Strong XOR Lemma for Communication with Bounded Rounds

(extended abstract)

Huacheng Yu
Department of Computer Science
Princeton University
Princeton, NJ, USA
yuhch123@gmail.com

Abstract—In this paper, we prove a strong XOR lemma for bounded-round two-player randomized communication. For a function \( f : \mathcal{X} \times \mathcal{Y} \to \{0,1\} \), the \( n \)-fold XOR function \( f^\oplus_n : \mathcal{X}^n \times \mathcal{Y}^n \to \{0,1\} \) maps \( n \) input pairs \( (X_1, \ldots, X_n, Y_1, \ldots, Y_n) \) to the XOR of the \( n \) output bits \( f(X_1, Y_1) \oplus \cdots \oplus f(X_n, Y_n) \). We prove that if every \( r \)-round communication protocol that computes \( f \) with probability \( 2/3 \) uses at least \( C \) bits of communication, then any \( r \)-round protocol that computes \( f^\oplus_n \) with probability \( 1/2 + \exp(-O(n)) \) must use \( n \cdot (r - O(r) \cdot C - 1) \) bits. When \( r \) is a constant and \( C \) is sufficiently large, this is \( \Omega(nc) \) bits. It matches the communication cost and the success probability of the trivial protocol that computes the \( n \) bits \( f(X_i, Y_i) \) independently and outputs their XOR, up to a constant factor in \( n \).

A similar XOR lemma has been proved for \( f \) whose communication lower bound can be obtained via bounding the discrepancy [17]. By the equivalence between the discrepancy and the correlation with 2-bit communication protocols [19], our new XOR lemma implies the previous result.

Index Terms—communication complexity, xor lemma

I. INTRODUCTION

In computational complexity, XOR lemmas study the relation between the complexity of a \( \{0,1\} \)-valued function \( f(x) \) and the complexity of the \( n \)-fold XOR function \( f^\oplus_n \) where

\[
f^\oplus_n(x_1, \ldots, x_n) = f(x_1) \oplus \cdots \oplus f(x_n)
\]

and \( \oplus \) is the XOR. A classic example is Yao’s XOR lemma for circuits [20], which states if \( f \) cannot be computed with probability \( 2/3 \) on a random input by size-\( s \) circuits, then \( f^\oplus_n \) cannot be computed with probability \( 1/2 + \exp(-\Omega(n)) \) on a random input by size-\( s' \) circuits for some \( s' < s \) (and small \( n \)). Such lemmas can be used to create very hard functions in a blackbox way, which can only be computed barely better than random guessing, from functions that are “just” hard to compute with constant probability. This approach of hardness amplification has been used in one-way functions [20], [16], pseudorandom generators [13], [14], and more recently, streaming lower bounds [1], [10].

In general, suppose computing a function \( f \) with probability \( 2/3 \) requires resource \( s \) in some model of computation (e.g., circuit size, running time, query complexity, communication cost, etc). Then the trivial way to compute \( f^\oplus_n \) is to compute each \( f(x_i) \) using resource \( s \) independently, and output their XOR. It uses resource \( n \cdot s \) in total, and each instance is correct with probability \( 2/3 \), hence, their XOR is correct with probability \( 1/2 + \exp(-\Theta(n)) \): For two independent random bits \( b_1, b_2 \), if \( \Pr[b_1 = 0] = 1/2 + \alpha_1/2 \) and \( \Pr[b_2 = 0] = 1/2 + \alpha_2/2 \), then

\[
\Pr[b_1 \oplus b_2 = 0] = (1/2 + \alpha_1/2)(1/2 + \alpha_2/2) + (1/2 - \alpha_1/2)(1/2 - \alpha_2/2)
= 1/2 + \alpha_1 \alpha_2/2;
\]

let \( b_i = 0 \) if and only if \( f(x_i) \) is computed correctly, applying the above calculation inductively gives the claimed probability. A strong XOR lemma asserts that to achieve \( 1/2 + \exp(-\Omega(n)) \) success probability, one must use \( \Omega(ns) \) resource — the trivial solution is essentially optimal.

Now suppose that we are given \( n \cdot s/2 \) resource in total, and we want to compute \( f^\oplus_n \). If we try to solve the \( n \) copies independently, then no matter how we distribute the resource among the \( n \) copies, at least half of them will get no more than \( s \). The function \( f^\oplus_n \) is computed correctly with probability at most \( 1/2 + \exp(-\Omega(n)) \). Of course, since all \( n \) inputs \( (x_1, \ldots, x_n) \) are given together, we can potentially process them jointly. This may correlate the \( n \) copies, and in particular, it may correlate the correctness of computing each \( f(x_i) \). Hence, one difficulty in proving the strong XOR lemma from the technical point of view is that in the above calculation of the probability of XOR of two independent bits, the linear terms perfectly cancel only because \( b_1 \) and \( b_2 \) are independent; when they are not independent, we may get a linear term remaining, and do not reduce the probability bias as desired. In computational models where one cannot expect the independence between the copies throughout the computation, a success probability lower bound of \( 1/2 + \exp(-\Omega(n)) \) (hence, a strong XOR lemma) is generally difficult to prove.

In this paper, we prove a strong XOR lemma for the two-player randomized communication complexity with bounded rounds: Alice and Bob receive \( X \) and \( Y \) respectively, they
alternatively send a total of \( r \) messages to each other with the goal of computing \( f(X,Y) \). For \( f^{\oplus n} \), Alice receives \((X_1,\ldots,X_n)\) and Bob receives \((Y_1,\ldots,Y_n)\), and they wish to compute \( f(X_1,Y_1) \oplus \cdots \oplus f(X_n,Y_n) \) after \( r \) rounds of communication. Each player has half of the inputs for all copies, and can send messages that arbitrarily depend on them, which can nontrivially correlate the \( n \) instances. Nevertheless, we show that one cannot do much better than simply solving all \( n \) copies in parallel.

Let \( R_x^{(r)}(f) \) be the minimum number of bits of communication needed in \( r \) messages in order to compute \( f(X,Y) \) correctly with probability \( p \). We prove the following theorem.

**Theorem 1.** For any \( \{0,1\} \)-valued function \( f \), we have

\[
R_x^{1/2+2^{-n}}(f^{\oplus n}) \geq n \cdot \left( r^{-O(r)} \cdot R_x^{(r)}(f) - 1 \right).
\]

In particular, when \( r \) is a constant, it implies that \( R_x^{1/2+2^{-n}}(f^{\oplus n}) = \Omega \left( n \cdot \left( R_x^{(r)}(f) - O(1) \right) \right) \).

To the best of our knowledge, such an XOR lemma was not known even for one-way communication and without the factor of \( n \).

As pointed in [3], the “\( -O(1) \)” term is needed. This is because for \( f(X,Y) = X \oplus Y \), we have \( R_x^{(r)}(f) = 2 \). On the other hand, \( f^{\oplus n} \) can also be computed with 2 bits of communication by simply (locally) computing \( \bigoplus_{i=1}^n X_i \) and \( \bigoplus_{i=1}^n Y_i \) and exchanging the values.

We obtain Theorem 1 via the following distributional strong XOR lemma. Let \( \text{suc}_\mu(f;C,A,C_B,r) \) be the maximum success probability of an \( r \)-round protocol \( \pi \) computing \( f(X,Y) \) where:

- Alice sends at most \( C_A \) bits in every odd round,
- Bob sends at most \( C_B \) bits in every even round, and
- \((X,Y)\) is sampled from \( \mu \).

**Theorem 2.** Let \( c > 0 \) be a sufficiently large constant. Fix \( \alpha \in (0, r^{-\omega(r)}) \) and \( C_A, C_B \geq 2c \log(r/\alpha) \). Let \( f : X \times Y \rightarrow \{0,1\} \) be a function, and \( \mu \) be a distribution over \( X \times Y \). Suppose \( f \) satisfies

\[
\text{suc}_\mu(f;C,A,C_B,r) < 1/2 + \alpha/2,
\]

then for any integer \( n \geq 2 \), we have

\[
\text{suc}_\mu(f^{\oplus n};2^{-8}r^{-1}n \cdot C_A,2^{-8}r^{-1}n \cdot C_B,r) \leq \frac{1}{2} + \frac{2^{-12}n^2}{2}.
\]

This distributional strong XOR lemma states that for any fixed input distribution \( \mu \) and function \( f \), to compute \( f^{\oplus n} \) when the \( n \) inputs are sampled independently from \( \mu \), either the advantage is exponentially small in \( \Omega(n) \), or one of the players need to communicate at least \( \Omega(n/r) \) times more than one copy. This also gives a strong XOR lemma in the asymmetric communication, where we separately count how many bits Alice and Bob send.

It is worth noting that Shaltiel [17] proved a similar strong XOR lemma for functions whose communication lower bound can be obtained via bounding the discrepancy. By the equivalence between the discrepancy and the correlation with 2-bit protocols [19], Theorem 2 implies their result. See the full version for a more detailed argument.

Note that a simple argument shows that Theorem 2 implies Theorem 1. Therefore, we will focus on the distributional version, and assume that the \( n \) input pairs are sampled independently from some distribution \( \mu \).

Our proof of the distributional version is inspired by the information complexity [9]. We define a new complexity measure for protocols, the \( \chi^2 \)-cost, which is related to the internal information cost [2], [3]. Roughly speaking, it replaces the KL-divergence in the internal information cost with the \( \chi^2 \)-divergence, which can be viewed as the “exponential” version of KL. This provides better concentration, which is needed in our argument. Throughout the proof, we will also work with distributions that are “close” to communication protocols, i.e., the speaker’s message may slightly depend on the receiver’s input. Such distributions have also been studied in the proof of direct product theorems [15], [5], [6].

We provide an overview in Section II, and the detailed proof can be found in the full version.

A. Related work

As we mentioned earlier, Shaltiel [17] proved a strong XOR lemma for functions whose communication lower bound can be obtained via bounding the discrepancy. Sherstov [18] extended this bound to generalized discrepancy and quantum communication complexity.

Barak, Braverman, Chen and Rao [3] obtained an XOR lemma for the information complexity and then an XOR lemma for communication (with worst parameters) via information compression. However, their XOR lemma does not give exponentially small advantage. They proved that if \( f \) is hard to compute with information cost \( C \), then \( f^{\oplus n} \) is hard to compute with information cost \( O(n \cdot C) \). In fact, the starting point of our proof is an alternative view of their argument, which we will outline in Section II-A.

Viola and Wigderson [19] proved a strong XOR lemma for multi-player \( c \)-bit communication for small \( c \). As pointed out in their paper, it implies the XOR lemma by Shaltiel [17]. XOR lemmas have also been proved in circuit complexity [20], [16], [13], [14], [12], query complexity [17], [18], [11], [8], streaming [1] and for low degree polynomials [19].

Direct product and direct sum theorems, which are results of similar types, have also been studied in the literature. They ask to return the outputs of all \( n \) copies instead of their XOR. Direct sum theorems state that the problem cannot be solved with the same probability unless \( \Omega(n) \) times more resource is used, while direct product theorems state that the problem can only be solved with probability exponentially small in \( \Omega(n) \) times more resource is used. The direct sum theorem for information complexity is known [9], [2], [3]. A direct sum theorem for communication complexity
with suboptimal parameters can be obtained via information compression [3]. A direct sum theorem for bounded-round communication has been proved [4], and we use a similar argument in one component of the proof (see Section II-F). Direct product theorems for communication complexity (with suboptimal parameters via information compression), bounded-round communication and from information complexity to communication complexity have also been studied [15], [6], [7].

II. TECHNICAL OVERVIEW

A. An alternative view of [3]

The starting point of our proof is an alternative view of the XOR lemma in [3] for information complexity, which does not give an exponentially small advantage. Running a protocol on an input pair sampled from some fixed input distribution defines a joint distribution over the input pairs and the transcripts. Information complexity studies that in this joint distribution, how much information the transcript reveals about the inputs. The (internal) information cost is defined as

\[ I(X; M | Y, M_0) + I(Y; M | X, M_0), \]

where \( M = (M_0, M_1, \ldots, M_n) \) is the transcript and \( M_0 \) is the public random bits.\(^2\) We assume that Alice sends all the odd \( M_i \) and Bob sends all the even \( M_i \). The internal information cost of a protocol is always at most its communication cost. It is also known that for bounded-round communication, the internal information complexity is roughly equal to the communication complexity [4] (up to some additive error probability).

For the XOR lemma for information complexity, we consider input pair \( X = (X_1, \ldots, X_n) \) and \( Y = (Y_1, \ldots, Y_n) \) sampled from \( \mu^n \). Suppose there is a protocol \( \pi \) computing \( f^{\oplus n} \) with information cost \( I \), we want to show that \( f \) can be computed with information cost \( \approx I/n \).

To this end, we show that \( \pi \) can be “decomposed” into a protocol \( \pi_{<n} \) computing \( f^{\oplus n-1} \) with information cost \( I_1 \) and a protocol \( \pi_n \) computing \( f \) with information cost \( I_2 \) such that \( I_1 + I_2 \approx I \), as follows (see also Figure 1).

- For \( \pi_{<n} \), given \( n-1 \) input pairs, the players view them as \( X_{<n} \) and \( Y_{<n} \) as part of the inputs for \( \pi \), where \( X_{<n} \) denotes \( (X_1, \ldots, X_{n-1}) \) and \( Y_{<n} \) denotes \( (Y_1, \ldots, Y_{n-1}) \); then the players publicly sample \( X_n \sim \mu_X \) and Bob privately samples \( Y_{<n} \) conditioned on \( X_n \); the players run \( \pi \) to compute \( f^{\oplus n}(X, Y) \); Bob sends one extra bit indicating \( f(X_n, Y_n) \).
- For \( \pi_n \), given one input pair, the players view it as \( X_n \) and \( Y_n \); then the players publicly sample \( Y_{<n} \sim \mu_Y^{n-1} \), and Alice privately samples \( X_{<n} \) conditioned on \( Y_{<n} \); the players run \( \pi \) to compute \( f^{\oplus n}(X, Y) \); Alice sends one extra bit indicating \( \oplus_{i=1}^{n-1} f(X_i, Y_i) \).

\(^2\)In the usual definition, the public random string is not part of the transcript. We add it for simplicity of notations. This does not change the values of the mutual information terms as it is already in the condition.

If \( \pi \) computes \( f^{\oplus n} \) correctly, then the two protocols compute \( f^{\oplus n-1} \) and \( f \) correctly respectively. For the information cost of \( \pi_{<n} \) (if we exclude the last bit indicating \( f(X_n, Y_n) \)), the first term is equal to \( I(X_{<n}; M \mid Y_{<n}, X_n, M_0) \), since \( X_n \) is sampled using public random bits. It is also equal to \( I(X_{<n}; M \mid Y, X_n, M_0) \) due to the rectangle property of communication protocols. For the information cost of \( \pi_n \) (if we exclude the last bit), the first term is equal to \( I(X_n; M \mid Y, M_0) \) since \( Y_{<n} \) is sampled using public random bits. Therefore, the first terms sum up to exactly \( I(X; M \mid Y, M_0) \), the first term in the information cost of \( \pi \), by the chain rule of mutual information. Similarly, the second terms sum up to \( I(Y; M \mid X, M_0) \), the second term in the information cost of \( \pi \).

Hence, including the last bits in the protocols, we have \( I_1 + I_2 \leq I + O(1) \). Thus, by repeatedly applying this argument, we obtain a protocol for \( f \) with information cost \( I/n + O(1) \), as desired. Note that in this decomposition, the players do not need to sample the private parts explicitly. As long as they can send the messages from the same distribution (e.g., by directly sampling the messages conditioned on the previous messages and their own inputs), the information costs and correctness are not affected.

The original paper proves the same result by explicitly writing out the protocol for \( f \) obtained after applying the above decomposition \( i \) times for a random \( i \in [n] \), and proving the expected cost is as claimed. The two proofs are essentially equivalent for this statement.\(^3\) However, as we will see later, our new view is more flexible, allowing for more sophisticated manipulations when doing the decomposition.

B. Obtaining exponentially small advantage

The above decomposition preserves the success probability. However, if we start from a protocol for \( f^{\oplus n} \) with exponentially small advantage, then we will not be able to obtain a

\(^3\)The original proof embeds the input to \( f \) into a random coordinate \( i \) of \( f^{\oplus n} \), and samples \( X_{>i} \) and \( Y_{<i} \) using public random bits.
protocol for $f$ with success probability $2/3$, which is required in order to prove the strong XOR lemma. Let $\text{adv}(f(X, Y) \mid R)$ denote the advantage for $f(X, Y)$ conditioned on $R$, which is defined as

$$2 \Pr[f(X, Y) = 1 \mid R] - 1,$$

i.e., the advantage is $\alpha$ if the conditional probability is either $1/2 + \alpha/2$ or $1/2 - \alpha/2$.

Now let us take a closer look at the two protocols $\pi_{<n}$ and $\pi_n$ (see Figure 2). For $\pi_{<n}$, in Bob’s view at the end of the communication, he knows his input $Y_{<n}$, the publicly sampled $X_n$ and the transcript $M$. Hence, he is able to predict $f_{\oplus n-1}(X_{<n}, Y_{<n})$ with advantage $\text{adv}(f_{\oplus n-1}(X_{<n}, Y_{<n}) \mid X_n, Y_{<n}, M)$. By letting Bob send one extra bit indicating his prediction, the advantage of the protocol achieves the same. For $\pi_n$, in Alice’s view at the end of the communication, she knows her input $X_n$, the publicly sampled $Y_{<n}$ and the transcript $M$. Hence, she is able to predict $f(X_n, Y_n)$ with advantage $\text{adv}(f(X_n, Y_n) \mid X_n, Y_{<n}, M)$. By letting Alice send one extra bit indicating her prediction, the advantage of the protocol achieves the same.

Now an important observation is that $X_{<n}$ and $Y_n$ are independent conditioned on $(X_n, Y_{<n}, M)$, by the rectangle property of communication protocols. Hence, $f_{\oplus n-1}(X_{<n}, Y_{<n})$ and $f(X_n, Y_n)$ are also independent conditioned on $(X_n, Y_{<n}, M)$. Since $f_{\oplus n}(X, Y) = f_{\oplus n-1}(X_{<n}, Y_{<n}) \oplus f(X_n, Y_n)$, by the probability of XOR of two independent bits, we have

$$\text{adv}(f_{\oplus n}(X, Y) \mid X_n, Y_{<n}, M)$$

$$= \text{adv}(f_{\oplus n-1}(X_{<n}, Y_{<n}) \mid X_n, Y_{<n}, M)$$

$$\times \text{adv}(f(X_n, Y_n) \mid X_n, Y_{<n}, M).$$

This suggests the following strategy for the decomposition:

- if the information cost of $\pi_n$ is large, then the information cost of $\pi_{<n}$ must be much smaller than that of $\pi$;
- if the information cost of $\pi_n$ is small and its advantage for $f$ is large, then we have obtained a good protocol for $f$;
- if the information cost of $\pi_n$ is small and its advantage for $f$ is small, then by (1), the advantage of $\pi_{<n}$ must be larger than that of $\pi$ by some factor.

Hence, in each decomposition, if we don’t already obtain a good protocol for $f$, then when decrementing $n$ to $n - 1$, we must either significantly decrease the information cost, or increase the advantage by a multiplicative factor. If we start with a protocol with a low cost and a mild exponentially small advantage for $f_{\oplus n}$, then we must obtain a good protocol for $f$ by applying this decomposition iteratively.

It turns out that the main difficulty in applying the above strategy is to formalize the last bullet point. Note that the expected advantage of $\pi_n$ (after Alice sending the one extra bit indicating her prediction) is $\mathbb{E}[\text{adv}(f(X_n, Y_n) \mid X_n, Y_{<n}, M)]$, the expected advantage of $\pi_{<n}$ (after Bob sending the one extra bit indicating his prediction) is $\mathbb{E}[\text{adv}(f_{\oplus n-1}(X_{<n}, Y_{<n}) \mid X_n, Y_{<n}, M)]$, and the expected advantage of $\pi$ is $\mathbb{E}[\text{adv}(f_{\oplus n}(X, Y) \mid M)]$, which is at most $\mathbb{E}[\text{adv}(f_{\oplus n}(X, Y) \mid X_n, Y_{<n}, M)]$.

When we say that the advantage of $\pi_n$ is small in the last bullet point, we can only guarantee that this expectation is small. Equation (1), which is a pointwise equality, does not directly give any useful bounds on the expectations. For example, it is possible that both $\mathbb{E}[\text{adv}(f(X_n, Y_n) \mid X_n, Y_{<n}, M)]$ and $\mathbb{E}[\text{adv}(f_{\oplus n-1}(X_{<n}, Y_{<n}) \mid X_n, Y_{<n}, M)]$ are very small, but $\text{adv}(f(X_n, Y_n) \mid X_n, Y_{<n}, M)$ and $\text{adv}(f_{\oplus n-1}(X_{<n}, Y_{<n}) \mid X_n, Y_{<n}, M)$ are always equal to zero or one at the same time, both concentrated on a small probability set. Then we have $\mathbb{E}[\text{adv}(f_{\oplus n}(X, Y) \mid X_n, Y_{<n}, M)] = \mathbb{E}[\text{adv}(f_{\oplus n-1}(X_{<n}, Y_{<n}) \mid X_n, Y_{<n}, M)]$, the advantage may not increase at all. In this case, the advantage $\text{adv}(f_{\oplus n}(X, Y) \mid X_n, Y_{<n}, M)$ is also concentrated on the same small probability set.

On the other hand, observe that if $\text{adv}(f_{\oplus n}(X, Y) \mid X_n, Y_{<n}, M)$ takes roughly the same value (say, $e$) most of the time, then we do obtain an advantage increase:

$$\mathbb{E}[	ext{adv}(f_{\oplus n-1}(X_{<n}, Y_{<n}) \mid X_n, Y_{<n}, M)]$$

$$= \mathbb{E}[e/\text{adv}(f(X_n, Y_n) \mid X_n, Y_{<n}, M)]$$

$$\geq e/\mathbb{E}[\text{adv}(f(X_n, Y_n) \mid X_n, Y_{<n}, M)]$$

by the convexity of $1/x$.

This motivates us to consider the following two extreme cases:

1) $\text{adv}(f_{\oplus n}(X, Y) \mid X_n, Y_{<n}, M)$ is roughly uniformly distributed among all $(X_n, Y_{<n}, M)$;

2) $\text{adv}(f_{\oplus n}(X, Y) \mid X_n, Y_{<n}, M)$ is concentrated on a tiny fraction of the triples $(X_n, Y_{<n}, M)$.

Basically following what we just argued, the above strategy directly applies in the first case. The second case is related to
the direct product theorems, where we also want to analyze
protocols that is correct with exponentially small probability.
This is because one possible strategy for the players is to
calculate all \( f(X_i, Y_i) \) correctly with some probability \( \varepsilon \) and
output a random bit otherwise. We must at least show that in
this case, \( \varepsilon \leq \exp(-\Omega(n)) \).

C. Generalized protocols

For the second case above, we follow one strategy for direct
product theorems [6]. When the advantage \( \text{adv}(f^{\exists n}(X, Y) \mid X_n, Y_n, M) \)
is concentrated on a small set \( U \) of triples \((X_n, Y_n, M)\), we restrict our attention to \( U \) by conditioning
\( \pi \) on \( U \). However, this immediately creates two issues.

The first issue is that although \( \pi \mid U \) is a well-defined
distribution, it is not necessarily a protocol, since conditioning
on an arbitrary event may break the independence between a
message and the receiver’s input, e.g., \( M_1 \) may no longer be
independent of \( Y \) conditioned on \( X \).

This issue was also encountered in the direct product
theorem proofs. Instead of studying standard protocols, we
focus on generalized protocols, where we allow each message
to depend on both player’s inputs, and we wish to restrict the
correlation between the odd \( M_i \) and Bob’s input and the
correlation between the even \( M_i \) and Alice’s input. In the
previous work, it bounds
\[
\theta(\pi) := \sum_{\text{odd } i} I(M_i; Y \mid X, M_{<i}) + \sum_{\text{even } i} I(M_i; Y \mid Y, M_{<i}),
\]
the mutual information between the message and the receiver’s
input.

Intuitively, the \( \theta \)-value measures how close to a standard
protocol a generalized protocol is. It turns out that the \( \theta \)-value
of a standard protocol conditioned on a not-too-small probabil-
ity event is small; on the other hand, when the \( \theta \)-value is small,
it is statistically close to a standard protocol. Furthermore,
an important feature of \( \theta(\pi) \) is that the decomposition of \( \pi \)
into \( \pi_{<n} \) and \( \pi_n \) also satisfies that \( \theta(\pi) = \theta(\pi_{<n}) + \theta(\pi_n) \).
Hence, when doing the decomposition, we hope to obtain a
generalized protocol for \( f \) that is very close to a standard
protocol.

The second issue is that conditioning on a small probability
event \( U \) could greatly increase the information cost, from \( I \) to
\( \Omega(I/\Pr[U]) \). Since \( I \) is close to the communication cost, such a
multiplicative loss in each step of decomposition is unafford-
able. Such a loss occurs because the mutual information is an
average measure (an expectation), which does not provide any
concentration (also recall the counterexample in footnote 4
where the communication cost and the advantage are both
concentrated on an \( \varepsilon \)-probability event, when we condition
on this event, both the expected communication cost and the
advantage increase by a factor of \( 1/\varepsilon \)). More specifically,
consider the first term in the information cost, \( I(X; M \mid Y) \)

\footnote{Conditioning on an event also distorts the input distribution, which needs
to be handled. But for simplicity, we omit it in the overview.}

(omit the public random bits for now). For standard protocols,
it is equal to
\[
\sum_{x,y,m} \pi(x, y, m) \log \left( \frac{\pi(x \mid m, y)}{\pi(x \mid y)} \right)
= \mathbb{E}_\pi \left[ \log \left( \frac{\pi(X \mid M, Y)}{\pi(X \mid Y)} \right) \right]
= \mathbb{E}_\pi \left[ \log \left( \frac{\pi(X \mid M, Y)}{\mu(X \mid Y)} \right) \right].
\]

If we only have a bound on this expectation, then inevitably
its value can greatly increase after conditioning on a small
probability event, not to say that the logarithm inside the
expectation is not nonnegative, so it can get worse than what
Markov’s inequality gives.

We also note that the argument in the previous subsections
 crucially uses the rectangle property of the communication
protocols, which does not necessarily hold for generalized
protocols. This turns out not to be a real issue, since throughout
the argument, we will maintain the rectangle property at all
leaves, which is sufficient for the argument to go through (see
also Section II-F).

D. \( \theta \)-cost and \( \chi^2 \)-costs

Our novel solution to the second issue above is to focus on the
“exponential version” of the information cost, i.e., for the
first term,
\[
\chi^2_{\mu, A}(\pi) := \mathbb{E}_\pi \left[ \frac{\pi(X \mid M, Y)}{\mu(X \mid Y)} \right],
\]
which we call the \( \chi^2 \)-cost by Alice. The \( \chi^2 \)-cost by Bob,
\( \chi^2_{\mu, B}(\pi) \), is defined similarly for the second term in
the information cost.

This notion of the cost has the following benefits.

- For a (deterministic) standard protocol with \( C \) bits of
  communication, \( \chi^2_{\mu, A}(\pi) \leq 2^C \). Hence, it corresponds to
  the exponential of the communication cost.
- When conditioning on a small probability event \( U \), we can
  essentially ensure that it increases by a factor of
  \( O(1/\pi(U)) \). Effectively, this only adds \( \log(1/\pi(U)) \) to the
  communication cost, which becomes affordable.

Note that the mutual information is the expected KL-
divergence, and the \( \chi^2 \)-cost is the expected \( \chi^2 \)-divergence
(plus one). Similarly, we also define an “exponential version”
of \( \theta(\pi) \), which we call the \( \theta \)-cost of \( \pi \). It also ensures that
the value does not increase significantly when conditioning on
a small probability event.

On the other hand, going from mutual information to its
“exponential version” loses many of its good properties, most
importantly, the chain rule. The next crucial observation is that
the chain rule for mutual information in fact holds pointwisely,
which enables us to work with the \( \chi^2 \)-costs.

More specifically, let \( X, Y, Z \) be three random variables
with joint distribution \( \pi \), the chain rules says
\[
I(X; Y, Z) =
\]
\[ I(X; Y) + I(X; Z \mid Y) \]. By writing the mutual information as an expectation, this is
\[
\mathbb{E} \left[ \log \left( \frac{\pi(Y, Z \mid X)}{\pi(Y, Z)} \right) \right] = \mathbb{E} \left[ \log \left( \frac{\pi(Y \mid X)}{\pi(Y)} \right) \right] + \mathbb{E} \left[ \log \left( \frac{\pi(Z \mid X, Y)}{\pi(Z \mid Y)} \right) \right].
\]

This equality holds pointwisely in the sense that for any concrete values \((x, y, z)\), the equality holds for the logarithms inside the expectation
\[
\log \left( \frac{\pi(y, z \mid x)}{\pi(y, z)} \right) = \log \left( \frac{\pi(y \mid x)}{\pi(y)} \right) + \log \left( \frac{\pi(z \mid x, y)}{\pi(z \mid y)} \right)
\]
by the definition of conditional probability.

Therefore, the “exponential version” also holds pointwisely:
\[
\frac{\pi(y, z \mid x)}{\pi(y, z)} = \frac{\pi(y \mid x)}{\pi(y)} \cdot \frac{\pi(z \mid x, y)}{\pi(z \mid y)}.
\]
This is what we use in replacement of the chain rule for mutual information. See the next subsection for more details.

E. Proof outline

We now give an outline of the proof of the following statement: Given an \(r\)-round standard protocol \(\pi\) for \(f^{\geq n}\) with communication cost \(o(n \cdot C)\) that succeeds with advantage \(\alpha^{o(n)}\) on the inputs sampled from \(\mu^n\), we can obtain an \(r\)-round generalized protocol \(\rho\) for \(f\) with \(\chi^2\)-costs \(\approx 2^C\), \(\theta\)-cost \(\approx 1/\alpha\), and advantage \(\approx \alpha\). We will then discuss how to convert such a generalized protocol to a standard protocol with low communication cost in the next subsection.

We first show that \(\pi\) is also a generalized protocol with \(\chi^2\)-cost \(2^{\Omega(nC)}\) and \(\theta\)-cost \(1\). Next, we decompose \(\pi\) into \(\pi_{\leq n}\) and \(\pi_n\) for \(f^{\geq n-1}\) and \(f\), and prove that the product of the \(\theta\)-cost [resp. \(\chi^2\)-costs] of \(\pi_{\leq n}\) and \(\pi_n\) is that of \(\pi\) pointwisely. Now if the advantage of \(\pi\) is not roughly evenly distributed, we will identify an event \(U\) such that the advantage conditioned on \(U\) is much higher than the average advantage, and more importantly, the advantage within \(U\) becomes roughly evenly distributed (not concentrated on any small probability event in \(U\)). Conditioning on \(U\) increases the advantage while also increases the \(\theta\)-cost and \(\chi^2\)-costs, it turns out that they all increase by about the same factor. Next, we partition the sample space of \(\pi\) into \(S_{\text{high-cost}}, S_{\text{low-cost}}\) and \(S_{\text{low-prob}}\) such that

- in \(S_{\text{high-cost}}\), \(\pi_n\) has high \(\theta\)-cost or high \(\chi^2\)-cost (excluding some corner cases), say \(\geq 1/\alpha\) for \(\theta\)-cost or \(\geq 2^C\) for \(\chi^2\)-cost,
- in \(S_{\text{low-cost}}\), \(\pi_n\) has low \(\theta\)-cost and low \(\chi^2\)-cost (also excluding some corner cases),
- \(S_{\text{low-prob}}\) is the rest, which will happen with very low probability.

Since the advantage is not concentrated on any small probability in \(U\), then (at least) one of \(S_{\text{high-cost}}\) or \(S_{\text{low-cost}}\) will have advantage about as high as the advantage of \(U\). If \(S_{\text{high-cost}}\) has the advantage as high as \(U\), then we prove that by the pointwise equality for the costs, \(\pi_{\leq n} \mid S_{\text{high-cost}}\) must have a much smaller cost than \(\pi \mid U\), while they have roughly the same advantage. If \(S_{\text{low-cost}}\) has the advantage as high as \(U\), then if \(\pi_n \mid S_{\text{low-cost}}\) has high advantage, then we obtain a desired generalized protocol for \(f\) with low costs and high advantage; otherwise we prove that \(\pi_{\leq n} \mid S_{\text{low-cost}}\) has a much higher advantage than \(\pi \mid U\) (as the advantage of \(\pi\) is roughly evenly distribution within \(U\)), while they have roughly the same costs.

To summarize the above argument, if we don’t already find a desired generalized protocol for \(f\), then when decrementing \(n\) to \(n-1\), we first condition on an event \(U\), increasing costs and advantage simultaneously by about the same factor, then either we reduce the \(\theta\)-cost by a factor of \(\geq 1/\alpha\), or we reduce the \(\chi^2\)-costs by a factor of \(\geq 2^C\), or we increase the advantage by a factor of \(\geq 1/\alpha\). Since we start with \(\chi^2\)-costs \(2^{\Omega(nC)}\), \(\theta\)-cost 1 and advantage \(\alpha^{o(n)}\), we cannot repeat this for \(n\) steps without finding a desired protocol for \(f\). More formally, we will measure the progress by using a potential function that depends on the costs and advantage of the current protocol, and show that each time we decrement from \(k\) to \(k-1\), how much the potential must decrease.

F. Convert a generalized protocol to a standard protocol

Finally, we need to show that the existence of a good generalized protocol implies the existence of a good standard protocol. We prove that if an \(r\)-round generalized protocol \(\rho\) has \(\chi^2\)-costs \(2^C\), \(\theta\)-cost \(1/\alpha\) and advantage \(\alpha\), then there is an \(r\)-round standard protocol \(\pi\) with communication cost \(\approx C\) and advantage \(\approx \alpha^2\). Together with what we summarized in the last subsection, we obtain the strong XOR lemma for \(r\)-round communication.

[4] converts a standard protocol \(\rho\) with constant rounds to a standard protocol with communication matching the internal information cost of \(\rho\). Using a similar argument, we can convert \(\rho\) to a standard protocol with communication \(\approx C\). By the convexity of \(2^x\), \(\chi^2\)-cost of \(2^C\) implies internal information cost of at most \(C\). It turns out that the (almost) same argument applies in our case, for generalized protocol \(\rho\).

Then the next crucial observation is that we can ensure the generalized protocol \(\rho\) that we obtain from the arguments in the previous subsection has the rectangle property with respect to \(\mu\). Roughly speaking, it means that for all transcripts \(M\), if we look at the ratio of the probabilities \(\frac{\rho(X, Y | M)}{\mu(X, Y)}\), it is a product function of \(X\) and \(Y\), i.e., it is equal to \(g_A(X) \cdot g_B(Y)\) for some functions \(g_A, g_B\) that may depend on \(M\). Note that a standard protocol has the rectangle property, since each message depends only on either \(X\) or \(Y\), and the same property holds even conditioned on any prefix of the transcript \(M_{\leq i}\). A generalized protocol may not have this property in general, but we can ensure that the protocol we obtain has this product structure conditioned on any complete transcript \(M\).

After generating a transcript \(M\) using [4], the rectangle property allows the players to locally “re-adjust” the probabilities (via rejection sampling) so that after the readjustment, the probability of a triple \((X, Y, M)\) is proportional to the “right”
probability $\rho(X,Y,M)$, which in turn, gives the advantage proportional to that of $\rho$.

The probability that is sacrificed in the rejection sampling depends on how far $\rho$ is from a standard protocol, i.e., the $\theta$-cost of $\rho$. It turns out that the above argument gives an overall advantage of at least $\alpha^2$ divided by the $\theta$-cost of $\rho$.

REFERENCES

[1] Sepehr Assadi and Vishvajeet N. Graph streaming lower bounds for parameter estimation and property testing via a streaming XOR lemma. In Samir Khuller and Virginia Vassilevska Williams, editors, STOC ’21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021, pages 612–625. ACM, 2021.

[2] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. J. Comput. Syst. Sci., 68(4):702–732, 2004.

[3] Boaz Barak, Mark Braverman, Xi Chen, and Anup Rao. How to compress interactive communication. SIAM J. Comput., 42(3):1327–1363, 2013.

[4] Mark Braverman and Anup Rao. Information equals amortized communication. In Rafail Ostrovsky, editor, IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011, pages 748–757. IEEE Computer Society, 2011.

[5] Mark Braverman, Anup Rao, Omri Weinstein, and Amir Yehudayoff. Direct product via round-preserving compression. In Fedor V. Fomin, Rusins Freivalds, Marta Z. Kwiatkowska, and David Peleg, editors, Automata, Languages, and Programming - 40th International Colloquium, ICALP 2013, Riga, Latvia, July 8-12, 2013. Proceedings, Part I, volume 7965 of Lecture Notes in Computer Science, pages 232–243. Springer, 2013.

[6] Mark Braverman, Anup Rao, Omri Weinstein, and Amir Yehudayoff. Direct products in communication complexity. In 54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA, pages 746–755. IEEE Computer Society, 2013.

[7] Mark Braverman and Omri Weinstein. An interactive information odometer and applications. In Rocco A. Servedio and Ronitt Rubinfeld, editors, Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015, pages 341–350. ACM, 2015.

[8] Joshua Brody, Jae Tae Kim, Peem Lerdputtipongporn, and Hariharan Srinivasulu. A strong XOR lemma for randomized query complexity. CoRR, abs/2007.05580, 2020.

[9] Amit Chakrabarti, Yaojun Shi, Anthony Wirth, and Andrew Chi-Chih Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In 42nd Annual Symposium on Foundations of Computer Science, FOCS 2001, 14-17 October 2001, Las Vegas, Nevada, USA, pages 270–278. IEEE Computer Society, 2001.

[10] Lijie Chen, Gillat Kol, Dmitry Paramonov, Raghuvansh R. Satija, Zhao Song, and Huacheng Yu. Almost optimal super-constant-pass streaming lower bounds for reachability. In Samir Khuller and Virginia Vassilevska Williams, editors, STOC ’21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021, pages 570–583. ACM, 2021.

[11] Andrew Drucker. Improved direct product theorems for randomized query complexity. Comput. Complex., 21(2):197–244, 2012.

[12] Oded Goldreich, Noam Nisan, and Avi Wigderson. On yao’s xor-lemma. In Oded Goldreich, editor, Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation - In Collaboration with Lidor Avigd, Mihir Bellare, Zvika Brakerski, Shafi Goldwasser, Shai Halevi, Tali Kaufman, Leonid Levin, Noam Nisan, Dana Ron, Madhu Sudan, Luca Trevisan, Salil Vadhan, Avi Wigderson, David Zuckerman, volume 6650 of Lecture Notes in Computer Science, pages 273–301. Springer, 2011.

[13] Russell Impagliazzo. Hard-core distributions for somewhat hard problems. In 36th Annual Symposium on Foundations of Computer Science, Milwaukee, Wisconsin, USA, 23-25 October 1995, pages 538–545. IEEE Computer Society, 1995.

[14] Russell Impagliazzo and Avi Wigderson. $P = BPP$ if $E$ requires exponential circuits: Derandomizing the XOR lemma. In Frank Thomson Leighton and Peter W. Shor, editors, Proceedings of the Twenty-Ninth Annual ACM Symposium on the Theory of Computing, El Paso, Texas, USA, May 4-6, 1997, pages 220–229. ACM, 1997.

[15] Rahul Jain, Attila Pereszlenyi, and Penghui Yao. A direct product theorem for the two-party bounded-round public-coin communication complexity. In 53rd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2012, New Brunswick, NJ, USA, October 20-23, 2012, pages 167–176. IEEE Computer Society, 2012.

[16] Leonid A. Levin. One-way functions and pseudorandom generators. Comb., 7(4):357–363, 1987.

[17] Ronen Shaltiel. Towards proving strong direct product theorems. Comput. Complex., 12(1-2):1-22, 2003.

[18] Alexander A. Sherstov. Strong direct product theorems for quantum communication and query complexity. In Lance Fortnow and Salil P. Vadhan, editors, Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011, pages 41–50. ACM, 2011.

[19] Emanuele Viola and Wigderson. Norms, XOR lemmas, and lower bounds for polynomials and protocols. Theory Comput., 4(1):137–168, 2008.

[20] Andrew Chi-Chih Yao. Theory and applications of trapdoor functions (extended abstract). In 23rd Annual Symposium on Foundations of Computer Science, Chicago, Illinois, USA, 3-5 November 1982, pages 80–91. IEEE Computer Society, 1982.