On the communication complexity of sparse set disjointness and exists-equal problems

Mert Sağlam
University of Washington
saglam@uw.edu

Gábor Tardos
Simon Fraser University and
Alfréd Rényi Institute of Mathematics
tardos@renyi.hu

***Abstract*—In this paper we study the two player randomized communication complexity of the sparse set disjointness and the exists-equal problems and give matching lower and upper bounds (up to constant factors) for any number of rounds for both of these problems. In the sparse set disjointness problem, each player receives a k-subset of $[m]$ and the goal is to determine whether the sets intersect. For this problem, we give a protocol that communicates a total of $O(k \log^{(r)} k)$ bits over r rounds and errs with very small probability. Here we can take $r = \log^* k$ to obtain a $O(k)$ total communication $\log^* k$-round protocol with exponentially small error probability, improving on the $O(k)$-bits $O(\log k)$-round constant error probability protocol of Håstad and Wigderson from 1997.**

In the exists-equal problem, the players receive vectors $x, y \in [t]^n$ and the goal is to determine whether there exists a coordinate i such that $x_i = y_i$. Namely, the exists-equal problem is the OR of n equality problems. Observe that exists-equal is an instance of sparse set disjointness with $k = n$, hence the protocol above applies here as well, giving an $O(n \log^{(r)} n)$ upper bound. Our main technical contribution in this paper is a matching lower bound: we show that when $t = \Omega(n)$, any r-round randomized protocol for the exists-equal problem with error probability at most $1/3$ should have a message of size $\Omega(n \log^{(r)} n)$. Our lower bound holds even for super-constant $r \leq \log^* n$, showing that any $O(n)$ bits exists-equal protocol should have $\log^* n - O(1)$ rounds. Note that the protocol we give errs only with less than polynomially small probability and provides guarantees on the total communication for the harder set disjointness problem, whereas our lower bound holds even for constant error probability protocols and for the easier exists-equal problem with guarantees on the max-communication. Hence our upper and lower bounds match in a strong sense.

Our lower bound on the constant round protocols for exists-equal shows that solving the OR of n instances of the equality problems requires strictly more than n times the cost of a single instance. To our knowledge this is the first example of such a *super-linear* increase in complexity.

***Keywords*-communication complexity; direct-sum; round-elimination; isoperimetric inequality**

I. Introduction

In a two player communication problem the players, named Alice and Bob, receive separate inputs, x and y, and they communicate in order to compute the value $f(x, y)$ of a function f. In an r-round protocol, the players can take at most r turns alternately sending each other a message and the last player to receive a message declares the output of the protocol. A protocol can be *deterministic* or *randomized*; in the latter case the players can base their actions on a common random source and we measure the *error probability*: the maximum over inputs (x, y), of the probability that the output of the protocol differs from $f(x, y)$.

A. Sparse set disjointness

Set disjointness is perhaps the most studied problem in communication complexity. In the most standard version Alice and Bob receive a subset of $[m] := \{1, \ldots, m\}$ each, with the goal of deciding whether their sets intersect or not. The primary question is whether the players can improve on the trivial deterministic protocol, where the first player sends the entire input to the other player, thereby communicating m bits. The first lower bound on the randomized complexity of this problem was given in [3] by Babai et al., who showed that any ϵ-error protocol for disjointness must communicate $\Omega(\sqrt{m})$ bits. The tight bound of $\Omega(m)$-bits was first given by Kalyanasundaram and Schnitger [29] and was later simplified by Razborov [42] and Bar-Yossef et al. [4].

In the sparse set disjointness problem DISJ_k^m, the sets given to the players are guaranteed to have at most k elements. The deterministic communication complexity of this problem is well understood. The trivial protocol, where Alice sends her entire input to Bob solves the problem in one round using $O(k \log(2m/k))$ bits. On the other hand, an $\Omega(k \log(2m/k))$ bit total communication lower bound can be shown even for protocols with an arbitrary number of rounds, say using the rank method; see [32], page 175.

The randomized complexity of the problem is far more subtle. The results cited above immediately imply a $\Omega(k)$ lower bound for this version of the problem. The folklore 1-round protocol solves the problem using $O(k \log k)$ bits, wherein Alice sends $O(\log k)$-bit hashes for each element of her set. Håstad and Wigderson [24] gave a protocol that matches the $\Omega(k)$ lower bound mentioned above. Their $O(k)$-bit randomized protocol runs in $O(\log k)$-rounds and errs with a small constant probability. In Section II, we improve this protocol to run in $\log^* k$ rounds, still with $O(k)$ total communication, but with exponentially small error in k. We also present an r-round protocol for any $r < \log^* k$ with total communication $O(k \log^{(r)} k)$ and error probability well below $1/k$; see Theorem 1. (Here $\log^{(r)}$ denotes the iterated logarithm function, see Section I-E.) As the exists-equal problem with parameters t and n (see below) is a special case of DISJ_n^{tn}, our lower bounds for the exists-equal problem (see below) show that complexity of this algorithm is optimal for any number $r \leq \log^* k$ of rounds,

0272-5428/13 $26.00 © 2013 IEEE
DOI 10.1109/FOCS.2013.78

even if we allow the much larger error probability of $1/3$. Buhrman et al. [13] and Woodruff [45] (as presented in [40]) show an $\Omega(k \log k)$ lower bound for 1-round complexity of DISJ_k^m by a reduction from the indexing problem (this reduction was also given in [17]). We note that these lower bounds do not apply to the exists-equal problem, as the input distribution they use generates instances inherently specific to the disjointness problem; furthermore this distribution admits a $O(\log k)$ bits protocol in two rounds.

B. The exists-equal problem

In the equality problem Alice and Bob receive elements x and y of a universe $[t]$ and they have to decide whether $x = y$. We define the two player communication game exists-equal with parameters t and n as follows. Each player is given an n-dimensional vector from $[t]^n$, namely x and y. The value of the game is one if there exists a coordinate $i \in [n]$ such that $x_i = y_i$, zero otherwise. Clearly, this problem is the OR of n independent instances of the equality problem.

The direct sum problem in communication complexity is the study of whether n instances of a problem can be solved using less than n times the communication required for a single instance of the problem. This question has been studied extensively for specific communication problems as well as some class of problems [14], [26], [27], [7], [19], [25], [22], [5]. The so called direct sum approach is a very powerful tool to show lower bounds for communication games. In this approach, one expresses the problem at hand, say as the OR of n instances of a simpler function and the lower bound is obtained by combining a lower bound for the simpler problem with a direct sum argument. For instance, the two-player and multi-player disjointness bounds of [4], the lopsided set disjointness bounds [41], and the lower bounds for several communication problems that arise from streaming algorithms [28], [34] are a few examples of results that follow this approach.

Exists-equal with parameters t and n is a special case of DISJ_n^{tn}, so our protocols in Section II solve exists-equal. We show that when $t = \Omega(n)$ these protocols are optimal, namely every r-round randomized protocol ($r \leq \log^* n$) with at most $1/3$ error error probability needs to send at least one message of size $\Omega(n \log^{(r)} n)$ bits. See Theorem 4. Our result shows that computing the OR of n instances of the equality problem requires *strictly more* than n times the communication required to solve a single instance of the equality problem when the number of rounds is smaller than $\log^* n - O(1)$. Recall that the equality problem admits an ϵ-error $\log(1/\epsilon)$-bit one-round protocol in the common random source model.

For $r = 1$, our result implies that to compute the OR of n instances of the equality problem with *constant probability*, no protocol can do better than solving each instance of the equality problem with *high probability* so that the union bound can be applied when taking the OR of the computed results. The single round case of our lower bound also generalizes the $\Omega(n \log n)$ lower bound of Molinaro et al. [37] for the one round communication problem, where the players have to find all the answers of n equality problems, outputting an n bit string.

C. Lower bound techniques

We obtain our general lower bound via a round elimination argument. In such an argument one assumes the existence of a protocol P that solves a communication problem, say f, in r rounds. By suitably modifying the internals of P, one obtains another protocol P' with $r - 1$ rounds, which typically solves smaller instances of f or has larger error than P. Iterating this process, one obtains a protocol with zero rounds. If the protocol we obtain solves non-trivial instances of f with good probability, we conclude that we have arrived at a contradiction, therefore the protocol we started with, P, cannot exist. Although round elimination arguments have been used for a long time, our round elimination lemma is the first to prove a *super-linear* communication lower bound in the number of primitive problems involved, obtaining which requires new and interesting ideas.

The general round elimination presented in Section V is very involved, but the lower bound on the one-round protocols can also be obtained in a more elementary way. As the one round case exhibits the most dramatic super-linear increase in the communication cost and also generalizes the lower bound in [37], we include this combinatorial argument separately in Section III, see Theorem 2.

At the heart of the general round elimination lemma is a new isoperimetric inequality on the discrete cube $[t]^n$ endowed with the Hamming distance. We present this result, Theorem 3, in Section IV. To the best of our knowledge, the first isoperimetric inequality on this metric space was proven by Lindsey in [33], where the subsets of $[t]^n$ of a certain size with the so called minimum induced-edge number were characterized. This result was rediscovered in [31] and [16] as well. See [2] for a generalization of this inequality to universes which are n-dimensional boxes with arbitrary side lengths. In [9], Bollobás et al. study isoperimetric inequalities on $[t]^n$ endowed with the ℓ_1 distance. For the purposes of our proof we need to find sets S that minimize a substantially more complicated measure. This measure also captures how spread out S is and can be described roughly as the average over points $x \in [t]^n$ of the logarithm of the number of points in the intersection of S and a Hamming ball around x.

D. Related work

In [36], a round elimination lemma was given, which applies to a class of problems with certain self-reducibility properties. The lemma is then is used to get lower bounds for various problems including the greater-than and the predecessor problems. This result was later tightened in [44] to get better bounds for the aforementioned problems. Different round elimination arguments were also used in [30], [20], [38], [35], [18], [6] for various communication complexity lower bounds and most recently in [10] and [12] for obtaining lower bounds for the gapped Hamming distance problem.

In parallel and independent of the present form of this paper Brody et al. [11] have also established an $\Omega(n \log^{(r)} n)$ lower bound for the r-round communication complexity of the exists-equal problem with parameter n. Their result applies for protocols with a polynomially small error probability like $1/n$. This stronger assumption on the protocol allows for simpler proof techniques, namely the information complexity based direct sum technique developed in several papers including [1], [14], but it is not enough to create an example where solving the OR of n communication problems requires more than n times the communication of solving a single instance. Indeed, even in the shared random source model one needs $\log n$ bits of communication (independent of the number of rounds) to achieve $1/n$ error in a single equality problem.

E. Notation

For a positive integer t, we write $[t]$ for the set of positive integers not exceeding t. For two n-dimensional vectors x, y, let $\mathrm{Match}(x, y)$ be the number of coordinates where x and y agree. Notice that $n - \mathrm{Match}(x, y)$ is the Hamming distance between x and y. For a vector $x \in [t]^n$ we write x_i for its ith coordinate. We denote the distribution of a random variable X by $\mathrm{dist}(X)$ and the support set of it by $\mathrm{supp}(X)$. We write $\Pr_{x\sim\nu}[\cdot]$ and $\mathbb{E}_{x\sim\nu}[\cdot]$ for the probability and expectation, respectively, when x is distributed according to a distribution ν. We write μ for the uniform distribution on $[t]^n$. For instance, for a set $S \subseteq [t]^n$, we have $\mu(S) = |S|/t^n$.

For $x, y \in [t]^n$ we denote the value of the exists-equal game by $\mathrm{EE}^t_n(x, y)$. Recall that it is zero if and only if x and y differ in each coordinate. Whenever we drop t from the notation we assume $t = 4n$. Often we will also drop n and simply denote the game value by $\mathrm{EE}(x, y)$ if n is clear from the context.

All logarithms in this paper are to the base 2. Analogously, throughout this paper we take $\exp(x) = 2^x$. We will also use the iterated versions of these functions:

$$\begin{aligned} \log^{(0)} x &:= x, & \exp^{(0)} x &:= x, \\ \log^{(r)} x &:= \log(\log^{(r-1)} x), & \exp^{(r)} x &:= \exp(\exp^{(r-1)} x) \end{aligned}$$

for $r \geq 1$. Moreover we define $\log^* x$ to be the smallest integer r for which $\log^{(r)} x < 2$.

Throughout the paper we ignore divisibility problems, e.g., in Lemma 2 in Section III we assume that $t^n/2^{c+1}$ is an integer. Dealing with rounding issues would complicate the presentation but does not add to the complexity of the proofs.

F. Information theory

Here we briefly review some definitions and facts from information theory that we use in this paper. For a random variable X, we denote its binary Shannon entropy by $\mathrm{H}(X)$. We will also use conditional entropies $\mathrm{H}(X \mid Y) = \mathrm{H}(X, Y) - \mathrm{H}(Y)$. Let μ and ν be two probability distributions, supported on the same set S. We denote the binary Kullback-Leibler divergence between μ and ν by $\mathbf{D}(\mu \,\|\, \nu)$. A random variable with Bernoulli distribution with parameter p takes the value 1 with probability p and the value 0 with probability $1-p$. The entropy of this variable is denoted by $\mathrm{H}_2(p)$. For two reals $p, q \in (0, 1)$, we denote by $\mathbf{D}_2(p \,\|\, q)$ the divergence between the Bernoulli distributions with parameters p and q.

If $X \in [t]^n$ and $L \subseteq [n]$, then the projection of X to the coordinates in L is denoted by X_L. Namely, X_L is obtained from $X = (X_1, \ldots, X_n)$ by keeping only the coordinates X_i with $i \in L$. The following lemma of Chung et al. [15] relates the entropy of a variable to the entropy of its projections.

Lemma 1. (Chung et al. [15]) *Let* $\mathrm{supp}(X) \subseteq [t]^n$. *We have* $\frac{l}{n}\mathrm{H}(X) \leq \mathbb{E}_L[\mathrm{H}(X_L)]$, *where the expectation is taken for a uniform random l-subset L of* $[n]$.

II. The upper bound

Recall that in the communication problem DISJ^m_k, each of the two players is given a subset of $[m]$ of size at most k and they communicate in order to determine whether their sets are disjoint or not. In 1997, Håstad and Wigderson [39], [24] gave a probabilistic protocol that solves this problem with $O(k)$ bits of communication and has constant one-sided error probability. The protocol takes $O(\log k)$ rounds. Let us briefly review this protocol as this is the starting point of our protocol.

Let $S, T \subseteq [m]$ be the inputs of Alice and Bob. Observe that if they find a set Z satisfying $S \subseteq Z \subseteq [m]$, then Bob can replace his input T with $T' = T \cap Z$ as $T' \cap S = T \cap S$. The main observation is that if S and T are disjoint, then a random set $Z \supseteq S$ will intersect T in a uniform random subset, so one can expect $|T'| \approx |T|/2$. In the Håstad-Wigderson protocol the players alternate in finding a random set that contains the current input of one of them, effectively halving the other player's input. If in this process the input of one of the players becomes empty, they know the original inputs were disjoint. If, however, the sizes of their inputs do not show the expected exponential decrease in time, then they declare that their inputs intersect. This introduces a small one sided error. Note that one of the two outcomes happens in $O(\log k)$ rounds. An important observation is that Alice can describe a random set $Z \supseteq S$ to Bob using an expected $O(|S|)$ bits by making use of the joint random source. This makes the total communication $O(k)$.

In our protocol proving the next theorem, we do almost the same, but we choose the random sets $Z \supseteq S$ not uniformly, but from a biased distribution favoring ever smaller sets. This makes the size of the input sets of the players decrease much more rapidly, but describing the random set Z to the other player becomes more costly. By carefully balancing the parameters we optimize for the total communication given any number of rounds. When the number of rounds reaches $\log^* k - O(1)$ the communication reaches its minimum of $O(k)$ and the error becomes exponentially small.

Theorem 1. *For any $r \leq \log^* k$, there is an r-round probabilistic protocol for DISJ_k^m with $O(k \log^{(r)} k)$ bits total communication. There is no error for intersecting input sets, and the probability of error for disjoint sets can be made $O(1/\exp^{(r)}(c \log^{(r)} k) + \exp(-\sqrt{k})) \ll 1/k$ for any constant $c > 1$*[1]

For $r = \log^* k - O(1)$ rounds this means an $O(k)$-bit protocol with error probability $O(\exp(-\sqrt{k}))$.

Proof: We start with the description of the protocol. Let S_0 and S_1 be the input sets of Alice and Bob, respectively. For $1 \leq i \leq r$, i even Alice sends a message describing a set $Z_i \supset S_i$ based on her "current input" S_i and Bob updates his "current input" S_{i-1} to $S_{i+1} := S_{i-1} \cap Z_i$. In odd numbered rounds the same happens with the role of Alice and Bob reversed. We depart from the Håstad-Wigderson protocol in the way we choose the sets Z_i: Using the shared random source the players generate l_i random subsets of $[m]$ containing each element of $[m]$ independently and with probability p_i. We will set these parameters later. The set Z_i is chosen to be the first such set containing S_i. Alice or Bob (depending on the parity of i) sends the index of this set or ends the protocol by sending a special error signal if none of the generated sets contain S_i. The protocol ends with declaring the inputs disjoint if the error signal is never sent and we have $S_{r+1} = \emptyset$. In all other cases the protocol ends with declaring "not disjoint".

This finishes the description of the protocol except for the setting of the parameters. Note that the error of the protocol is one-sided: $S_0 \cap S_1 = S_i \cap S_{i+1}$ for $i \leq r$, so intersecting inputs cannot yield $S_{r+1} = \emptyset$.

We set the parameters (including k_i used in the analysis) as follows:

$$\begin{aligned}
u &= (c+1) \log^{(r)} k, && \\
p_i &= \frac{1}{\exp^{(i)} u} && \text{for } 1 \leq i \leq r, \\
l_1 &= k \exp(ku), && \\
l_i &= k 2^{k/2^{i-4}} && \text{for } 2 \leq i \leq r, \\
k_0 &= k_1 = k, && \\
k_i &= \frac{k}{2^{i-4} \exp^{(i-1)} u} && \text{for } 2 \leq i \leq r, \\
k_{r+1} &= 0. &&
\end{aligned}$$

The message sent in round $i > 1$ has length $\lceil \log(l_i + 1) \rceil < k/2^{i-4} + \log k + 1$, thus the total communication in all rounds but the first is $O(k)$. The length of the first message is $\lceil \log(l_1 + 1) \rceil \leq ku + \log k + 1$. The total communication is $O(ku) = O(ck \log^{(r)} k)$ as claimed (recall that c is a constant).

Let us assume the input pair is disjoint. To estimate the error probability we call round i *bad* if an error message is sent or a set S_{i+1} is created with $|S_{i+1}| > k_{i+1}$. If no bad round exists we have $S_{r+1} = \emptyset$ and the protocol makes no error. In what follows we bound the probability that round i is bad assuming the previous rounds are not bad and therefore having $|S_j| \leq k_j$ for $0 \leq j \leq i$.

The probability that a random set constructed in round i contains S_i is $p_i^{-|S_i|} \geq p_i^{-k_i}$. The probability that none of the l_i sets contains S_i and thus an error message is sent is therefore at most $(1 - p_i^{k_i})^{l_i} < e^{-k}$.

If no error occurs in the first bad round i, then $|S_{i+1}| > k_{i+1}$. Note that in this case $S_{i+1} = S_{i-1} \cap Z_i$ contains each element of S_{i-1} independently and with probability p_i. This is because the choice of Z_i was based on it containing S_i, so it was independent of its intersection with S_{i-1} (recall that $S_i \cap S_{i-1} = S_1 \cap S_0 = \emptyset$). For $i < r$ we use the Chernoff bound. The expected size of S_{i+1} is $|S_{i-1}| p_i \leq k_{i-1} p_i \leq k_{i+1}/2$, thus the probability of $|S_{i+1}| > k_{i+1}$ is at most $2^{-k_{i+1}/4}$. Finally for the last round $i = r$ we use the simpler estimate $p_r k_{r-1} \leq k / \exp^{(r)} u$ for $|S_{r+1}| > k_{r+1} = 0$.

Summing over all these estimates we obtain the following error bound for our protocol:

$$\Pr[\text{error}] \leq r e^{-k} + \frac{k}{\exp^{(r)} u} + \sum_{i=2}^{r} 2^{-k_i/4}.$$

In case $k_r \geq 4\sqrt{k}$ this error estimate proves the theorem. In case $k_r < 4\sqrt{k}$ we need to make a minor adjustments in the setting of our parameters. We take j to be the smallest value with $k_j < 4\sqrt{k}$, modify the parameters for round j and stop the protocol after this round declaring "disjoint" if $S_{j+1} = \emptyset$ and "intersecting" otherwise. The new parameters for round j are $k'_j = 4\sqrt{k}$, $p'_j = 2^{-2\sqrt{k}}$, $l'_j = k2^{8k}$. This new setting of the parameters makes the message in the last round linear in k, while both the probability that round $j-1$

[1] In an earlier manuscript of this paper we shared in 2010 with the first two authors of [11], we gave a protocol with identical communication cost but polynomially small error probability of k^{-c}. In [11] a version of this protocol with polynomially small error that works for EE_k^t was presented and it was mentioned that the error bound can be improved to $\exp(-\prod_{i=1}^{r} \log^{(i)} k)$. This work inspired us to explore and lower the error further and led us to obtain the tighther $1/\exp^{(r)}(c \log^{(r)} k) + \exp(-\sqrt{k})$ bound.

is bad because it makes $|S_j| > k'_j$, or the probability that round j is bad for any reason (error message or $S_{j+1} \neq \emptyset$) is $O(2^{-\sqrt{k}})$. This finishes the analysis of our protocol. ■

III. Lower bound for single round protocols

In this section we give an combinatorial proof that any single round randomized protocol for the exists-equal problem with parameters n and $t = 4n$ has complexity $\Omega(n \log n)$ if its error probability is at most $1/3$. As pointed out in the Introduction, to our knowledge this is the fist established case when solving the OR of n instances of a communication problem requires strictly more than n times the complexity needed to solve a single such instance.

We start with with a simple and standard reduction from the randomized protocol to the deterministic one and further to a large set of inputs that makes the first (and in this case only) message fixed. These steps are also used in the general round elimination argument therefore we state them in general form.

Let $\epsilon > 0$ be a small constant and let P be an $1/3$-error randomized protocol for the exists-equal problem with parameters n and $t = 4n$. We repeat the protocol P in parallel taking the majority output, so that the number of rounds does not change, the length of the messages is multiplied by a constant and the error probability decreases below ϵ. Now we fix the coins of of this ϵ-error protocol in a way to make the resulting deterministic protocol err on at most ϵ fraction of the possible inputs. Denote the deterministic protocol we obtain by Q.

Lemma 2. *Let Q be a deterministic protocol for the EE_n problem that makes at most ϵ error on the uniform distribution. Assume Alice sends the first message of length c. There exists an $S \subset [t]^n$ of size $\mu(S) = 2^{-c-1}$ such that the first message of Alice is fixed when $x \in S$ and we have $\Pr_{y\sim\mu}[Q(x,y) \neq \mathrm{EE}(x,y)] \leq 2\epsilon$ for all $x \in S$.*

Proof: Note that the quantity $e(x) = \Pr_{y\sim\mu}[Q(x,y) \neq \mathrm{EE}(x,y)]$, averaged over all x, is the error probability of Q on the uniform input, hence is at most ϵ. Therefore for at least half of x, we have $e(x) \leq 2\epsilon$. The first message of Alice partitions this half into at most 2^c subsets. We pick S to consist of $t^n/2^{c+1}$ vectors of the same part: at least one part must have this many elements. ■

We fix a set S as guaranteed by the lemma. We assume we started with a single round protocol, so $Q(x,y) = Q(x',y)$ whenever $x, x' \in S$. Indeed, Alice sends the same message by the choice of S and then the output is determined by Bob, who has the same input in the two cases.

We call a pair (x,y) *bad* if $x \in S$, $y \in [t]^n$ and Q errs on this input, i.e., $Q(x,y) \neq \mathrm{EE}(x,y)$. Let b be the number of bad pairs. By Lemma 2 each $x \in |S|$ is involved in at most $2\epsilon t^n$ bad pairs, so we have

$$b \leq 2\epsilon |S| t^n.$$

We call a triple (x, x', y) *bad* if $x, x' \in S$, $y \in [t]^n$, $\mathrm{EE}(x,y) = 1$ and $\mathrm{EE}(x',y) = 0$. The proof is based on double counting the number z of bad triples. Note that for a bad triple (x,x',y) we have $Q(x,y) = Q(x',y)$ but $\mathrm{EE}(x,y) \neq \mathrm{EE}(x',y)$, so Q must err on either (x,y) or (x',y) making one of these pairs bad. Any pair (bad or not) is involved in at most $|S|$ bad triples, so we have

$$z \leq b|S| \leq 2\epsilon |S|^2 t^n.$$

Let us fix arbitrary $x, x' \in S$ with $\mathrm{Match}(x,x') \leq n/2$. We estimate the number of $y \in [t]^n$ that makes (x,x',y) a bad triple. Such a y must have $\mathrm{Match}(x,y) > \mathrm{Match}(x',y) = 0$. To simplify the calculation we only count the vectors y with $\mathrm{Match}(x,y) = 1$. The match between y and x can occur at any position i with $x_i \neq x'_i$. After fixing the coordinate $y_i = x_i$ we can pick the remaining coordinates y_j of y freely as long as we avoid x_j and x'_j. Thus we have

$$\begin{aligned}|\{y \mid (x,x'y) \text{ is bad}\}| &\geq (n - \mathrm{Match}(x,x'))(t-2)^{n-1}\\ &\geq (n/2)(t-2)^{n-1} > t^n/14,\end{aligned}$$

where in the last inequality we used $t = 4n$. Let s be the size of the Hamming ball $B_{n/2}(x) = \{y \in [t]^n \mid \mathrm{Match}(x,y) > n/2\}$. By the Chernoff bound we have $s < t^n/n^{n/2}$ (using $t = 4n$ again). For a fixed x we have at least $|S| - s$ choices for $x' \in S$ with $\mathrm{Match}(x,x') \leq n/2$ when the above bound for triples apply. Thus we have

$$z \geq |S|(|S| - s)t^n/14.$$

Combining this with the lower bound on the number of bad triples we get

$$28\epsilon|S| \geq |S| - s.$$

Therefore we conclude that we either have large error $\epsilon > 1/56$ or else we have $|S| \leq 2s < 2t^n/n^{n/2}$. As we have $|S| = t^n/2^{c+1}$ the latter possibility implies

$$c \geq n \log n/2 - 2.$$

Summarizing we have the following.

Theorem 2. *A single round probabilistic protocol for EE_n with error probability $1/3$ has complexity $\Omega(n \log n)$.*

A single round deterministic protocol for EE_n that errs on at most $1/56$ fraction of the inputs has complexity at least $n \log n/2 - 2$.

IV. An isoperimetric inequality on the discrete grid

The isoperimetric problem on the Boolean cube $\{0,1\}^n$ proved extremely useful in theoretical computer science. The problem is to determine the set $S \subseteq \{0,1\}^n$ of a fixed cardinality with the smallest "perimeter", or more generally, to establish connection between the size of a set and the size of its boundary. Here the boundary can be defined in several ways. Considering the Boolean cube as a graph where vertices of Hamming distance 1 are connected, the *edge boundary* of a set S is defined as the set of edges connecting S and its complement, while the *vertex boundary* consists of the vertices outside S having a neighbor in S.

Harper [21] showed that the vertex boundary of a Hamming ball is smallest among all sets of equal size, and the same holds for the edge boundary of a subcube. These results can be generalized to other cardinalities [23]; see the survey by Bezrukov [8].

Consider the metric space over the set $[t]^n$ endowed with the Hamming distance. Let f be a concave function on the nonnegative integers and $1 \le M < n$ be an integer. We consider the following value as a generalized perimeter of a set $S \subseteq [t]^n$:

$$\mathop{\mathbb{E}}_{x \sim \mu}[f(|B_M(x) \cap S|)],$$

where $B_M(x) = \{y \in [t]^n \mid \mathrm{Match}(x, y) \ge M\}$ is the radius $n - M$ Hamming ball around x. Note that when $M = n - 1$ and f is the counting function given as $f(0) = 0$ and $f(l) = 1$ for $l > 0$ (which is concave), the above quantity is exactly the normalized size of the vertex boundary of S. For other concave functions f and parameters M this quantity can still be considered a measure of how "spread out" the set S is.

We start the technical part of this section by introducing the notation we will use. For $x, y \in [t]^n$ and $i \in [n]$ we write $x \sim_i y$ if $x_j = y_j$ for $j \in [n] \setminus \{i\}$. Observe that $\sim_i$ is an equivalence relation. A set $K \subseteq [t]^n$ is called an *i-ideal* if $x \sim_i y$, $x_i < y_i$ and $y \in K$ implies $x \in K$. We call a set $K \subseteq [t]^n$ an *ideal* if it is an i-ideal for all $i \in [n]$.

For $i \in [n]$ and $x \in [t]^n$ we define $\mathrm{down}_i(x) = (x_1, \ldots, x_{i-1}, x_i - 1, x_{i+1}, \ldots, x_n)$. We have $\mathrm{down}_i(x) \in [t]^n$ whenever $x_i > 1$. Let $K \subseteq [t]^n$ be a set, $i \in [n]$ and $2 \le a \in [t]$. For $x \in K$, we define $\mathrm{down}_{i,a}(x, K) = \mathrm{down}_i(x)$ if $x_i = a$ and $\mathrm{down}_i(x) \notin K$ and we set $\mathrm{down}_{i,a}(x, K) = x$ otherwise. We further define $\mathrm{down}_{i,a}(K) = \{\mathrm{down}_{i,a}(x, K) \mid x \in K\}$. For $K \subseteq [t]^n$ and $i \in [n]$ we define

$$\mathrm{down}_i(K) = \bigl\{y \in [t]^n \mid y_i \le |\{z \in K \mid y \sim_i z\}|\bigr\}.$$

Finally for $K \subseteq [t]^n$ we define

$$\mathrm{down}(K) = \mathrm{down}_1(\mathrm{down}_2(\ldots \mathrm{down}_n(K) \ldots)).$$

The following lemma states few simple observations about these down operations.

Lemma 3. *Let $K \subseteq [t]^n$ be a set and let $i, j \in [n]$ be integers. The following hold.*

(i) $\mathrm{down}_i(K)$ *can be obtained from K by applying several operations* $\mathrm{down}_{i,a}$.

(ii) $|\mathrm{down}_{i,a}(K)| = |K|$ *for each* $2 \le a \le t$, $|\mathrm{down}_i(K)| = |K|$ *and* $|\mathrm{down}(K)| = |K|$.

(iii) $\mathrm{down}_i(K)$ *is an i-ideal and if K is a j-ideal, then* $\mathrm{down}_i(K)$ *is also a j-ideal.*

(iv) $\mathrm{down}(K)$ *is an ideal. For any* $x \in \mathrm{down}(K)$ *we have* $P := [x_1] \times [x_2] \times \cdots \times [x_n] \subseteq \mathrm{down}(K)$ *and there exists a set $T \subseteq K$ with* $P = \mathrm{down}(T)$.

Proof: For statement (i) notice that as long as K is not an i-ideal one of the operations $\mathrm{down}_{i,a}$ will not fix K and hence will decrease $\sum_{x \in K} x_i$. Thus a finite sequence of these operations will transform K into an i-ideal. It is easy to see that the operations $\mathrm{down}_{i,a}$ preserve the number of elements in each equivalence class of $\sim_i$, thus the i-ideal we arrive at must indeed be $\mathrm{down}_i(K)$.

Statement (ii) follows directly from the definitions of each of these down operations.

The first claim of statement (iii), namely that $\mathrm{down}_i(K)$ is an i-ideal, is trivial from the definition. Now assume $j \ne i$ and K is a j-ideal, $y \in \mathrm{down}_i(K)$ and $y_j > 1$. To see that $\mathrm{down}_i(K)$ is a j-ideal it is enough to prove that $\mathrm{down}_j(y) \in \mathrm{down}_i(K)$. Since $y \in \mathrm{down}_i(K)$, there are y_i distinct vectors $z \in K$ that satisfy $z \sim_i y$. Considering the vectors $\mathrm{down}_j(z) \sim_i \mathrm{down}_j(y)$ and using that these distinct vectors are in the j-ideal K proves that $\mathrm{down}_j(y)$ is indeed contained in $\mathrm{down}_i(K)$.

By statement (iii), $\mathrm{down}(K)$ is an i-ideal for each $i \in [n]$. Therefore $\mathrm{down}(K)$ is an ideal and the first part of statement (iv), that is, $P \subseteq K'$ follows. We prove the existence of suitable T by induction on the dimension n. The base case $n = 0$ (or even $n = 1$) is trivial. For the inductive step consider $K' = \mathrm{down}_2(\mathrm{down}_3(\ldots \mathrm{down}_n(K) \ldots))$. As $x \in \mathrm{down}(K) = \mathrm{down}_1(K')$, we have distinct vectors $x^{(k)} \in K'$ for $k = 1, \ldots, x_1$, satisfying $x^{(k)} \sim_1 x$. Notice that the construction of K' from K is performed independently on each of the $(n-1)$-dimensional "hyperplanes" $S^l = \{y \in [t]^n \mid y_1 = l\}$ as none of the operations $\mathrm{down}_2, \ldots, \mathrm{down}_n$ change the first coordinate of the vectors. We apply the inductive hypothesis to obtain the sets $T^{(k)} \subseteq S^{x_1^{(k)}} \cap K$ such that $\mathrm{down}_2(\ldots \mathrm{down}_n(T^{(k)}) \ldots) = \{x_1^{(k)}\} \times [x_2] \times \cdots \times [x_n]$. Using again that these sets are in distinct hyperplanes and the operations $\mathrm{down}_2, \ldots, \mathrm{down}_n$ act separately on the hyperplanes S^l, we get for $T := \cup_{k=1}^{x_1} T^{(k)}$ that

$$\mathrm{down}_2(\ldots \mathrm{down}_n(T) \ldots) = \{x_1^{(k)} \mid k \in [x_1]\} \times [x_2] \times \cdots \times [x_n].$$

Applying down_1 on both sides finishes the proof of this last part of the lemma. ■

For sets $x \in [t]^n$, $I \subseteq [n]$, and integer $M \in [n]$ we define $B_{I,M}(x) = \{y \in [t]^n \mid \mathrm{Match}(x_I, y_I) \ge M\}$. The projection of $B_{I,M}$ to the coordinates in I is the Hamming ball of radius $|I| - M$ around the projection of x.

Lemma 4. *Let $I \subseteq [n]$, $M \in [n]$ and let f be a concave function on the nonnegative integers. For arbitrary $K \subseteq [t]^n$*

we have

$$\mathop{\mathbb{E}}_{x\sim\mu}[f(|B_{I,M}(x)\cap \operatorname{down}(K)|)] \le \mathop{\mathbb{E}}_{x\sim\mu}[f(|B_{I,M}(x)\cap K|)].$$

Proof: By Lemma 3(i), the set $\operatorname{down}(K)$ can be obtained from K by a series of operations $\operatorname{down}_{i,a}$ with various $i\in[n]$ and $2\le a\le t$. Therefore, it is enough to prove that the expectation in the lemma does not increase in any one step. Let us fix $i\in[n]$ and $2\le a\le t$. We write $N_x = B_{I,M}(x)\cap K$ and $N'_x = B_{I,M}(x)\cap \operatorname{down}_{i,a}(K)$ for $x\in[t]^n$. We need to prove that

$$\mathop{\mathbb{E}}_{x\sim\mu}[f(|N_x|)] \ge \mathop{\mathbb{E}}_{x\sim\mu}[f(|N'_x|)].$$

Note that $|N_x| = |N'_x|$ whenever $i\notin I$ or $x_i\notin\{a,a-1\}$. Thus, we can assume $i\in I$ and concentrate on $x\in[t]^n$ with $x_i\in\{a,a-1\}$. It is enough to prove $f(|N_x|)+f(|N_y|)\ge f(|N'_x|)+f(|N'_y|)$ for any pair of vectors $x,y\in[t]^n$, satisfying $x_i=a$, and $y=\operatorname{down}_i(x)$.

Let us fix such a pair x,y and set $C=\{z\in K\setminus \operatorname{down}_{i,a}(K)\mid \operatorname{Match}(x_I,z_I)=M\}$. Observe that $N_x = N'_x\cup C$ and $N'_x\cap C=\emptyset$. Similarly, observe that $N'_y = N_y\cup \operatorname{down}_{i,a}(C)$ and $N_y\cap \operatorname{down}_{i,a}(C)=\emptyset$. Thus we have $|N'_x| = |N_x|-|C|$ and $|N'_y| = |N_y| + |\operatorname{down}_{i,a}(C)| = |N_y|+|C|$.

The inequality $f(|N_x|)+f(|N_y|)\ge f(|N'_x|)+f(|N'_y|)$ follows now from the concavity of f, the inequalities $|N'_x|\le |N_y|\le |N'_y|$ and the equality $|N_x|+|N_y| = |N'_x|+|N'_y|$. Here the first inequality follows from $\operatorname{down}_{i,a}(N'_x)\subseteq \operatorname{down}_{i,a}(N_y)$, the second inequality and the equality comes from the observations of the previous paragraph. ■

Lemma 5. *Let $K\subseteq[t]^n$ be arbitrary. There exists a vector $x\in K$ having at least $n/5$ coordinates that are greater than $k:=\frac{t}{2}\mu(K)^{5/(4n)}$.*

Proof: See the full version [43]. ■

Theorem 3. *Let S be an arbitrary subset of $[t]^n$. Let $k=\frac{t}{2}\mu(S)^{5/(4n)}$ and $M=nk/(20t)$. There exists a subset $T\subset S$ of size $k^{n/5}$ and $I\subset[n]$ of size $n/5$ such that, defining $N_x=\{x'\in T\mid \operatorname{Match}(x_I,x'_I)\ge M\}$, we have*

(i) $\Pr_{x\sim\mu}[N_x=\emptyset]\le 5^{-M}$ *and*

(ii) $\mathbb{E}_{x\sim\mu}[\log|N_x|] \ge (n/5-M)\log k - n\log k/5^M$, *where we take* $\log 0=-1$ *to make the above expectation exist.*

Proof: By Lemma 3(ii), we have $|\operatorname{down}(S)|=|S|$. By Lemma 5, there exists an $x\in\operatorname{down}(S)$ having at least $n/5$ coordinates that are greater than k. Let $I\subset[n]$ be a set of $n/5$ coordinates such that $x_i\ge k$ for a fixed $x\in\operatorname{down}(S)$. By Lemma 3(iv), $\operatorname{down}(S)$ is an ideal and thus it contains the set $P=\prod_i P_i$, where $P_i=[k]$ for $i\in I$ and $P_i=\{1\}$ for $i\notin I$. Also by Lemma 3(iv), there exists a $T\subseteq S$ such that $P=\operatorname{down}(T)$. We fix such a set T. Clearly, $|T|=k^{n/5}$.

For a vector $x\in[t]^n$, let $h(x)$ be the number of coordinates $i\in I$ such that $x_i\in[k]$. Note that $\mathbb{E}_{x\sim\mu}[h(x)]=4M$ and $h(x)$ has a binomial distribution. By the Chernoff bound we have $\Pr_{x\sim\mu}[h(x)<M]<5^{-M}$. For x with $h(x)\ge M$ we have $|B_{I,M}(x)\cap P|\ge k^{n/5-M}$, but for $h(x)<M$ we have $B_{I,M}(x)\cap P=\emptyset$. With the unusual convention $\log 0=-1$ we have

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mu}&[\log|B_{I,M}(x)\cap P|]\\ &\ge \Pr[h(x)\ge M](n/5-M)\log k - \Pr[h(x)<M]\\ &> (n/5-M)\log k - n\log k/5^M\end{aligned}$$

We have $\operatorname{down}(T)=P$ and our unusual $\log$ is concave on the nonnegative integers, so Lemma 4 applies and proves statement (ii):

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mu}[\log|N_x|] &\ge \mathop{\mathbb{E}}_{x\sim\mu}[\log|B_{I,M}(x)\cap P|]\\ &\ge (n/5-M)\log k - n\log k/5^M.\end{aligned}$$

To show statement (i), we apply Lemma 4 with the concave function f defined as $f(0)=-1$ and $f(l)=0$ for all $l>0$. We obtain that

$$\begin{aligned}\Pr_{x\sim\mu}[N_x=\emptyset] &= -\mathop{\mathbb{E}}_{x\sim\mu}[f(|N_x|)]\\ &\le -\mathop{\mathbb{E}}_{x\sim\mu}[f(|B_{I,M}(x)\cap P|)]\\ &= \Pr_{x\sim\mu}[B_{I,M}(x)\cap P=\emptyset] < 5^{-M}.\end{aligned}$$

This completes the proof. ■

V. Lower bound for multiple round protocols

In this section we prove our main lower bound result:

Theorem 4. *For any $r\le\log^* n$, an r-round probabilistic protocol for EE_n with error probability at most $1/3$ sends at least one message of size $\Omega(n\log^{(r)}n)$.*

Note that the $r=1$ round case of this theorem was proved as Theorem 2 in Section III. The other extreme, which immediately follows from Theorem 4, is the following.

Corollary 1. *Any probabilistic protocol for EE_n with maximum message size $O(n)$ and error $1/3$ has at least $\log^* n - O(1)$ rounds.*

Theorem 4 is a direct consequence of the corresponding statement on deterministic protocols with small distributional error on uniform distribution; see Theorem 5 at the end of this section. Indeed, we can decrease the error of a randomized protocol below any constant $\epsilon>0$ for the price of increasing the message length by a constant factor, then we can fix the coins of this low error protocol in a way that makes the resulting deterministic protocol Q err in at most ϵ fraction of the possible inputs. Applying Theorem 5 to the protocol Q proves Theorem 4.

In the rest of this section we use round-elimination to prove Theorem 5, that is, we will use Q to solve smaller instances of the exists-equal problem in a way that the first message is always the same, and hence can be eliminated.

Suppose Alice sends the first message of c bits in Q. By Lemma 2, there exists a $S \subset [t]^n$ of size $\mu(S) = 2^{-c-1}$ such that the first message of Alice is fixed when $x \in S$ and we have $\Pr_{y\sim\mu}[Q(x,y) \neq \mathrm{EE}(x,y)] \leq 2\epsilon$ for all $x \in S$. Fix such a set S and let $k := t/2^{\frac{5(c+1)}{4n}+1}$ and $M := nk/(20t)$. By Theorem 3, there exists a $T \subset S$ of size $k^{n/5}$ and $I \subset [n]$ of size $n/5$ such that defining

$$N_x = \{y \in T \mid \mathrm{Match}(x_I, y_I) \geq M\}$$

we have $\Pr_{x\sim\mu}[N_x = \emptyset] \leq 5^{-M}$ and $\mathbb{E}_{x\sim\mu}[\log|N_x|] \geq (n/5 - M)\log k - n\log k/5^M$. Let us fix such sets T and I. Note also that Theorem 3 guarantees that T is a strict subset of S. Designate an arbitrary element of $S \setminus T$ as x'_e.

A. Embedding the smaller problem

The players embed a smaller instance $u, v \in [t']^{n'}$ of the exists-equal problem in EE_n concentrating on the coordinates I determined above. We set $n' := M/10$ and $t' := 4n'$. Optimally, the same embedding should guarantee low error probability for all pairs of inputs, but for technical reasons we need to know the number of coordinate agreements $\mathrm{Match}(u,v)$ for the input pairs (u,v) in the smaller problem having $\mathrm{EE}_{n'}(u,v) = 1$. Let $R \geq 1$ be this number, so we are interested in inputs $u, v \in [t']^{n'}$ with $\mathrm{Match}(u,v) = 0$ or R. We need this extra parameter so that we can eliminate a non-constant number of rounds and still keep the error bound a constant. For results on constant round protocols one can concentrate on the $R = 1$ case.

In order to solve the exist-equal problem with parameters t' and n' Alice and Bob use the joint random source to turn their input $u, v \in [t']^{n'}$ into longer random vectors $X', Y \in [t]^n$, respectively, and apply the protocol Q above to solve this exists-equal problem for these larger inputs. Here we informally list the main requirements on the process generating X' and Y. We require these properties for the random vectors $X', Y \in [t]^n$ generated from a fixed pair $u, v \in [t']^{n'}$ satisfying $\mathrm{Match}(u,v) = 0$ or R.

(P1) $\mathrm{EE}(X', Y) = \mathrm{EE}(u,v)$ with large probability,
(P2) $\mathrm{supp}(X') = T \cup \{x'_e\}$ and
(P3) for most $x' \sim X'$, we have $\mathrm{dist}(Y \mid X' = x')$ is close to uniform distribution on $[t]^n$.

Combining these properties with the fact that $\Pr_{y\sim\mu}[Q(x,y) \neq \mathrm{EE}(x,y)] \leq 2\epsilon$ for each $x \in S$, we will argue that for the considered pairs of inputs $Q(X', Y)$ equals $\mathrm{EE}(u,v)$ with large probability, thus the combined protocol solves the small exists-equal instance with small error, at least for input pairs with $\mathrm{Match}(u,v) = 0$ or R. Furthermore, by Property (P2) the first message of Alice will be fixed and hence does not need to be sent, making the combined protocol one round shorter.

The random variables X' and Y are constructed as follows. Let $m := 2n/(MR)$ be an integer. Each player repeats his or her input (u and v, respectively) m times, obtaining a vector of size $n/(5R)$. Then using the shared randomness, the players pick $n/(5R)$ uniform random maps $m_i : [t'] \to [t]$ independently and apply m_i to ith coordinate. Furthermore, the players pick a uniform random 1-1 mapping $\pi : [n/(5R)] \to I$ and use it to embed the coordinates of the vectors they constructed among the coordinates of the vectors X and Y of length n. The remaining $n - n/(5R)$ coordinates of X is picked uniformly at random by Alice and similarly, the remaining $n - n/(5R)$ coordinates of Y is picked uniformly at random by Bob. Note that the marginal distribution of both X and Y are uniform on $[t]^n$. If $\mathrm{Match}(u,v) = 0$ the vectors X and Y are independent, while if $\mathrm{Match}(u,v) = R$, then Y can be obtained by selecting a random subset of I of cardinality mR, copying the corresponding coordinates of X and filling the rest of Y uniformly at random.

This completes the description of the random process for Bob. However Alice generates one more random variable X' as follows. Recall that $N_x = \{z \in T \mid \mathrm{Match}(z_I, x_I) \geq M\}$. The random variable X' is obtained by drawing $x \sim X$ first and then choosing a uniform random element of N_x. In the (unlikely) case that $N_x = \emptyset$, Alice chooses $X' = x'_e$.

Note that X' either equals x'_e or takes values from T, hence Property (P2) holds. In the next lemma we quantify and prove Property (P1) as well.

Lemma 6. *Assume $n \geq 3$, $M \geq 2$ and $u, v \in [t']^{n'}$. We have*

(i) if $\mathrm{Match}(u,v) = 0$ then $\Pr[\mathrm{EE}(X', Y) = 0] > 0.77$;
(ii) if $\mathrm{Match}(u,v) = R$, then $\Pr[\mathrm{EE}(X', Y) = 1] \geq 0.80$.

Proof: For the first claim, note that when $\mathrm{Match}(u,v) = 0$, the random variables X and Y are independent and uniformly distributed. We construct X' based on X, so its value is also independent of Y. Hence $\Pr[\mathrm{EE}(X', Y) = 0] = (1 - 1/t)^n$. This quantity goes to $e^{-1/4}$ since $t = 4n$ and is larger than 0.77 when $n \geq 3$. This establishes the first claim.

For the second claim let $J = \{i \in I \mid X_i = Y_i\}$ and $K = \{i \in I \mid X'_i = X_i\}$. By construction, $|J| = \mathrm{Match}(X_I, Y_I) \geq mR$ and $|K| = \mathrm{Match}(X'_I, X_I) \geq M$ unless $N_X = \emptyset$. By our construction, each $J \subset I$ of the same size is equally likely by symmetry, even when we condition on a fix value of X and X'. Thus we have $\mathbb{E}[|J \cap K| \mid N_X \neq \emptyset] \geq mRM/|I| = 10$ and $\Pr[J \cap K = \emptyset \mid N_X \neq \emptyset] < e^{-10}$. Note that X is distributed uniformly over $[t]^n$, therefore by Theorem 3(i) the probability that $N_X = \emptyset$ is at most 5^{-M}. Note that $\mathrm{Match}(X', Y) \geq |J \cap K|$ and thus $\Pr[\mathrm{EE}(X', Y) = 0] \leq \Pr[J \cap K = \emptyset] \leq \Pr[J \cap K = \emptyset \mid N_X \neq \emptyset] + \Pr[N_X = \emptyset] \leq e^{-10} + 5^{-M}$. This completes the proof. ■

We measure "closeness to uniformity" in Property (P3) by simply calculating the entropy. This entropy argument is postponed to the next subsection; here we show how such a bound to the entropy implies that the error introduced by Q is small.

Lemma 7. *Let $x' \in S$ be fixed and let γ be a probability in the range $2\epsilon \leq \gamma < 1$. If $\mathrm{H}(Y \mid X' = x') \geq n \log t - \mathbf{D}_2(\gamma \,\|\, 2\epsilon)$ then $\Pr_{y \sim Y \mid X' = x'}[Q(x', y) \neq \mathrm{EE}(x', y)] \leq \gamma$.*

Proof: For a distribution ν over $[t]^n$, let $e(\nu) = \Pr_{y \sim \nu}[Q(x', y) \neq \mathrm{EE}(x', y)]$. We prove the contrapositive of the statement of the lemma, that is assuming $\Pr_{y \sim Y \mid X' = x'}[Q(x', y) \neq \mathrm{EE}(x', y)] > \gamma$ we prove $\mathrm{H}(Y \mid X' = x') < n \log t - \mathbf{D}_2(\gamma \,\|\, 2\epsilon)$:

$$\begin{aligned} n \log t - \mathrm{H}(Y \mid X' = x') &= \mathbf{D}(\mathrm{dist}(Y \mid X' = x') \,\|\, \mu) \\ &\geq \mathbf{D}_2(e(\mathrm{dist}(Y \mid X' = x')) \,\|\, e(\mu)) \\ &\geq \mathbf{D}_2(\gamma \,\|\, 2\epsilon), \end{aligned}$$

where the first inequality follows from the chain rule for the Kullback-Leibler divergence. ■

B. Establishing Property (P3)

We quantify Property (P3) using the conditional entropy $\mathrm{H}(Y \mid X')$. If $\mathrm{Match}(u, v) = R$ our process generates X and Y with the expected number $\mathbb{E}[\mathrm{Match}(X_I, Y_I)]$ of matches only slightly more than the minimum mR. We lose most of these matches with Y when we replace X by X' and only an expected constant number remains. A constant number of forced matches with X' within I restricts the number of possible vectors Y but it only decreases the entropy by $O(1)$. The calculations in this subsection make this intuitive argument precise.

Lemma 8. *Let X', Y be as constructed above. The following hold.*

(i) If $\mathrm{Match}(u, v) = 0$ we have $\mathrm{H}(Y \mid X') = n \log t$.

(ii) If $M > 100 \log n$ and $\mathrm{Match}(u, v) = R$ we have $\mathrm{H}(Y \mid X') = n \log t - O(1)$.

Proof: Part (i) holds as Y is uniformly distributed and independent of X' whenever $\mathrm{EE}(u, v) = 0$.

For part (ii) recall that if $\mathrm{Match}(u, v) = R$ one can construct X and Y by uniformly selecting a size mR set $L \subseteq I$ and selecting X and Y uniformly among all pairs satisfying $X_L = Y_L$. Recall that L is the set of coordinates the mR matches between u^m and v^m were mapped. These are the "intentional matches" between X_I and Y_I. Note that there may be also "unintended matches" between X_I and Y_I, but not too many: their expected number is $(n/5 - mR)/t < 1/20$. As given any fixed L, the marginal distribution of both X and Y are still uniform, so in particular X is independent of L and so is X' constructed from X. Therefore we have

$$\mathrm{H}(Y \mid X') = \mathrm{H}(Y \mid X', L) + \mathrm{H}(L) - \mathrm{H}(L \mid Y, X').$$

We treat the terms separately. First we split the first term:

$$\mathrm{H}(Y \mid X', L) = \mathrm{H}(Y_L \mid X', L) + \mathrm{H}(Y_{[n] \setminus L} \mid X', L, Y_L)$$

and use that $Y_{[n] \setminus L}$ is uniformly distributed for any fixed L, X' and Y_L, making

$$\mathrm{H}(Y_{[n] \setminus L} \mid X', L, Y_L) = (n - mR) \log t.$$

We have $X_L = Y_L$, thus

$$\begin{aligned} \mathrm{H}(Y_L \mid X', L) &= \mathrm{H}(X_L \mid X', L) \\ &\geq \frac{mR}{n/5} \mathrm{H}(X_I \mid X') \\ &\geq mR \log t - 10 \log k - \frac{MR}{5^{M-1}} \log k, \end{aligned}$$

where the first inequality follows by Lemma 1 as L is a uniform and independent of X and X' and the second inequality follows from Lemma 9 that we will prove shortly and the formula defining m.

The next term, $\mathrm{H}(L)$ is easy to compute as L is a uniform subset of I of size mR:

$$\mathrm{H}(L) = \log \binom{n/5}{mR}$$

It remains to bound the term $\mathrm{H}(L \mid Y, X')$. Let $Z = \{i \mid i \in I \text{ and } X'_i = Y_i\}$. Note that Z can be derived from X', Y (as I is fixed) hence $\mathrm{H}(L \mid Y, X') \leq \mathrm{H}(L \mid Z)$. Further, let $C = |Z \setminus L|$. We obtain

$$\begin{aligned} \mathrm{H}(L \mid Y, X') &\leq \mathrm{H}(L \mid Z) \leq \mathrm{H}(L \mid Z, C) + \mathrm{H}(C) \\ &< \mathop{\mathbb{E}}_{Z,C}\left[\log \binom{n/5 - |Z| + C}{mR - |Z| + C}\right] + \mathop{\mathbb{E}}_{Z,C}\left[\log \binom{|Z|}{C}\right] + 2 \end{aligned}$$

where we used $\mathrm{H}(C) < 2$. Note that for any fixed $x' \in T$ and $x \in \mathrm{supp}(X \mid X' = x')$, we have

$$\mathbb{E}[|Z| - C \mid X = x, X' = x'] = \mathrm{Match}(x_I, x'_I) mR / (n/5) \geq 10$$

as $\mathrm{Match}(x_I, x'_I) \geq M$ by definition.

Hence we have

$$\log \binom{n/5}{mR} - \log \binom{n/5 - |Z| + |C|}{mR - |Z| + |C|} \geq 10 \log \frac{n}{5m} - O(1),$$

$$\mathop{\mathbb{E}}_{Z,C}\left[\log \binom{|Z|}{C}\right] \leq \mathbb{E}[|Z|] < 20.$$

Summing the estimates above for the various parts of $\mathrm{H}(Y \mid X')$ the statement of the lemma follows. ■

It remains to prove the following simple lemma that "reverses" the conditional entropy bound in Theorem 3(ii):

Lemma 9. *For any $u, v \in [t']^{n'}$ we have $\mathrm{H}(X_I \mid X') \geq \frac{n}{5} \log t - M \log k - n \log k / 5^M$.*

Proof: Using the fact that $\mathrm{H}(A, B) = \mathrm{H}(A \mid B) + \mathrm{H}(B) = \mathrm{H}(B \mid A) + \mathrm{H}(A)$ we get

$$\begin{aligned} \mathrm{H}(X_I \mid X') &= \mathrm{H}(X' \mid X_I) + \mathrm{H}(X_I) - \mathrm{H}(X') \\ &\geq \frac{n}{5} \log t + \mathrm{H}(X' \mid X_I) - \frac{n}{5} \log k, \end{aligned}$$

where in the last step we used $\mathrm{H}(X') \leq \log|\operatorname{supp}(X')| = \log|T| = \frac{n}{5}\log k$ and $\mathrm{H}(X_I) = (n/5)\log t$ as X is uniformly distributed.

Observe that $\mathrm{H}(X' \mid X_I) = \mathrm{H}(X' \mid X) = \mathbb{E}_{x\sim\mu}[\log|N_x|]$, where $\log 0$ is now taken to be 0. From Theorem 3(ii) we get $\mathrm{H}(X' \mid X) \geq \frac{n}{5}\log k - M\log k - n\log k/5^M$ finishing the proof of the lemma. ■

C. The round elimination lemma

Let ν_n be the uniform distribution on $[t]^n \times [t]^n$, where we set $t = 4n$. The following lemma gives the base case of the round elimination argument.

Lemma 10. *Any 0-round deterministic protocol for* EE_n *has at least 0.22 distributional error on* ν_n*, when* $n \geq 1$.

Proof: The output of the protocol is decided by a single player, say Bob. For any given input $y \in [t]^n$ we have $3/4 \leq \Pr_{x\sim\mu}[\mathrm{EE}(x,y) = 0] < e^{-1/4} < 0.78$. Therefore the distributional error is at least 0.22 for any given y regardless of the output Bob chooses, thus the overall error is also at least 0.22. ■

The next is the statement of our full round elimination lemma followed by the main theorem of this section. The proofs are omitted due to space reasons; see the full version [43].

Lemma 11. *Let* $r > 0, c, n$ *be an integers such that* $c < (n\log n)/2$*. There is a constant* $0 < \epsilon_0 < 1/200$ *such that if there is an* r*-round deterministic protocol with* c*-bit messages for* EE_n *that has* ϵ_0 *error on* ν_n*, then there is an* $(r-1)$*-round deterministic protocol with* $O(c)$*-bit messages for* $\mathrm{EE}_{n'}$ *that has* ϵ_0 *error on* $\nu_{n'}$*, where* $n' = \Omega(n/2^{\frac{5c}{4n}})$.

Theorem 5. *There exists a constant* ϵ_0 *such that for any* $r \leq \log^* n$*, an* r*-round deterministic protocol for* EE_n *which has* ϵ_0 *error on* ν_n *sends at least one message of size* $\Omega(n\log^{(r)} n)$.

Acknowledgments

We would like to thank Hossein Jowhari for many stimulating discussions during the early stages of this work.

References

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

[2] M. Azizoğlu and Ö. Öğecioğlu. Extremal sets minimizing dimension-normalized boundary in Hamming graphs. *SIAM Journal on Discrete Mathematics*, 17(2):219–236, 2003.

[3] László Babai, Peter Frankl, and Janos Simon. Complexity classes in communication complexity theory (preliminary version). In *FOCS*, pages 337–347. IEEE Computer Society, 1986.

[4] 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.

[5] Boaz Barak, Mark Braverman, Xi Chen, and Anup Rao. How to compress interactive communication. In *STOC*, pages 67–76, 2010.

[6] Paul Beame and Faith E. Fich. Optimal bounds for the predecessor problem and related problems. *Journal of Computer and System Sciences*, 65:2002, 2001.

[7] Avraham Ben-Aroya, Oded Regev, and Ronald de Wolf. A hypercontractive inequality for matrix-valued functions with applications to quantum computing and ldcs. In *FOCS*, pages 477–486. IEEE Computer Society, 2008.

[8] Sergei Bezrukov. Isoperimetric problems in discrete spaces. In *Bolyai Soc. Math. Stud*, pages 59–91, 1994.

[9] Béla Bollobás and Imre Leader. Edge-isoperimetric inequalities in the grid. *Combinatorica*, 11(4):299–314, 1991.

[10] Joshua Brody and Amit Chakrabarti. A multi-round communication lower bound for gap Hamming and some consequences. In *IEEE Conference on Computational Complexity*, pages 358–368. IEEE Computer Society, 2009.

[11] Joshua Brody, Amit Chakrabarti, and Ranganath Kondapally. Certifying equality with limited interaction. *Electronic Colloquium on Computational Complexity (ECCC)*, 19:153, 2012.

[12] Joshua Brody, Amit Chakrabarti, Oded Regev, Thomas Vidick, and Ronald de Wolf. Better Gap-Hamming lower bounds via better round elimination. In Maria J. Serna, Ronen Shaltiel, Klaus Jansen, and José D. P. Rolim, editors, *APPROX-RANDOM*, volume 6302 of *Lecture Notes in Computer Science*, pages 476–489. Springer, 2010.

[13] Harry Buhrman, David Garcia-Soriano, Arie Matsliah, and Ronald de Wolf. The non-adaptive query complexity of testing k-parities, 2012.

[14] Amit Chakrabarti, Yaoyun Shi, Anthony Wirth, and Andrew Chi-Chih Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In *FOCS*, pages 270–278. IEEE Computer Society, 2001.

[15] F. R. Chung, R. L. Graham, P. Frankl, and J. B. Shearer. Some intersection theorems for ordered sets and graphs. *J. Comb. Theory Ser. A*, 43(1):23–37, September 1986.

[16] G. F. Clements. Sets of lattice points which contain a maximal number of edges. *Proc. Amer. Math. Soc.*, 27:13–15, 1971.

[17] Anirban Dasgupta, Ravi Kumar, and D. Sivakumar. Sparse and lopsided set disjointness via information theory. In *APPROX-RANDOM*, pages 517–528, 2012.

[18] Pavol Duris, Zvi Galil, and Georg Schnitger. Lower bounds on communication complexity. *Inf. Comput.*, 73(1):1–22, 1987.

[19] Dmitry Gavinsky. On the role of shared entanglement. *Quantum Information & Computation*, 8(1):82–95, 2008.

[20] Bernd Halstenberg and Rüdiger Reischuk. On different modes of communication (extended abstract). In Janos Simon, editor, *STOC*, pages 162–172. ACM, 1988.

[21] L.H. Harper. Optimal assignment of numbers to vertices. *J. Soc. Ind. Appl. Math.*, 12:131–135, 1964.

[22] Prahladh Harsha, Rahul Jain, David A. McAllester, and Jaikumar Radhakrishnan. The communication complexity of correlation. *IEEE Transactions on Information Theory*, 56(1):438–449, 2010.

[23] Sergiu Hart. A note on the edges of the n-cube. *Discrete Mathematics*, 14(2):157 – 163, 1976.

[24] Johan Håstad and Avi Wigderson. The randomized communication complexity of set disjointness. *Theory of Computing*, 3(1):211–219, 2007.

[25] Rahul Jain, Hartmut Klauck, and Ashwin Nayak. Direct product theorems for classical communication complexity via subdistribution bounds: extended abstract. In Cynthia Dwork, editor, *STOC*, pages 599–608. ACM, 2008.

[26] Rahul Jain, Jaikumar Radhakrishnan, and Pranab Sen. A direct sum theorem in communication complexity via message compression. In Jos C. M. Baeten, Jan Karel Lenstra, Joachim Parrow, and Gerhard J. Woeginger, editors, *ICALP*, volume 2719 of *Lecture Notes in Computer Science*, pages 300–315. Springer, 2003.

[27] Rahul Jain, Jaikumar Radhakrishnan, and Pranab Sen. Prior entanglement, message compression and privacy in quantum communication. In *IEEE Conference on Computational Complexity*, pages 285–296. IEEE Computer Society, 2005.

[28] T. S. Jayram and David P. Woodruff. The data stream space complexity of cascaded norms. In *FOCS*, pages 765–774. IEEE Computer Society, 2009.

[29] Bala Kalyanasundaram and Georg Schnitger. The probabilistic communication complexity of set intersection. *SIAM J. Discrete Math.*, 5(4):545–557, 1992.

[30] Mauricio Karchmer and Avi Wigderson. Monotone circuits for connectivity require super-logarithmic depth. *SIAM J. Discrete Math.*, 3(2):255–265, 1990.

[31] D. L. Kleitman, M. M. Krieger, and B. L. Rotschild. Configurations maximizing the number of pairs of Hamming-adjacent lattice points. *Studies in Appl. Math.*, 50:115–119, 1971.

[32] Eyal Kushilevitz and Noam Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[33] John H. Lindsey. Assignment of numbers to vertices. *The American Mathematical Monthly*, 71(5):508–516, 1964.

[34] Frédéric Magniez, Claire Mathieu, and Ashwin Nayak. Recognizing well-parenthesized expressions in the streaming model. In *STOC*, pages 261–270, 2010.

[35] Peter Bro Miltersen. Lower bounds for union-split-find related problems on random access machines, 1994.

[36] Peter Bro Miltersen, Noam Nisan, Shmuel Safra, and Avi Wigderson. On data structures and asymmetric communication complexity. *J. Comput. Syst. Sci.*, 57(1):37–49, 1998.

[37] Marco Molinaro, David Woodruff, and Grigory Yaroslavtsev. Beating the direct sum theorem in communication complexity with implications for sketching. In *Proceedings of 24th ACM-SIAM Symposium on Discrete Algorithms*, pages 1486–1502, 2013.

[38] Noam Nisan and Avi Wigderson. Rounds in communication complexity revisited. *SIAM J. Comput.*, 22(1):211–219, 1993.

[39] Itzhak Parnafes, Ran Raz, and Avi Wigderson. Direct product results and the gcd problem, in old and new communication models. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, STOC '97, pages 363–372, New York, NY, USA, 1997. ACM.

[40] Mihai Pătraşcu. Cc4: One-way communication and a puzzle. http://infoweekly.blogspot.com/2009/04/cc4-one-way-communication-and-puzzle.html. Accessed: 31/03/2013.

[41] Mihai Pătraşcu. Unifying the landscape of cell-probe lower bounds. *SIAM J. Comput.*, 40(3):827–847, 2011.

[42] Alexander A. Razborov. On the distributional complexity of disjointness. *Theor. Comput. Sci.*, 106(2):385–390, 1992.

[43] Mert Saglam and Gábor Tardos. On the communication complexity of sparse set disjointness and exists-equal problems. *CoRR*, abs/1304.1217, 2013.

[44] Pranab Sen. Lower bounds for predecessor searching in the cell probe model. In *IEEE Conference on Computational Complexity*, pages 73–83. IEEE Computer Society, 2003.

[45] David P. Woodruff. personal communication, 2008.