide. This leads to a direct sum theorem for zero-error average-case communication complexity over product distributions. Further discussion on the conjecture can be found in [7].

Multi-party information complexity. Generalizing further out, it would be very interesting to develop the "right" notions for information complexity in the multi-party communication complexity setting. There are examples where information-theoretic methods were successfully applied to multiparty number-in-hand communication [12]. However, it is not clear whether (and how) similar techniques can apply to the number-on-the-forehead model. One obstacle here is the existence of *private multi-party protocols* that allow three or more parties to evaluate a function of their inputs while only learning the value of the function [6].

2 Our main results

Let π be a communication protocol attempting to solve some two-party function $f(x, y)$ with zero error where inputs are sampled according to a joint distribution μ. Our first contribution is a characterization of the zero-error information cost function $IC_\mu(f, 0)$ in terms of certain local concavity constrains. A related – but more abstract – characterization was given in the information theory literature by Ma and Ishwar [29].

Lemma 2.1. *For any function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ there exist a family $\mathfrak{C}(f)$ of functions $C : \Delta(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}^+$ satisfying certain local concavity constraints, such that for any distribution μ, and any protocol π solving f with zero error under μ, it holds that*

$$\forall\ C \in \mathfrak{C}(f) \quad C(\mu) \leq \mathrm{IC}_\mu(\pi).$$

Furthermore, $IC_\mu(f, 0)$ is the point-wise maximum of $\mathfrak{C}(f)$.

This lemma gives a very general technique for proving information-complexity lower bounds, and plays a central role in one of our main results: The exact information complexity of the 2-bit AND function $f(x, y) = x \wedge y$. Since the inputs of the parties consist of only 2 bits, the information complexity of this function is trivially bounded by 2. By fixing $x = 1$, it is also easy to see that 1 is a lower bound on the information complexity. We present a zero-error "clocked" protocol which has an infinite number of rounds and computes the AND function, under any input distribution μ, with information cost at most $C_\wedge \approx 1.4923$. The maximum external information cost of our protocol is $\log_2 3 \approx 1.58496$. While the analysis itself is nontrivial, the main bulk of effort is proving this protocol is in fact optimal, both in the internal and external sense:

Theorem 2.2.

$$\mathrm{IC}(AND, 0) = C_\wedge \approx 1.4923$$

Theorem 2.3.

$$\mathrm{IC}^{ext}(AND, 0) = \log_2 3 \approx 1.58496$$

We also analyze the rate of convergence to the optimal information cost, as the number r of permitted rounds increases. We view this result as a step towards proving that information complexity of functions is computable.

Theorem 2.4. *For all $\mu \in \Delta(\{0, 1\}^2)$ with full support we have*

$$\mathrm{IC}^r_\mu(AND, 0) = \mathrm{IC}_\mu(AND, 0) + \Theta_\mu\left(\frac{1}{r^2}\right).$$

Moreover, the lower bound holds even for μ such that $\mu(1, 1) = 0$.

In the second part of our work we show how tight information bounds may lead to exact communication bounds. We first prove a theorem which characterizes the exact randomized communication complexity of "$\bigvee$"-type functions with error tending to zero, in terms of an informational quantity $\mathsf{IC}^0(f,0)$ which informally measures the information cost required to solve f under the "worst" distribution supported on $f^{-1}(0)$[1].

Theorem 2.5. *For any Boolean function $f:\{0,1\}^k\times\{0,1\}^k\to\{0,1\}$, let $g_n(X,Y):=\vee_{i=1}^n f(x_i,y_i)$, where $X=\{x_i\}_{i=1}^n, Y=\{y_i\}_{i=1}^n$ and $x_i,y_i\in\{0,1\}^k$. Then for all $\epsilon>0$, there exists $\delta=\delta(f,\epsilon)>0$ such that $\delta\to 0$ as $\epsilon\to 0$ and*

$$(\mathsf{IC}^0(f,0)-\delta)\cdot n\leq\mathcal{R}_\epsilon(g_n)\leq\mathsf{IC}^0(f,0)\cdot n+o(n)\cdot k.$$

Finally, we tie in all of our results to prove the *exact* randomized communication complexity of the $Disj_n$ function, with error tending to zero. For the general disjointness function we get:

Theorem 2.6. *For all $\epsilon>0$, there exists $\delta=\delta(\epsilon)>0$ such that $\delta\to 0$ as $\epsilon\to 0$ and*

$$(C_{DISJ}-\delta)\cdot n\leq\mathcal{R}_\epsilon(g_n)\leq C_{DISJ}\cdot n+o(n).$$

where $C_{DISJ}\approx 0.4827$ bits.

For the case of disjointness $DISJ_n^k$ of sets of size $\leq k$ we get

Theorem 2.7. *Let n,k be such that $k=\omega(1)$ and $n/k=\omega(1)$. Then for all constant $\epsilon>0$,*

$$\left(\frac{2}{\ln 2}-O(\sqrt{\epsilon})\right)\cdot k-o(k)\leq R_\epsilon(DISJ_n^k)\leq\frac{2}{\ln 2}\cdot k+o(k).$$

Our results rely on new insights for understanding communication protocols from an informational point of view, as functionals on the space of distributions. This requires further development of some non-trivial properties of the information cost function. One such property is the continuity of the information complexity function at $\epsilon=0$:

Theorem 2.8. *For all $f:\mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$ and $\mu\in\Delta(\mathcal{X}\times\mathcal{Y})$ we have*

$$\lim_{\epsilon\to 0}\mathrm{IC}_\mu(f,\epsilon)=\mathrm{IC}_\mu(f,0),\tag{1}$$

$$\lim_{\epsilon\to 0}\mathrm{IC}_\mu^{ext}(f,\epsilon)=\mathrm{IC}_\mu^{ext}(f,0).\tag{2}$$

Acknowledgments

We would like to thank Paul Beame, Prakash Ishwar, Nan Ma, Anup Rao, Michael Saks, David Xiao, and Avi Wigderson for enlightening discussions. We would like to thank Grigory Yaroslavtsev for bringing to our attention the small set intersection problem (Remark 9.15). We would like to thank Prakash Ishwar and Nan Ma for sharing the code used to generate Figure 1.

3 Organization

The paper is organized as follows. Following preliminaries, in Sections 5 and 6 we develop new tools and properties of the IC function, which will be essential to the proof of the main results of this paper. We begin by stating and proving our local characterization of the zero error information cost function in Section 5, and then use this new view to prove the continuity of the zero error IC function (Section 6). In Section 7 we analyze the zero error external and internal information complexity of the AND function. We begin by presenting a zero-error continuous protocol and analyzing its internal and external information cost, and then

[1] Note that this quantity is not zero, since we will range only over protocols which solve f under *any* input.

use the machinery developed in preceding sections to prove the optimality of our protocol. In section 7.9 we analyze the rate of convergence of the r-round information complexity of AND to $IC(AND, 0)$, and present a natural discertization of our "clocked" protocol, which achieves the aforementioned rate. In section 8 we prove Theorem 2.5. Finally, in Sections 8.3 and 9, we tie in all our results to prove the exact communication complexity of $DISJ_n$ (Theorem 2.6) and $DISJ_n^k$ (Theorem 2.7).

4 Preliminaries

4.1 Notation

Capital letters are reserved for random variables (e. g., A, B, C), calligraphic letters for sets (e. g., $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \ldots$), and small letters for elements of sets (e. g., $a, b, c, \ldots$). For typographical purposes we shall write $A_1A_2\cdots A_n$ to denote the random variable $(A_1, A_2, \ldots, A_n)$ and *not* the random variable that is the product of the A_i, unless otherwise specified.

We use $[n]$ to denote the set $\{1, \ldots, n\}$.

For random variables A and B_i ($i \in [n]$) and elements $b_i \in \operatorname{range} B_i$ ($i \in [n]$) we write $A_{b_1b_2\cdots b_n}$ to denote the random variable A conditioned on the event "$B_1 = b_1, B_2 = b_2, \ldots, B_n = b_n$".

Whenever convenient we shall view a probability distribution μ on a sample space $\mathcal{X} \times \mathcal{Y}$ as a $|\mathcal{X}| \times |\mathcal{Y}|$ matrix, where the rows are indexed by elements of $\mathcal{X}$ and columns are indexed by elements of $\mathcal{Y}$ in some standard order (e. g., lexicographic order when $\mathcal{X}$ and $\mathcal{Y}$ are sets of binary strings). For example, we shall often write distribution μ on $\{0,1\} \times \{0,1\}$ as $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & \delta \\ \hline \end{array}$ meaning that $\mu(0,0) = \alpha, \mu(0,1) = \beta, \mu(0,1) = \gamma$, and $\mu(1,1) = \delta$.

For a distribution μ on $\mathcal{X} \times \mathcal{Y}$ we use μ^T to denote the probability distribution on $\mathcal{Y} \times \mathcal{X}$ that is given by the transpose of the matrix representation of μ.

4.2 Communication Complexity

The two-party communication model was introduced by Yao [43] in 1979. In this model, two parties, traditionally called Alice and Bob, are trying to collaboratively compute a known Boolean function $f : \mathcal{X} \times \mathcal{Y}$. Each party is computationally unbounded; however, Alice is only given input $x \in \mathcal{X}$ and Bob is only given $y \in \mathcal{Y}$. In order to compute $f(x, y)$, Alice and Bob communicate in accordance with an agreed-upon communication protocol π. Protocol π specifies as a function of transmitted bits only whether the communication is over and, if not, who sends the next bit. Moreover, π specifies as a function of the transmitted bits and x the value of the next bit to be sent by Alice. Similarly for Bob. The communication is over when *both parties* know the value of $f(x, y)$. The cost of the protocol π is the number of bits exchanged on the worst input. *The transcript* of a protocol is a concatenation of all the bits exchanged during the execution of the protocol.

There are several ways in which the deterministic communication model can be extended to include randomization. In the *public-coin model*, Alice and Bob have access to a shared random string r chosen according to some probability distribution. The only difference in the definition of a protocol is that now the protocol π specifies the next bit to be sent by Alice as a function of x, the already transmitted bits, and a random string r. Similarly for Bob. This process can also be viewed as the two players having an agreed-upon distribution on deterministic protocols. Then the players jointly sample a protocol from this distribution. In the *private-coin model*, Alice has access to a random string r_A hidden from Bob, and Bob has access to a random string r_B hidden from Alice.

Definition 4.1 (Randomized Communication Complexity)**.** For a function $f : \mathcal{X} \times \mathcal{Y} \to Z$ and a parameter $\epsilon > 0$, $R_\epsilon(f)$ denotes the cost of the best randomized public coin protocol for computing f with error at most ϵ on *every* input.

Observe that for the purpose of the communication complexity, once we allow public randomness, it makes no difference whether we permit the players to have private random strings or not. This is because the private random strings can be simulated by parts of the public random string, which is infinite. However, for information complexity it is crucial to consider protocols that permit *both private and public* randomness. Thus for a protocol π we use $\Pi(x, y)$ to denote the concatenation of the transcript of π and the public randomness when the protocol runs on inputs (x, y). The worst-case number of bits transmitted in π is denoted by $\mathrm{CC}(\pi)$. For $i \in [\mathrm{CC}(\pi)]$ we write $\Pi_i(x, y)$ to denote the ith bit transmitted in Π on input (x, y) if it exists.

For the pre-1997 results on communication complexity see the excellent monograph by Kushilevitz and Nisan [27].

4.3 Information Theory

In this section we briefly provide the essential information-theoretic concepts required to understand the rest of the paper. For a thorough introduction to the area of information theory, the reader should consult a classical monograph by Cover and Thomas [17]. Unless stated otherwise, all log's in this paper are base-2.

We use $\Delta(\mathcal{X})$ to denote *the family of all probability distributions* on $\mathcal{X}$.

Definition 4.2. Let μ be a probability distribution on sample space Ω. *Shannon entropy* (or just *entropy*) of μ, denoted by $H(\mu)$, is defined as $H(\mu) := \sum_{\omega \in \Omega} \mu(\omega) \log \frac{1}{\mu(\omega)}$.

For a random variable A we shall write $H(A)$ to denote the entropy of the induced distribution on the range of A. The same also holds for other information-theoretic quantities appearing later in this section.

For the Bernoulli distribution with probability of success p we write $H(p) = -p \log p - (1-p) \log(1-p)$.

Definition 4.3. *Conditional entropy* of a random variable A conditioned on B is defined as

$$H(A|B) = \mathbb{E}_b(H(A|B = b)).$$

Fact 4.4. $H(AB) = H(A) + H(B|A)$.

Definition 4.5. The *mutual information* between two random variable A and B, denoted by $I(A; B)$ is defined as

$$I(A; B) := H(A) - H(A|B) = H(B) - H(B|A).$$

The *conditional mutual information* between A and B given C, denoted by $I(A; B|C)$, is defined as

$$I(A; B|C) := H(A|C) - H(A|BC) = H(B|C) - H(B|AC).$$

Fact 4.6 (Chain Rule)**.** *Let A_1, A_2, B, C be random variables. Then*

$$I(A_1A_2; B|C) = I(A_1; B|C) + I(A_2; B|A_1C).$$

Fact 4.7. *Let A, B, C, D be four random variables such that $I(B; D|AC) = 0$. Then*

$$I(A; B|C) \geq I(A; B|CD)$$

Definition 4.8. Given two probability distributions μ_1 and μ_2 on the same sample space Ω such that $(\forall \omega \in \Omega)(\mu_2(\omega) = 0 \Rightarrow \mu_1(\omega) = 0)$, the *Kullback-Leibler Divergence* between is defined as

$$\mathbb{D}(\mu_1||\mu_2) = \sum_{\omega \in \Omega} \mu_1(\omega) \log \frac{\mu_1(\omega)}{\mu_2(\omega)}.$$

The connection between the mutual information and the Kullback-Leibler divergence is provided by the following fact.

Fact 4.9. *For random variables $A, B,$ and C we have*

$$I(A; B|C) = \mathbb{E}_{b,c}(\mathbb{D}(A_{bc}||A_c)).$$

Definition 4.10. Let μ_1 and μ_2 be two probability distributions on the same sample space Ω. *Total variation distance* is defined as

$$\|\mu_1 - \mu_2\| := \frac{1}{2} \sum_{\omega \in \Omega} |\mu_1(\omega) - \mu_2(\omega)|.$$

Observe that $\|\mu_1 - \mu_2\| = \max_{\mathcal{S} \subseteq \Omega} |\mu_1(\mathcal{S}) - \mu_2(\mathcal{S})|$.

Fact 4.11 (Data Processing Inequality)**.** *Let A, B, C be random variables on the same sample space, and let D be a probabilistic function of B only. Then we have*

$$I(A; D|C) \leq I(A; B|C).$$

The above concepts were defined for the *discrete probability distributions*. In this paper we shall also work with continuous probability distributions. There are a lot of subtleties in going from discrete case to continuous case in the area of information theory; however, we shall not encounter those subtleties. For our purposes, the above definitions and facts generalize to the continuous case in a straightforward way.

For instance, Kullback-Leibler divergence between two continuous distributions over $\mathbb{R}$ given by their probability density functions (PDFs) p and q is defined as

$$\mathbb{D}(p||q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}\, dx.$$

4.4 Information Cost (IC)

Definition 4.12. The *internal information cost* of a protocol π with respect to a distribution μ on inputs from $\mathcal{X} \times \mathcal{Y}$ is defined as

$$\mathrm{IC}_\mu(\pi) := I(\Pi(X, Y); X|Y) + I(\Pi(X, Y); Y|X).$$

The *external information cost* of π with respect to μ is

$$\mathrm{IC}^{\mathrm{ext}}_\mu(\pi) := I(\Pi(X, Y); XY).$$

Lemma 4.13. *[10] For any distribution μ, $\mathsf{IC}_\mu(\pi) \leq CC(\pi)$.*

The *information complexity* of f with respect to μ is

$$\mathrm{IC}_\mu(f, \epsilon) := \inf_\pi \mathrm{IC}_\mu(\pi),$$

where the infimum ranges over all (randomized) protocols π solving f with error at most ϵ when inputs are sampled according to μ. Note that we cannot replace the above quantifier with a min, since the information complexity of a function may not be achievable by any fixed (finite-round) protocol[2].

Similarly, the *external information complexity* of f with respect to μ is defined as

$$\mathsf{IC}^{\mathsf{ext}}_\mu(f, \epsilon) := \inf_\pi \mathrm{IC}^{\mathrm{ext}}_\mu(\pi).$$

The *prior-free* information complexity of a function f (or simply, the *information complexity* of f) with error ϵ is defined as

$$IC(f, \epsilon) := \inf_\pi \max_{\mu \text{ a distribution on } \mathcal{X} \times \mathcal{Y}} IC_\mu(\pi).$$

where the infimum is over protocols that work correctly for each input, except with probability ϵ. The *external prior-free* information complexity is defined analogously.

The special case $\mathrm{IC}(f, 0)$ is referred to as the *zero error* information complexity of f, and will be of primary interest in this paper. It turns out that for this special case ($\epsilon = 0$), we may reverse the order of quantifiers:

[2]In fact, we shall see that this is the case for the AND function whose information complexity is analyzed in this paper.

Theorem 4.14. *[8]*

$$IC(f,0) = \max_{\mu} \inf_{\pi \text{ correct on support of } \mu} \mathsf{IC}_\mu(\pi),$$

i.e, we can choose the protocol dependent on the distribution and yet the information cost doesn't decrease.

For $r \in \mathbb{N}$, the *r-round information complexity* of a function f is defined as

$$\mathrm{IC}^r_\mu(f,\epsilon) := \inf_\pi \mathrm{IC}_\mu(\pi),$$

where the infimum ranges over all r-round protocols π solving f with error at most ϵ when inputs are sampled according to μ. The *r-round external information complexity* is defined analogously.

5 Characterization of IC via Local Concavity Constraints

In this section we prove Lemma 2.1, a local characterization of the zero-error information cost function. More precisely, for an arbitrary function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ we shall define a family $\mathfrak{C}(f)$ of functions $\Delta(\mathcal{X} \times \mathcal{Y}) \to \mathcal{Z}$ satisfying certain local concavity constraints, and demonstrate that each member of $\mathfrak{C}(f)$ is a lower bound on the zero-error information cost function $I(\mu) := \mathrm{IC}_\mu(f,0)$ of f. It will be evident that $I(\mu)$ itself satisfies the local concavity constraints, i. e., $I(\mu) \in \mathfrak{C}(f)$. Thus we obtain a new characterization of the zero-error information cost of a function f as a point-wise maximum over all functions in the family $\mathfrak{C}(f)$.

It turns out that the number of local concavity constraints used to define $\mathfrak{C}(f)$ can be greatly reduced if we assume that every bit sent in a protocol π, nearly achieving the information cost of f, is uniformly distributed from an external point of view. We say that such a protocol is in *normal form*. In Section 5.1 we show that the normal form assumption can be made without loss of generality.

In Section 5.2 we describe the concavity constraints and demonstrate that they lead to the characterization of the information cost described above.

5.1 Normal Form of a Protocol

Definition 5.1. We say that a protocol π is in *normal form* if for each fixing r of public randomness and for each node u in the protocol π_r

$$P(\text{owner of } u \text{ sends } 0 | \Pi_r \text{ reaches } u) = 1/2.$$

Lemma 5.2. *Let π be a protocol on inputs from $\mathcal{X} \times \mathcal{Y}$ and let $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$. For every $\delta > 0$, there exists a protocol π_δ in normal form such that*

1. *π_δ δ-simulates π i.e. for all x, y, $Pr[\pi(x,y) \neq \pi_\delta(x,y)] \leq \delta$. $\pi(x,y)$ denotes the random variable for the transcript of π on inputs x, y.*
2. *$IC_\mu(\pi_\delta) \leq IC_\mu(\pi)$.*

Proof. Let ℓ be such that $CC(\pi) \cdot 2^{-\ell} \leq \delta$. In π_δ, Alice and Bob try to simulate execution of π on (x,y). Suppose that the players reached node u of π and it is Alice's turn to speak. Let $p_x = P(\text{Alice sends } 0 \mid \pi \text{ reaches } u, X = x)$ $(X, Y \sim \mu^u$, where μ^u is the distribution conditioned on reaching node $u)$. Also let $p := P(\text{Alice sends } 0 \mid \pi \text{ reaches } u)$. Note that $p = \sum_x \mu^u_x \cdot p_x$. To simulate sending the next bit, Alice first uses private randomness to decide whether to send 0 or 1, just like in π. That is Alice samples a bit that is 0 w.p. p_x and 1 w.p. $1 - p_x$. If the outcome is 0, Alice samples a random number v uniformly at random from the interval $[0, p]$. If the outcome is 1, Alice samples v uniformly at random from interval $(p, 1]$. Then Alice sends the first ℓ bits of the binary expansion of v to Bob. Lastly, Bob uses these bits to check if $v < p$ or $v > p$ and decode the transmitted bit. If it turns out that the first ℓ bits do not suffice to decide whether $v < p$ or $v > p$ then the players end the simulation and decide to abort (and incur an error). Note that both Alice and Bob know p, so they know when to abort.

Now the pdf of v, is $q(v) = \sum_x \mu_x^u \cdot (p_x \cdot (1/p))$ for $v \le p$ and $q(v) = \sum_x \mu_x^u \cdot (1-p_x) \cdot 1/(1-p)$ for $v > p$. Since $\sum_x \mu_x^u \cdot p_x = p$, $q(v) = 1$ for all $v \in [0,1]$. Therefore each bit in its binary expansion is uniform. This proves that π_δ is in normal form.

Simulation of a bit of protocol π fails if the first ℓ bits of v equal the first ℓ bits of p, which happens with probability $2^{-\ell}$. Taking the union bound we get that the simulation fails only with probability $CC(\pi) \cdot 2^{-\ell} \le \delta$.

Let $B_{\ell,u}$ denote the random variable for the ℓ bits transmitted by Alice to Bob after reaching node u in π_δ. Also let B_u denote the random variable for the bit to be sent by Alice to Bob in π after reaching node u. Then the information cost corresponding to node u for π is $I(B_u; X|Y)$ and for π_δ is $I(B_{\ell,u}; X|Y)$. Also let R_u denote the private randomness used by Alice to sample the ℓ bits in π_δ. Since R_u and B_u determine $B_{\ell,u}$,

$$I(B_{\ell,u}; X|Y) \le I(B_u R_u; X|Y) = I(B_u; X|Y) + I(R_u; X|Y B_u) = I(B_u; X|Y)$$

$I(R_u; X|Y B_u) = 0$, since conditioned on the bit B_u to be sent, R_u consists of sampling a uniform random number from $[0,p]$ or from $(p,1]$, which is independent of X. We can similarly get that for all nodes v owned by Bob, $I(B_{\ell,v}; X'|Y') \le I(B_v; X'|Y')$, where $X', Y' \sim \mu^v$. Hence $IC_\mu(\pi_\delta) \le IC_\mu(\pi)$. □

Remark 5.3. Similar result holds for the *external information cost*.

5.2 The Characterization

Definition 5.4. Let $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a given function. Define a family $\mathfrak{C}(f)$ of all functions $C : \Delta(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}^+$ satisfying the following constraints:

- for all $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$ if $f|_{\text{supp}(\mu)}$ is a constant function then $C(\mu) = 0$,
- for all $\mu, \mu_0^A, \mu_1^A \in \Delta(\mathcal{X} \times \mathcal{Y})$ if there exists a signal B that Alice can send starting from μ such that $P(B=0) = P(B=1) = 1/2$, $\mu_0^A(x,y) = P(X=x, Y=y|B=0)$, and $\mu_1^A(x,y) = P(X=x, Y=y|B=1)$ then
$$C(\mu) \le C(\mu_0^A)/2 + C(\mu_1^A)/2 + I(X; B|Y),$$
- for all $\mu, \mu_0^B, \mu_1^B \in \Delta(\mathcal{X} \times \mathcal{Y})$ if there exists a signal B that Bob can send starting from μ such that $P(B=0) = P(B=1) = 1/2$, $\mu_0^B(x,y) = P(X=x, Y=y|B=0)$, and $\mu_1^B(x,y) = P(X=x, Y=y|B=1)$ then
$$C(\mu) \le C(\mu_0^B)/2 + C(\mu_1^B)/2 + I(Y; B|X).$$
- for all μ, $C(\mu) \le \log(|\mathcal{X}| \cdot |\mathcal{Y}|)$.

Remark 5.5. The notation $f|_{\text{supp}(\mu)} \equiv$ *Constant* means that both parties can determine the function's output under μ by looking at their own input - We do not consider the player's output as part of the protocol transcript, so the latter condition need not imply that the function is determined under μ from an *external* point of view. The example $f(0,0) = 0$, $f(1,1) = 1$, $\mu(0,0) = \mu(1,1) = 1/2$ illustrates this point.

Note that each protocol induces a distribution over the leaves. For a protocol π, let Π denote the transcript of the protocol when the inputs $X, Y \sim \mu$. Also let μ_t denote the distribution conditioning on reaching leaf t i.e. $\mu_t(x,y) = Pr[X=x, Y=y|\Pi = t]$.

Lemma 5.6. *Let $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a given function. Let π be a protocol in* normal form. *Then for all $C \in \mathfrak{C}(f)$ and all $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$ we have $C(\mu) \le \text{IC}_\mu(\pi) + \mathbb{E}_{t \sim \Pi} C(\mu_t)$.*

Proof. [Proof by induction on $c := \text{CC}(\pi)$] When $c = 0$ the claim is clearly true, since there is only one leaf t and $\mu_t = \mu$.

Assume the claim holds for all c-bit protocols where $c \ge 0$. Consider a $c+1$-bit protocol π. Assume wlog that Alice sends the first bit B. If this bit is 0 then Alice and Bob end up with a new distribution on the

inputs μ_0^A, otherwise they end up with distribution μ_1^A. After the first bit, the protocol π reduces to a c-bit protocol π^0 if 0 was sent and π^1 if 1 was sent. Since Alice's bit is uniformly distributed we have

$$\begin{aligned} I(\Pi; X|Y) &= I(\Pi_1; X|Y) + I(\Pi_{\geq 2}; X|Y\Pi_1) \\ &= I(B; X|Y) + I(\Pi^0; X|Y)/2 + I(\Pi^1; Y|X)/2. \end{aligned}$$

Similarly for $I(\Pi; Y|X)$. Let Π^0 denote the random variable for transcript of π^0 and Π^1 for π^1. Thus we obtain

$$\begin{aligned} \mathrm{IC}_\mu(\pi) &= \mathrm{IC}_{\mu_0^A}(\pi_0)/2 + \mathrm{IC}_{\mu_1^A}(\pi_1)/2 + I(X; B|Y) \\ &\geq C(\mu_0^A)/2 - 1/2 \cdot \mathbb{E}_{t^0 \sim \Pi^0} C(\mu_{0t^0}) + C(\mu_0^A)/2 - 1/2 \cdot \mathbb{E}_{t^1 \sim \Pi^1} C(\mu_{1t^1}) + I(X; B|Y) \text{ (by induction)} \\ &= C(\mu_0^A)/2 + C(\mu_0^A)/2 + I(X; B|Y) - \mathbb{E}_{t \sim \Pi} C(\mu_t) \\ &\geq C(\mu) - \mathbb{E}_{t \sim \Pi} C(\mu_t) \text{ (by properties of } C) \end{aligned}$$

□

Lemma 5.7. *Let $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a given function. Let τ be a protocol that solves f correctly on all inputs. Then for all $C \in \mathfrak{C}(f)$ and all $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$ we have $C(\mu) \leq \mathrm{IC}_\mu(\tau)$*

Proof. Let G_π denote the set of leaves t of π such that $f|_{\mathrm{supp}(\mu_t)}$ is constant. By Lemma 5.2, for all $\delta > 0$, there exists a protocol π_δ in normal form that δ-simulates τ and $IC_\mu(\pi_\delta) \leq IC_\mu(\tau)$. Then we have $\sum_{t \in G_{\pi_\delta}} Pr[\Pi_\delta = t] \geq (1 - \delta)$. Moreover, by definition of C we have $C(\mu_t) = 0$ for constant $f|_{\mathrm{supp}(\mu_t)}$ and $C(\mu) \leq \log(|\mathcal{X} \times |\mathcal{Y}|)$ for all μ. Thus by Lemma 5.6 it follows that that $C(\mu) \leq IC_\mu(\pi_\delta) + \delta \cdot \log(|\mathcal{X}| \cdot |\mathcal{Y}|)$. As it holds for all $\delta > 0$, we have $C(\mu) \leq IC_\mu(\tau)$. □

Corollary 5.8. *For all $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ we have*

1. $\mathrm{IC}_\mu(f, 0) \in \mathfrak{C}(f)$,
2. *for all $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$ and for all $C \in \mathfrak{C}(f)$ we have $\mathrm{IC}_\mu(f, 0) \geq C(\mu)$.*
3. *for all $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$ we have $\mathrm{IC}_\mu(f, 0) = \max_{C \in \mathfrak{C}(f)} C(\mu)$.*

Remark 5.9. The above definitions and claims can be repeated for *the external information cost*. In the concavity constraints, replacing $I(X; B|Y)$ and $I(Y; B|X)$ with $I(XY; B)$, we obtain a class $\mathfrak{C}^{\mathrm{ext}}(f)$ of all lower bounds on the $I^{\mathrm{ext}}(\mu) := \mathrm{IC}_\mu^{\mathrm{ext}}(f)$. Repeating the steps of Lemma 5.6 and Lemma 5.7 but replacing internal information cost with external information cost, we arrive at similar conclusions as in Corollary 5.8:

1. $\mathrm{IC}_\mu^{\mathrm{ext}}(f, 0) \in \mathfrak{C}^{\mathrm{ext}}(f)$,
2. for all $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$ and for all $C \in \mathfrak{C}^{\mathrm{ext}}(f)$ we have $\mathrm{IC}_\mu^{\mathrm{ext}}(f, 0) \geq C(\mu)$.
3. for all $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$ we have $\mathrm{IC}_\mu^{\mathrm{ext}}(f, 0) = \max_{C \in \mathfrak{C}^{\mathrm{ext}}(f)} C(\mu)$.

6 On Continuity of IC

In [8] it was shown that the information cost function $\mathrm{IC}_\mu(f, \epsilon)$ is convex in ϵ on the interval $[0, 1]$. An immediate corollary is that the information cost is continuous in ϵ on the open interval $(0, 1)$. Note that the information cost is trivially continuous at $\epsilon = 1$. However, this still left open a question whether $\mathrm{IC}_\mu(f, \epsilon)$ is continuous at $\epsilon = 0$. In this section we prove that it is. This property turns out to be essential for our work. We arrive at the result in two major steps: in Section 6.1 we take the matrix view of message transmission in communication protocols, which in Section 6.2 lets us exploit the rectangular nature of protocols. We show that protocols solving f with small probability of error must terminate with a distribution that with high probability has all but a negligible weight on monochromatic rectangles. To turn such a protocol into a zero-error protocol, the players may simply verify that their inputs belong to such a rectangle, and if so they know the answer, otherwise they exchange the inputs.

6.1 Matrix View of Message Transmission

Since information complexity only makes sense in the distributional communication model, we assume that the players' inputs come from a publicly known prior distribution μ. Let π be a communication protocol under μ. Transmitting a bit in π is equivalent to updating the prior μ: As the players exchange bits, they keep track of a sequence of priors $\mu_1, \ldots, \mu_{CC(\pi)}$. The protocol dictates the rules for how prior μ_{i+1} is obtained from μ_i. More specifically, if Alice talks at step i then μ_{i+1} is obtained from μ_i by multiplying rows of μ_i by certain numbers, and if Bob talks at step i then μ_{i+1} is obtained from μ_i by multiplying columns. This equivalence of bit transmission and changes to the prior is formalized in the following lemma.

Lemma 6.1. *Let $\mu, \mu_0, \mu_1 \in \Delta(\mathcal{X} \times \mathcal{Y})$. The following two statements are equivalent:*

1. *There exists signal B that Bob can send such that $\mu_i(x,y) = P(X = x, Y = y \mid B = i)$ for $i \in \{0,1\}$.*
2. *There exists $t \in (0,1)$ and $\delta_0^y \in [0, 1/t], \delta_1^y \in [0, 1/(1-t)]$ $(y \in \mathcal{Y})$ such that*
   - $\mu = t\mu_0 + (1-t)\mu_1$
   - $(\forall i \in \{0,1\})(\forall (x,y) \in \mathcal{X} \times \mathcal{Y})(\mu_i(x,y) = \delta_i^y \mu(x,y))$.

Similarly for Alice, but with rows.

Proof. ($\Rightarrow$) By definition, $\mu_i(x,y) = P(X = x, Y = y | B = i)$, which by Bayes' rule is equivalent to $\mu_i(x,y) = P(B = i | X = x, Y = y)P(X = x, Y = y)/P(B = i)$. Since Bob is the speaker, $P(B = i | X = x, Y = y) = P(B = i | Y = y)$. Thus we have

$$\mu_i(x,y) = \frac{P(B = i | Y = y)}{P(B = i)} \mu(x,y).$$

Defining $\delta_i^y = P(B = i | Y = y)/P(B = i)$ and $t = P(B = 0)$ finishes the proof of (1) $\Rightarrow$ (2).

($\Leftarrow$) Define signal B by $P(B = 0 | Y = y) := t\delta_0^y$ and $P(B = 1 | Y = y) := (1-t)\delta_1^y$. For each y this defines a valid distribution on $\{0,1\}$, because $\mu(x,y) = t\mu_0(x,y) + (1-t)\mu_1(x,y) = t\delta_0^y \mu(x,y) + (1-t)\delta_1^y \mu(x,y)$. Fix an $x \in \mathcal{X}$ such that $\mu(x,y) \neq 0$, then $t\delta_0^y + (1-t)\delta_1^y = 1$ (if no such x exists, y is never observed as an input, and $P(B = i | Y = y)$ can be defined to be whatever we want, e. g., 1/2).

Next, observe that $P(B = 0) = \sum_y P(B = 0 | Y = y)P(Y = y) = \sum_{x,y} t\delta_0^y \mu(x,y) = \sum_{x,y} t\mu_0(x,y) = t$. Thus, we have defined the signal B in such a way that $\delta_i^y = P(B = i | Y = y)/P(B = i)$, and consequently we have $\mu_i(x,y) = P(X = x, Y = y | B = i)$ (following the steps of ($\Rightarrow$) direction in reverse). □

Corollary 6.2. *For every protocol π, prior $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$ and a transcript $t \in \{0,1\}^{CC(\pi)}$, there exist vectors $V_r^t \in (\mathbb{R}^+)^{|\mathcal{X}|}$ and $V_c^t \in (\mathbb{R}^+)^{|\mathcal{Y}|}$ such that for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$ we have*

$$P(X = x, Y = y | \Pi(X,Y) = t) = V_r^t(x) V_c^t(y) \mu(x,y).$$

Here $\mathbb{R}^+$ is the set of non-negative reals.

6.2 Continuity of IC at 0-error

In this section we prove Theorem 2.8.
Clearly we have $\mathrm{IC}_\mu(f, \epsilon) \leq \mathrm{IC}_\mu(f, 0)$, since the infimum on the left-hand side of (1) ranges over a larger set of protocols. To prove the claim we show that the reverse inequality holds up to a small additive error, i.e., $\mathrm{IC}_\mu(f, 0) \leq \mathrm{IC}_\mu(f, \epsilon) + q(\epsilon)$ where $q(\epsilon) \to 0$ as $\epsilon \to 0$.

We first prove the theorem for full support distributions, and then show how to reduce the general case to it. To facilitate the proof for general distributions, it will be useful to prove the full support case for (complete) *relations* and not only functions. A relation R is *complete* if $\forall \ (x,y) \in \mathcal{X} \times \mathcal{Y} \ \exists \ z \in \mathcal{Z}$ *s.t* $(x,y,z) \in R$. We say a combinatorial rectangle G is z-monochromatic with respect to a relation R if there is some $z \in \mathcal{Z}$ such that $\forall \ (x,y) \in G$, $(x,y,z) \in R$. We define the *color* of a monochromatic rectangle G to be the first lexicographically ordered $z \in \mathcal{Z}$ for which G is z-monochromatic.

Lemma 6.3. *Let $R \subseteq \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ be a complete relation, and let μ be a full support distribution on pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Denote by $\rho := \min_{(x,y)} \mu(x, y)$ the minimum mass of an element under μ. Then for any $\epsilon > 0$ small enough, zero-error information complexity of R under μ, $IC_\mu(R, 0)$ is at most*

$$\mathrm{IC}_\mu(R, \epsilon) + 2\Big(H(1 - 2\frac{|\mathcal{X}||\mathcal{Y}|\epsilon^{1/4}}{\rho}) + 2(\log|\mathcal{X}| + \log|\mathcal{Y}| + 2)\frac{|\mathcal{X}||\mathcal{Y}|\epsilon^{1/4}}{\rho}\Big).$$

Proof. For $\beta > 0$, let π be a protocol with $\mathrm{IC}_\mu(\pi) \leq \mathrm{IC}_\mu(R, \epsilon) + \beta$ that solves R with probability of error ϵ on every input. That is, the probability (over $(x, y) \sim \mu$) that that the parties output an element $\pi(x, y) \in \mathcal{Z}$ such that $R(x, y, \pi(x, y)) \notin R$ is at most ϵ. For each transcript $t \in \{0, 1\}^{C(\pi)}$, the players end up with a shared posterior distribution μ_t on $\mathcal{X} \times \mathcal{Y}$, which is obtained by iterated multiplications of either the rows or columns of the previous distribution in each round of the protocol. Let V_c^t, V_r^t denote the column and row multiplying coefficient vectors for the transcript t. That is, for all (x, y), $\mu_t(x, y) = V_r^t(x)V_c^t(y)\mu(x, y)$. Such vectors exist by Corollary 6.2. Note that for any t, V_c^t, V_r^t are known to *both* Alice and Bob. Protocol 1 is the 0-error protocol τ constructed out of π.

1. Players run π. Let t be the resulting transcript.
2. Let $L = \{x \mid V_r^t(x) > \epsilon^{1/4}/\rho^{1/2}\}$ and $M = \{y \mid V_c^t(y) > \epsilon^{1/4}/\rho^{1/2}\}$. Note that L, M are computable by both parties.
3. Alice and Bob (privately) check whether $L \times M$ is a monochromatic rectangle, and if not they exchange inputs.
4. Otherwise Alice sends a bit indicating if her input is in L.
5. Similarly, Bob sends a bit indicating if his input is in M.
6. If inputs belong to $L \times M$, the players output the color of $L \times M$.
7. Otherwise, they exchange inputs.

Protocol 1: 0-error protocol τ for f constructed out of π.

The intuition behind Protocol 1 relies on the fact that Alice's and Bob's communicated bits simply multiply rows and columns (respectively) of the original distribution μ. It follows that since the protocol's error is small on most transcripts, the final distributions of most transcripts must be concentrated on monochromatic rectangles. Verifying that players' inputs lie in such a rectangle reveals negligible amount of information.

We now turn to formalize this intuition. Let $\mathcal{E}$ be the event that π makes a mistake, and let $\mathcal{E}_t$ denote the event that π makes a mistake given that transcript is t. We have $P(\mathcal{E}) = \mathbb{E}_t(P(\mathcal{E}_t)) \leq \epsilon$ and by Markov's inequality it follows that

$$P_t(P(\mathcal{E}_t) > \epsilon^{1/2}) \leq \epsilon^{1/2}.$$

For the remainder of the argument, consider a transcript t such that $P(\mathcal{E}_t) \leq \epsilon^{1/2}$. We begin with the following claim which upper bounds the maximal entry in V_c^t, V_r^t:

Claim 6.4. *(1) w.l.o.g, $\|V_r^t(x))\|_\infty = \|V_c^t(y))\|_\infty$.*
(2) $\|V_r^t(x))\|_\infty, \|V_c^t(y))\|_\infty < 1/\sqrt{\rho}$.

Proof. Let m_c be the max entry of V_c^t, and m_r be the max entry of V_r^t, and suppose that $m_c > m_r$. Then we can divide V_r^t and multiply V_c^t by $d = \sqrt{m_r/m_c}$, without affecting the distribution μ_t (indeed, recall that $\mu_t(x, y) = V_r^t(x)V_c^t(y)\mu(x, y)$). the resulting vectors $d \cdot V_r^t$ and $(1/d) \cdot V_c^t$ satisfy (1).
(2) Recall that $\rho := min_{(x,y)}\mu(x, y)$. Let (x^*, y^*) be the maximum entries in V_r^t, V_c^t respectively (if these entries aren't unique, let x^*, y^* be the first such entries). By (1), $V_c^t(y*) = V_r^t(x*)$. Thus, if these numbers are larger than $1/\sqrt{\rho}$, then $\mu_t(x*, y*) \geq V_c^t(y*)V_r^t(x*)\mu(x*, y*) > 1$, Contradiction . (Note that here we crucially use the assumption that μ is full support by assuming $\mu(x*, y*) > 0$). □

The following claim asserts that all but a negligible mass of μ_t lies on a single monochromatic rectangle.

Claim 6.5. *The rectangle $L \times M$ defined in step 3 of protocol 1 satisfies the following properties:*
(1) $L \times M$ is monochromatic.
(2) $\mu_t(L), \mu_t(M) \geq 1 - \frac{|\mathcal{Y}| \cdot |\mathcal{X}| \cdot \epsilon^{1/4}}{\rho}$.

Proof. (1) Suppose not, then there exists some input $(x_0, y_0) \in L \times M$ such that $(x_0, y_0, \tau_t(x_0, y_0)) \notin R$, and therefore the error probability of τ_t is at least $\mu_t(x_0, y_0) = V_r^t(x_0) V_c^t(y_0) \mu(x_0, y_0) > (\sqrt{\epsilon}/\rho) \cdot \rho = \sqrt{\epsilon}$ (by definition of $L \times M$), contradicting the assumption that $P(\mathcal{E}_t) \leq \epsilon^{1/2}$.
(2) let $x^* \notin L$. Then

$$\mu_t(x^*) = \sum_y \mu_t(x^*, y) \leq |\mathcal{Y}| \cdot (\epsilon^{1/4}/\rho^{1/2}) \cdot (1/\rho^{1/2}) \cdot 1 = \frac{|\mathcal{X}| \cdot \epsilon^{1/4}}{\rho}$$

since by claim 6.4, $V_c^t(y) < 1/\rho^{1/2}$ for all y. By a union bound over all $x \notin L$, we get that $\mu_t(\bar{L}) \leq \frac{|\mathcal{Y}| \cdot |\mathcal{X}| \cdot \epsilon^{1/4}}{\rho}$. A similar proof holds for M.

□

Now, let τ_1 denote the part of transcript of τ that corresponds to running π and τ_2 the remaining part of transcript of τ. Let S be an indicator random variable of the event "players do not exchange inputs in τ_2". We have

$$\begin{aligned}
&P(S = 1) \geq P(P(\mathcal{E}_t) \leq \epsilon^{1/2})) P((X, Y) \in L \times M | P(\mathcal{E}_t) \leq \epsilon^{1/2}) \\
&\geq (1 - \epsilon^{1/2})(1 - |\mathcal{X}||\mathcal{Y}|\epsilon^{1/4}/\rho)^2 \\
&\geq 1 - 2|\mathcal{X}||\mathcal{Y}|\epsilon^{1/4}/\rho
\end{aligned}$$

for all small enough ϵ. Since S is determined by τ_2 we have

$$\begin{aligned}
H(\tau_2) = H(\tau_2 S) &= H(S) + H(\tau_2 | S) \\
&= H(S) + H(\tau_2 | S = 0) p(S = 0) + H(\tau_2 | S = 1) p(S = 1) \\
&= H(S) + H(\tau_2 | S = 0) p(S = 0) \\
&\leq H(1 - 2\frac{|\mathcal{X}||\mathcal{Y}|\epsilon^{1/4}}{\rho}) + 2(\log|\mathcal{X}| + \log|\mathcal{Y}| + 2)\frac{|\mathcal{X}||\mathcal{Y}|}{\rho}\epsilon^{1/4}
\end{aligned}$$

where (1) $H(\tau_2 | S = 1) = 0$, since when players do not exchange inputs $\tau_2 =$"11", (2) $H(p)$ is a decreasing function for $p \in [1/2, 1]$, and (3) $H(\tau_2 | S = 0) \leq \log|\mathcal{X}| + \log|\mathcal{Y}| + 2$, since when players exchange inputs $range(\tau_2) = \mathcal{X} \times \mathcal{Y} \times \{0, 1\}^2$.

Now, we can relate information cost of τ to that of π.

$$\begin{aligned}
&\mathrm{IC}_\mu(\tau) = I(\tau; X|Y) + I(\tau; Y|X) \\
&= I(\tau_1\tau_2; X|Y) + I(\tau_1\tau_2; Y|X) \\
&= I(\tau_1; X|Y) + I(\tau_2; X|Y\tau_1) + I(\tau_1; Y|X) + I(\tau_2; Y|X\tau_1) \\
&= \mathrm{IC}_\mu(\pi) + I(\tau_2; X|Y\tau_1) + I(\tau_2; Y|X\tau_1) \\
&\leq \mathrm{IC}_\mu(f, \epsilon) + \beta + 2H(\tau_2).
\end{aligned}$$

The above inequality holds for all $\beta > 0$ and therefore the 0-error information complexity of R is

$$\begin{aligned}
&\mathrm{IC}_\mu(R, 0) \leq \mathrm{IC}_\mu(R, \epsilon) + 2H(\tau_2) \\
&\leq \mathrm{IC}_\mu(R, \epsilon) + 2\Big(H(1 - 2\frac{|\mathcal{X}||\mathcal{Y}|\epsilon^{1/4}}{\rho}) + 2(\log|\mathcal{X}| + \log|\mathcal{Y}| + 2)\frac{|\mathcal{X}||\mathcal{Y}|\epsilon^{1/4}}{\rho}\Big).
\end{aligned}$$

as claimed.

□

With Lemma 6.3 at hand, we are finally ready to prove Theorem 2.8.

Proof. To prove part (1) of the theorem, let μ be any distribution over $\mathcal{X} \times \mathcal{Y}$ (not necessarily full support). Let Π be an ϵ-error protocol for f under μ, and denote $I := \mathsf{IC}_\mu(\Pi)$. Denote by U the uniform distribution on $\mathcal{X} \times \mathcal{Y}$, and define the distribution $\mu' := p \cdot U + (1-p) \cdot \mu$ for $p = \epsilon^{1/8}$. Note that μ' is full support, and that for small enough ϵ, $\rho = \min_{(x,y)} \mu'(x,y) = p = \epsilon^{1/8}$.
Define the relation $R_f \subseteq \mathcal{X} \times \mathcal{Y} \times \{0,1\}$ so that $(x,y,f(x,y)) \in R_f$ for all $(x,y) \in Supp(\mu)$, and otherwise $(x,y,z) \in R_f$ for $z = \{0,1\}$ (So R_f is trivially satisfied outside the support of μ, and inside it, it agrees with f). Clearly, Π is an ϵ-error protocol for R_f under μ, and since R_f is always satisfied outside $Supp(\mu)$, Π is also an ϵ-error[3] protocol for R_f under μ'. By Lemma 6.3, there is a zero-error protocol τ for R_f under μ', whose information cost is at most

$$\mathrm{IC}_{\mu'}(\tau) \leq \mathrm{IC}_{\mu'}(\Pi) + \alpha,$$

for $\alpha := 2\Big(H(1 - 2\frac{|\mathcal{X}||\mathcal{Y}|\epsilon^{1/4}}{\rho}) + 2(\log|\mathcal{X}| + \log|\mathcal{Y}| + 2)\frac{|\mathcal{X}||\mathcal{Y}|}{\rho}\epsilon^{1/4}\Big)$.
Since $\|\mu - \mu'\| \leq p$, Lemma B.1 implies that

$$\mathrm{IC}_{\mu'}(\Pi) \leq \mathrm{IC}_{\mu}(\Pi) + 2p(\log|\mathcal{X}| \cdot |\mathcal{Y}|) + 2H(2p)$$

and therefore

$$\mathrm{IC}_{\mu'}(\tau) \leq \mathrm{IC}_{\mu}(\Pi) + 2p(\log|\mathcal{X}| \cdot |\mathcal{Y}|) + 2H(2p) + \alpha.$$

But by definition of R_f, τ is clearly a zero-error protocol for f under μ. Using Lemma B.1 again, we have

$$\begin{aligned}
\mathrm{IC}_\mu(\tau) &\leq \mathrm{IC}_{\mu'}(\tau) + 2p(\log|\mathcal{X}| \cdot |\mathcal{Y}|) + 2H(2p) \\
&\leq \mathrm{IC}_\mu(\Pi) + 4p(\log|\mathcal{X}| \cdot |\mathcal{Y}|) + 4H(2p) + \alpha \\
&\leq I + 4\epsilon^{1/8}(\log|\mathcal{X}| \cdot |\mathcal{Y}|) + 4H(2\epsilon^{1/8}) + 2\Big(H(1 - 2\frac{|\mathcal{X}||\mathcal{Y}|\epsilon^{1/4}}{\epsilon^{1/8}}) + 2(\log|\mathcal{X}| + \log|\mathcal{Y}|)\frac{|\mathcal{X}||\mathcal{Y}|}{\epsilon^{1/8}}\epsilon^{1/4}\Big) \\
&= I + 4\epsilon^{1/8}(\log|\mathcal{X}| \cdot |\mathcal{Y}|) + 4H(2\epsilon^{1/8}) + 2\Big(H(1 - 2 \cdot |\mathcal{X}||\mathcal{Y}| \cdot \epsilon^{1/8}) + 2(\log|\mathcal{X}| + \log|\mathcal{Y}|)|\mathcal{X}||\mathcal{Y}|\epsilon^{1/8}\Big),
\end{aligned}$$

and clearly all the terms except I in the above quantity tend to 0 when $\epsilon \to 0$.

To prove part (2) of the theorem we follow exactly the same steps as in part (1).

□

7 The 0-error Information Cost of AND

In this section we shall compute the *exact internal and the external information cost* of the 2-bit AND function. We summarize our findings about the AND function in Section 7.1. In Section 7.3 we present a *clocked protocol* π for the AND function, in which the parties use a continuously increasing clock in an asynchronous fashion (this will become clearer soon). The protocol π is infeasible in the sense that no finite-round protocol can simulate it; however, we may still analyze its information cost as a function of the input distribution μ. We use the machinery developed in the previous sections to demonstrate that the information cost function of π gives a lower bound on the IC (Sections 7.6 and 7.8) of AND. Hence, the information cost of π is precisely the information cost of the AND *function*. Thus, the infeasibility of π is an expected side effect - the information cost of a function is the infimum over protocols, and thus may not be achievable by any finite-round protocol. In Section 7.9 we describe a natural finite-round discretization of π and analyze

[3] In fact, Π has error at most $(1-p)\epsilon$ under μ'.

its rate of convergence to the true (unbounded-round) information cost of AND, as a function of the number of rounds.

The protocol π suggests that the space of distributions on $\{0,1\} \times \{0,1\}$ is partitioned into three regions - "Alice's region", "Bob's region", and a "diagonal" region (corresponding to symmetric distributions). Section 7.4 describes the regions and how together with the results from Section 7.2 they reduce the number of cases necessary to consider in the analysis of the information cost function of π.

7.1 Summary of Results for AND

In Sections 7.5, 7.6, 7.7, and 7.8 we shall derive exact closed-form formulas for the distributional internal and external 0-error information costs of the AND function. In this section we present the main results.

Theorem 7.1 (Theorem 2.2 restated)**.**

$$\mathrm{IC}(AND, 0) = C_{\wedge} = 1.49238\ldots$$

Proof. The prior-free information cost of a function is just a maximum over distributions of the distributional information cost. The precise number $C_{\wedge}$ was obtained via numerical optimization of the formulas obtained in Sections 7.5 and 7.6, using Wolfram Mathematica. The distribution that achieves this maximum is

$$\mu = \begin{array}{|c|c|}\hline 0.0808931\ldots & 0.264381\ldots \\ \hline 0.264381\ldots & 0.390346\ldots \\ \hline \end{array}.$$

□

Remark 7.2. Observe that there is a symmetric distribution that achieves the maximum of $\mathrm{IC}(\mathrm{AND}, 0)$. This holds for all symmetric functions. Let f be a symmetric function and μ be an arbitrary distribution on the inputs of f. Then $\mathrm{IC}_{\mu}(f, 0) = \mathrm{IC}_{\mu^T}(f, 0)$ and it is easy to see that the information complexity is a concave function in μ (Lemma A.1). Thus for $\mu' = \mu/2 + \mu^T/2$, which is symmetric, we have $\mathrm{IC}_{\mu'}(f, 0) \geq \mathrm{IC}_{\mu}(f, 0)/2 + \mathrm{IC}_{\mu^T}(f, 0)/2 = \mathrm{IC}_{\mu}(f, 0)$. The same holds for the external information cost.

Theorem 7.3 (Theorem 2.3 restated)**.**

$$\mathrm{IC}^{ext}(AND, 0) = \log 3 = 1.58396\ldots$$

Proof. Even the external information complexity is concave, so the distribution that achieves the maximum has to be symmetric. We first show an upper bound. That is for every distribution $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \beta & \delta \\ \hline \end{array}$ we have $\mathrm{IC}^{\mathrm{ext}}_{\mu}(\mathrm{AND}, 0) \leq \log 3$. Consider a trivial protocol, in which Alice sends her bit X. Then if X turns out to be 1, Bob sends his bit. The information cost of this protocol is

$$\begin{aligned} &H(X) + p(X=1)H(Y|X=1) = (\alpha+\beta)\log\frac{1}{\alpha+\beta} + (\beta+\delta)\log\frac{1}{\beta+\delta} + \\ &\quad + (\beta+\delta)\left(\frac{\beta}{\beta+\delta}\log\frac{\beta+\delta}{\beta} + \frac{\delta}{\beta+\delta}\log\frac{\beta+\delta}{\delta}\right) \\ &= (\alpha+\beta)\log\frac{1}{\alpha+\beta} + \beta\log\frac{1}{\beta} + \delta\log\frac{1}{\delta} = H(\mu'), \end{aligned}$$

where μ' is a distribution on a sample space with three elements $1, 2, 3$ and $\mu'(1) = \alpha+\beta, \mu'(2) = \beta, \mu'(3) = \delta$. Since Shannon entropy is maximized for a uniform distribution, we immediately get that the information cost of the above protocol is at most $\log 3$.

Now we turn to the lower bound. Consider the distribution

$$\mu = \begin{array}{|c|c|}\hline 0 & 1/3 \\ \hline 1/3 & 1/3 \\ \hline \end{array}.$$

Let π be a 0-error protocol for solving AND and let $X, Y \sim \mu$. Let Π denote the transcript of π on inputs X, Y. Because of the rectangle property of protocols, it follows that Π determines X, Y. Thus $I(\Pi; XY) = H(XY) = \log 3$.

□

Remark 7.4. Although the trivial protocol is optimal for the worst distribution (and for distributions with $\alpha = 0$), it isn't optimal for distributions with $\alpha \neq 0$. The protocol we present is optimal for the case $\alpha \neq 0$.

The following theorem plays a crucial role in providing the exact communication complexity of the disjointness function in Section 8.

Theorem 7.5.

$$\lim_{\epsilon \to 0} \max_{\mu : \mu(1,1) \leq \epsilon} \mathsf{IC}_\mu(AND, 0) = 0.482702\ldots.$$

The above result is obtained from the formulas from Section 7.5 using Wolfram Mathematica.

When in later sections we consider the disjointness function, distributions μ that place 0 mass on $(1, 1)$ entry will play a crucial role. Note that for such distributions we still insist that the protocol solving AND has 0 error on *all* inputs. The following two claims describe the information cost of such distributions.

Claim 7.6. *For symmetric distributions*

$$\mu = \begin{array}{|c|c|} \hline \alpha & \beta \\ \hline \beta & 0 \\ \hline \end{array}$$

we have

$$\mathrm{IC}_\mu(AND, 0) = \frac{\beta}{\ln 2} + \frac{\beta^2}{\alpha} \log \frac{\beta}{\alpha + \beta} + \alpha \log \frac{\alpha + \beta}{\alpha}.$$

Proof. Immediate from formulas from Section 7.6. Note that although we measure the information cost w.r.t a distribution that has zero mass on $(1, 1)$, we still require the protocol to be correct for *all* inputs. □

Claim 7.7. *For distributions*

$$\mu = \begin{array}{|c|c|} \hline \alpha & \beta \\ \hline \gamma & 0 \\ \hline \end{array}$$

we have

$$\mathrm{IC}_\mu(AND, 0) = (\alpha + \beta) H\left(\frac{\beta}{\gamma} \frac{\alpha + \gamma}{\alpha + \beta}\right) - \alpha H\left(\frac{\beta}{\gamma}\right) + t\, \mathrm{IC}_\nu(AND, 0),$$

where

$$t = 2\beta + \frac{\alpha\beta}{\gamma}$$

and

$$\nu = \begin{array}{|c|c|} \hline \frac{\beta\alpha}{\gamma t} & \frac{\beta}{t} \\ \hline \frac{\beta}{t} & 0 \\ \hline \end{array}$$

.

Proof. Immediate from formulas from Section 7.6. □

We will also need the following claim about the information cost of symmetric distributions with non-zero mass on $(1, 1)$.

Claim 7.8. *For a symmetric distribution* $\mu = \begin{array}{|c|c|} \hline \alpha & \beta \\ \hline \beta & \delta \\ \hline \end{array}$ *we have*

$$\mathrm{IC}_\mu(\pi) = \frac{\beta}{\ln 2} + 2\delta \log \frac{\beta + \delta}{\delta} + +2\beta \log \frac{\beta + \delta}{\beta} + \frac{\beta^2}{\alpha} \log \frac{\beta}{\beta + \alpha} + \alpha \log \frac{\alpha + \beta}{\alpha}.$$

Proof. Immediate from formulas from Section 7.6. □

In Section 7.9 we prove Theorem 2.4 - a tight bound on the rate of convergence of the r-round information cost of the AND function to the unbounded-round information cost:

Theorem 7.9 (Theorem 2.4 restated)**.** *For all $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$ with full support we have*

$$\mathrm{IC}^r_\mu(AND, 0) = \mathrm{IC}_\mu(AND, 0) + \Theta_\mu\left(\frac{1}{r^2}\right).$$

Moreover, the lower bound holds even for μ such that $\mu(1,1) = 0$.

7.2 Distribution on Distributions and IC

A natural question which arises when we take the "informational view" of a protocol as a random walk on $\Delta(\mathcal{X} \times \mathcal{Y})$, is whether the amount of information revealed in a single step of a protocol depends on how that step was performed. Each step of a protocol can be viewed as follows: starting from a commonly known prior distribution μ on the inputs (x, y), the current speaker transmits a message $M \in_R \{0,1\}^t$ where t is the largest length of a message for this step. When a certain instance m of the message is communicated, the players update their common prior knowledge to $\mu_m(x,y) \stackrel{\text{def}}{=} P(X = x, Y = y | M = m)$. Observe that different messages $\widetilde{m}$ may lead to the same distribution $\mu_{\widetilde{m}} = \mu_m$.

Definition 7.10. For a message M we define the *distribution on distributions for M* as follows: the sample space is $\Omega = \{\mu_m \mid m \in \mathrm{range}(M)\}$ and the distribution p on Ω is $p(\mu_m) = P(\mu_M = \mu_m) = \sum_{\widetilde{m}:\mu_{\widetilde{m}}=\mu_m} P(M = \widetilde{m})$.

We shall use notation $(\{\mu_1, \mu_2, \ldots\}, \{p_1, p_2, \ldots\})$ to denote a particular distribution on distributions.

In this section we show that the information cost of a step depends only on the distribution on distributions, and not on the message itself. In other words, the player may transmit an arbitrary message M' instead of M, and it will reveal the same information as M, as long as M' induces the same Ω and p.

The tools developed in the current section shall be used later to reduce the number of cases necessary to consider in the analysis of the information cost of AND function. One such tool is the distribution on distributions. Another tool is the Splitting Lemma: if a player can "split" prior μ into μ_0 and μ_1 by transmitting a bit, then the same player can split any prior ρ into any $\rho_0, \rho_1 \in [\mu_0, \mu_1]$ satisfying $\rho \in [\rho_0, \rho_1]$ by transmitting a bit. Essentially it says splitting cares only about the direction.

The proof of the Splitting Lemma uses the matrix view of message transmission (Lemma 6.1). Since the transmitted bit B satisfies the assumptions of Lemma 6.1, we may express μ_0 and μ_1 as μ with its columns scaled by certain *scaling coefficients* (direction (1) ⇒ (2) of Lemma 6.1). Every distribution in the interval $[\mu_0, \mu_1]$ is a linear combination of "column-scaled" versions of μ, and thus is a "column-scaled" μ itself. Finding scaling coefficients for ρ_0, ρ_1 and ρ we observe that ρ_0 and ρ_1 are, in fact, "column-scaled" versions of ρ. Applying direction (2) ⇒ (1) of Lemma 6.1 we arrive at the desired conclusion.

Lemma 7.11 (Splitting Lemma)**.** *Suppose that starting with $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$ Bob sends signal B such that $\mu_i(x,y) = P(X = x, Y = y \mid B = i)$. Let $\rho_0, \rho_1 \in [\mu_0, \mu_1]$ and $\rho \in [\rho_0, \rho_1]$. Then there exists signal B' that Bob can send starting at distribution ρ such that $\rho_i(x,y) = P(X = x, Y = y \mid B' = i)$. Similarly, when Alice sends bit B.*

Proof. Since $\rho_0, \rho_1 \in [\mu_0, \mu_1]$ there exist numbers $t_0, t_1 \in [0,1]$ such that $\rho_0 = t_0\mu_0 + (1 - t_0)\mu_1$ and $\rho_1 = t_1\mu_0 + (1 - t_1)\mu_1$. Also, since $\rho \in [\rho_0, \rho_1]$ we have $\rho = t\rho_0 + (1 - t)\rho_1$ for some $t \in [0,1]$. By direction (1) ⇒ (2) of Lemma 6.1 we have $\mu_i(x,y) = \delta^y_i \mu(x,y)$ for some δ^y_i, $i \in \{0,1\}, y \in \mathcal{Y}$. Then we can express ρ_0 and ρ_1 in terms of μ as follows:

$$\begin{aligned} \rho_0(x,y) &= (t_0\delta^y_0 + (1 - t_0)\delta^y_1)\mu(x,y), \\ \rho_1(x,y) &= (t_1\delta^y_0 + (1 - t_1)\delta^y_1)\mu(x,y). \end{aligned}$$

Define $C_0^y := t_0\delta_0^y + (1-t_0)\delta_1^y$ and $C_1^y := t_1\delta_0^y + (1-t_1)\delta_1^y$. Then we have

$$\rho(x,y) = (tC_0^y + (1-t)C_1^y)\mu(x,y).$$

Now, it is easy to see that ρ_0 and ρ_1 are "column-scaled" versions of ρ with scaling coefficients defined by

$$\widetilde{\delta}_i^y := \frac{C_i^y}{tC_0^y + (1-t)C_1^y}.$$

Overall, we have

1. $\rho = t\rho_0 + (1-t)\rho_1$,
2. $\rho_i(x,y) = \widetilde{\delta}_i^y \rho(x,y)$,
3. $\widetilde{\delta}_0^y = \frac{C_0^y}{tC_0^y + (1-t)C_1^y}$,
4. $\widetilde{\delta}_1^y = \frac{C_1^y}{tC_0^y + (1-t)C_1^y}$

Thus by Lemma 6.1 there exists a signal B' with the desired properties. □

Lemma 7.12 (Distribution on Distributions Lemma)**.** *Let μ be a prior on inputs $\mathcal{X} \times \mathcal{Y}$. Suppose that in one protocol starting with μ Bob transmits B such that $P(B=0) = P(B=1) = 1/2$ and $\mu_b(x,y) = P(X = x, Y = y \mid B = b)$ for $b \in \{0,1\}$. Suppose that in another protocol starting with μ Bob transmits a sequence of bits M such that*

- $\mu_m(x,y) := P(X = x, Y = y | M = m)$,
- $(\forall m \in \text{range}(M))(\mu_m \in \{\mu_0, \mu_1\})$,
- $P(\mathcal{M}_b) = P(B = b) = 1/2$, *where* $\mathcal{M}_b = \{m | \mu_m = \mu_b\}$ *for* $b \in \{0,1\}$.

Then we have

$$I(Y; M|X) = I(Y; B|X).$$

Proof. For all $b \in \{0,1\}$ and for all $m \in \mathcal{M}_b$ we have $\mu_m = \mu_b$, i. e., $P(X = x, Y = y | M = m) = P(X = x, Y = y | B = b)$. Hence $P(X = x | M = m) = P(X = x | B = b)$ and consequently $P(Y = y | X = x, M = m) = P(Y = y | X = x, B = b)$. We have

$$\begin{aligned}
&I(Y; M|X) = \\
&= \mathbb{E}_{x,m}(\mathbb{D}(Y_{xm} || Y_x) \\
&= \textstyle\sum_{x,y,m} P(X = x, Y = y, M = m) \log \frac{P(Y=y|X=x,M=m)}{P(Y=y|X=x)} \\
&= \textstyle\sum_{x,y,b} \sum_{m \in \mathcal{M}_b} P(X = x, Y = y, M = m) \log \frac{P(Y=y|X=x,B=b)}{P(Y=y|X=x)} \\
&= \textstyle\sum_{x,y,b} \sum_{m \in \mathcal{M}_b} \mu_m(x,y) P(M = m) \log \frac{P(Y=y|X=x,B=b)}{P(Y=y|X=x)} \\
&= \textstyle\sum_{x,y,b} \mu_b(x,y) P(\mathcal{M}_b) \log \frac{P(Y=y|X=x,B=b)}{P(Y=y|X=x)} \\
&= \textstyle\sum_{x,y,b} P(X = x, Y = y | B = b) P(B = b) \log \frac{P(Y=y|X=x,B=b)}{P(Y=y|X=x)} \\
&= \mathbb{E}_{x,b}(\mathbb{D}(Y_{xb} || Y_x)) \\
&= I(Y; B|X).
\end{aligned}$$

□

1. If $\beta < \gamma$ then Bob sends bit B as follows

$$B = \begin{cases} 1 & \text{if } y = 1 \\ 0 & \text{with probability } 1 - \beta/\gamma \text{ if } y = 0 \\ 1 & \text{with probability } \beta/\gamma \text{ if } y = 0 \end{cases}$$

   If $B = 0$ the protocol terminates, the players output 0.

2. If $\beta > \gamma$ then Alice sends bit B as follows

$$B = \begin{cases} 1 & \text{if } x = 1 \\ 0 & \text{with probability } 1 - \gamma/\beta \text{ if } x = 0 \\ 1 & \text{with probability } \gamma/\beta \text{ if } x = 0 \end{cases}$$

   If $B = 0$ the protocol terminates, the players output 0.

3. If $x = 0$ then Alice samples $N^A \in_R [0, 1)$ uniformly at random. If $x = 1$ then Alice sets $N^A = 1$.

4. If $y = 0$ then Bob samples $N^B \in_R [0, 1)$ uniformly at random. If $y = 1$ then Bob sets $N^B = 1$.

5. Alice and Bob monitor the clock C, which starts at value 0.

6. The clock continuously increases to 1. If $\min(N^A, N^B) < 1$, when the clock reaches $\min(N^A, N^B)$ the corresponding player sends 0 to the other player, the protocol ends, the players output 0. If $\min(N^A, N^B) = 1$, once the clock reaches 1, Alice sends 1 to Bob, the protocol ends, and the players output 1.

Protocol 2: Protocol π for the AND-function

7.3 The Protocol

In this section we present a zero-error protocol π for the function AND : $\{0,1\}^2 \to \{0,1\}$ (see Protocol 2), which achieves both the internal and the external information costs of AND. The inputs (X, Y) to AND are distributed according to $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & \delta \\ \hline \end{array}$.

Protocol 2 consists of two parts. In the first part (steps 1 and 2), Alice and Bob check to see if their prior is symmetric, and if it is not they communicate "a bit" to make it symmetric. We shall refer to the first part of π as its **non-symmetric part**. In the second part (steps $3 - 6$), Alice and Bob start with a symmetric prior and observe the clock as it increases from 0 to 1. As the time passes, the prior gets modified, but it remains symmetric. Also as the time passes, each player becomes more and more convinced that the other player has 1 as an input. The presence of this clock and this "continuous leakage" of information is precisely what makes this protocol infeasible - no finite-round protocol can simulate it: a finite-round protocol necessarily leaks bounded-from-zero amount of information in each round. In π when a player's private number (N^A or N^B) is reached by the clock, the player "raises the flag" to indicate the end of a protocol. The rules for picking the private numbers N^A and N^B can be intuitively justified by the following two observations:

1. When a player has input 0, that player does not need to know the other player's input. However, the other player must become aware that the first player has input 0, so that both players agree on the output of AND being 0.

2. When both players have 0 as input, their roles are completely symmetric, because AND is a symmetric function.

We shall refer to the second part of π as its **symmetric part**.

The intuition as to why π reveals little information is as follows: Since the protocol is zero error, the players must learn, with absolute certainty, either that they both have 1's, or that *at least one* of them has a 0 input. The "savings" of the protocol come from the latter case - Suppose that the protocol terminates with Alice announcing that the counter has reached her number N^A. In this case, Bob learns that Alice has a 0. But what does Alice know about Bob's input? Granted, Alice is now slightly more inclined to believe that Bob has a 1 (since $N^B > N^A$), but it is of course still quite probable that Bob has a 0, since in that case the numbers N^A and N^B are chosen independently at random and the latter event happens with probability 1/2. Thus, when the protocol terminates there is still much uncertainty left to Alice's knowledge about Bob's input.

Remark 7.13. In a well-defined protocol, the order in which the players communicate should depend solely on the partial transcript. For our "clocked" protocol, it is natural to require the order depend on the partial transcript and the value of the clock. This presents a small problem: in case $N^A = N^B < 1$ the players both transmit 0 simultaneously. However, this event "$N^A = N^B < 1$" happens with probability 0, thus we may pretend that it never happens.

From the definition of π, it is clear that it correctly solves AND on all inputs. Analyzing its information cost, on the other hand, requires careful calculations. This is what the remaining part of this section is devoted to. The division of π into **non-symmetric** and **symmetric** parts makes the calculations more modular and will appear throughout the rest of this section.

7.4 Regions of $\Delta(\{0,1\} \times \{0,1\})$ for the AND Function

Protocol 2 suggests that the space of distributions $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & \delta \\ \hline \end{array}$ on $\{0,1\} \times \{0,1\}$ is partitioned into three regions for the AND function:

1. **Bob's region** consisting of all distributions μ with $\beta < \gamma$,
2. **Alice's region** consisting of all distributions μ with $\beta > \gamma$,
3. **Diagonal region** consisting of all symmetric distributions μ with $\beta = \gamma$.

Bob's regions consists of all priors, for which Bob is more likely to have 0 as an input than Alice, i. e.,

$$P(Y = 0) = \alpha + \gamma > \alpha + \beta = P(X = 0).$$

Intuitively, if the players start with a prior in Bob's region then to achieve minimum leakage of information Bob should speak first until Alice is more likely to have 0 as an input. This happens when the prior moves into the Alice's region. Hence the names of the regions. If in some protocol Alice speaks in Bob's region, then that particular step releases non-optimal amount of information and may be improved by changing the speaker (see Sections 7.6 and 7.8).

Ideally, the players should try to keep the prior in the diagonal region - this corresponds to increasing Alice's and Bob's probabilities of having 1 as an input **simultaneously**. In a feasible (i. e., finite-round) protocol, once the prior is on the diagonal, the next bit of communication necessarily moves the prior off the diagonal with probability 1/2 (assuming normal form) into Alice's region and probability 1/2 into Bob's region, making that step non-optimal no matter who the speaker is. If the players could transmit infinitesimal amount of information at each step, they would be able to maintain the prior on the diagonal. This is exactly what the clock in Protocol 2 achieves.

In Sections 7.6 and 7.8 we shall demonstrate that the information cost function of Protocol 2 is also a lower bound on the information cost of AND *function* for each distribution by showing that the information cost of Protocol 2 satisfies the constraints of Definition 5.4.

We claim that among all possible signals that either Alice or Bob can send in Definition 5.4, it suffices to consider just three cases. Assume that the players start with a prior μ. The three cases are as follows:

1. the prior μ is in Bob's region, Bob sends a bit, the resulting distributions *remains* in Bob's region,

2. the prior μ is in Alice's region, Bob sends a bit, the resulting distributions *remains* in Alice's region,

3. the prior μ is in the diagonal region, Bob sends a bit, the resulting distributions fall in Alice's and Bob's regions.

The cases missing above are when Bob sends a bit and one of the resulting distribution "crosses the diagonal" (i. e., if we start in Bob's region and end up in Alice's region or start in Alice's region and end up in Bob's region). We refer to such bits as **crossing bits**, and bits of one of the forms above (1-3) as **non-crossing bits**. The following claim shows that we can replace every crossing bit with a sequence of non-crossing bits without changing the information carried by B.

Claim 7.14. *Any crossing bit B sent by Bob in an execution of a normal-form protocol may be replaced by a sequence $(B_1, B_2, \ldots)$ of non-crossing bits (in normal form) such that the distribution on distributions of $(B_1, B_2, \ldots)$ is the same as the distribution on distributions of B.*

Proof. Suppose that Bob's signal B starts at μ and has a distribution on distributions $(\{\mu_0, \mu_1\}, \{1/2, 1/2\})$ and moreover $[\mu_0, \mu_1]$ contains a symmetric distribution μ_D. We shall replace B with a sequence $(B_1, B_2, \ldots)$ representing the random walk on $[\mu_0, \mu_1]$ where each step is as large as possible under a constraint of not crossing μ_D, μ_0, and μ_1. If the random walk reaches μ_0 or μ_1 it terminates. Formally this simulation is described in Protocol 3.

Set $\mu_c \leftarrow \mu$

Set $i \leftarrow 1$

Repeat until $\mu_c = \mu_0$ or $\mu_c = \mu_1$

    If $(2\mu_c - \mu_D) \in [\mu_0, \mu_1]$ then

        Bob sends signal B_i (by Splitting Lemma 7.11) splitting μ_c into $2\mu_c - \mu_D$ and μ_D

    Else if $(2\mu_c - \mu_0) \in [\mu_0, \mu_1]$ then

        Bob sends signal B_i (by Splitting Lemma 7.11) splitting μ_c into $2\mu_c - \mu_0$ and μ_0

    Else

        Bob sends signal B_i (by Splitting Lemma 7.11) splitting μ_c into $2\mu_c - \mu_1$ and μ_1

    Update μ_c to the current distribution

    $i \leftarrow i + 1$

Protocol 3: simulating crossing bit B by a sequence of non-crossing bits.

Each bit sent in Protocol 3 is in normal form, hence the random walk on $[\mu_0, \mu_1]$ is unbiased. The optional stopping theorem from the theory of martingales implies that the probability of random walk reaching μ_0 is $1/2$. Hence the distribution on distributions is preserved. □

By Distribution on Distributions Lemma 7.12, the message $(B_1, B_2, \ldots)$ in Protocol 3 carries exactly the same information as the crossing bit B. Protocol 3 may not terminate, but this happens with probability 0. It can be overcome by a standard argument - truncating the protocol after a sufficiently large number of steps have been performed.

So far we have only considered Bob as a speaker. Observe that since the roles of Alice and Bob are completely symmetric, we do not have to consider the case when Alice sends a signal separately.

The same reasoning holds for the external information cost.

We can reduce the number of inequalities necessary to verify that the information cost function of Protocol 2 satisfies Definition 5.4 even further. We claim that case 1 above is automatically satisfied. Suppose that starting from μ in Bob's regions Bob sends a non-crossing bit B and then executes Protocol 2.

The information about inputs revealed by these two steps is exactly the same as if the players executed Protocol 2 from μ right away. We prove this in the rest of this section. Let π denote Protocol 2. First we need a simple lemma.

Lemma 7.15. *Let μ be a non-symmetric distribution $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & \delta \\ \hline \end{array}$ such that at least one symmetric distribution is reachable from μ if only Bob speaks. Then there exists a unique symmetric distribution μ_D such that for any message M that Bob can send we have $(\forall m \in \text{range}(M))(\mu_m \text{is symmetric} \Rightarrow \mu_m = \mu_D)$.*

Proof. Suppose that $\gamma < \beta$. Bob sending a message is equivalent to multiplying the columns of the matrix for μ by nonnegative numbers c_0, c_1. In order for Bob to arrive at a symmetric distribution he must achieve $c_0\gamma = c_1\beta$. There are two possibilities:

1. $\gamma = 0$ then $\beta \neq 0$ (μ is not symmetric). There is only one possibility for the resulting symmetric distribution $\begin{array}{|c|c|}\hline 1 & 0 \\ \hline 0 & 0 \\ \hline \end{array}$, which uniquely determines $(c_0, c_1) = (1/\alpha, 0)$.
2. $\gamma > 0$ then we must have $c_0 > 1$. But since the resulting matrix still has to correspond to a valid distribution we have $c_0(\alpha + \gamma) + c_1(\beta + \delta) = 1$. This forces $c_1 < 1$. Moreover, as c_0 decreases, c_1 increases. Thus, by continuity there is only one solution (c_0, c_1) satisfying $c_0\gamma = c_1\beta$.

□

Claim 7.16. *If Bob sends a non-crossing signal B in normal form starting from prior μ in Bob's region and having a distribution on distributions $(\{\mu_0, \mu_1\}, \{1/2, 1/2\})$ then*

$$\text{IC}_\mu(\pi) = \text{IC}_{\mu_0}(\pi)/2 + \text{IC}_{\mu_1}(\pi)/2 + I(B; Y|X).$$

In particular, constraint in Definition 5.4 is satisfied for such signals.

Proof. Define τ to be the following protocol:

1. Bob sends signal B as in the statement of the claim, resulting in distribution μ_B
2. The players run π starting at μ_B

Observe that expanding the information cost of τ after step 1 above we obtain

$$\text{IC}_\mu(\tau) = \text{IC}_{\mu_0}(\pi)/2 + \text{IC}_{\mu_1}(\pi)/2 + I(Y; B|X)$$

It is left to show that $\text{IC}_\mu(\tau) = \text{IC}_\mu(\pi)$.

Let π_1 denote the non-symmetric part (lines 1-2 in Protocol 2) of π when it is executed on μ and π_2 denote the remaining part of π. Let τ_1 denote the part of τ corresponding to step 1 above together with the non-symmetric part of π from step 2. Let τ_2 denote the remaining part of τ. To finish the proof it suffices to show that Π_1 and T_1 have the same distribution on distributions, because then the information content of messages Π_1 and T_1 would be the same by Lemma 7.12, and $\Pi_2|\Pi_1$ would have the same distribution as $T_2|T_1$ implying that $\text{IC}_\mu(\tau) = \text{IC}_\mu(\pi)$.

Suppose that the message Π_1 has a distribution on distributions $(\{\nu_0, \nu_1\}, \{t, 1-t\})$, i. e., $\mu = t\nu_0 + (1-t)\nu_1$, where ν_0 is the distribution after Bob sent 0 in the non-symmetric part of π (note: $P_{\nu_0}(Y = 1) = 0$) and ν_1 is the distribution on the diagonal.

Define random variables $X_0 = \mu$, $X_1 = \mu_B$ - the updated distribution after Bob sent bit B, and $X_2 = \mu_{T_1}$ - the updated distribution after τ_1 was executed. We have

1. $\mathbb{E}(X_2) = X_0 = \mu$, because X_0, X_1, X_2 is a martingale by definition of μ_B and μ_{T_1}, and
2. $X_2 \in \{\nu_0, \nu_1\}$ by Lemma 7.15 and a simple observation that there is a unique distribution $\widetilde{\nu}$ reachable by Bob from μ such that $P_{\widetilde{\nu}}(Y = 1) = 0$.

The above two facts imply that $P(X_2 = \nu_0) = t$. So T_1 has the same distribution on distributions as Π_1. □

In other words, we've shown that the information cost of a protocol is locally optimal with respects to steps that are "aligned" with the steps of the protocol.

7.5 Internal Information Cost: Upper Bound

We start by analyzing **the symmetric part** of Protocol 2, i. e., we shall compute $\mathrm{IC}_\nu(\pi)$ where

$$\nu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \beta & 1-\alpha-2\beta \\ \hline \end{array}$$

is a symmetric distribution.

Since ν is symmetric and the roles of Alice and Bob in Protocol 2 are symmetric, we have

$$\mathrm{IC}_\nu(\pi) = I(X;\Pi|Y) + I(Y;\Pi|X) = 2I(X;\Pi|Y).$$

Working from first-principles, we obtain

$$\begin{aligned} I(X;\Pi|Y) &= (\alpha+\beta)I(X;\Pi|Y=0) + (1-\alpha-\beta)I(X;\Pi|Y=1) \\ &= (\alpha+\beta)I(X;\Pi|Y=0) + (1-\alpha-\beta)H(X|Y=1) \\ &= \alpha\mathbb{D}(\Pi_{X=0,Y=0}||\Pi_{Y=0}) + \beta\mathbb{D}(\Pi_{X=1,Y=0}||\Pi_{Y=0}) \\ &\quad + (1-\alpha-\beta)H(X|Y=1). \end{aligned}$$

The second step follows from

$$I(X;\Pi|Y=1) = H(X|Y=1) - H(X|\Pi, Y=1)$$

and $H(X|\Pi, Y=1) = 0$, since given $Y=1$ the transcript Π determines X.

A transcript of Π on x, y can be represented uniquely by the value $c \in [0,1]$ of the clock when the protocol is terminated together with a name of a player $\mathcal{P} \in \{A, B\}$, whose random number is reached by a counter first. For $x, y \in \{0,1\}$ we have

$$\mathbb{D}(\Pi_{xy}||\Pi_y) = \sum_{\mathcal{P}\in\{A,B\}} \int_0^1 f_{x,y}(c,\mathcal{P}) \log \frac{f_{x,y}(c,\mathcal{P})}{f_y(c,\mathcal{P})} dc,$$

where $f_{x,y}(c,\mathcal{P})$ is the PDF for Π_{xy} and $f_y(c,\mathcal{P})$ is the PDF for Π_y.

We have

- $f_{0,0}(c,A) = f_{0,0}(c,B) = 1-c$ for $c \in [0,1]$
- $f_{1,0}(c,A) = 0$ for $c \in [0,1)$ and $f_{1,0}(c,B) = 1$ for $c \in [0,1)$
- $f_0(c,A) = \frac{\alpha}{\alpha+\beta}(1-c)$ for $c \in [0,1]$ and $f_0(c,B) = \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\beta+\alpha}(1-c)$ for $c \in [0,1)$

Overall we obtain

$$\begin{aligned} I(X;\Pi \mid Y) = \alpha \int_0^1 (1-c) \log \frac{\alpha+\beta}{\alpha} + (1-c) \log \frac{(1-c)(\alpha+\beta)}{\beta+(1-c)\alpha}\, \mathrm{d}c + \\ + \beta \int_0^1 \log \frac{\alpha+\beta}{\beta+(1-c)\alpha}\, \mathrm{d}c + (1-\alpha-\beta) H\left(\frac{\beta}{1-\alpha-\beta}\right). \end{aligned}$$

After using Wolfram Mathematica to simplify the expressions, we obtain:

$$\begin{aligned} \mathrm{IC}_\nu(\pi) = \frac{\beta}{\ln 2} + 2(1-\alpha-2\beta)\log\frac{1-\alpha-\beta}{1-\alpha-2\beta} + \\ + 2\beta\log\frac{1-\alpha-\beta}{\beta} + \frac{\beta^2}{\alpha}\log\frac{\beta}{\beta+\alpha} + \alpha\log\frac{\alpha+\beta}{\alpha} \end{aligned} \tag{3}$$

Now, we consider **the non-symmetric part** of Protocol 2 for the prior $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & 1-\alpha-\beta-\gamma \\ \hline \end{array}$, where $\beta < \gamma$. Recall that Bob sends bit B with distribution

$$B = \begin{cases} 1 & \text{if } y = 1 \\ 0 & \text{with probability } 1 - \beta/\gamma \text{ if } y = 0 \\ 1 & \text{with probability } \beta/\gamma \text{ if } y = 0 \end{cases}$$

The contribution of this bit to the *internal* information cost is

$$\begin{aligned} I(Y;B|X) &= H(B|X) - H(B|XY) \\ &= (\alpha+\beta)H\left(\tfrac{\beta}{\alpha+\beta} + \tfrac{\beta}{\gamma}\cdot\tfrac{\alpha}{\alpha+\beta}\right) + (\gamma+\delta)H\left(\tfrac{\delta}{\gamma+\delta} + \tfrac{\beta}{\gamma}\cdot\tfrac{\gamma}{\gamma+\delta}\right) - \\ &- (\alpha+\gamma)H\left(\tfrac{\beta}{\gamma}\right). \end{aligned} \tag{4}$$

Bob sends bit 1 with probability $t = 1 - \alpha - \gamma + \beta + \alpha\beta/\gamma$. In that case the protocol continues on distribution $\widetilde{\nu} = \begin{array}{|c|c|}\hline \frac{\beta\alpha}{\gamma t} & \frac{\beta}{t} \\ \hline \frac{\beta}{t} & \frac{1-\alpha-\beta-\gamma}{t} \\ \hline \end{array}$. If Bob sends 0 the protocol terminates. Thus the overall internal information cost of π for the case $\beta \le \gamma$ is

$$\mathrm{IC}_\mu(\pi) = I(Y;B|X) + t\,\mathrm{IC}_{\widetilde{\nu}}(\pi). \tag{5}$$

Closed-form formula for the above equation may be obtained from (3) and (4). Since the roles of Alice and Bob are symmetric, we have

$$\mathrm{IC}_\mu(\pi) = \mathrm{IC}_{\mu^T}(\pi).$$

This completes the analysis of $\mathrm{IC}_\mu(\pi)$ for all three cases $\beta < \gamma, \beta = \gamma, \beta > \gamma$.

7.6 Internal Information Cost: Lower Bound

In this section we shall show that Expression (5) is a lower bound on $\mathrm{IC}_\mu(\mathrm{AND}, 0)$. Let

$$\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & 1-\alpha-\beta-\gamma \\ \hline \end{array}$$

and suppose that Bob sends signal B with properties

- $P(B = 1) = P(B = 0) = 1/2$,
- $P(B = 1 \mid Y = 1) = 1/2 + \epsilon_1/2$,
- $P(B = 0 \mid Y = 0) = 1/2 + \epsilon_0/2$.

The resulting distributions are

- $\mu_0 = \begin{array}{|c|c|}\hline (1+\epsilon_0)\alpha & (1-\epsilon_1)\beta \\ \hline (1+\epsilon_0)\gamma & (1-\epsilon_1)(1-\alpha-\beta-\gamma) \\ \hline \end{array}$ if Bob sends 0, and
- $\mu_1 = \begin{array}{|c|c|}\hline (1-\epsilon_0)\alpha & (1+\epsilon_1)\beta \\ \hline (1-\epsilon_0)\gamma & (1+\epsilon_1)(1-\alpha-\beta-\gamma) \\ \hline \end{array}$ if Bob sends 1.

Also note that $\epsilon_1 = \epsilon_0 \frac{\alpha+\gamma}{1-\alpha-\gamma}$.

Corollary 5.8 says that to demonstrate that $\mathrm{IC}_\mu(\pi)$ is a lower bound on $\mathrm{IC}_\mu(\mathrm{AND}, 0)$ it suffices to prove the following concavity constraint:

$$\mathrm{IC}_\mu(\pi) \le \mathrm{IC}_{\mu_0}(\pi)/2 + \mathrm{IC}_{\mu_1}(\pi)/2 + I(B;Y|X),$$

where

$$\begin{aligned}
I(B;Y|X) &= H(B|X) - H(B|XY) \\
&= (\alpha+\beta)H(B \mid X=0) + (\gamma+\delta)H(B \mid X=1) - \sum_{i,j\in\{0,1\}} H(B \mid X=i, Y=j) \\
&= (\alpha+\beta)H(\frac{\alpha}{\alpha+\beta}(1/2-\epsilon_0/2) + \frac{\beta}{\alpha+\beta}(1/2+\epsilon_1/2)) + (\gamma+\delta)H(\frac{\gamma}{\gamma+\delta}(1/2-\epsilon_0/2) + \frac{\delta}{\gamma+\delta}(1/2+\epsilon_1/2)) - \\
&\quad - (\alpha+\gamma)H(1/2+\epsilon_0/2) - (\beta+\delta)H(1/2+\epsilon_1/2)
\end{aligned}$$

By Claims 7.14 and 7.16, to demonstrate that $\mathrm{IC}_\mu(\pi)$ is a lower bound on $I(\mathrm{AND}) := \mathrm{IC}_\mu(AND, 0)$ it suffices to consider only two types of *non-crossing* signals B that are sent by Bob:

1. The prior μ is in Alice's region, i. e., $\beta > \gamma$. Using Wolfram Mathematica we obtain

$$\mathrm{IC}_{\mu_0}(\pi)/2 + \mathrm{IC}_{\mu_1}(\pi)/2 + I(Y;B|X) - \mathrm{IC}_\mu(\pi) = \frac{\alpha(\beta-\gamma)}{(\alpha+\beta)(1-\alpha-\gamma)^2 \ln 4}\epsilon_0^2 + O(\epsilon_0^3),$$

   which is > 0 for small enough ϵ_0.

2. The prior μ is in the diagonal region, i. e., $\beta = \gamma$. Using Wolfram Mathematica we obtain

$$\mathrm{IC}_{\mu_0}(\pi)/2 + \mathrm{IC}_{\mu_1}(\pi)/2 + I(Y;B|X) - \mathrm{IC}_\mu(\pi) = \frac{\alpha\beta}{12(\alpha+\beta)(1-\alpha-\beta)^3 \ln 2}\epsilon_0^3 + O(\epsilon_0^4),$$

   which is > 0 for small enough ϵ_0.

Also, note that trivially $\mathrm{IC}_\mu(\pi) \le 2$, as the players learn at most each others bits during the execution of π. Hence Expression (5) satisfies all the constraints of Definition 5.4 and thus is a lower bound on $\mathrm{IC}_\mu(\mathrm{AND}, 0)$ by Corollary 5.8.

7.7 External Information Cost: Upper Bound

We start by analyzing **the symmetric part** of Protocol 2, i. e., we shall compute $\mathrm{IC}^{\mathrm{ext}}_\nu(\pi)$ where

$$\nu = \begin{array}{|c|c|} \hline \alpha & \beta \\ \hline \beta & 1-\alpha-2\beta \\ \hline \end{array}$$

is a symmetric distribution.

Working from first-principles, we obtain

$$\begin{aligned}
\mathrm{IC}^{\mathrm{ext}}_\nu(\pi) &= I(XY;\Pi) \\
&= \mathbb{E}_{x,y}(\mathbb{D}(\Pi_{xy}||\Pi)) \\
&= \alpha\mathbb{D}(\Pi_{X=0,Y=0}||\Pi) + \beta\mathbb{D}(\Pi_{X=0,Y=1}||\Pi) + \\
&\quad + \beta\mathbb{D}(\Pi_{X=1,Y=0}||\Pi) + (1-\alpha-2\beta)\mathbb{D}(\Pi_{X=1,Y=1}||\Pi).
\end{aligned}$$

A transcript of Π on x, y is determined by the value $c \in [0,1]$ of the clock when the protocol is terminated together with a name of a player $\mathcal{P} \in \{A, B\}$, whose random number is reached by a counter first. For $x, y \in \{0,1\}$ we have

$$\mathbb{D}(\Pi_{xy}||\Pi) = \sum_{\mathcal{P}\in\{A,B\}} \int_0^1 f_{x,y}(c,\mathcal{P}) \log \frac{f_{x,y}(c,\mathcal{P})}{f(c,\mathcal{P})} dc,$$

where $f_{x,y}(c,\mathcal{P})$ is the pdf for Π_{xy} and $f(c,\mathcal{P})$ is the PDF for Π.

We have

- $f_{0,0}(c,A) = f_{0,0}(c,B) = 1-c$ for $c \in [0,1]$
- $f_{0,1}(c,A) = 1$ for $c \in [0,1)$ and $f_{0,1}(c,B) = 0$ for $c \in [0,1]$
- $f_{1,1}(c,A) = f_{1,1}(c,B) = 0$ for $c \in [0,1)$ and $P(\Pi_{X=1,Y=1} = (1,A)) = 1$
- $f(c,A) = f(c,B) = \alpha(1-c)+\beta$ for $c \in [0,1)$ and $P(\Pi = (1,A)) = 1-\alpha-2\beta$

After plugging in the above PDFs in the expression for $\mathrm{IC}^{\mathrm{ext}}_{\nu}(\pi)$ and using Wolfram Mathematica to simplify the expressions, we obtain:

$$\begin{aligned}
&\mathrm{IC}^{\mathrm{ext}}_{\nu}(\pi) \\
&= 2\alpha \int_0^1 (1-c)\log\frac{(1-c)}{\alpha(1-c)+\beta}dc + 2\beta\int_0^1 \log\frac{1}{\alpha(1-c)+\beta}dc + \\
&\quad + (1-\alpha-2\beta)\log\frac{1}{1-\alpha-2\beta} \\
&= (1-\alpha-2\beta)\log\frac{1}{1-\alpha-2\beta} + \frac{\beta}{\ln 2} + \frac{\beta^2}{\alpha}\log\beta - \frac{(\alpha+\beta)^2}{\alpha}\log(\alpha+\beta).
\end{aligned}$$

Now, we consider **the non-symmetric part** of Protocol 2 for the prior $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & 1-\alpha-\beta-\gamma \\ \hline \end{array}$, where
$\beta < \gamma$. Bob sends bit B with distribution

$$B = \begin{cases} 1 & \text{if } y = 1 \\ 0 & \text{with probability } 1-\beta/\gamma \text{ if } y = 0 \\ 1 & \text{with probability } \beta/\gamma \text{ if } y = 0 \end{cases}$$

The contribution of this bit to the external information cost is

$$\begin{aligned}
&I(XY;B) \\
&= H(B) - H(B \mid XY) \\
&= H(B) - H(B \mid Y) \\
&= H((1-\alpha-\gamma) + (\beta/\gamma)(\alpha+\gamma)) - (\alpha+\gamma)H(\beta/\gamma).
\end{aligned}$$

Bob sends bit 1 with probability $t = 1-\alpha-\gamma+\beta+\alpha\beta/\gamma$. In that case the protocol continues on distribution

$$\widetilde{\nu} = \begin{array}{|c|c|}\hline \dfrac{\beta\alpha}{\gamma t} & \dfrac{\beta}{t} \\ \hline \dfrac{\beta}{t} & \dfrac{1-\alpha-\beta-\gamma}{t} \\ \hline \end{array}.$$

If Bob sends 0 the protocol terminates. Thus the overall external information cost of π for the case $\beta \le \gamma$ is as follows (once again, Wolfram Mathematica was used to simplify the expressions):

$$\begin{aligned}
&\mathrm{IC}^{\mathrm{ext}}_{\mu}(\pi) \\
&\quad = I(XY;B) + t\,\mathrm{IC}^{\mathrm{ext}}_{\widetilde{\nu}}(\pi) \\
&\quad = \tfrac{\beta}{\ln 2} + \beta\log\tfrac{1}{\beta} + (1-\alpha-\beta-\gamma)\log\tfrac{1}{1-\alpha-\beta-\gamma} + \\
&\qquad + \tfrac{\beta(\alpha+\gamma)}{\alpha}\log\gamma + \tfrac{(\alpha+\beta)(\alpha+\gamma)}{\alpha}\log\tfrac{1}{\alpha+\gamma}.
\end{aligned} \tag{6}$$

Since the roles of Alice and Bob are symmetric, we have

$$\mathrm{IC}^{\mathrm{ext}}_{\mu}(\pi) = \mathrm{IC}^{\mathrm{ext}}_{\mu^T}(\pi).$$

This completes the analysis of $\mathrm{IC}^{\mathrm{ext}}_{\mu}(\pi)$ for all three cases $\beta < \gamma, \beta = \gamma, \beta > \gamma$.

Remark 7.17. Observe that if $\alpha = 0$, i. e.,

$$\mu = \begin{array}{|c|c|}\hline 0 & \beta \\ \hline \gamma & 1-\beta-\gamma \\ \hline \end{array},$$

the expression of $\mathrm{IC}^{\mathrm{ext}}_{\mu}(\pi)$ simplifies to

$$\mathrm{IC}^{\mathrm{ext}}_{\mu}(\pi) = \beta \log \frac{1}{\beta} + \gamma \log \frac{1}{\gamma} + (1-\beta-\gamma) \log \frac{1}{1-\beta-\gamma} = H(\mu).$$

7.8 External Information Cost: Lower Bound

In this section we shall show that Expression (6) is a lower bound on $\mathrm{IC}^{\mathrm{ext}}_{\mu}(\mathrm{AND}, 0)$. Let

$$\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & 1-\alpha-\beta-\gamma \\ \hline \end{array}$$

and suppose that Bob sends signal B with properties

- $P(B=1) = P(B=0) = 1/2$,
- $P(B=1 \mid Y=1) = 1/2 + \epsilon_1/2$,
- $P(B=0 \mid Y=0) = 1/2 + \epsilon_0/2$.

The resulting distributions are

- $\mu_0 = \begin{array}{|c|c|}\hline (1+\epsilon_0)\alpha & (1-\epsilon_1)\beta \\ \hline (1+\epsilon_0)\gamma & (1-\epsilon_1)(1-\alpha-\beta-\gamma) \\ \hline \end{array}$ if Bob sends 0, and
- $\mu_1 = \begin{array}{|c|c|}\hline (1-\epsilon_0)\alpha & (1+\epsilon_1)\beta \\ \hline (1-\epsilon_0)\gamma & (1+\epsilon_1)(1-\alpha-\beta-\gamma) \\ \hline \end{array}$ if Bob sends 1.

Also note that $\epsilon_1 = \epsilon_0 \frac{\alpha+\gamma}{1-\alpha-\gamma}$.

Remark 5.9 says that to demonstrate that $\mathrm{IC}^{\mathrm{ext}}_{\mu}(\pi)$ is a lower bound on $\mathrm{IC}^{\mathrm{ext}}_{\mu}(\mathrm{AND}, 0)$ it suffices to prove the following concavity constraint:

$$\mathrm{IC}^{\mathrm{ext}}_{\mu}(\pi) \le \mathrm{IC}^{\mathrm{ext}}_{\mu_0}(\pi)/2 + \mathrm{IC}^{\mathrm{ext}}_{\mu_1}(\pi)/2 + I(XY; B),$$

where

$$\begin{aligned} & I(XY;B) \\ & = H(B) - H(B \mid XY) \\ & = H(B) - H(B \mid Y) \\ & = 1 - ((\alpha+\gamma) H(1/2 + \epsilon_0/2) + (1-\alpha-\gamma) H(1/2+\epsilon_1/2)) \,. \end{aligned}$$

By Claims 7.14 and 7.16, to demonstrate that $\mathrm{IC}^{\mathrm{ext}}_{\mu}(\pi)$ is a lower bound on $I^{\mathrm{ext}}(\mathrm{AND}) := \mathrm{IC}^{\mathrm{ext}}_{\mu}(AND, 0)$ it suffices to consider only two types of *non-crossing* signals B that are sent by Bob:

1. The prior μ is in Alice's region, i. e., $\beta > \gamma$. Using Wolfram Mathematica we obtain

$$\mathrm{IC}^{\mathrm{ext}}_{\mu_0}(\pi)/2 + \mathrm{IC}^{\mathrm{ext}}_{\mu_1}(\pi)/2 + I(XY;B) - \mathrm{IC}^{\mathrm{ext}}_{\mu}(\pi) = \frac{\alpha(\beta-\gamma)}{(\alpha+\beta)(1-\alpha-\gamma)^2 \ln 4} \epsilon_0^2 + O(\epsilon_0^3),$$

which is > 0 for small enough ϵ_0.

2. The prior μ is in the diagonal region, i. e., $\beta = \gamma$. Using Wolfram Mathematica we obtain

$$\mathrm{IC}^{\mathrm{ext}}_{\mu_0}(\pi)/2 + \mathrm{IC}^{\mathrm{ext}}_{\mu_1}(\pi)/2 + I(XY;B) - \mathrm{IC}^{\mathrm{ext}}_{\mu}(\pi) = \frac{\alpha\beta}{12(\alpha+\beta)(1-\alpha-\beta)^3 \ln 2}\epsilon_0^3 + O(\epsilon_0^4),$$

which is > 0 for small enough ϵ_0.

Also, note that trivially $\mathrm{IC}^{\mathrm{ext}}_{\mu}(\pi) \leq 2$, as the players learn at most each others bits during the execution of π. Hence Expression (6) satisfies all the constraints of Definition 5.4 and thus is a lower bound on $\mathrm{IC}_\mu(\mathrm{AND}, 0)$ by Remark 5.9.

7.9 Rate of Convergence

In this section we prove that for most distributions μ the rate at which $\mathrm{IC}^r_\mu(\mathrm{AND}, 0)$ converges to $\mathrm{IC}_\mu(\mathrm{AND}, 0)$ is $\Theta(1/r^2)$. The empirical evidence that the rate of convergence is $\Theta(1/r^2)$ has appeared in the information theory literature prior to our work. In [30], Ma and Ishwar consider the task f of computing AND when only Bob is required to learn the answer. They derive an explicit formula for $\mathrm{IC}_\mu(f)$ for product distributions μ and design an algorithm that computes $\mathrm{IC}^r_\mu(f)$ to within a desired accuracy. Ishwar and Ma generously provided their scripts, which we used to generate Figure 1 (it is a variant of Figure 4(a) from [30]). Figure 1 demonstrates that $\max_{\mu \text{ - product}} \mathrm{IC}^r_\mu(f) - \mathrm{IC}_\mu(f)$ asymptotically behaves like $\Theta(1/r^2)$.

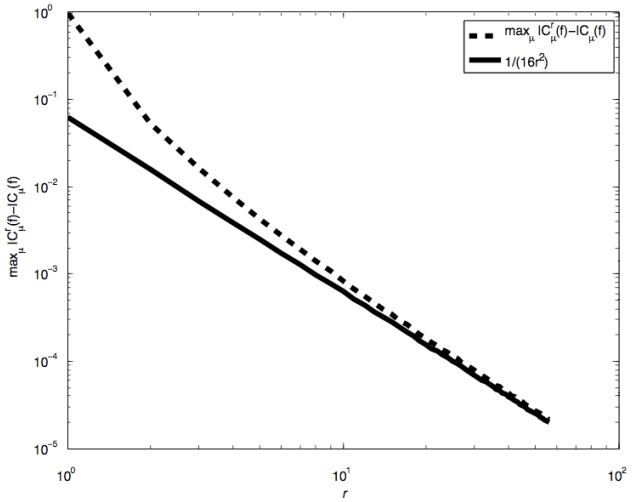

Figure 1: Empirical evidence that rate of convergence is $\Theta(1/r^2)$. The log-log scale figure shows the graph of $\max_{\mu \text{ - product}} \mathrm{IC}^r_\mu(f) - \mathrm{IC}_\mu(f)$ for a range of values r together with the line $1/(16r^2)$. The x-axis is the number of rounds r. The y-axis is the change in the information cost $\max_{\mu \text{ - product}} \mathrm{IC}^r_\mu(f) - \mathrm{IC}_\mu(f)$.

Our proof consists of two main parts: (1) the lower bound $\Omega(1/r^2)$ on the rate of convergence and (2) a matching upper bound $O(1/r^2)$. The high-level idea for the lower bound is to show that any r-round protocol, when viewed as a random walk on $\Delta(\mathcal{X} \times \mathcal{Y})$, has to travel a lot in the wrong region. In other words, Alice often speaks in Bob's region, and Bob often speaks in Alice's region. Then we can use formulas from Section 7.6 to conclude that each such step wastes a lot of information as compared to the optimal protocol. Aggregating this wastage over all rounds, $\Omega(1/r^2)$ information has to be wasted overall. The upper bound is obtained by carefully analyzing a discretized version of our infeasible protocol for AND from Section 7.3. Both upper and lower bounds require a number of technical lemmas, which we also include in the text.

The rest of this section is organized as follows. In Subsection 7.9.1 we prove the lower bound on the rate of convergence modulo two technical lemmas. Subsection 7.9.2 contains the proof of the first lemma, which quantifies how much information is wasted by a feasible protocol versus an optimal infeasible one in terms of the distance traveled in the wrong region. Subsection 7.9.3 proves the second technical lemma from the lower bound on the rate of convergence. The second lemma gives a lower bound on the distance traveled in the wrong region by a protocol that solves the AND function. Finally, in Subsection 7.9.4 we prove the upper bound on the rate of convergence.

In this section it will be easier for us to work with general protocols and forgo the normal-form assumption.

7.9.1 Lower Bound on the Rate of Convergence

We say that a message M *crosses the diagonal* if this message starts at prior μ, has distribution on distributions $(\{\mu_m\}, \{p_m\})$, and there exists m such that the interval $[\mu, \mu_m]$ intersects the diagonal region, i. e., the interval $[\mu, \mu_m]$ contains a symmetric distribution.

We begin by showing that we can split a message that crosses the diagonal into two that do not cross the diagonal.

Lemma 7.18. *Let M be a message sent by one of the players that crosses the diagonal. There exists two messages M_1 and M_2 such that neither M_1, nor M_2 crosses the diagonal, and (M_1, M_2) has the same distribution on distributions as M.*

Proof. The idea of the proof is that each message M is simply a sequence of bits, so the player can generate M bit by bit until there is a danger of the next bit crossing the diagonal. If the player is about to generate a crossing bit, the player will instead split that bit into two using the Splitting Lemma (Lemma 7.11). The split happens in such a way that after the first bit is sent the player either ends up on the diagonal, or moves away from the diagonal. If the player does not jump to the diagonal, then the process continues in the same way. If the player happens to jump to the diagonal that signifies the end of message M_1 and beginning of M_2.

All that is left to show is that a crossing bit may be split into two non-crossing bits while preserving the distribution on distributions. Suppose that the player sends a bit B starting at prior μ and splitting μ into μ_0 and μ_1, such that $[\mu, \mu_1]$ contains a symmetric distribution μ_D. Since $\mu \in [\mu_0, \mu_D]$ there is a signal B_1 that splits μ into μ_0 and μ_D (by the Splitting Lemma). Also, since $\mu_D \in [\mu_0, \mu_1]$ there is a signal B_2 that splits μ_D into μ_0 and μ_1. Now instead of sending bit B, the player first sends B_1. If $B_1 = 0$ the message is terminated, otherwise the player sends B_2. This new message induces the same distribution on distributions as B, because (B_1, B_2) and B express μ as a convex combination of μ_0, μ_1, which is unique. Note that we allow B, B_1 and B_2 be biased. □

Theorem 7.19. *For all* $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & \delta \\ \hline \end{array}$ *with* $\{\alpha, \beta, \gamma\} \subseteq \operatorname{supp}(\mu)$ *we have*

$$\mathrm{IC}^r_\mu(AND, 0) = \mathrm{IC}_\mu(AND, 0) + \Omega_\mu\left(\frac{1}{r^2}\right).$$

Proof. Fix an arbitrary r-round protocol π that solves AND with 0-error and distribution $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & \delta \\ \hline \end{array}$ with $\alpha, \beta, \gamma \neq 0$. Using Lemma 7.18 we obtain a protocol π' with $m \leq 2r$ messages, such that no message crosses the diagonal and $\mathrm{IC}_\mu(\pi') = \mathrm{IC}_\mu(\pi)$. We shall view π' as a random walk on the set of distributions $\Delta(\{0,1\}^2)$. For technical reasons we shall restrict this random walk to the subset S of $\Delta(\{0,1\}^2)$ defined as follows

$$S := \{\mu' \mid \alpha' \geq 0.01\alpha \text{ and } \min(\beta', \gamma') \geq 0.01 \min(\beta, \gamma)\}.$$

Using ideas from the proof of Lemma 7.18 we can always impose a constraint that π' does not make steps that cross from S into $\overline{S}$ without stopping at the boundary of S. We let π'' denote such a modification of

π'. Clearly, $\mathrm{IC}_\mu(\pi'') = \mathrm{IC}_\mu(\pi')$ and the number of messages in the first part of π'' that proceeds only until the boundary of S is at most m.

We shall show that $\mathrm{IC}_\mu(\pi'') = \mathrm{IC}_\mu(\mathrm{AND}, 0) + \Omega_\mu\left(\frac{1}{r^2}\right)$ by showing that the part of π'' until the boundary of S already wastes $\Omega_\mu\left(\frac{1}{r^2}\right)$ amount of information as compared to the optimal protocol.

Let T_i denote the ith message of π'' for $i \leq m$. The whole transcript T until the boundary of S is then simply $T_1 T_2 \cdots T_m$. We shall write $T_{\leq i} = T_1 T_2 \cdots T_i$. Similarly, we shall write $T_{>i}$ meaning $T_{i+1} T_{i+2} \cdots T_m$.

A transcript t gives rise to $m+1$ distributions $\mu_0^{t\leq 0}, \mu_1^{t\leq 1}, \ldots, \mu_m^{t\leq m}$ traced out by the protocol π when viewed as a random walk on $\Delta(\mathcal{X} \times \mathcal{Y})$. Observe that $\mu_0^{t\leq 0} = \mu$. We define the central object of this proof: $\delta_i^{t\leq i}$ - the distance traveled by a player in the wrong region during the ith round. More formally

$$\delta_i^{t\leq i} = \begin{cases} \|[\mu_{i-1}^{t\leq i-1}, \mu_i^{t\leq i}] \cap \Delta_A\| & \text{if the } i\text{th message is} \\ & \text{transmitted by Bob,} \\ \|[\mu_{i-1}^{t\leq i-1}, \mu_i^{t\leq i}] \cap \Delta_B\| & \text{if the } i\text{th message is} \\ & \text{transmitted by Alice.} \end{cases}$$

The lower bound $\Omega_\mu(1/r^2)$ on the overall wastage of protocol π'' follows from two crucial observations:

Lemma 7.20.

$$\mathrm{IC}_\mu(\pi'') - \mathrm{IC}_\mu(AND, 0) = \Omega_\mu\left(\sum_{i=1}^{m} (\mathbb{E}_t \delta_i^{t\leq i})^3\right).$$

Lemma 7.21.

$$\mathbb{E}_t\left(\sum_{i=1}^{m} \delta_i^{t\leq i}\right) = \Omega_\mu(1).$$

We prove the above lemmas later in Subsections 7.9.2 and 7.9.3. Now, by Hölder's inequality we have

$$\sum_{i=1}^{m} \left(\mathbb{E}_t \delta_i^{t\leq i}\right)^3 \geq \left(\mathbb{E}_t\left(\sum_{i=1}^{m} \delta_i^{t\leq i}\right)\right)^3 / m^2 = \Omega_\mu(1/r^2),$$

where the last step follows from Lemma 7.21 and the fact that $m \leq 2r$. This finishes the proof by Lemma 7.20. □

7.9.2 Informational Wastage

The goal of the current subsection is to prove Lemma 7.20 that appears in the proof of the lower bound on the rate of convergence. For definitions of relevant mathematical objects see Subsection 7.9.1. Recall that Lemma 7.20 asserts that the information wasted by an m-message protocol as compared to the optimal infeasible protocol is roughly the sum of the cubed distances traveled in the wrong region. The proof of this lemma consists of a sequence of reductions. We start with analyzing how much information is wasted by a single bit and gradually build up the result to the entire protocol.

We start by formally defining what it means for a particular step in a protocol, which consists of one of the players sending a message, to waste information.

Definition 7.22. Suppose that Bob sends message M with distribution on distributions $(\{\mu_m\}, \{p_m\})$ from prior μ. Then the informational wastage of M is defined as

$$\mathrm{IW}(\mu, M) := \sum_{m \in \mathrm{range}(M)} p_m \,\mathrm{IC}_{\mu_m}(\mathrm{AND}, 0) + I(M; Y|X) - \mathrm{IC}_\mu(\mathrm{AND}, 0).$$

For Alice's messages it can be defined similarly.

The information wasted is how much extra information is revealed by a protocol that sends message M and then plays optimally versus the protocol that plays optimally from the start.

When the message is a single bit B sent by Bob from a symmetric prior μ, the above definition simplifies to

$$\mathrm{IW}(\mu, B) = p\,\mathrm{IC}_{\mu_0}(\mathrm{AND}, 0) + (1-p)\,\mathrm{IC}_{\mu_1}(\mathrm{AND}, 0) + I(B; Y|X) - \mathrm{IC}_{\mu}(\mathrm{AND}, 0),$$

where $\mu_0 := P(X = x, Y = y|B = 0)$ belongs to Bob's region and $\mu_1 := P(X = x, Y = y|B = 1)$ belongs to Alice's region.

Observe that formulas from Section 7.6 simply say that for a *uniform bit* B and symmetric prior μ we have

$$\mathrm{IC}(\mu, B) \geq C(\mu)||\mu_1 - \mu||^3 = \Omega(||\mu_1 - \mu||^3), \tag{7}$$

where $C(\mu) = \frac{\alpha\beta}{12(\alpha+\beta)(1-\alpha-\beta)^3}$ is a continuous positive function of μ. In other words, the information wasted is roughly the cube of the distance traveled in the wrong region.

Remark 7.23. In what follows we only consider the information wasted from a symmetric prior, because the information wasted when a player speaks starting in the wrong region is strictly larger (see formulas at the end of Section 7.6).

Now we extend this result to *nonuniform* bits. As expected, for a nonuniform bit the cube of the distance traveled in the wrong region gets scaled by the probability of jumping into the wrong region.

Lemma 7.24. *Suppose that Bob sends bit B from symmetric prior μ with distribution on distributions $(\{\mu_0, \mu_1\}, \{p, 1-p\})$. If $\mu_1 + 2p(\mu_0 - \mu_1) \in \Delta(\{0,1\} \times \{0,1\})$ then*

$$\mathrm{IW}(\mu, B) \geq C(\mu)(1-p)||\mu_1 - \mu||^3 = \Omega((1-p)||\mu_1 - \mu||^3),$$

where $C(\mu) = \frac{\alpha\beta}{12(\alpha+\beta)(1-\alpha-\beta)^3}$. Similarly for Alice.

Proof. **Case** $p \leq 1/2$**.** Let $\mu_0' := \mu_1 + 2p(\mu_0 - \mu_1) \in [\mu_0, \mu_1]$. Then we have $\mu = (1/2)\mu_1 + (1/2)\mu_0'$, so there exists signal B' that Bob can send with distribution on distributions $(\{\mu_0', \mu_1\}, \{1/2, 1/2\})$. Clearly we have $\mathrm{IW}(\mu, B) \geq \mathrm{IW}(\mu, B')$. Finally, from Equation (7) we obtain $\mathrm{IW}(\mu, B') \geq C(\mu)||\mu_1 - \mu||^3 \geq C(\mu)(1-p)||\mu_1 - \mu||^3$.

Case $p > 1/2$**.** Let $\mu_0' := \mu_1 + 2p(\mu_0 - \mu_1)$. By conditions of the lemma, μ_0' is a valid distribution. Then we have $\mu_0 := ((1-p)/p)\mu_0' + ((2p-1)/p)\mu$, so there exists bit B' that Bob can send from prior μ_0 with distribution on distributions $(\{\mu_0', \mu\}, \{(1-p)/p, (2p-1)/p\})$. By Claim 7.16 we have $\mathrm{IW}(\mu_0, B') = 0$ thus

$$\mathrm{IW}(\mu, B) = \mathrm{IW}(\mu, B) + p\,\mathrm{IW}(\mu_0, B') = \mathrm{IW}(\mu, M),$$

where M is message (B, B') that has distribution on distributions $(\{\mu_0', \mu, \mu_1\}, \{1-p, 2p-1, 1-p\})$. Since $\mu = (1/2)\mu_0' + (1/2)\mu_1$ there exists signal B'' with distribution on distributions $(\{\mu_0', \mu_1\}, \{1/2, 1/2\})$.
Define $I(\nu) := \mathrm{IC}_\nu(\mathrm{AND}, 0)$. Then we have

$$\begin{aligned}
&\mathrm{IW}(\mu, M)\\
&= (1-p)I(\mu_0') + (1-p)I(\mu_1) + (2p-1)I(\mu) + I(M; Y|X) - I(\mu)\\
&= 2(1-p)[(1/2)I(\mu_0') + (1/2)I(\mu_1) + I(M; Y|X)/(2(1-p)) - I(\mu)]\\
&= 2(1-p)[(1/2)I(\mu_0') + (1/2)I(\mu_1) + I(B''; Y|X) - I(\mu)]\\
&= 2(1-p)\,\mathrm{IW}(\mu, B'')\\
&\geq 2(1-p)C(\mu)||\mu_1 - \mu||^3,
\end{aligned}$$

$I(M; Y|X)/(2(1-p)) = I(B''; Y|X)$, since sending M and *staying at prior μ with probability* $2(1-p)$ *and otherwise sending* B'' induce the same distribution on distributions.The last step follows from Equation (7). □

The next step is to extend our lower bound on the information wasted in a single step of a protocol to messages. Suppose that Bob sends a message M with distribution on distributions $(\{\mu_m\}, \{p_m\})$ from symmetric prior μ. Define sets $\mathcal{M}_1 := \{m \mid \mu_m(0,1) > \mu_m(1,0)\}$ and $\mathcal{M}_2 := \{m \mid \mu_m(0,1) \leq \mu_m(1,0)\}$. The set $\mathcal{M}_1$ contains all the messages that lead to Alice's region and $\mathcal{M}_2$ contains all the messages that lead to Bob's region. Let μ_1 be the average of $\mu_m \in \mathcal{M}_1$ and μ_0 to be the average of μ_m in $\mathcal{M}_2$. If $\mu_1 + 2p(\mu_0 - \mu_1) \in \Delta(\{0,1\}^2)$ then the information wasted by sending M is at least $\Omega(P(m \in \mathcal{M}_1)||\mu_1 - \mu||^3) = \Omega((\mathbb{E}_m||[\mu_m, \mu] \cap \Delta_A||)^3)$.

Lemma 7.25. *Suppose that the conditioned specified in the above paragraph hold for a message M sent by Bob, then we have*

$$\mathrm{IW}(\mu, M) \geq C(\mu)(\mathbb{E}_m||[\mu_m, \mu] \cap \Delta_A||)^3,$$

where $C(\mu) = \frac{\alpha\beta}{12(\alpha+\beta)(1-\alpha-\beta)^3}$. Similar inequality holds for Alice.

Proof. Define an indicator random variable Z as follows

$$Z = \begin{cases} 1 & \text{if } \mu_M(0,1) > \mu_M(1,0) \\ 0 & \text{otherwise} \end{cases}$$

In other words, Z indicates if after sending M the players end up in Alice's region.

Consider the two protocols:

1. π_1 - Bob first sends M and then players play optimally.
2. π_2 - Bob first sends Z, then $M|Z$ and then players play optimally.

Clearly, sending Z followed by $M|Z$ produces the same distribution on distributions as simply sending M, thus π_1 and π_2 have the same information cost. Therefore they have the same informational wastage. Observe that if Bob sends $Z = 1$ then the players update their distribution μ to distribution $\mu_1 = \mathbb{E}_{m \sim M|Z=1}(\mu_m)$. It is easy to see that $||\mu_1 - \mu|| = \mathbb{E}_{m \sim M|Z=1}(||\mu_m - \mu||)$ (note that this matches the definition of μ_1 we gave in a paragraph prior to the statement of this lemma). Now we are in a position to apply Lemma 7.24 to the bit Z. All in all, we have $\mathrm{IW}(\mu, (Z, M|Z)) \geq \mathrm{IW}(\mu, Z) \geq P(Z = 1)C(\mu)||\mu_1 - \mu||^3 \geq C(\mu)(P(Z = 1)\mathbb{E}_{m \sim M|Z=1}(||\mu_m - \mu||))^3 \geq C(\mu)(\mathbb{E}_m||[\mu_m, \mu] \cap \Delta_A||)^3$. □

Now we are in a position to prove Lemma 7.20.

Lemma (7.20 restated)**.**

$$\mathrm{IC}_\mu(\pi'') - \mathrm{IC}_\mu(\mathrm{AND}, 0) = \Omega_\mu\left(\sum_{i=1}^{m} (\mathbb{E}_t \delta_i^{t_{\leq i}})^3\right).$$

Proof of Lemma 7.20. By Lemma 7.25 the informational waste of the ith message T_i given a *fixed partial transcript* $t_{\leq i-1}$ is at least

$$C(\mu_{i-1}^{t_{\leq i-1}})(\mathbb{E}_{t_i \sim T_i | t_{\leq i-1}}(\delta_i^{t_{\leq i}}))^3 \geq \frac{(0.01\alpha)(0.01\min(\beta, \gamma))}{12}(\mathbb{E}_{t_i \sim T_i | t_{\leq i-1}}(\delta_i^{t_{\leq i}}))^3,$$

where the last step follows from the fact that $\mu_{i-1}^{t_{\leq i-1}} \in S$ and $C(\mu') \geq \frac{\alpha'\beta'}{12}$ (see proof of Theorem 7.19 for the relevant definitions). Aggregating this over all messages T_i we finish the proof of the lemma

$$\mathrm{IC}_\mu(\pi'') - \mathrm{IC}_\mu(\mathrm{AND}, 0) \geq \frac{(0.01\alpha)(0.01\min(\beta, \gamma))}{12}\left(\sum_{i=1}^{m} (\mathbb{E}_t \delta_i^{t_{\leq i}})^3\right).$$

□

7.9.3 Distance Traveled in the Wrong Region

The goal of the current subsection is to prove Lemma 7.21 appearing in the proof of the lower bound on the rate of convergence. See Subsection 7.9.1 for the relevant definitions. Lemma 7.21 asserts that when a protocol is viewed as a random walk on the space of distributions, the protocol has to spend non-trivial amount of time in the wrong region if it solves the AND function.

The proof relies on the following observation. Consider a protocol that solves AND correctly on all inputs. We view it as a random walk on the space of distributions. Recall that a single move multiplies rows or columns of the current distribution. Thus if the random walk starts from a non-trivial distribution (i. e., we cannot derive the answer to AND from it immediately), the protocol would have to multiply some row or column by 0. This immediately implies that a protocol solving AND correctly has to travel statistical distance at least $\min(\beta, \gamma)$ overall. A more careful analysis reveals that in fact such a protocol has to travel $\min(\beta, \gamma)$ in the "wrong region". This is proved in this section via an invariant argument (see Lemma 7.27). We start by proving the following lemma, which shows that a certain process defined by a random walk of the protocol is a supermartingale.

Lemma 7.26. *Let π be a protocol that starts at prior μ. For a (partial) transcript t, let $\mu_t = \begin{array}{|c|c|}\hline \alpha_t & \beta_t \\ \hline \gamma_t & \delta_t \\ \hline \end{array}$ denote the resulting distribution arising from t. Then $\beta_T \gamma_T$ is a supermartingale.*

Proof. Let B be a bit sent by Bob from μ. Then $\mu_i(x, y) = P(X = x, Y = y | B = i)$ for $i \in \{0, 1\}$. We need to show that

$$\mathbb{E}_{b \sim B}(\beta_b \gamma_b) \leq \beta\gamma.$$

Let $p := P(B = 0)$. Recall that the jth column of μ_i is simply a multiple of the jth column of μ. We can write $\mu_0 = \begin{array}{|c|c|}\hline C_0\alpha & C_1\beta \\ \hline C_0\gamma & C_1\delta \\ \hline \end{array}$ and $\mu_1 = \begin{array}{|c|c|}\hline D_0\alpha & D_1\beta \\ \hline D_0\gamma & D_1\delta \\ \hline \end{array}$, where $C_i = P(B = 0 | Y = i)/P(B = 0)$ and $D_i = P(B = 1 | Y = i)/P(B = 1)$. Observe that $D_i = (1 - C_i p)/(1 - p)$. Therefore we have

$$\begin{aligned}
&\mathbb{E}_{b \sim B}(\beta_b \gamma_b) \\
&= pC_0C_1\beta\gamma + (1-p)D_0D_1\beta\gamma \\
&= \beta\gamma(pC_0C_1 + (1 - C_0p)(1 - C_1p)/(1-p)) \\
&= \beta\gamma((1-p)pC_0C_1 + (1 - C_0p)(1 - C_1p))/(1-p) \\
&= \beta\gamma(1 - p + (C_1 - 1)(C_0 - 1)p)/(1-p) \\
&= \beta\gamma(1 + (C_1 - 1)(C_0 - 1)p/(1-p)) \\
&\leq \beta\gamma,
\end{aligned}$$

where the last step follows from the fact that $C_i \leq 1 \iff C_{1-i} \geq 1$, so $(C_1 - 1)(C_0 - 1) \leq 0$. □

The next lemma proves an invariant of a protocol solving the AND function. The lemma says that in order for a protocol to decrease the value of $\min(\beta, \gamma)$ by a certain amount, the protocol has to spend an equivalent amount of time in the wrong region.

Lemma 7.27.

$$\mathbb{E}_t \left(\min(\beta_m^t, \gamma_m^t) - \min(\beta, \gamma) + \sum_{i=1}^{m} \delta_i^{t \leq i} \right) \geq 0.$$

Proof. We prove the claim for all m-message protocols π and for all distributions μ by induction on m.

Base case is obvious, because it happens when $m = 0$ and we have $\min(\beta_0^t, \gamma_0^t) = \min(\beta, \gamma)$.

Now, consider the inductive step. We have

$$\begin{aligned}
&\mathbb{E}_t(\min(\beta_m^t, \gamma_m^t) - \min(\beta,\gamma) + \sum_{i=1}^{m} \delta_i^{t_{\le i}}) \\
&= \mathbb{E}_t(\min(\beta_m^t, \gamma_m^t) - \min(\beta_1^{t_1}, \gamma_1^{t_1}) + \sum_{i=2}^{m} \delta_i^{t_{\le i}} + \min(\beta_1^{t_1}, \gamma_1^{t_1}) - \min(\beta,\gamma) + \delta_1^{t_1}) \\
&\geq \mathbb{E}_{t_1}(\min(\beta_1^{t_1}, \gamma_1^{t_1}) - \min(\beta,\gamma) + \delta_1^{t_1}),
\end{aligned}$$

where the last step follows by induction. To complete the inductive step it is left to show that $\mathbb{E}_{t_1}(\min(\beta_1^{t_1}, \gamma_1^{t_1}) - \min(\beta,\gamma) + \delta_1^{t_1}) \geq 0$. We shall assume that the first message is sent by Bob. The case when Alice sends the first message is similar.

There are two possibilities, which we analyze separately.

First possibility is that μ is not a symmetric prior. So μ either belongs to Bob's region, or Alice's region. Consider the case when μ belongs to Bob's region ($\gamma > \beta$). Then $\min(\beta,\gamma) = \gamma$. Moreover, since the messages in our protocol do not cross the diagonal, we have that $\gamma_1^{t_1} \geq \beta_1^{t_1}$ for all $t_1 \in T_1$. Consequently $\min(\beta_1^{t_1}, \gamma_1^{t_1}) = \gamma_1^{t_1}$. Since $\gamma_i^{T_i}$ is a martingale, we have

$$\mathbb{E}_{t_1}(\min(\beta_1^{t_1}, \gamma_1^{t_1}) - \min(\beta,\gamma)) = 0.$$

Adding $\mathbb{E}_{t_1}(\delta_1^{t_1})$ to the above only increases the right-hand side. Similar calculation works for the case when μ belongs to Alice's region.

Second possibility is that μ is a symmetric prior, i. e., $\gamma = \beta$. Recall that the prior gets modified by multiplying the columns:

$$\begin{aligned}
\mu_1^{t_1}(x,y) &= P(X=x, Y=y | T_1 = t_1) \\
&= \tfrac{P(T_1=t_1|X=x,Y=y)}{P(T_1=t_1)} P(X=x, Y=y) \\
&= \tfrac{P(T_1=t_1|Y=y)}{P(T_1=t_1)} \mu(x,y).
\end{aligned}$$

Thus on the first message t_1 the first column of μ gets multiplied by $C_0^{t_1} := P(T_1 = t_1 | Y = 0)/P(T_1 = t_1)$ and the second column gets multiplied by $C_1^{t_1} := P(T_1 = t_1 | Y = 1)/P(T_1 = t_1)$. Next we define two sets of messages $\mathcal{S} := \{t_1 | C_0^{t_1} < C_1^{t_1}\}$ and $\mathcal{R} := \{t_1 | C_0^{t_1} \geq C_1^{t_1}\}$. Observe that $C_0^{t_1} P(Y=0) + C_1^{t_1} P(Y=1) = 1$. Hence if $C_0^{t_1} < C_1^{t_1}$ then $C_0^{t_1} < 1$ and $C_1^{t_1} > 1$; similarly, if $C_0^{t_1} > C_1^{t_1}$ then $C_0^{t_1} > 1$ and $C_1^{t_1} < 1$,

Observe that

- $(\forall t_1 \in \mathcal{R})(\delta_1^{t_1} = 0)$,
- $(\forall t_1 \in \mathcal{S})(\delta_1^{t_1} = (1 - C_0^{t_1})(\alpha + \beta) + (C_1^{t_1} - 1)(\beta + \delta))$,
- $(\forall t_1 \in \mathcal{R})(\min(\beta_1^{t_1}, \gamma_1^{t_1}) = C_1^{t_1}\beta)$, and
- $(\forall t_1 \in \mathcal{S})(\min(\beta_1^{t_1}, \gamma_1^{t_1}) = C_0^{t_1}\beta)$.

Introduce notation $p_{t_1} := P(T_1 = t_1)$. Then we have

$$\begin{aligned}
&\mathbb{E}_{t_1}\left(\min(\beta_1^{t_1}, \gamma_1^{t_1}) + \delta_1^{t_1}\right) \\
&= \textstyle\sum_{t_1 \in \mathcal{S}} p_{t_1} C_0^{t_1}\beta + \sum_{t_1 \in \mathcal{R}} p_{t_1} C_1^{t_1}\beta + \sum_{t_1 \in \mathcal{S}} p_{t_1}((1 - C_0^{t_1})(\alpha+\beta) + (C_1^{t_1} - 1)(\beta+\delta)) \\
&\geq \textstyle\sum_{t_1 \in \mathcal{S}} p_{t_1} C_0^{t_1}\beta + \sum_{t_1 \in \mathcal{R}} p_{t_1} C_1^{t_1}\beta + \sum_{t_1 \in \mathcal{S}} p_{t_1}(1 - C_0^{t_1})\beta + \sum_{t_1 \in \mathcal{S}} p_{t_1}(C_1^{t_1} - 1)\beta \\
&= \textstyle\sum_{t_1 \in \mathcal{S}} p_{t_1} C_1^{t_1}\beta + \sum_{t_1 \in \mathcal{R}} p_{t_1} C_1^{t_1}\beta \\
&= \beta.
\end{aligned}$$

□

Finally we are in a position to prove Lemma 7.21.

Lemma (7.21 restated)**.**

$$\mathbb{E}_t\left(\sum_{i=1}^{m}\delta_i^{t\le i}\right)=\Omega_\mu(1).$$

Proof of Lemma 7.21. By Lemma 7.26 $\beta_i^{T_i}\gamma_i^{T_i}$ is a supermartingale. Therefore $-2\beta_i^{T_i}\gamma_i^{T_i}=(\beta_i^{T_i}-\gamma_i^{T_i})^2-(\beta_i^{T_i})^2-(\gamma_i^{T_i})^2$ is a submartingale. By optional stopping theorem we have

$$\mathbb{E}_t\left((\beta_m^T-\gamma_m^T)^2-(\beta_m^T)^2-(\gamma_m^T)^2\right)\ge(\beta-\gamma)^2-\beta^2-\gamma^2.$$

Rearranging we get

$$\mathbb{E}_t\left((\beta_m^T-\gamma_m^T)^2-(\beta-\gamma)^2\right)\ge\mathrm{Var}(\beta_m^T)+\mathrm{Var}(\gamma_m^T).$$

By definition of S, when transcript t is observed exactly one of the following three cases happens:

1. $\beta_m^t=0.01\min(\beta,\gamma)$
   This transcript contributes at least $(0.99\min(\beta,\gamma))^2$ to $\mathrm{Var}(\beta_m^T)$.
2. $\gamma_m^t=0.01\min(\beta,\gamma)$
   This contributes at least $(0.99\min(\beta,\gamma))^2$ to $\mathrm{Var}(\gamma_m^T)$.
3. $\alpha_m^t=0.01\alpha$
   We do not have a guarantee on the contribution to $\mathrm{Var}(\beta_m^T)$ or $\mathrm{Var}(\gamma_m^T)$, but since $\alpha_m^{T_i}$ is a martingale we have $\mathbb{E}_t(\alpha_m^t)=\alpha$. In addition, $\alpha_m^t\le 1$. Thus $P(\alpha_m^T>0.01\alpha)\ge 0.99\alpha$.

From the above it follows that

$$\mathrm{Var}(\beta_m^T)+\mathrm{Var}(\gamma_m^T)\ge(0.99\alpha)(0.99\min(\beta,\gamma))^2=:c_\mu.$$

Consequently

$$\mathbb{E}_t\left((\beta_m^t-\gamma_m^t)^2-(\beta-\gamma)^2\right)\ge c_\mu. \tag{8}$$

Observe that $|\beta_m^t-\gamma_m^t|+|\beta-\gamma|\le 2$. Thus

$$|\beta_m^t-\gamma_m^t|-|\beta-\gamma|\ge((\beta_m^t-\gamma_m^t)^2-(\beta-\gamma)^2)/2.$$

Taking expectation of both sides and using inequality (8) we obtain

$$\mathbb{E}_t(|\beta_m^t-\gamma_m^t|-|\beta-\gamma|)\ge c_\mu/2. \tag{9}$$

By Lemma 7.27 we have

$$\mathbb{E}_t\left(\min(\beta_m^t,\gamma_m^t)-\min(\beta,\gamma)+\sum_{i=1}^{m}\delta_i^{t\le i}\right)\ge 0.$$

Using $\min(a,b)=(a+b)/2-|a-b|/2$ we derive

$$\begin{aligned}&\mathbb{E}_t\left(\frac{\beta_m^t+\gamma_m^t}{2}-\frac{|\beta_m^t-\gamma_m^t|}{2}-\frac{\beta+\gamma}{2}+\frac{|\beta-\gamma|}{2}+\sum_{i=1}^{m}\delta_i^{t\le i}\right)\\&=\mathbb{E}_t\left(-\frac{|\beta_m^t-\gamma_m^t|}{2}+\frac{|\beta-\gamma|}{2}+\sum_{i=1}^{m}\delta_i^{t\le i}\right)\ge 0,\end{aligned}$$

where the first step follows from the fact that $\beta_i^{T_i}$ and $\gamma_i^{T_i}$ are martingales. Rearranging and applying inequality (9) we finally arrive at the conclusion of the statement.

$$\mathbb{E}_t\left(\sum_{i=1}^{m}\delta_i^{t\le i}\right)\ge(1/2)\mathbb{E}_t(|\beta_m^t-\gamma_m^t|-|\beta-\gamma|)\ge c_\mu/4.$$

□

7.9.4 Upper Bound on the Rate of Convergence

In this subsection we present an r-round discretization (see Protocol 4) of our optimal protocol (see Protocol 2) for AND. We shall prove that the discretized AND protocol achieves $O(1/r^2)$ upper bound on the rate of convergence. This matches the lower bound on the rate of convergence proven in Subsection 7.9.1.

If $\beta < \gamma$ then Bob sends bit B as follows

$$B = \begin{cases} 1 & \text{if } y = 1 \\ 0 & \text{with probability } 1 - \beta/\gamma \text{ if } y = 0 \\ 1 & \text{with probability } \beta/\gamma \text{ if } y = 0 \end{cases}$$

If $B = 0$ the protocol terminates, the players output 0.

If $\beta > \gamma$ then Alice sends bit B as follows

$$B = \begin{cases} 1 & \text{if } x = 1 \\ 0 & \text{with probability } 1 - \gamma/\beta \text{ if } x = 0 \\ 1 & \text{with probability } \gamma/\beta \text{ if } x = 0 \end{cases}$$

If $B = 0$ the protocol terminates, the players output 0.

If $x = 0$ then Alice samples $N^A \in \{0, 1, \ldots, r-1\}$ with $P(N^A = i) = \frac{2r-2i-1}{r^2}$.

If $x = 1$ then Alice sets $N^A = r$.

If $y = 0$ then Bob samples $N^B \in \{0, 1, \ldots, r-1\}$ with $P(N^B = i) = \frac{2r-2i-1}{r^2}$.

If $y = 1$ then Bob sets $N^B = r$.

For $C = 0$ to $r - 1$

- If $C = N^A$ then Alice sends 1 to Bob, protocol terminates, players output 0
- Else Alice sends 0 to Bob
- If $C = N^B$ then Bob sends 1 to Alice, protocol terminates, players output 0
- Else Bob sends 0 to Alice

Protocol terminates, players output 1

Protocol 4: Discretized r-round protocol π_r for the AND-function

Recall that the "informational wastage" (or "information wasted") is how much extra information a particular bit, message, or protocol reveals when compared to the optimal protocol.

The most natural way to discretize our continuous AND protocol would be to sample numbers N^A and N^B uniformly at random from the set $\{0, \ldots, r-1\}$ when the corresponding player(s) have 0 as input. While analysing this option, we discovered that this discretization wastes increasing amounts information in later rounds as the counter C approaches r. This leads to a total information wasted $\approx \frac{1}{r^2} \sum_{i=1}^{r} \frac{1}{i} = \Theta\left(\frac{\log r}{r^2}\right)$. A natural remedy is to select numbers N^A and N^B non-uniformly, assigning less probability mass to the later rounds. Indeed, Protocol 4 assigns probability $\frac{2r-2i-1}{r^2}$ to the ith value of N^A and N^B leading to the correct $O(\frac{1}{r^2})$ bound on the total information wasted. In the rest of this section we prove this claim formally.

We start with two technical lemmas.

Lemma 7.28. *Suppose that Alice sends bit B distributed as follows*

$$B = \begin{cases} 1 & \text{if } X = 1 \\ 0 & \text{with probability } \psi \text{ if } X = 0 \\ 1 & \text{with probability } 1-\psi \text{ if } X = 0 \end{cases}$$

from prior $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \beta & \delta \\ \hline \end{array}$. *Then the informational wastage of B is*

$$O\left(\frac{\alpha\beta}{\alpha+\beta}\psi^3 + \frac{2\alpha\beta(\beta^2+3\alpha\beta+2\alpha^2)}{(1-\psi)^3\beta^3}\psi^4\right).$$

Proof. The informational wastage of bit B is

$$IW(\alpha,\beta,\psi) = I(B;X|Y) + P(B=1)\,\mathrm{IC}'_\mu(\mathrm{AND},0) - \mathrm{IC}_\mu(\mathrm{AND},0),$$

where

$$\mu' = \begin{array}{|c|c|}\hline (1-\psi)\alpha/t & (1-\psi)\beta/t \\ \hline \beta/t & \delta/t \\ \hline \end{array},$$

and $t = (1-\psi)\alpha + (2-\psi)\beta + \delta$. Furthermore, we have

$$I(B;X|Y) = H(B|Y) - H(B|XY) = (\alpha+\beta)H\left(\frac{\alpha}{\alpha+\beta}\psi\right) + (\beta+\delta)H\left(\frac{\beta}{\beta+\delta}\psi\right) - (\alpha+\beta)H(\psi).$$

Writing Taylor series for $IW(\alpha,\beta,\psi)$ for ψ around 0 we obtain

$$\exists\zeta \in [0,\psi] \text{ s.th. } IW(\alpha,\beta,\psi) = \frac{\alpha\beta}{(\alpha+\beta)\ln 64}\psi^3 + R(\zeta)\frac{\psi^4}{24},$$

where $R(\zeta) = \frac{2\alpha\beta(\beta^2+3\alpha\beta(1-\zeta)+\alpha^2(2-3\zeta+\zeta^3))}{(1-\zeta)^3(\alpha+\beta-\alpha\zeta)^3\ln 2}$.

The above expressions were obtained with help from Wolfram Mathematica. The statement of the lemma follows immediately. □

Suppose that players start with a symmetric prior $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \beta & \delta \\ \hline \end{array}$. Observe that the counter C in Protocol 4 can be viewed as a discrete implementation of a continuous clock from Protocol 2. The hand of our clock now moves in discrete steps from position z to position $z+\phi$ where $z = 1-\left(\frac{r-i}{r}\right)^2$ and $\phi = \frac{2r-2i-1}{r^2}$ for $i \in \{0,\ldots,r-1\}$. We now analyse how a single such move is accomplished by Alice and Bob in our protocol and how much information is wasted during this move.

At time z the prior μ becomes $\mu_z = \begin{array}{|c|c|}\hline (1-z)^2\alpha & (1-z)\beta \\ \hline (1-z)\beta & \delta \\ \hline \end{array}$ normalized. Thus, when the players move from time z to time $z+\phi$ it is equivalent to first Alice sending bit B as in Lemma 7.28 with $\psi = \frac{\phi}{1-z}$ followed by a similar bit B' sent by Bob. Note that after Alice sends bit B, the prior moves into Bob's region. In the optimal protocol, Bob would send *exactly* bit B'. Hence Bob's bit wastes no information. Therefore the total informational wastage incurred while moving clock hand from time z to time $z+\phi$ in 2 rounds of communication comes from bit B.

Lemma 7.29. *Let* $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \beta & \delta \\ \hline \end{array}$ *be a distribution with full support. When Alice and Bob in 2 rounds of communication advance the clock from position z to $z+\phi$ with $\frac{\phi}{1-z} \le \frac{2}{3}$ they waste a total of $O_\mu\left(\frac{\phi^3}{1-z}\right)$ information.*

Proof. As discussed in the paragraph before the statement of the lemma, we simply have to apply Lemma 7.28 to Alice's signal B with $\psi = \frac{\phi}{1-z}$ and distribution

$$\mu_z = \begin{array}{|c|c|}\hline (1-z)^2\alpha/n & (1-z)\beta/n \\ \hline (1-z)\beta/n & \delta/n \\ \hline \end{array},$$

where $n = (1-z)^2\alpha + 2(1-z)\beta + \delta$. Note that by assumptions of the lemma we have $\psi \le 2/3$, therefore we have $1/(1-\psi)^3 \le 27$. Furthermore we have $n \ge \delta$ and $\phi \le 1-z$. Plugging all this in Lemma 7.28 and simplifying we obtain that the total information wasted is

$$O\left(\frac{(1-z)^3\alpha\beta}{n((1-z)^2\alpha+(1-z)\beta)}\frac{\phi^3}{(1-z)^3} + 27\frac{2(1-z)^3\alpha\beta n^3}{(1-z)^3\beta^3n^2}\frac{(1-z)^2\beta^2+3(1-z)^3\alpha\beta+2(1-z)^4\alpha}{n^2}\frac{\phi^4}{(1-z)^4}\right) =$$

$$= O\left(\frac{\alpha}{\delta}\frac{\phi^3}{1-z} + \frac{\alpha}{\delta\beta^2}\frac{\phi^4}{(1-z)^2}\right) = O_\mu\left(\frac{\phi^3}{1-z}\right).$$

□

Now we are in a position to prove the main result of this subsection.

Theorem 7.30. *For distributions* $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & \delta \\ \hline \end{array}$ *with full support we have*

$$\mathrm{IC}^r_\mu(AND, 0) - \mathrm{IC}_\mu(AND, 0) = O_\mu\left(\frac{1}{r^2}\right).$$

Proof. Let π_r denote Protocol 4 and π denote Protocol 2. In the first stage of protocol π_r the player who is more likely to have 0 sends a bit that either terminates the protocol or moves the prior to the diagonal. This stage is exactly the same in protocol π. Thus the difference in the information cost of the two protocols arises only from the second (which we previously called symmetric) stage of π and π_r. Thus for the rest of the proof we shall assume that μ is symmetric, i. e., $\beta = \gamma$.

Observe that for the ith jump of the clock we have $\phi_i = \frac{2r-2i-1}{r^2}$ and $z_i = 1 - \left(\frac{r-i}{r}\right)^2$. Therefore $\frac{\phi_i}{1-z_i} = \frac{2r-2i-1}{(r-i)^2} \le \frac{2}{r-i}$. Hence $\frac{\phi_i}{1-z_i} \le 2/3$ for all i except $i \in \{r-2, r-1\}$. The later event happens with probability $O(1/r^2)$ conditioned on Alice having 0 as input. In addition if $X = 1$, Alice learns the entire Bob's bit. Hence the difference in the information cost of π_r and π arises from the events when Alice or Bob have 0 as input. Consequently we may ignore the last two movements of the clock as they contribute at most $O(1/r^2)$ to the informational wastage. For the rest of the clock movements we may apply Lemma 7.29 which says that the information wasted in the ith movement is $O_\mu\left(\frac{\phi_i^3}{1-z_i}\right)$. Aggregating it over the first $r-2$ movements of the clock we get the total amount of information wasted is

$$O_\mu\left(\sum_{i=0}^{r-3}\frac{(2r-2i-1)^3r^2}{r^6(r-i)^2}\right) = O_\mu\left(\sum_{i=0}^{r-3}\frac{(r-i)}{r^4}\right) = O_\mu\left(\frac{1}{r^2}\right).$$

□

8 The Communication Complexity of $\bigvee$-type Functions

The main result of this section is a characterization of the (randomized) communication complexity required to solve $\bigvee$-type functions, that is, functions of the form $g_n(X,Y)_n = \vee_{i=1}^n f(x_i, y_i)$, in terms of an informational quantity of the function f. The following definitions will be central to our analysis. Call a protocol π

good for f if π solves f correctly on all inputs. Let $\mathcal{U}_0, \mathcal{U}_1$ be the set of distributions supported on $f^{-1}(0)$, $f^{-1}(1)$ respectively. Define

$$IC^0(f,0) := \inf_{\pi \text{ good for } f} \max_{\mu \in \mathcal{U}_0} IC(\pi,\mu)$$
$$IC^1(f,0) := \inf_{\pi \text{ good for } f} \max_{\mu \in \mathcal{U}_1} IC(\pi,\mu)$$

By a minimax argument similar to the one in [8], it can be shown that the above definitions are equivalent to the following :

$$IC^0(f,0) := \max_{\mu \in \mathcal{U}_0} \inf_{\pi \text{ good for } f} IC(\pi,\mu)$$
$$IC^1(f,0) := \max_{\mu \in \mathcal{U}_1} \inf_{\pi \text{ good for } f} IC(\pi,\mu)$$

It means that instead of needing a single protocol to be good for all distributions, we can choose a protocol based on the distribution. We formally prove this in the Appendix.

Theorem 8.1. *Let $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ be a function. Then*

$$\inf_{\pi \text{ good for } f} \max_{\mu \in \mathcal{U}_0} IC(\pi,\mu) = \max_{\mu \in \mathcal{U}_0} \inf_{\pi \text{ good for } f} IC(\pi,\mu)$$

Similarly

$$\inf_{\pi \text{ good for } f} \max_{\mu \in \mathcal{U}_1} IC(\pi,\mu) = \max_{\mu \in \mathcal{U}_1} \inf_{\pi \text{ good for } f} IC(\pi,\mu)$$

These quantities measure the zero-error information cost of the function with respect to the restricted family of distributions having zero mass on the pre-image of 1 (0)[4] . But we require that the protocol be correct for *all* inputs. Also note that $\mathsf{IC}^1(f,0) = \mathsf{IC}^0(\bar{f},0)$.
We are now ready to prove Theorem 2.5. For convenience, we restate it below.

Theorem 8.2 (Theorem 2.5 restated)**.** *For any Boolean function $f : \{0,1\}^k \times \{0,1\}^k \to \{0,1\}$, let $g_n(X,Y) := \vee_{i=1}^n f(x_i,y_i)$, where $X = \{x_i\}_{i=1}^n, Y = \{y_i\}_{i=1}^n$ and $x_i, y_i \in \{0,1\}^k$. Then for all $\epsilon > 0$, there exists $\delta = \delta(f,\epsilon) > 0$ such that $\delta \to 0$ as $\epsilon \to 0$ and*

$$(\mathsf{IC}^0(f,0) - \delta) \cdot n \leq \mathcal{R}_\epsilon(g_n) \leq \mathsf{IC}^0(f,0) \cdot n + o(n)k.$$

The hardness of $R_\epsilon(g_n)$ of $\bigvee$-type functions g_n in the view of Yao's mini-max argument is captured by the distributions for f that put negligible mass on 1-entries (i.e. x, y such that $f(x,y) = 1$), for otherwise a trivial and small-communication sampling protocol would succeed with high probability. Interestingly, this phenomenon enters our proof in both the upper bound and the lower bound. In the upper bound, we use the aforementioned "small-communication sampling protocol" as a preprocessing step to either solve g_n or conclude that the prior puts negligible mass on 1-entries. In the lower bound, we extract the f function from the g_n function by feeding inputs to g_n sampled from a prior μ. We require that the value of g_n depends only on the extracted function f, hence the restriction that the inputs do not evaluate to 1.

Remark 8.3. Theorem 2.5 is stated for $\bigvee$-type functions only, but an equivalent result holds for $\bigwedge$-type functions, with $\mathsf{IC}^0(f,0)$ replaced by $\mathsf{IC}^1(f,0)$. Can be easily proved via De-Morgan's rule.

8.1 The Lower Bound

We begin with the '$\geq$' direction of theorem 2.5.

[4]Note that taking a maximum in the above definitions is allowed since the corresponding set of distributions is compact and information is continuous

Lemma 8.4. *For all $\epsilon > 0$, there exists $\delta = \delta(f, \epsilon) > 0$ such that $\delta \to 0$ as $\epsilon \to 0$ and $\mathcal{R}_\epsilon(g) \geq (\mathsf{IC}^0(f, 0) - \delta) \cdot n$*

At the heart of the proof lies a reduction from computing g_n function to computing f with n times less information cost with respect to the restricted distributions. This kind of a reduction was first introduced in [3] and since then has been used in [4], [10] and [8] in the context of direct sums for information complexity. We also need continuity of information cost to go from ϵ-error information cost (w.r.t the restricted distributions) to 0-error information cost.

Proof. Denote $\rho := \mathsf{IC}^0(f, 0)$. Suppose that there is a protocol Π for computing $g_n(X, Y)_n$ with error probability at most ϵ and communication cost $|\Pi|$. Now for all $\nu \in \mathcal{U}_0$, we will use Π to generate a protocol π which computes $f(x, y)$ with error probability at most ϵ for *any* $(x, y) \in \{0, 1\}^k$ with information cost smaller than $|\Pi|/n$ under ν. This protocol is obtained by restricting Π to a single (random) coordinate, where the rest of the coordinates are publicly sampled according to ν. The protocol is described in Figure 5 (cf. [10]). It can be proved that information cost of the restricted protocol π under ν is (see Proof of Theorem 3.17 in [10]):

$$\mathsf{IC}_\nu(\pi) \leq \frac{\mathsf{IC}_{\nu^n}(\Pi)}{n}$$

Furthermore, since ν is supported on $f^{-1}(0)$ and Π has error $\leq \epsilon$, we have that except with probability ϵ, for any $x, y \in \{0, 1\}^k$ it holds that:

$$\pi(x, y) = \Pi((X_1, \ldots, X_{J-1}, x, X_{J+1}, \ldots, X_n), (Y_1, \ldots, Y_{J-1}, y, Y_{J+1}, \ldots, Y_n)) = f(x, y)$$

Note that we measure the information cost of π with respect to ν only, while π computes $f(x, y)$ for *all* inputs $(x, y) \in \{0, 1\}^k$, except with probability ϵ (over the randomness of π).

Let $IC^g_\nu(f, 0) := \inf_{\pi \text{ good for } f} IC(\pi, \nu)$. Continuity of information cost at error $= 0$(Theorem 2.8) implies that $IC(\pi, \nu) \geq IC^g_\nu(f, 0) - \delta(f, \nu, \epsilon)$ for some $\delta(f, \nu, \epsilon) \to 0$ as $\epsilon \to 0$[5]. It is easy to observe from the proof of the continuity that we can take the rate of convergence independent of distribution. So $\max_{\nu \in \mathcal{U}_0} IC(\pi, \nu) \geq IC^0(f, 0) - \delta(f, \epsilon)$. Hence

$$\mathcal{R}_\epsilon(g) \geq \max_{\nu \in \mathcal{U}_0} \mathsf{IC}_{\nu^n}(\Pi) \geq n \cdot (\max_{\nu \in \mathcal{U}_0} IC(\pi, \nu)) \geq n \cdot (IC^0(f, 0) - \delta(f, \epsilon))$$

□

1. The parties jointly and publicly sample a uniformly selected index $J \in \{1, \ldots, n\}$.
2. The parties publicly sample $X_1, \ldots, X_{J-1}, Y_{J+1}, \ldots, Y_n$ independently according to ν.
3. The first party privately samples $X_{J+1}, \ldots, X_n$ and the second party privately samples $Y_1, \ldots, Y_{J-1}$ conditioned on the corresponding publicly sampled variables, so that each (X_i, Y_i) is distributed according to ν.
4. The parties run $\Pi((X_1, \ldots, X_{J-1}, x, X_{J+1}, \ldots, X_n),$ $(Y_1, \ldots, Y_{J-1}, y, Y_{J+1}, \ldots, Y_n))$ and output its output.

Protocol 5: The protocol $\pi(x, y)$, $x, y \in \{0, 1\}$

8.2 Upper Bound

In this section we prove an upper bound on the communication complexity of OR-type functions. Recall that we call a protocol π good for f if π solves f correctly on all inputs. We first bound the information complexity of g_n.

[5] In the proof of continuity, we get good protocols in the end

Lemma 8.5. *Let $f : \{0,1\}^k \times \{0,1\}^k \to \{0,1\}$ be a function and let $I = IC^0(f,0)$. Then for all n and for all distributions μ on $\{0,1\}^{nk} \times \{0,1\}^{nk}$, $IC_\mu(g_n,0) \leq nI + o(n)k$, where $g_n(x_1,x_2,\dots x_n,y_1,y_2,\dots y_n) = \vee_{i=1}^n f(x_i,y_i)$ and $x_i,y_i \in \{0,1\}^k$ for all i. More precisely $IC_\mu(g_n,0) \leq nI + O(n^{2/3}\log(n)\cdot k)$*

The intuition behind the proof is that the hardest distributions for g_n are the ones in which the marginal distributions on almost all copies have negligible mass on $f^{-1}(1)$, otherwise Alice and Bob could just sample a small number of coordinates and either find a coordinate (x_i,y_i) such that $f(x_i,y_i)=1$.

Proof. Let π be a protocol that is good for f such that $\max_{\mu\in\mathcal{U}_0} IC(\pi,\mu) \leq I+\delta$, for some $\delta > 0$. Consider the protocol π_n for computing g_n (Figure 6) :

1. Alice and Bob exchange(with replacement using public randomness) $n^{2/3}$ random coordinates $(x_i,y_i) \in \{0,1\}^k$ (using $n^{2/3}\cdot k$ bits). Denote the (multi) set of random coordinates by J. If for some $i \in J$ $f(x_i,y_i)=1$ the parties output 1 and terminate.
2. On each coordinate (other than those in J), Alice and Bob run the protocol π and output 1 if π outputs 1 on some coordinate.

Protocol 6: Protocol $\pi_n(x,y)$

Correctness of the above protocol follows from the fact that π is good for f. We now analyze its information cost. Let $X,Y \sim \mu$ denote the random variables for the input. Let Π be a random variable for the transcript of the protocol π_n. Then $\Pi = J\Pi_1\Pi_2$, where J denotes the random coordinates sampled, Π_1 denotes the contents of the random coordinates and Π_2 denotes the random variable for Step 2 of the protocol. Let E denote the indicator random variable for the event that for some $i \in J$ $f(x_i,y_i)=1$. Note that if $E=1$, then Π_2 is empty. Now

$$\begin{aligned} I(\Pi;X|Y) &= I(\Pi_1\Pi_2;X|YJ) \\ &= I(\Pi_1;X|YJ) + I(\Pi_2;X|YJ\Pi_1) \\ &\leq n^{2/3}\cdot 2k + I(\Pi_2;X|YJ\Pi_1E) \\ &\leq n^{2/3}\cdot 2k + I(\Pi_2;X|YE) \end{aligned}$$

First inequality follows from the fact that we are exchanging at most $n^{2/3}\cdot 2k$ bits (can be easily done in $n^{2/3}\cdot k+1$ bits) and the fact that J and Π_1 determine E. The second inequality is true because conditioned on E and X,Y, Π_2 is independent of J and Π_1.

$$I(\Pi_2;X|YE) = Pr[E=0]\cdot I(\Pi_2;X|Y,E=0)$$

Let $N_f(x,y) = |\{x_i,y_i \text{ s.t. } f(x_i,y_i)=1\}|$. Here $x,y \in \{0,1\}^{nk}$ and $x_i,y_i \in \{0,1\}^k$ are blocks of x,y. We slightly abuse notation and let $\mu(d)$ denote $Pr[N(X,Y)=d]$, where $X,Y\sim\mu$. If $Pr[E=0] \leq 1/n^{1/3}$, then $I(\Pi_2;X|YE) \leq n^{2/3}\cdot k$. Hence we can assume that $Pr[E=0] \geq 1/n^{1/3}$.
Let $\mu' = \mu|(E=0)$. If we have x,y such that $N_f(x,y)=d$, then the probability of sampling $n^{2/3}$ coordinates and not getting any (x_j,y_j) such that $f(x_j,y_j)=1$ is bounded by $e^{-2d^2/n^{4/3}}$ by Chernoff bounds. Thus

$$\mu'(d) \leq \frac{\mu(d)\cdot e^{-2d^2/n^{4/3}}}{Pr[E=0]} \leq \mu(d)\cdot e^{-2d^2/n^{4/3}}\cdot n^{1/3}$$

Thus for $d \geq n^{2/3}\log(n)$, $\mu'(d)$ is very small. Hence

$$E_{X,Y\sim\mu'}(N_f(X,Y)) \leq O(n^{2/3}\log(n))$$

Let μ'_i denote the marginal distribution of μ' on the i^{th} block. Let $\epsilon_i = Pr[f(X_i,Y_i)=1]$, where $(X_i,Y_i)\sim\mu'_i$. We will need the following lemma from [8] (see Proof of Theorem 4.2). It says that the information cost of a protocol Π, that runs a protocol π independently on many copies, is less than the sum of information costs of the protocol π on different copies w.r.t the marginal distributions.

Lemma 8.6. *Let μ be a distribution on $\{0,1\}^{nk}$. Divide the input into n blocks of size k each and let μ_i denote the marginal distribution on i^{th} block. Let τ be a protocol that runs on $2k$ sized inputs. Then $IC(\tau^n, \mu) \leq \sum_{i=1}^{n} IC(\tau, \mu_i)$.*

Now using the above lemma, we get that

$$I(\Pi_2; X|Y, E=0) + I(\Pi_2; Y|X, E=0) \leq \sum_{i=1}^{n} IC(\pi, \mu_i') \leq n \cdot IC(\pi, (\sum_{i=1}^{n} \mu_i')/n)$$

where the last inequality follows from concavity of information cost (Lemma A.1). By linearity of expectation

$$\sum_{i=1}^{n} \epsilon_i = E_{X,Y\sim\mu'}(N_f(X,Y)) \leq O(n^{2/3}\log(n))$$

Thus $\nu = (\sum_{i=1}^{n} \mu_i')/n$ has $O(\log(n)/n^{1/3})$ mass on $f^{-1}(1)$ and hence is $O(\log(n)/n^{1/3})$ close to distribution ν' in $\mathcal{U}_0$. Using Lemma B.1 along with the fact that $IC(\pi, \nu') \leq I + \delta$ gives us that $IC(\pi, \nu) \leq I + \delta + O(\log(n)/n^{1/3} \cdot k) + H(O(\log(n)/n^{1/3}))$. Hence

$$I(\Pi_2; X|Y, E=0) + I(\Pi_2; Y|X, E=0) \leq n(I+\delta) + O(n^{2/3}\log(n) \cdot k)$$

Thus we get that $IC(g_n, 0) \leq n(I+\delta) + O(n^{2/3}\log(n) \cdot k)$, $\forall \delta > 0$. Hence $IC(g_n, 0) \leq nI + O(n^{2/3}\log(n) \cdot k)$. □

The next theorem proves an upper bound on the communication complexity of g_M.

Theorem 8.7. *For any constant $\epsilon > 0$, $R_\epsilon(g_M) \leq M \cdot IC^0(f) + o(M)k$*

We will need the following non-distributional version of "information equals amortized communication" from [8].

Theorem 8.8. *Let $g : X \times Y \to \{0,1\}$ be a function, and let $IC(g, 0) = I$. Then for each $\delta_1, \delta_2 > 0$, there is an $C = C(g, \delta_1, \delta_2)$ such that for each $N \geq C$, there exists a protocol $\pi_N = \pi_N((x_1, x_2, \ldots, x_N), (y_1, y_2, \ldots, y_N))$ for computing N instances of g. The protocol has communication complexity $< N \cdot I \cdot (1 + \delta_1)$ and answers on all coordinates correctly except with probability δ_2.*

Proof. (of Theorem 8.7) The proof utilizes the self-reducible structure of g_M. This kind of a self-reducibility trick was used in [8] to analyze the information complexity of the Disjointness function. Consider a sufficiently large M. Choose n to be the largest such that $n \cdot C(g_n, 1/n, \epsilon) \leq M$. Now let $N \geq C(g_n, 1/n, \epsilon)$ be the largest such that $n \cdot N \leq M$. We can assume that $n \cdot N = M$. By Theorem 8.8, there exists a protocol π_N for solving N instances of g_n with communication $< N \cdot IC(g_n, 0) \cdot (1 + 1/n)$ and solving all instances correctly except with probability ϵ. Now consider the protocol π_M for solving g_M (Figure 7) :

1. Divide the input into N blocks of size n each and run π_N to solve these N instances of g_n.
2. Output 1 if π_N outputs 1 on some instance.

Protocol 7: Protocol π_M

Clearly the protocol has error $\leq \epsilon$. The communication cost of the protocol is :

$$\begin{aligned} N \cdot IC(g_n, 0) \cdot (1 + 1/n) &\leq N \cdot (n \cdot IC^0(f, 0) + O(n^{2/3}\log(n) \cdot k)) \cdot (1 + 1/n) \\ &\leq M \cdot IC^0(f) + o(M)k \end{aligned}$$

□

We give an overview of the whole proof and protocol for solving g_M just for clarity :

1. Pick $n << M$, let $N = M/n$.
2. Construct a low information cost protocol τ for g_n by:
   - Sampling $n^{2/3}$ coordinates and terminating if we know output of g_n.
   - Apply the protocol that achieves $IC^0(f, 0)$ to the remaining coordinates.
3. Use "information equals amortized communication" to compress the low information cost protocol τ^N into a low communication protocol for g_M.

8.3 The exact communication cost of $DISJ_n$

In this section we show how our previous results easily imply Theorem 2.6. Since Non-Disjointness is a $\bigvee$-type function, with the inner function "f" being the AND function, and since the tools obtained in section 7 enable us to compute $\mathsf{IC}^0(AND, 0)$, we can use Theorem 2.5 to obtain the exact randomized communication complexity of $DISJ_n$ with error tending to 0.

Corollary 8.9 (Theorem 2.6 restated)**.** *For all $\epsilon > 0$, there exists $\delta = \delta(f, \epsilon) > 0$ such that $\delta \to 0$ as $\epsilon \to 0$ and*

$$(C_{DISJ} - \delta) \cdot n \leq \mathcal{R}_\epsilon(g_n) \leq C_{DISJ} \cdot n + o(n)k.$$

where $C_{DISJ} \approx 0.4827$ bits.

Note that the reductions in the proof of the upper bound and lower bound of Theorem 2.5 preserve the number or rounds. Hence an r-round protocol for $DISJ_n$ will be suboptimal by at least $\Omega(n/r^2)$, because of Theorem 2.4.

Proof. Theorem 7.5 says that

$$\lim_{\epsilon \to 0} \max_{\mu : \mu(1,1) \leq \epsilon} \mathsf{IC}_\mu(\text{AND}, 0) \approx 0.4827\ldots.$$

Note that here we have distributions which have negligible mass on $(1, 1)$ rather than 0 mass which we require in the definition of $IC^0(AND, 0)$. But that is because definition of $IC^0(AND, 0)$ deals only with protocols that work correctly for each input, whereas the definition of $IC_\mu(AND, 0)$ required protocols to be correct only on the support of μ. So the fact that $IC^0(AND, 0) \approx 0.4827$ requires a small proof.

Claim 8.10. *For all functions $f : \{0, 1\}^k \times \{0, 1\}^k \to \{0, 1\}$,*

$$IC^0(f, 0) = \lim_{\epsilon \to 0} \max_{\mu : \mu(f^{-1}(1)) \leq \epsilon} \mathsf{IC}_\mu(f, 0)$$

Proof. By the definition of $IC^0(f, 0)$, for all $\delta > 0$, there exists a protocol π that solves f correctly on all inputs and $\max_{\mu \in \mathcal{U}_0} IC(\pi, \mu) \leq IC^0(f, 0) + \delta$. Let $\epsilon > 0$ (recall that $\mathcal{U}_0$ is the set of distributions supported on $f^{-1}(0)$) and let μ_ϵ be a distribution such that $\mu_\epsilon(f^{-1}(1)) \leq \epsilon$ and let μ be the distribution obtained by restricting μ_ϵ to $f^{-1}(0)$. Then by Lemma B.1, $|IC(\pi, \mu) - IC(\pi, \mu_\epsilon)| \leq t(\epsilon)$, where $t(\epsilon) \to 0$ as $\epsilon \to 0$. Also since π solves f correctly on all inputs, π has 0-error w.r.t every distribution. Thus

$$IC_{\mu_\epsilon}(f, 0) \leq IC^0(f, 0) + \delta + t(\epsilon)$$

and since this is true for all $\epsilon, \delta > 0$

$$\lim_{\epsilon \to 0} \max_{\mu : \mu(f^{-1}(1)) \leq \epsilon} \mathsf{IC}_\mu(f, 0) \leq IC^0(f, 0)$$

For the other direction, we use the other definition for $IC^0(f, 0)$ i.e.

$$IC^0(f, 0) = \max_{\mu \in \mathcal{U}_0} \inf_{\pi \text{ good for } f} IC(\pi, \mu)$$

Let $\mu \in \mathcal{U}_0$ be the distribution that achieves the maximum in the above definition. Perturb μ by ϵ to obtain μ_ϵ i.e. $\mu_\epsilon = \epsilon \cdot \mathcal{U}_k + (1-\epsilon) \cdot \mu$ ($\mathcal{U}_k$ is the uniform distribution over $\{0,1\}^k \times \{0,1\}^k$). Then μ_ϵ has full support and $\mu_\epsilon(f^{-1}(1)) \leq \epsilon$. Let π be a protocol that has 0-error w.r.t μ_ϵ and $IC(\pi, \mu_\epsilon) \leq IC_{\mu_\epsilon}(f,0) + \delta$. Since μ_ϵ has full support, π works correctly for all inputs. Also by Lemma B.1, $IC(\pi,\mu) \leq IC(\pi,\mu_\epsilon) + t(\epsilon) \leq IC_{\mu_\epsilon}(f,0) + \delta + t(\epsilon)$. Since this is true for all $\delta > 0$,

$$\inf_{\pi \text{ good for } f} IC(\pi,\mu) \leq IC_{\mu_\epsilon}(f,0) + t(\epsilon)$$

and hence

$$IC^0(f,0) \leq \lim_{\epsilon \to 0} \max_{\mu:\mu(f^{-1}(1)) \leq \epsilon} \mathsf{IC}_\mu(f,0)$$

□

Now Theorem 2.6 follows from Theorem 2.5 and the fact that

$$C_{DISJ} = \mathsf{IC}^0(AND, 0) \approx 0.4827$$

□

9 Exact Complexity of $DISJ$ with small sets

We also study the $DISJ_n$ problem with the promise that both Alice and Bob have sets of size $\leq k$. Lets denote this by $DISJ_n^k$. This problem was studied in [20]. It is also one of the problems that give a separation between deterministic communication complexity and average-case 0-error communication complexity (e.g. see [27]). There they proved the following theorem:

Theorem 9.1. *$R_\epsilon(DISJ_n^k) \leq O(k)$, for all constant $\epsilon > 0$. Moreover the error is one-sided i.e. when the sets intersect, the protocol always outputs intersect.*

A lower bound of $\Omega(k)$ is immediate from the $\Omega(n)$ lower bound on the communication complexity of $DISJ_n$. We are able to determine the exact communication complexity of this problem except for some regimes.

Theorem 9.2. *Let n, k be such that $k = \omega(1)$ and $n/k = \omega(1)$. Then for all constant $\epsilon > 0$, $(\frac{2}{\ln 2} - O(\sqrt{\epsilon})) \cdot k - o(k) \leq R_\epsilon(DISJ_n^k) \leq \frac{2}{\ln 2} \cdot k + o(k)$.*

We start by proving a lower bound. In this section when we talk about a 0-error protocol, we will mean a protocol that is correct for *all* inputs. So we will use $IC_\mu(AND, 0)$ to denote the information cost of AND w.r.t the best protocol that works correctly not only for the support of μ but for *all* inputs.

9.1 Lower Bound

Lemma 9.3. *Let n, k be such that $k = \omega(1)$ and $n/k = \omega(1)$. Then $R_\epsilon(DISJ_n^k) \geq (\frac{2}{\ln 2} - O(\sqrt{\epsilon})) \cdot k - o(k)$.*

Proof. Once again, the idea is to show that a low communication protocol for $DISJ_n^k$ can be used to devise a low information protocol for a single copy of AND under some distribution. Consider the following distribution for the AND function.

$\mu =$

$1 - 2(k - k^{2/3})/n$	$(k - k^{2/3})/n$
$(k - k^{2/3})/n$	0

Let Π be a protocol for $DISJ_n^k$ with error probability at most ϵ and communication $|\Pi|$. We will design a protocol π for AND which works correctly for *all* inputs with high probability and which has information cost $\leq |\Pi|/n$ w.r.t. μ. The protocol is the same as Protocol 5, except that we sample the remaining coordinates according to μ. Let x, y be the inputs to π.

As before

$$\mathsf{IC}_{\mu}(\pi) \leq \frac{\mathsf{IC}_{\mu^n}(\Pi)}{n} \leq \frac{|\Pi|}{n}$$

By multiplicative Chernoff bounds, except with probability $2e^{-(k-k^{2/3})^{-2/3}(k-k^{2/3})/3}$, Alice and Bob both have sets of size (not counting the embedded coordinate)

$$\leq (k-k^{2/3})(1+(k-k^{2/3})^{-1/3}) \leq k-1$$

Furthermore, since the distribution μ has zero mass on $(1,1)$, the answer of π is determined by $x \wedge y$, so by the guarantee on Π, except with probability $\epsilon + e^{-k^{\Omega(1)}}$ (probability over the internal randomness of π), π correctly computes $x \wedge y$. However, to complete the proof we shall use previous analysis which only has guarantees for 0-error computation of AND. Thus we shall need a continuity argument to argue that π can be extended to a 0-error protocol with a "tiny" overhead in the information cost. Unfortunately the convergence rate in Theorem 2.8 is not good enough here (because the information cost of AND w.r.t μ is a sub-constant and the deviation from the information cost is constant for constant ϵ in Theorem 2.8). So we get a stronger convergence rate for this particular case.

Lemma 9.4. *Let* $\nu = \begin{array}{|c|c|}\hline 1-2k/n & k/n \\ \hline k/n & 0 \\ \hline \end{array}$. *Then if there is protocol* π *for AND such that for* all *inputs,* π *outputs the correct answer with probability* $\geq \epsilon$*, and* $IC(\pi,\nu) = I$*. Then for all* $\delta > 0$*, there is protocol* π' *for AND such that it is correct for all inputs and* $IC(\pi',\nu) \leq I + O(\frac{k}{n} \cdot \sqrt{\epsilon}) + \delta$.

First lets see what this lemma gives us. We get a protocol π' for AND with information cost $\leq \frac{|\Pi|}{n} + O(\frac{(k-k^{2/3})}{n} \cdot \sqrt{\epsilon + e^{-k^{\Omega(1)}}})$. However the information cost of AND with respect to $\begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \beta & 0 \\ \hline \end{array}$ (w.r.t. restricted protocols that work for each input) is (by Claim 7.6)

$$= \frac{\beta}{\ln 2} + \frac{\beta^2}{\alpha}\log\frac{\beta}{\beta+\alpha} + \alpha\log\frac{\alpha+\beta}{\alpha}$$

For μ, $\alpha = 1-2(k-k^{2/3})/n$ and $\beta = (k-k^{2/3})/n$. For these parameters, we get that the information cost is $\frac{2}{\ln 2} \cdot \frac{k}{n} \pm o(\frac{k}{n})$. Thus we get that

$$\begin{aligned} |\Pi| &\geq n \cdot \left(\frac{2}{\ln 2} \cdot \frac{k}{n} - o\left(\frac{k}{n}\right) - O\left(\frac{(k-k^{2/3})}{n} \cdot \sqrt{\epsilon + e^{-k^{\Omega(1)}}} \right) \right) \\ &= \left(\frac{2}{\ln 2} - O\left(\sqrt{\epsilon}\right) \right) \cdot k - o(k) \end{aligned}$$

We will need the following proposition:

Proposition 9.5. *Let* $\chi = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \gamma & 0 \\ \hline \end{array}$ *be a distribution for AND. Also let* $\beta \leq \gamma$*. Then* $IC_{\chi}(AND) \leq O(\beta \cdot \log(2\gamma/\beta))$

Proof. Let us first look at the information cost of a symmetric distribution. Consider the distribution $\begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \beta & 0 \\ \hline \end{array}$ By Claim 7.6 the information cost is

$$\frac{\beta}{\ln 2} + \frac{\beta^2}{\alpha}\log\frac{\beta}{\beta+\alpha} + \alpha\log\frac{\alpha+\beta}{\alpha} \leq \frac{2\beta}{\ln 2}$$

Now for χ, the information cost by Claim 7.7 is (cost of the symmetrization step) + t·(cost of remaining symmetric distribution). t·(cost of remaining symmetric distribution) $\leq t \cdot O(\beta/t) \leq O(\beta)$. For the symmetrization step, the cost is

$$
\begin{aligned}
&(\alpha+\beta)H\left(\frac{\beta}{\alpha+\beta}+\frac{\beta}{\gamma}\cdot\frac{\alpha}{\alpha+\beta}\right)-\alpha H\left(\frac{\beta}{\gamma}\right)\\
&=(\alpha+\beta)H\left(\frac{\beta}{\gamma}+\frac{\beta}{\alpha+\beta}\left(1-\frac{\beta}{\gamma}\right)\right)-\alpha H\left(\frac{\beta}{\gamma}\right)\\
&\leq(\alpha+\beta)\left(H\left(\frac{\beta}{\gamma}\right)+\frac{\beta}{\alpha+\beta}\left(1-\frac{\beta}{\gamma}\right)H'\left(\frac{\beta}{\gamma}\right)\right)-\alpha H\left(\frac{\beta}{\gamma}\right)\\
&=\beta H\left(\frac{\beta}{\gamma}\right)+\beta\left(1-\frac{\beta}{\gamma}\right)\left(\log\left(1-\frac{\beta}{\gamma}\right)-\log\left(\frac{\beta}{\gamma}\right)\right)\\
&=\beta\log\left(\frac{\gamma}{\beta}\right)
\end{aligned}
$$

The first inequality follows from the concavity of the entropy function. This completes the proof of Proposition 9.5 □

Proof. (**Of Lemma 9.4**) The main idea behind the proof is that if π is a ϵ-error protocol, then the distributions on the leaves should be "easy to solve". Hence, to finish the job, when reaching a leaf ℓ, Alice and Bob run the the optimal protocol for the distribution at leaf ℓ. We formalize this idea below (Figure 8):

1. Run the protocol π.
2. Upon reaching a leaf l, run a protocol τ such that $IC(\tau,\nu_l)\leq IC_{\nu_l}(AND,0)+\delta$ and output according to the output of τ.

Protocol 8: 0-error protocol

The novel idea here is that upon reaching a leaf, Alice and Bob run the optimal 0-error protocol for AND for the particular distribution reached in the corresponding leaf, rather than run a simpler protocol as we do in the proof of continuity of IC for general functions.

Alice and Bob start with the distribution ν. Let l be a leaf in the protocol tree. Let p_l be the probability of reaching this leaf (probability over the randomness of the protocol and the distribution ν). Let $X,Y\sim\nu$. Let ν_l be the distribution conditioned on reaching leaf l. Let $\nu_l=$

α	β
γ	0

We consider two cases. The first case is when the protocol outputs 0 upon reaching the leaf l.

Claim 9.6. *$Pr[\pi$ reaches leaf l on input $(1,1)]=p_l\cdot\frac{\beta}{k/n}\cdot\frac{\gamma}{k/n}\cdot\frac{1-2k/n}{\alpha}$*

Proof. We use the rectangular structure of a protocol. Particularly since π is a protocol, it holds that for all leaves l, there exists functions $p_A:\{0,1\}\to[0,1]$ and $p_B:\{0,1\}\to[0,1]$ such that

$$Pr[\pi \text{ reaches leaf } l \text{ on input } (a,b)]=p_A(a)\cdot p_B(b)$$

p_A and p_B are basically the products of probabilities at Alice's and Bob's nodes, respectively. Now

$$p_A(1)p_B(1)=\frac{p_A(1)p_B(0)\cdot p_A(0)p_B(1)}{p_A(0)p_B(0)} \tag{10}$$

Also

$$
\begin{aligned}
p_A(1)p_B(0)=Pr[\pi \text{ reaches leaf } l \text{ on input } (1,0)]&=Pr[\pi \text{ reaches leaf } l|X=1,Y=0]\\
&=\frac{Pr[\pi \text{ reaches leaf } l]\cdot\nu_l(1,0)}{\nu(1,0)}=p_l\cdot\frac{\gamma}{k/n}
\end{aligned}
$$

Similarly we get that

$$p_A(0)p_B(1) = p_l \cdot \frac{\beta}{k/n}$$
$$p_A(0)p_B(0) = p_l \cdot \frac{\alpha}{1-2k/n}$$

Using (10), we get Claim 9.6 □

Now since we output 0 on this leaf l, this leaf contributes error $p_l \cdot \frac{\beta}{k/n} \cdot \frac{\gamma}{k/n} \cdot \frac{1-2k/n}{\alpha}$ to the error for $(1,1)$. Thus

$$\sum_{\text{leaves } l \text{ that output } 0} p_l \cdot \frac{\beta}{k/n} \cdot \frac{\gamma}{k/n} \cdot \frac{1-2k/n}{\alpha} \leq \epsilon \implies \sum_{\text{leaves } l \text{ that output } 0} p_l \cdot \frac{\beta}{k/n} \cdot \frac{\gamma}{k/n} \leq 2\epsilon$$

Now if $\beta \leq \gamma$, the contribution of extra information cost from leaf l is $\leq p_l \cdot \beta \log(2\gamma/\beta)$ (we will ignore the δ term because it is not important), otherwise $\leq p_l \cdot \gamma \log(2\beta/\gamma)$. So wlog assume that $\beta \leq \gamma$. Let $\beta_0 = \frac{\beta}{k/n}$ and $\gamma_0 = \frac{\gamma}{k/n}$. We make the following simple claim.

Claim 9.7. *Either* $\beta_0 \log(2\gamma_0/\beta_0) < 4\sqrt{2\epsilon}$ *or* $\beta_0 \log(2\gamma_0/\beta_0) < \frac{\beta_0\gamma_0}{\sqrt{2\epsilon}}$.

Note that the above claim proves that (after considering the other similar region where $\gamma \leq \beta$)

$$\sum_{\text{leaves } l \text{ that output } 0} \text{extra info from leaf } l \leq O(k/n \cdot \sqrt{\epsilon})$$

□

Proof. (Of Claim 9.7) Assume on the contrary and let $\frac{2\gamma_0}{\beta_0} = 2^l$, where $l \geq 1$. Then

$$\beta_0 \cdot l > 4\sqrt{2\epsilon} > 2\gamma_0/l \implies l^2 > 2^{l+1}$$

a contradiction □

Now consider a leaf l which outputs 1. Again assume wlog that $\beta \leq \gamma$. Then

$$Pr[\pi \text{ reaches leaf } l \text{ on input } (1,0)] = p_l \cdot \frac{\gamma}{k/n}$$

So the contribution of this leaf l to the error for $(1,0)$ is $p_l \cdot \frac{\gamma}{k/n}$. Thus we get that

$$\sum_{\text{leaves } l \text{ that output } 1} p_l \cdot \frac{\gamma}{k/n} \leq \epsilon$$

The contribution of this leaf to the extra info cost is $\leq p_l \cdot O(\beta \log(2\gamma/\beta))$. Now since $\beta_0 \log(2\gamma_0/\beta_0) < 2\gamma_0$, hence (after considering the other region $\gamma \leq \beta$)

$$\sum_{\text{leaves } l \text{ that output } 1} \text{extra info from leaf } l \leq O(k/n \cdot \epsilon)$$

This completes the proof of Lemma 9.4. □

9.2 Upper Bound

Now we prove the upper bound on the communication complexity of $DISJ_n^k$.

Theorem 9.8. *Let N, K be such that $K = \omega(1)$ and $N/K = \omega(1)$. Then for all constant $\epsilon > 0$, $R_\epsilon(DISJ_N^K) \leq \frac{2}{\ln 2} \cdot K + o(K)$.*

We start with the following proposition :

Proposition 9.9. *Let $\nu = \begin{array}{|c|c|}\hline 1-2k/n+l/n & (k-l)/n \\ \hline (k-l)/n & l/n \\ \hline \end{array}$, where $n \geq k \geq l$. Then $IC_\nu(AND) \leq \frac{2}{\ln 2} \cdot \frac{k}{n} + O\left(\frac{l}{n} \cdot \log\left(\frac{2k}{l}\right)\right)$.*

Proof. By Claim 7.8, the information cost is $\leq$

$$\frac{\beta}{\ln 2} + 2\gamma \log \frac{\beta+\gamma}{\gamma} + 2\beta \log \frac{\beta+\gamma}{\beta} + \alpha \log \frac{\alpha+\beta}{\alpha}$$

Here $\alpha = 1 - 2k/n + l/n$, $\beta = (k-l)/n$ and $\gamma = l/n$. Plugging in the values we get that $IC_\nu(AND) \leq$

$$\frac{2}{\ln 2} \cdot \frac{k}{n} + O\left(\frac{l}{n}\right) + O\left(\frac{l}{n} \cdot \log\left(\frac{k}{l} + 1\right)\right) \leq \frac{2}{\ln 2} \cdot \frac{k}{n} + O\left(\frac{l}{n} \cdot \log\left(\frac{2k}{l}\right)\right)$$

□

For an upper bound on the communication complexity of set-disjointness, we will also need to study the complexity of set-intersection for some regime of parameters. Consider the distribution ν where $n = k^2$. Then $IC_\nu(AND, 0) \leq \frac{2}{\ln 2} \cdot 1/k + O\left(\frac{l}{k^2} \cdot \log\left(\frac{2k}{l}\right)\right)$. Let $r(k)$ be the number of rounds of a protocol π for AND such that $IC(\pi, \nu) \leq \frac{2}{\ln 2} \cdot 1/k + O\left(\frac{l}{k^2} \cdot \log\left(\frac{2k}{l}\right)\right) + 1/k^{3/2}$. Let $t(N)$ be the largest k such that $k^2 r(k) \cdot \log(k^4 r(k)) \leq N$. Then we have the following lemma :

Lemma 9.10. *Let K, N be such that $K = \omega(1)$, $N/K = t(N)$. Then for all $L = o(K)$, there is a randomized protocol Π_L such that if $x, y \in \{0,1\}^N \times \{0,1\}^N$ and $|x| \leq K, |y| \leq K$, $|x \wedge y| \leq L$, then $\Pi(x, y)$ returns the intersecting coordinates in x, y (i.e. solves set-intersection), except with probability $1/t(N)^{7/2}$. Π_L has expected communication cost $\leq \frac{2}{\ln 2} \cdot K + o(K)$, for all x, y such that $|x| \leq K, |y| \leq K$, $|x \wedge y| \leq L$ and has maximum communication $O(N \cdot t(N))$ for all x, y.*

Proof. The central idea of the proof is to run an optimal information-cost protocol for AND on each coordinate and then compress the resulting protocol using "information equals amortized communication". Note that we can assume that Alice and Bob both have sets of size exactly K, otherwise Alice and Bob can have K dummy elements each which are distinct and they can complete their sets using these dummy elements. The universe size increases from N to $N + 2K$, but the universe size doesn't matter anyways.
Choose $k = t(N)$ and $n = k^2$ (in the distribution ν of proposition 9.9). Let μ be any distribution on $\{0,1\}^N \times \{0,1\}^N$. Then consider the protocol π^N in which Alice and Bob run the protocol π on each coordinate (π being the protocol that has $r(k)$ number of rounds and information cost $\frac{2}{\ln 2} \cdot 1/k + O\left(\frac{l}{k^2} \cdot \log\left(\frac{2k}{l}\right)\right) + 1/k^{3/2}$ w.r.t ν). Let μ_i be the marginal distribution on coordinate i.

$$IC(\pi^N, \mu) \leq \sum_{i=1}^{N} IC(\pi, \mu_i) \leq N \cdot IC(\pi, (\sum_{i=1}^{N} \mu_i)/N)$$

The first inequality follows from Lemma 8.6 and the second follows from concavity of information cost (Lemma A.1). Let $\sum_{i=1}^{N} \mu_i)/N = \bar{\mu}$. For $x, y \in \{0,1\}^N$, let $N_{a,b}(x, y) = |\{i \text{ s.t. } x_i = a, y_i = b\}|$. Then $\bar{\mu}(a, b) = \mathbb{E}_{X,Y \sim \mu}[N_{a,b}(X, Y)]$. Let L be the expected intersection size of $X, Y \sim \mu$. Then $\bar{\mu}(1,1) = L/N$ and $\bar{\mu}(1,0) = \bar{\mu}(0,1) = (K - L)/N$. Let $Lk/K = l$. Then $\bar{\mu}(1,1) = l/k^2$ and $\bar{\mu}(1,0) = \bar{\mu}(0,1) = (k-l)/k^2$. The number of rounds of π^N is $r(k)$ and information cost $\leq$

$$N \cdot \left(\frac{2}{\ln 2} \cdot 1/k + O\left(\frac{l}{k^2} \cdot \log\left(\frac{2k}{l} \right) \right) + 1/k^{3/2} \right) = \frac{2}{\ln 2} \cdot K + O\left(L \cdot \log\left(\frac{2K}{L} \right) \right) + K/\sqrt{k}$$
$$= \frac{2}{\ln 2} \cdot K + O\left(L \cdot \log\left(\frac{2K}{L} \right) \right) + o(K)$$

Now we compress the protocol π^N using the following round-by-round compression lemma from [10] :

Lemma 9.11. *Let $X, Y \sim \mu$ be inputs to a r round communication protocol $\bar{\pi}$ whose internal information cost is I. Then for every $\delta_1 > 0$, there exists a protocol τ such that at the end of the protocol, each party outputs a transcript for $\bar{\pi}$. Furthermore, there is an event G with $P[G] > 1 - r\delta_1$ such that conditioned on G, the expected communication of τ is $I + O(\sqrt{rI}) + 2r\log(1/\delta_1)$, and both parties output the same transcript distributed exactly according to $\pi(X, Y)$.*

Now for $\delta_2 = 1/k^2$, consider the protocol τ' obtained by compressing π^N with $\delta_1 = \frac{1}{k^4 r(k)}$ and exchanging inputs after communicating K/δ_2 bits.

$$\mathbb{E}_{x,y\sim\mu}[|\tau'(x,y)|] \leq \frac{2}{\ln 2} \cdot K + O\left(L \cdot \log\left(\frac{2K}{L} \right) \right) + o(K) + Pr[\neg G] \cdot (K/\delta_2 + 2N) + O(\delta_2 \cdot N)$$
$$\leq \frac{2}{\ln 2} \cdot K + O\left(L \cdot \log\left(\frac{2K}{L} \right) \right) + o(K)$$

Note that $Pr[\neg G] \cdot (K/\delta_2 + 2N) = o(K)$ for all δ_2 such that $1/\delta_2 = o(k^4)$ and $O(\delta_2 \cdot N) = o(K)$ for all δ_2 such that $\delta_2 = o(1/k)$. The error of τ' w.r.t μ is $r(k) \cdot \epsilon = 1/k^4$. Note that if $L = o(K)$, then $\mathbb{E}_{x,y\sim\mu}[|\tau'(x,y)|] = \frac{2}{\ln 2} \cdot K + o(K)$. Now we apply a minimax argument similar to the arguments in [8] and the proof of Theorem 8.1 in order to produce a protocol which has low worst-case communication cost[6]. Let $\mathcal{U}_L$ be the set of distributions over $\{0,1\}^N \times \{0,1\}^N$ with expected intersection size $\leq L$.
Consider the following zero-sum game G. The first player M chooses a protocol a distribution $\mu \in \mathcal{U}_L$ and the second player T chooses a (randomized) protocol τ'. The payoff for player M is given by :

$$P_M(\mu, \tau') = (1 - 1/\sqrt{k}) \cdot \frac{\mathbb{E}_{x,y\sim\mu}[|\tau'(x,y)|]}{\frac{2}{\ln 2} \cdot K + O\left(L \cdot \log\left(\frac{2K}{L}\right)\right) + o(K)} + 1/\sqrt{k} \cdot \frac{\mathbb{E}_{x,y\sim\mu}[\text{error}(\tau'(x,y))]}{1/k^4}$$

We first establish that the value of the game is bounded away by 1.

Claim 9.12. $Val_G(M) \leq 1$

Proof. Let ν be any mixed strategy for player M. Denote by $\bar{\mu}$ the average distribution in ν : $\bar{\mu}(x,y) = \mathbb{E}_{\mu\sim\nu}\mu(x,y)$. Since the payoff function is calculated in terms of expectations over $(x,y) \sim \mu$, for any τ' we have:

$$\mathbb{E}_{\mu\sim\nu} P_M(\mu, \tau') = P_M(\bar{\mu}, \tau')$$

Since $\mathcal{U}_L$ is convex, $\bar{\mu} \in \mathcal{U}_L$. Then by the arguments above, there is a randomized protocol τ' such that $\mathbb{E}_{x,y\sim\bar{\mu}}[|\tau'(x,y)|] \leq \frac{2}{\ln 2} \cdot K + O\left(L \cdot \log\left(\frac{2K}{L}\right)\right) + o(K)$ and $\mathbb{E}_{x,y\sim\bar{\mu}}[\text{error}(\tau'(x,y))] \leq 1/k^4$. This implies that $P_M(\bar{\mu}, \tau') \leq 1$. □

By the Minimax Theorem, there is a distribution ν on protocols τ', such that for each distribution $\mu \in \mathcal{U}_L$, $\mathbb{E}_{\tau'\sim\nu} P_M(\mu, \tau') \leq 1$. This implies that the randomized protocol Π_L obtained by executing a protocol τ' that is distributed according to ν also satisfies $P_M(\mu, \Pi_L) \leq 1$ for all $\mu \in \mathcal{U}_L$. Also since the maximum communication for each τ' was $O(N \cdot k)$, the maximum communication for Π_L is $O(N \cdot k)$. $P_M(\mu, \Pi_L) \leq$ implies that for all $\mu \in \mathcal{U}_L$

[6] We need to apply minimax to infinite matrices, but we are dealing with convex sets of distributions as rows and continuous entities as matrix entries, so the justification follows along the same lines as proof of Theorem 8.1 in Appendix

$$\mathbb{E}_{x,y\sim\mu}[|\Pi_L(x,y)|] \leq \left(\frac{2}{\ln 2}\cdot K + O\left(L\cdot\log\left(\frac{2K}{L}\right)\right) + o(K)\right)\cdot(1+2/\sqrt{k})$$
$$= \frac{2}{\ln 2}\cdot K + O\left(L\cdot\log\left(\frac{2K}{L}\right)\right) + o(K)$$

Also for all $\mu \in \mathcal{U}_L$:

$$\mathbb{E}_{x,y\sim\mu}[\text{error}(\Pi_L(x,y))] \leq 1/k^{7/2}$$

Considering the singleton distribution $1_{(x,y)}$, where $|x| = |y| = K$ and $|x \wedge y| \leq L$, we get that $\mathbb{E}|\Pi_L(x,y)| \leq \frac{2}{\ln 2}\cdot K + O\left(L\cdot\log\left(\frac{2K}{L}\right)\right) + o(K) = \frac{2}{\ln 2}\cdot K + o(K)$, if $L = o(K)$ and $\text{error}(\Pi_L(x,y)) \leq 1/k^{7/2}$. This completes the proof of the lemma. □

In this lemma we get a protocol which has low expected communication but high maximum communication. However the bound on the maximum communication ensures that taking multiple copies of same problem would result in concentration and we can get a protocol which has low maximum communication. We formalize this in the lemma below :

Remark 9.13. Note that we can get a similar lemma for N, K such that $K = o(N)$ and $N/K < t(N)$ by choosing a smaller value of k and a larger value of n in the proof.

Lemma 9.14. *There exists a slowly growing function of N, $s(N)$ such that for N, K such that $\omega(1) \leq N/K \leq s(N)$, there exists a protocol Γ which solves set-intersection for x, y such that $|x| \leq K$, $|y| \leq K$ and $|x \wedge y| \leq K/\sqrt{s(N)}$ and has maximum communication $\frac{2}{\ln 2}\cdot K + o(K)$ and sub-constant error.*

Proof. Essentially we want to take $t(M)^3$ copies of $INT_M^{M/t(M)}$ for some M and use the previous lemma. Let M be the largest such that $M \cdot t(M)^3 \leq N$. We can assume $N = M \cdot t(M)^3$, since otherwise solving for $(M+1)t(M+1)^3$ doesn't cost us much more. Define $s(N) = t(M)$. We also assume wlog that $N/K = s(N)$, the case $N/K < s(N)$ corresponds to the previous lemma with $M/K_M < t(M)$. Let $k = t(M)$, so that $N = M \cdot k^3$.
Let that Alice has $x \in \{0,1\}^N$ and Bob has $y \in \{0,1\}^N$ such that $|x| = |y| = K$ and $|x \wedge y| \leq c \cdot K/\sqrt{s(N)}$. Consider the protocol in Figure 9 for finding the intersecting coordinates of x and y.

1. Alice and Bob (using public randomness) randomly divide their inputs into k^3 blocks of length M each.
2. They run the protocol from Lemma 9.11 with $K_M = M/k \cdot (1 + 5/k)$ and $L_M = M/k^{3/2} \cdot (1 + 5/k)$ on each block and terminate after exchanging $k^3 \cdot (\text{expected communication on single copy}) + k^{11/4} \cdot K_M$ bits.

Protocol 9: Protocol for big sets

Since $K = Mk^2$, the communication clearly is less than $\frac{2}{\ln 2}\cdot K + o(K)$. There are three different sources of error :

1. The error from the protocol of Lemma 9.11
2. In some block B, the number of elements of either Alice or Bob is more than K_M, or the number of intersecting elements is more than L_M.
3. The total communication is more than $k^3 \cdot (\text{expected communication on single copy}) + k^{11/4} \cdot K_M$ bits.

We show that the error from each source is sub-constant, hence completing the proof of the lemma.

1. Since the error of the protocol of Lemma 9.11 is $1/k^{7/2}$ on each copy, the total error by union-bound is $\leq 1/\sqrt{k}$.

2. Consider a particular block B. Let X_i denote the random variable that the i^{th} coordinate in the block has Alice's element. Also let $S_i = \sum_{j=1}^{i} X_j$. Then $Pr[X_i = 1|S_{i-1}] \leq K/(N-M) = k^2/(k^3-1) \leq 1/k \cdot (1+2/k^3)$. Then by Chernoff bounds, $Pr[S_M \geq M/k \cdot (1+2/k^3)(1+1/k)] \leq e^{-\frac{M/k \cdot (1+2/k^3)}{3k^2}}$, which is exponentially small since M is much larger than k. Similarly the probabilities that Bob has more than K_M elements in block B and that the number of intersecting elements in block B are more than L_M are exponentially small. The probability that there is some block in which these events happen is $k^3 \cdot$ (something exponentially small) and is again sub-constant.

3. Let T_i be the amount of communication needed for i^{th} block. Then the T_i's are independent. Also $Var[T_i] < E[T_i] \cdot \max T_i = E[T_i] \cdot O(Mk)$. Let $T = \sum_{i=1}^{k^3} T_i$. Then $E[T] = \sum_i E[T_i] = k^3 \cdot O(M/k) = O(Mk^2)$ and $Var[T] < O(M^2k^3)$. By Chebyshev's inequality, $Pr[T \geq E[T] + k^{11/4} \cdot K_M] \leq \frac{Var[T]}{k^{5.5}K_M^2} = O(1/\sqrt{k})$, which is again sub-constant.

□

Now we complete the proof of upper bound for all regimes of N, K.

Proof. (Of Theorem 9.8) The central idea of the proof is to reduce the size of the universe by hashing and then apply lemma 9.14. Note that for $N/K < s(N)$, Alice and Bob can solve $DISJ_N^K$ by first sampling enough elements and finding a common element and later solving set-intersection (if they didn't find a common element in the first step). So we assume $N/K >> s(N)$. Let R be such that $\frac{R}{K} = s(R)$. Consider the protocol in Figure 10.

1. Alice and Bob each sample $\frac{N}{\sqrt{K}}$ random coordinates (with replacement) and then they run the Håstad-Wigderson protocol (for sets of size $\leq 2\sqrt{K}$)(Theorem 9.1) in $O(\sqrt{K})$ communication. If the protocol outputs "intersecting", then output 0 else continue.

2. Alice and Bob choose a uniformly random hash function $H : [N] \to [R]$ and hash the universe into R bins. If Alice and Bob have sets X and Y, they run the protocol Π from Lemma 9.14 on $H(X)$ and $H(Y)$, but run the protocol Π only for $\frac{2}{\ln 2} \cdot K + o(K)$ bits (The performance guarantee of the Lemma for instances of small intersection size). If Π doesn't stop after communicating these many bits, output a random answer. If Π returns that the sets are disjoint, then they output 1. Else continue.

3. Now look at all the bins that have both Alice's and Bob's elements (returned by protocol Π in Step 2). Each bin can be viewed as a smaller DISJ problem. Alice and Bob run the Håstad-Wigderson protocol on each bin with probability of failing 1/2 if the bins are disjoint (The protocol has only one-sided error). They then run the protocol again on each bin that the protocol says is intersecting and keep doing this until the protocol declares all bins as disjoint, in which case they output 1, or they use communication more than $K \cdot \left(\frac{K}{R}\right)^{1/4}$ in this step, in which case they output 0.

Protocol 10: Protocol for K-DISJ

Lets first see why the protocol is correct with high probability.

1. In the first step, the expected sizes of Alice's and Bob's remaining sets is $\sqrt{K}$. Thus, except with exponentially small probability, both Alice and Bob have sets of size $\leq 2\sqrt{K}$ (by Chernoff bounds). If $|X \cap Y| > K^{3/4}$ then again by Chernoff bounds, except with exponentially small probability, Alice's and Bob's remaining sets intersect. Thus the parties output 0 in the first step except with an exponentially small probability (HW protocol has error only when the sets are disjoint) if $|X \cap Y| > K^{3/4}$.

2. For the further steps, lets analyze $|H(X) \cap H(Y)|$. Let X_{ij} be the indicator random variable for the event that Alice's i^{th} element a_i and Bob's j^{th} element b_j are mapped to the same bin. Then

if $a_i = b_j$, $Pr[X_{ij} = 1] = 1$ and if $a_i \neq b_j$, $Pr[X_{ij} = 1] = \frac{1}{R}$. Also let $X = \sum_{i,j=1}^{K} X_{ij}$. Then by linearity of expectation, $E[X] \leq |X \cap Y| + K^2/R \leq K^{3/4} + K^2/R \leq 2K^2/R$ ($s(R)$ is a slowly growing function). Let $B_1, \ldots, B_t$ be the bins that are common between Alice and Bob. Also let l_i be the number of elements of Alice or Bob in bin B_i (whoever has larger number of elements in bin B_i). Then the contribution to number of collisions from bin B_i is atleast l_i. Thus, except with probability $O(\sqrt{K/R})$, $\sum_i l_i \leq \frac{K\sqrt{K}}{\sqrt{R}}$. In particular the number of intersecting bins are atmost $\frac{K\sqrt{K}}{\sqrt{R}}$ and thus by Lemma 9.14, uses communication atmost $\frac{2}{\ln 2} \cdot K + o(K)$. Now since the error of protocol Π is sub-constant, the error in second step is sub-constant.

3. In the third step, we can have an error only when all the bins are disjoint (because of the one-sided error of HW protocol). In that case, the expected amount of communication required (until the HW protocol outputs disjoint on each bin) is $\leq O(\sum_i l_i) + \frac{1}{2} \cdot O(\sum_i l_i) + \left(\frac{1}{2}\right)^2 \cdot O(\sum_i l_i) + \ldots = O(\sum_i l_i) = O(\frac{K\sqrt{K}}{\sqrt{R}})$. This is because probability that the HW protocol says "intersecting" on a disjoint bin for i steps is $1/2^i$. Thus only with probability $O\left(\left(\frac{K}{R}\right)^{1/4}\right)$, communication more than $K \cdot \left(\frac{K}{R}\right)^{1/4}$ is required.

It remains to analyze the protocol's communication complexity. Step 1 has communication complexity $O(\sqrt{K}) = o(K)$. Step 3 has communication atmost $K \cdot \left(\frac{K}{R}\right)^{1/4} = o(K)$. The bulk of the communication occurs in step 2, where by Lemma 9.14 (and because of the check) it is bounded by $\frac{2}{\ln 2} \cdot K + o(K)$ □

Remark 9.15. By similar techniques, we can prove that the randomized communication complexity of SET-INTERSECTION for sparse sets (sets of size $\leq$ k) is essentially $\left(\frac{2}{\ln 2} \cdot \ln(1+e)\right) \cdot k \pm o(k) \approx 3.7893 \cdot k \pm o(k)$. This is because $\max_{\mu \in \mathcal{C}_{k,n}} IC_\mu(AND, 0) = \frac{2}{\ln 2} \cdot \ln(1+e) \cdot \frac{k}{n} + o\left(\frac{k}{n}\right)$, where $\mathcal{C}_{k,n}$ is the set of distributions μ such that $\mu = \begin{array}{|c|c|}\hline \alpha & \beta \\ \hline \beta & \delta \\ \hline \end{array}$ and $\beta + \delta = \frac{k}{n}$.

Appendix

Here we give all the pending proofs.

A Concavity of Information cost

We need the following lemma about concavity of information cost from [8]

Lemma A.1. *Let ν be a distribution on probability distributions μ over $X \times Y$, and denote $\bar{\mu}(x,y) := \mathbf{E}_{\mu \sim \nu} \mu(\mathrm{x}, \mathrm{y})$. Then for any protocol π it holds that*

$$\mathbf{E}_{\mu \sim \nu_{\mathrm{B}}} \left[\mathsf{IC}_\mu(\pi)\right] \leq \mathsf{IC}_{\bar{\mu}}(\pi).$$

In other words, the average amount of information revealed by π with respect to the different distributions $\mu \sim \nu$ is smaller or equal to the amount of information revealed with respect to $\bar{\mu}$.

To establish the statement of the theorem, consider the following four random variables. Let M be a random variable representing the distribution μ. Then M is distributed according to ν. Let X and Y be the inputs to the two parties in π such that (X,Y) is distributed according to μ. Finally, let $\Pi = \pi(X,Y)$ be the transcript of the protocol executed on X and Y. Π is randomized even conditioned on (X,Y) due to the public and private randomness used in the execution of the protocol. In this language, we have:

$$\mathbf{E}_{\mu \sim \nu} \left[\mathrm{I}_{(\mathrm{X},\mathrm{Y}) \sim \mu}(\pi(\mathrm{X},\mathrm{Y}); \mathrm{X} | \mathrm{Y})\right] = \mathrm{I}(\Pi; \mathrm{X} | \mathrm{YM}),$$

and

$$I_{(X,Y) \sim \bar{\mu}}(\pi(X,Y); X|Y) = I(\Pi; X|Y).$$

Since the distribution of Π only depends on X and Y, we have $I(\Pi; M|XY) = 0$. By substituting $A = X$, $B = \Pi$, $C = Y$, and $D = M$ into Proposition 4.7 we get

$$I(X; \Pi|Y) \geq I(X; \Pi|YM), \tag{11}$$

which proves that

$$\mathbf{E}_{\mu\sim\nu}\left[\mathrm{I}_{(\mathrm{X},\mathrm{Y})\sim\mu}(\pi(\mathrm{X},\mathrm{Y});\mathrm{X}|\mathrm{Y})\right] \leq \mathrm{I}_{(\mathrm{X},\mathrm{Y})\sim\bar{\mu}}(\pi(\mathrm{X},\mathrm{Y});\mathrm{X}|\mathrm{Y}).$$

Similarly, the following symmetric inequality is established: $\mathbf{E}_{\mu\sim\nu}\left[\mathrm{I}_{(\mathrm{X},\mathrm{Y})\sim\mu}(\pi(\mathrm{X},\mathrm{Y});\mathrm{Y}|\mathrm{X})\right] \leq \mathrm{I}_{(\mathrm{X},\mathrm{Y})\sim\bar{\mu}}(\pi(\mathrm{X},\mathrm{Y});\mathrm{Y}|\mathrm{X})$. Together, the last two inequalities imply

$$\mathbf{E}_{\mu\sim\nu}\left[\mathsf{IC}_{\mu}(\pi)\right] \leq \mathsf{IC}_{\bar{\mu}}(\pi). \tag{12}$$

B Proof of Theorem 9.1

Here we provide the proof of Theorem 8.1 (adapted from [8]). Actually we can replace $\mathcal{U}_0$ by any convex and compact subset of distributions. $\inf_{\pi \text{ good for } f} \max_{\mu\in\mathcal{U}_0} IC(\pi,\mu) \geq \max_{\mu\in\mathcal{U}_0} \inf_{\pi \text{ good for } f} IC(\pi,\mu)$ is easy to see. We prove the other inequality below. The proof is essentially a minimax argument.

We will need the following lemma (from another working paper by the same authors) :

Lemma B.1. *Let μ_1 and μ_2 be distributions on $\{0,1\}^N \times \{0,1\}^N$ such that $|\mu_1 - \mu_2| \leq \epsilon$. Also let $\epsilon < 1/2$. Let π be a protocol for solving a function (possibly partial) with domain $\{0,1\}^N \times \{0,1\}^N$. Then $|IC(\pi,\mu_1) - IC(\pi,\mu_2)| \leq 4N\epsilon + 2H(2\epsilon)$.*

Proof. We will design random variables X, Y, E such that $X, Y \in \{0,1\}^N$ and $E \in \{0,1,2\}$, $X, Y|E \in \{0,1\} \sim \mu_1$, $X, Y|E \in \{0,2\} \sim \mu_2$ and $Pr[E = 1] = Pr[E = 2] \leq \epsilon$. First let us see how this helps. Let Π denote the random variable for the transcript of the protocol when the inputs are X, Y. Let $X_1Y_1 \sim \mu_1$ and $X_2Y_2 \sim \mu_2$. Also let Π_1 and Π_2 denote the random variables for the transcript in these cases respectively.

$$\begin{aligned} I(\Pi; X|YE) &= Pr[E=0]\cdot I(\Pi; X|Y, E=0) + Pr[E=1]\cdot I(\Pi; X|Y, E=1) + Pr[E=2]\cdot I(\Pi; X|Y, E=2) \\ &= Pr[E \in \{0,1\}]\cdot I(\Pi; X|Y, E_{\{0,1\}}) + Pr[E=2]\cdot I(\Pi; X|Y, E=2) \end{aligned}$$

Here conditioning on $E_{\{0,1\}}$ means that $E \in \{0,1\}$ and that both Alice and Bob know the value of E. Now

$$I(\Pi; X|Y, E \in \{0,1\}) \leq I(\Pi; X|Y, E_{\{0,1\}}) + H(E|E \in \{0,1\}) = I(\Pi; X|Y, E_{\{0,1\}}) + C_1$$

, where $C_1 \leq H(\epsilon/(1-\epsilon)) \leq H(2\epsilon)$. Also $I(\Pi; X|Y, E=2) \leq N$ and $I(\Pi; X|Y, E \in \{0,1\}) = I(\Pi_1; X_1|Y_1)$. Thus

$$I(\Pi; X|YE) = (1 - Pr[E=2])\cdot(I(\Pi_1; X_1|Y_1) + C_1) + Pr[E=2]\cdot C_2$$

where $C_1 \leq 1$ and $C_2 \leq N$. Similarly

$$I(\Pi; X|YE) = (1 - Pr[E=1])\cdot(I(\Pi_2; X_2|Y_2) + C_3) + Pr[E=1]\cdot C_4$$

where $C_3 \leq H(2\epsilon)$ and $C_4 \leq N$. Equating the two we get that

$$(1 - Pr[E=1])\cdot(I(\Pi_1; X_1|Y_1) - I(\Pi_2; X_2|Y_2)) = Pr[E=1]\cdot(C_4 - C_3) + (1 - Pr[E=1])\cdot(C_2 - C1)$$

Since $Pr[E=1] \leq \epsilon \leq 1/2$, we get that

$$|I(\Pi_1; X_1|Y_1) - I(\Pi_2; X_2|Y_2)| \leq 2N\epsilon + H(2\epsilon)$$

and hence $|IC(\pi,\mu_1) - IC(\pi,\mu_2)| \leq 4N\epsilon + 2H(2\epsilon)$.

Now let us see how to design random variables X, Y, E satisfying the given conditions. Let U, V, P denote the random variables obtained by sampling uniformly from $\{0,1\}^N \times \{0,1\}^N \times [0,1]$. Let G denote the event that $P < \max(\mu_1(U,V), \mu_2(U,V))$. Let $X, Y = U, V|G$. Also define a random variable $F \in \{0,1,2\}$ as follows :

- $F = 0$, if $P < \min(\mu_1(U,V), \mu_2(U,V))$
- $F = 1$, if $\mu_2(U,V) \leq P < \mu_1(U,V)$
- $F = 2$, if $\mu_1(U,V) \leq P < \mu_2(U,V)$

Now define $E = F|G$. Let us verify that X, Y, E satisfy the conditions.

$$Pr[X = x, Y = y | E \in \{0,1\}] = \frac{Pr[U = x, V = y, F \in \{0,1\}, G]}{Pr[F \in \{0,1\}, G]} = \frac{\frac{1}{2^{2N}}\mu_1(x,y)}{\sum_{x,y}\frac{1}{2^{2N}}\mu_1(x,y)} = \mu_1(x,y)$$

Thus $X, Y | E \in \{0,1\} \sim \mu_1$. Similarly $X, Y | E \in \{0,2\} \sim \mu_2$. Also

$$\begin{aligned} Pr[E = 1] = Pr[F = 1|G] &= \sum_{x,y} Pr[U = x, V = y|G] Pr[F = 1 | G, U = x, V = y] \\ &= \sum_{x,y \text{ s.t. } \mu_1(x,y) > \mu_2(x,y)} \frac{\frac{1}{2^{2N}} \max(\mu_1(x,y), \mu_2(x,y))}{\frac{1}{2^{2N}} \sum_{x,y} \max(\mu_1(x,y), \mu_2(x,y))} \cdot \frac{\mu_1(x,y) - \mu_2(x,y)}{\max(\mu_1(x,y), \mu_2(x,y))} \\ &= \frac{\sum_{x,y \text{ s.t. } \mu_1(x,y) > \mu_2(x,y)} (\mu_1(x,y) - \mu_2(x,y))}{\sum_{x,y} \max(\mu_1(x,y), \mu_2(x,y))} \end{aligned}$$

Thus $Pr[E = 1] = \frac{|\mu_1 - \mu_2|}{\sum_{x,y} \max(\mu_1(x,y), \mu_2(x,y))} \leq |\mu_1 - \mu_2| \leq \epsilon$. Similarly $Pr[E = 2] = \frac{|\mu_1 - \mu_2|}{\sum_{x,y} \max(\mu_1(x,y), \mu_2(x,y))}$. Hence $Pr[E = 1] = Pr[E = 2] \leq \epsilon$. This completes the proof.

□

Note that this proves that $IC(\pi, \mu)$ is continuous as a function of μ for all protocols π.

Proof. Of Theorem 8.1 Since we are dealing with protocols and distributions (which are infinite sets), we also provide justification for why minimax applies here. Let G be the set of protocols that are good for f. Note that G is an infinite set. We first prove the following lemma.

Lemma B.2. *Let H be any finite subset of G. Then for any $\alpha \geq \max_{\mu \in \mathcal{U}_0} \min_{\pi \in H} IC(\pi, \mu)$, there exists a protocol $\tau \in G$ such that $IC(\tau, \mu) \leq \alpha$, $\forall \mu \in \mathcal{U}_0$.*

Note that $\mathcal{U}_0$ is the set of distributions supported on $f^{-1}(0)$.

We define the following zero-sum two player game $\mathcal{G}_0$. Player A will come up with a (randomized) two-party protocol $\pi \in H$. Player B will come up with a distribution $\mu \in \mathcal{U}_0$. Player B's payoff is given by:

$$P_B(\pi, \mu) := IC(\pi, \mu).$$

We first prove that the value of the game for player B is bounded by α.

Claim B.3. *The value $V_B(\mathcal{G}_0) \leq \alpha$.*

Proof. Let ν_B be a probability distribution representing a mixed strategy for player B. Thus ν_B is a distribution on probability distributions μ over $\mathcal{X} \times \mathcal{Y}$. We need to show that there is a zero-error protocol $\tau \in H$ such that $\mathbf{E}_{\mu \sim \nu_\mathrm{B}}[\mathrm{IC}(\tau, \mu)] \leq \alpha$. Let $\bar{\mu}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$ that is obtained by taking the average of $\mu \sim \nu_B$. Formally,

$$\bar{\mu}(x,y) := \mathbf{E}_{\mu \sim \nu_\mathrm{B}} \mu(\mathrm{x}, \mathrm{y}).$$

Note that since all distributions $\mu \in \mathcal{U}_0$, hence $\bar{\mu} \in \mathcal{U}_0$. Since $\alpha \geq \max_{\mu \in \mathcal{U}_0} \min_{\pi \in H} IC(\pi, \mu)$, there is a protocol $\tau \in H$ such that $IC(\tau, \bar{\mu}) \leq \alpha$. Now by Lemma A.1, it holds that

$$\mathbf{E}_{\mu \sim \nu_\mathrm{B}} [\mathrm{IC}(\tau, \mu)] \leq \mathrm{IC}(\tau, \bar{\mu}).$$

Thus the value of the game is bounded by α completing the proof of Claim B.3. □

The minimax theorem holds for our game by an ϵ-net argument and continuity of $IC(\pi, \mu)$. Applying the minimax theorem, we get that there is a mixed strategy for player A such that for each response by player B, the value of the game for player B is at most α. A mixed strategy for player A is a distribution ν_A on protocols. In other words,

$$\mathbf{E}_{\pi\sim\nu_{\mathrm{A}}}\mathrm{P_B}(\pi,\mu) \leq \alpha, \text{ for all } \mu. \tag{13}$$

Let $\bar{\pi}$ be the randomized protocol obtained by publicly sampling $\pi \sim \nu_A$, and then applying π to the inputs. We claim that $\bar{\pi}$ is the protocol we are looking for. In other words, the randomized protocol $\bar{\pi}$ has the desired payoff properties. Clearly $\bar{\pi} \in G$.

Claim B.4. *For each distribution μ, $IC(\bar{\pi}, \mu) \leq \alpha$.*

Proof. The proof proceeds similarly to the proof of Lemma A.1. We will prove that

$$I_{(X,Y)\sim\mu}(\bar{\pi}(X,Y);X|Y) \leq \mathbf{E}_{\pi\sim\nu_{\mathrm{A}}}\left[\mathrm{I}_{(\mathrm{X},\mathrm{Y})\sim\mu}(\pi(\mathrm{X},\mathrm{Y});\mathrm{X}|\mathrm{Y})\right]. \tag{14}$$

In other words, the amount of information revealed by $\bar{\pi}$ is bounded by the average amount of information revealed by π that is drawn according to ν_A.

To establish (14), consider the following four random variables. Let S be a "selector" random variable, that picks the protocol π to run according to the distribution ν_A. Let X and Y be inputs distributed according to μ independently of S. Finally, let $\Pi = \pi(X, Y)$ be the transcript of the selected protocol executed on X and Y. We have:

$$\mathbf{E}_{\pi\sim\nu_{\mathrm{A}}}\left[\mathrm{I}_{(\mathrm{X},\mathrm{Y})\sim\mu}(\pi(\mathrm{X},\mathrm{Y});\mathrm{X}|\mathrm{Y})\right] = \mathrm{I}(\Pi;\mathrm{X}|\mathrm{YS}),$$

and

$$I_{(X,Y)\sim\mu}(\bar{\pi}(X,Y);X|Y) = I(\Pi;X|Y).$$

Since the protocol π is selected independently of the inputs, we have $I(X; S|\Pi Y) = 0$. By substituting $A = \Pi$, $B = X$, $C = Y$, and $D = S$ into Proposition 4.7 we get

$$I(\Pi;X|Y) \leq I(\Pi;X|YS), \tag{15}$$

establishing (14). Similarly to (14) the following symmetric inequality is established:

$$I_{(X,Y)\sim\mu}(\bar{\pi}(X,Y);Y|X) \leq \mathbf{E}_{\pi\sim\nu_{\mathrm{A}}}\left[\mathrm{I}_{(\mathrm{X},\mathrm{Y})\sim\mu}(\pi(\mathrm{X},\mathrm{Y});\mathrm{Y}|\mathrm{X})\right]. \tag{16}$$

Together, (14) and (16) imply

$$IC(\bar{\pi},\mu) \leq \mathbf{E}_{\pi\sim\nu_{\mathrm{A}}}\left[\mathrm{IC}(\pi,\mu)\right]. \tag{17}$$

□

Now we use a compactness argument (adapted from [42]) to complete the proof. Choose any $\alpha > \max_{\mu\in\mathcal{U}_0} \inf_{\pi\in G} IC(\pi, \mu)$. Define

$$A(\pi) := \{\mu \in \mathcal{U}_0 : IC(\pi,\mu) \geq \alpha\}$$

Then $\cap_{\pi\in G} A(\pi) = \phi$. Since $\mathcal{U}_0$ is compact and the sets $A(\pi)$ are closed because of the continuity of $IC(\pi, \mu)$, we get that there is a finite set of protocols $H \subset G$ such that $\cap_{\pi\in H} A(\pi) = \phi$. Thus we have that $\min_{\pi\in H} IC(\pi, \mu) < \alpha$, $\forall \mu \in \mathcal{U}_0$. Then by Lemma $B.2$, there exists a protocol $\tau \in G$ such that $IC(\tau, \mu) \leq \alpha$, $\forall \mu \in \mathcal{U}_0$. Thus

$$\inf_{\pi\in G} \max_{\mu\in\mathcal{U}_0} IC(\pi,\mu) \leq \max_{\mu\in\mathcal{U}_0} \inf_{\pi\in G} IC(\pi,\mu)$$

which completes the proof. □

References

[1] F. Ablayev. Lower bounds for one-way probabilistic communication complexity and their application to space complexity. *Theoretical Computer Science*, 157(2):139–159, 1996.

[2] R. Ahlswede and N. Cai. On communication complexity of vector-valued functions. *Information Theory, IEEE Transactions on*, 40(6):2062–2067, 1994.

[3] Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68(4):702–732, 2004.

[4] B. Barak, M. Braverman, X. Chen, and A. Rao. How to compress interactive communication. In *Proceedings of the 42nd ACM symposium on Theory of computing*, STOC '10, pages 67–76, New York, NY, USA, 2010. ACM.

[5] R. Beigel and J. Tarui. On ACC [circuit complexity]. In *Foundations of Computer Science, 1991. Proceedings., 32nd Annual Symposium on*, pages 783–792. IEEE, 1991.

[6] M. Ben-Or, S. Goldwasser, J. Kilian, and A. Wigderson. Multi-prover interactive proofs: How to remove intractability assumptions. In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing*, pages 113–131, 1988.

[7] M. Braverman. Coding for interactive computation: progress and challenges. In *50th Annual Allerton Conference on Communication, Control, and Computing.*, 2012. to appear, available at `http://www.cs.princeton.edu/~mbraverm/`.

[8] M. Braverman. Interactive information complexity. In *Proceedings of the 44th symposium on Theory of Computing*, STOC '12, pages 505–524, New York, NY, USA, 2012. ACM.

[9] M. Braverman and A. Moitra. An information complexity approach to extended formulations. *ECCC*, 2012.

[10] M. Braverman and A. Rao. Information equals amortized communication. In *FOCS*, pages 748–757, 2011.

[11] J. Brody, A. Chakrabarti, and R. Kondapally. Certifying equality with limited interaction. *Electronic Colloquium on Computational Complexity (ECCC)*, 2012.

[12] A. Chakrabarti, S. Khot, and X. Sun. Near-optimal lower bounds on the multi-party communication complexity of set disjointness. In *Computational Complexity, 2003. Proceedings. 18th IEEE Annual Conference on*, pages 107–117. IEEE, 2003.

[13] A. Chakrabarti and O. Regev. An optimal lower bound on the communication complexity of gap-hamming-distance. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, pages 51–60. ACM, 2011.

[14] A. Chakrabarti, Y. Shi, A. Wirth, and A. Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In B. Werner, editor, *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*, pages 270–278, Los Alamitos, CA, Oct. 14–17 2001. IEEE Computer Society.

[15] A. Chattopadhyay and A. Ada. Multiparty communication complexity of disjointness. *arXiv preprint arXiv:0801.3624*, 2008.

[16] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley series in telecommunications. J. Wiley and Sons, New York, 1991.

[17] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.

[18] T. Feder, E. Kushilevitz, M. Naor, and N. Nisan. Amortized communication complexity. *SIAM Journal on Computing*, 24(4):736–750, 1995. Prelim version by Feder, Kushilevitz, Naor FOCS 1991.

[19] P. Harsha, R. Jain, D. A. McAllester, and J. Radhakrishnan. The communication complexity of correlation. In *IEEE Conference on Computational Complexity*, pages 10–23. IEEE Computer Society, 2007.

[20] J. Håstad and A. Wigderson. The randomized communication complexity of set disjointness. *Theory Of Computing*, 3:211–219, 2007.

[21] D. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

[22] P. Ishwar and N. Ma. Personal communication.

[23] T. Jayram. Hellinger strikes back: A note on the multi-party information complexity of and. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 562–573, 2009.

[24] T. Jayram, S. Kopparty, and P. Raghavendra. On the communication complexity of read-once ac^0 formulae. In *Computational Complexity, 2009. CCC'09. 24th Annual IEEE Conference on*, pages 329–340. IEEE, 2009.

[25] B. Kalyanasundaram and G. Schnitger. The probabilistic communication complexity of set intersection. *SIAM Journal on Discrete Mathematics*, 5(4):545–557, Nov. 1992.

[26] M. Karchmer and A. Wigderson. Monotone circuits for connectivity require super-logarithmic depth. *SIAM Journal on Discrete Mathematics*, 3(2):255–265, 1990.

[27] E. Kushilevitz and N. Nisan. *Communication complexity*. Cambridge University Press, Cambridge, 1997.

[28] N. Leonardos and M. Saks. Lower bounds on the randomized communication complexity of read-once functions. *Computational Complexity*, 19(2):153–181, 2010.

[29] N. Ma and P. Ishwar. Two-terminal distributed source coding with alternating messages for function computation. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pages 51–55. IEEE, 2008.

[30] N. Ma and P. Ishwar. Infinite-message distributed source coding for two-terminal interactive computing. In *Proc. of the 47th annual Allerton Conf. on Comm., Control, and Comp.*, Allerton'09, pages 1510–1517, Piscataway, NJ, USA, 2009. IEEE Press.

[31] N. Ma and P. Ishwar. Some results on distributed source coding for interactive function computation. *Information Theory, IEEE Transactions on*, 57(9):6180–6195, 2011.

[32] N. Nisan and A. Wigderson. Rounds in communication complexity revisited. *SIAM Journal on Computing*, 22(1):211–219, 1993.

[33] A. Orlitsky. Worst-case interactive communication. i. two messages are almost optimal. *Information Theory, IEEE Transactions on*, 36(5):1111–1126, 1990.

[34] A. Orlitsky. Worst-case interactive communication. ii. two messages are not optimal. *Information Theory, IEEE Transactions on*, 37(4):995–1005, 1991.

[35] Razborov. On the distributed complexity of disjointness. *TCS: Theoretical Computer Science*, 106, 1992.

[36] A. Razborov. Quantum communication complexity of symmetric predicates. *Izvestiya: Mathematics*, 67:145–159, 2003.

[37] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27, 1948. Monograph B-1598.

[38] A. Sherstov. The pattern matrix method (journal version). *arXiv preprint arXiv:0906.4291*, 2009.

[39] A. Sherstov. The communication complexity of gap hamming distance. *Theory of Computing*, 8:197–208, 2012.

[40] A. Sherstov. The multiparty communication complexity of set disjointness. In *Proceedings of the 44th symposium on Theory of Computing*, pages 525–548. ACM, 2012.

[41] D. Slepian and J. K. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 19(4):471–480, July 1973.

[42] F. Terkelsen. Some minimax theorems. *Mathematica Scandinavica*, 31:405–413, 1972.

[43] A. Yao. Some complexity questions related to distributive computing (preliminary report). In *Proceedings of the eleventh annual ACM symposium on Theory of computing*, pages 209–213. ACM, 1979.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de