THEORY OF COMPUTING, Volume 3 (2007), pp. 211–219
http://theoryofcomputing.org

The Randomized Communication Complexity of Set Disjointness

Johan Håstad[*] Avi Wigderson[†]

Received: July 2, 2007; published: October 15, 2007.

Abstract: We study the communication complexity of the disjointness function, in which each of two players holds a k-subset of a universe of size n and the goal is to determine whether the sets are disjoint. In the model of a common random string we prove that $O(k)$ communication bits are sufficient, regardless of n. In the model of private random coins $O(k+\log\log n)$ bits suffice. Both results are asymptotically tight.

1 Introduction

Communication complexity, introduced by Yao [13], is an extremely basic and useful model which has been widely studied [6]. Set disjointness is perhaps the most studied problem in this model, and its complexity has been used for such diverse applications as circuit complexity (e. g. [11]) and auction theory (e. g. [9]).

Here we give a simple protocol, showing that, in the model of common random coins, the probabilistic communication complexity of disjointness depends only on the sizes of the sets, and not on the

[*]Royal Institute of Technology, Stockholm.
[†]School of Mathematics, Institute for Advanced Study, supported by NSF grant CCR-0324906.

ACM Classification: F.2.2

AMS Classification: 68Q25

Key words and phrases: communication complexity, randomized protocol, set disjointness

 DOI: 10.4086/toc.2007.v003a011

size of the universe from which they are taken. This contrasts with the deterministic model, in which dependence on the universe size cannot be avoided.

We now briefly define the model; more details can be found in the excellent book [6].

Let $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$. Alice and Bob are given respectively x and y from $\{0,1\}^n$, and need jointly to compute $f(x,y)$. They alternate sending bits to each other according to a pre-specified protocol and then announce a bit. In a deterministic protocol, their answer must always be correct, i. e., equal to $f(x,y)$ for every input pair x,y. The deterministic communication complexity $D(f)$ is the minimum number of bits the players exchange on the worst case input in the best deterministic protocol for f.

In a probabilistic protocol we have the common randomness model where Alice and Bob share an infinite string of independent, unbiased coin tosses, and are required to give the correct answer with a probability bounded away from 1/2 on *every* input. As we mostly ignore constant factors, the exact probability of being correct is not important but for concreteness let us assume that the probability of being correct is at least $2/3$. The probabilistic communication complexity $R(f)$ is the minimum number of coin tosses used plus bits exchanged by the players on the worst case input and coin tosses in the best probabilistic protocol for f.

In the possibly more realistic private coins model each player can generate his/her own randomness. By a result of Newman [7], any problem with T different inputs can also be solved in this model, adding only $O(\log\log T)$ communication bits to that of the common randomness model. In view of this result we do our analysis in the model of common randomness.

We are also interested in what is commonly called the "Las Vegas"-type probabilistic algorithms where the answer is required always to be correct but the complexity measure is the expected number of bits exchanged. We denote this complexity by $R_0(f)$.

Let DISJ^n denote the disjointness function, namely $\text{DISJ}^n(S,T) = 1$ iff $S \cap T = \emptyset$ (the inputs S,T are given by their characteristic vectors). Let DISJ^n_k denote the restriction of this functions to inputs sets S,T which are both of size k.

It is not difficult to see that the deterministic complexity is lower bounded by the logarithm of the rank of corresponding game matrix [6] and the following lower bounds follow from lower bounds on the rank of the disjointness matrix. For a proof of the rank lower bound, see [4], page 175.

Theorem 1.1.

1. $D(\text{DISJ}^n) = \Theta(n)$.
2. $D(\text{DISJ}^n_k) = \Theta(\log \binom{n}{k})$ *for every* $k \le n/2$.

The probabilistic complexity is far more subtle. A first lower bound of $\Omega(\sqrt{n})$ when $k = n/3$ was proved by Babai et al. [1]. This bound was strengthened by Kalyanasundaram and Schnitger [5], simplified by Razborov [12], and further simplified by Bar-Yossef et al. [3], yielding the following theorem.

Theorem 1.2. *[5, 12, 3]*

1. $R(\text{DISJ}^n) = \Theta(n)$.
2. *For any* $c < 1/2$, $R(\text{DISJ}^n_k) = \Omega(k)$ *for every* $k \le cn$.

Proof. (Sketch) The lower bound $\Omega(n)$ for the first part is given in the quoted papers. The upper bound is trivial as either party can just send its input. For the same reason the lower bound happens for $k_0 = c_0 n$ for some n.

The lower bound for the second part is obtained for $k \le k_0$ by setting $n_0 = k/c_0$ and studying the problem when only the first n_0 elements are allowed to be in the set. For larger values of k we fix a value d and let the first d elements be in the first set and not in the second set while the following d elements are in the second set but not in the first. This reduces the original problem to a problem with $(k-d)$-element sets and a universe size of $n-2d$. Selecting d such $k-d = c_0(n-2d)$ makes it possible to apply the first lower bound. □

The gap between the deterministic (and thus probabilistic) upper bound of Theorem 1.1 and probabilistic lower bound of Theorem 1.2 for DISJ_k^n naturally raises the question what is the probabilistic complexity for $k = o(n)$.

In this paper we prove that the lower bound is tight for all k, and in particular the probabilistic complexity is independent of the universe size n.

Theorem 1.3. *In the model of common randomness, $R(\mathrm{DISJ}_k^n) = O(k)$ for all k.*

In the next section we prove this theorem for the very special case of constant size sets, i. e., $k = O(1)$. This will both give some motivation as well as the "base case" to the protocol and proof for general k, which we give in Section 4.

By applying the procedure of Newman [7] we get a result for the private coin model.

Theorem 1.4. *In the model of private randomness, $R(\mathrm{DISJ}_k^n) = O(k + \log\log n)$ for all k.*

In Section 6 we establish that the additive term $\log\log n$ is needed.

Finally, looking at Las Vegas protocols, it turns out that the complexity is different for positive and for negative instances. Informally what happens is that to be certain that two sets intersect, we need to know a point in the intersection and this gives an added complexity of $\log n$. The need for this extra term is formally argued in Section 5.

Theorem 1.5. *In the model of common randomness, $R_0(\mathrm{DISJ}_k^n) = O(k)$ for instances of disjoint sets and $R_0(\mathrm{DISJ}_k^n) = O(k + \log n)$ for non-disjoint sets.*

This protocol can also be transformed to the model of private randomness, adding a term $O(\log\log n)$.

As a side remark let us note that these results were proved over 10 years ago, and were since used and referred to in several papers (e. g. [10]). Writing them up was long overdue, but better late than never.

2 Notation

We use standard notation throughout the paper with the exception of the notation $\exp(k)$ which is a function of the form c^k for some constant $c > 1$ which is not specified and might change.

3 Simultaneous protocol for constant k

In this section we prove the following theorem, which holds for all k but will be used later only for a large constant value of k.

Theorem 3.1. *In the model of common randomness, $R(\text{DISJ}_k^n) = O(2^{2k})$ for all k.*

Proof. We actually prove a stronger theorem, namely, we give a *simultaneous* protocol [13, 8, 2] for DISJ_k^n. In such protocols each player sends only one message to a referee, who (even without access to the random string or any of the inputs) can determine the function value with high probability.

The players regard their shared random string as a sequence of $t = c2^{2k}$ vectors of length n, representing the random subsets $Z_1, Z_2, \cdots, Z_t$ of $[n]$. We describe Alice's message first. Assume her input is the subset $S \subseteq [n]$. Alice sends the bits $a_1, \cdots, a_t$, with $a_i = 1$ iff $S \subseteq Z_i$. Bob behaves similarly, only with respect to the complements of the Z_i. If his input is $T \subseteq [n]$, he sends the sequence of bits $b_1, \cdots, b_t$, with $b_i = 1$ iff $T \cap Z_i = \emptyset$. Now the referee answers 1 if for some i we have $a_i = b_i = 1$ and answers 0 otherwise.

It is clear that if $\text{DISJ}_k^n(S,T) = 0$, i. e., S and T intersect, no such index i exists regardless of the random string, and the referee will always give the correct answer. On the other hand, if $\text{DISJ}_k^n(S,T) = 1$, i. e., S and T are disjoint, we will see that the probability that no such index i exists is small.

First note that for a random set Z, the events $S \subseteq Z$ and $T \cap Z = \emptyset$ are independent, since S and T are disjoint, and membership in Z is decided by independent coin tosses for every element in $[n]$. Moreover, the two probabilities are exactly 2^{-k} each. We conclude that the probability that Z does not satisfy both events is $1 - 2^{-2k}$. Thus the probability that all $t = c2^{2k}$ independently chosen subsets Z_i fail to prove the disjointness of S and T is $(1 - 2^{-2k})^t < \exp(-c)$ which we can make arbitrarily small by choosing the constant c sufficiently large. □

4 Proof of Theorem 1.3

First, let us give an intuitive overview of the proof. Assume Alice and Bob are holding, respectively, the sets $S, T \in [n]$, each of size k. They will attempt to construct a proof that their sets are disjoint, in the form of a subset $Z \subseteq [n]$ with $S \subseteq Z$ and $T \subseteq \bar{Z}$. Clearly if they find such a set Z then their inputs are indeed disjoint. We will need to show that if they fail, then with high probability their sets intersect. Later in Section 5 we will modify the protocol never to make mistakes and where randomness is only used to bound the expected number of bits exchanged.

The protocol will proceed in phases, which can be viewed as a series of downward self-reductions of the problem, to the same problem on smaller size sets. More precisely, let $S_0 = S$ and $T_0 = T$ be the inputs to the first phase. Then after phase i the players will hold sets S_i and T_i, respectively, of total size $k_i = |S_i| + |T_i|$, which, unless the protocol has already halted, have the following properties for every $i \geq 1$:

1. S_i and T_i are disjoint iff S_{i-1} and T_{i-1} are;

2. $k_i \leq 7k_{i-1}/8$;

3. The communication used by the two players in phase i is $O(k_i)$.

The players will continue until a phase j for which $k_j < c$, for a constant c which we choose to be the same constant c as in Section 3, at which point they will resort to the protocol of Theorem 3.1 on S_j and T_j. By property (1) it is indeed equivalent to disjointness of the original S and T. If the protocol halts before this happens it will halt with the output "not disjoint." We maintain the property that if the sets are disjoint then the probability of halting with output "not disjoint" in phase i is $\exp(-k_i)$.

As the set sizes, by property (2), form a geometric progression, the probability of ever halting in an early phase with the incorrect answer "not disjoint" is bounded by $\exp(-c)$. Moreover, by property (3), the total communication is $O(\sum_i k_i)$, which is a geometric progression as well, bounded, up to constant factor, by the first term $k_0 = 2k$. Let us fill in the details.

We describe one phase. At the input to the phase, Alice holds S of size $|S| = s$ and Bob holds T of size $|T| = t$, with both s and t known to both players. At the end of the phase they hold sets S' and T' of sizes s' and t' respectively. Assume that $s \le t$, the other case being symmetric.

As before, we think of the random tape as a long sequence of random subsets $Z_1, Z_2, \dots$ of $[n]$. Alice finds the first index $a \le 2^{2s}$ (if any) such that $S \subseteq Z_a$. If there is no such index the protocol halts with answer "not disjoint."

Bob checks if $|T \cap Z_a| \le 3t/4$, in which case he sends Alice the integer $1 + |T \cap Z_a|$ and otherwise he sends 0. If he does not send 0, then the players set $S' = S = S \cap Z_a$ and $T' = T \cap Z_a$, and proceed to the next phase. If Bob sends 0, they halt and output that S and T intersect.

Lemma 4.1. *The following properties hold:*

1. *The communication complexity of a phase is* $O(s+t)$*;*

2. S' *and* T' *are disjoint iff* S *and* T *are;*

3. *If* S *and* T *are disjoint, then except with probability at most* $\exp(-t)$, $t' \le 3t/4$.

Proof. Properties (1) and (2) clearly hold. Property (3) holds due to the independence of $Z_a \cap T$ from the event that $S \subseteq Z_a$ and standard Chernoff bounds. □

The proof of the theorem follows from the lemma by induction on the phases, exactly along the lines of the overview.

5 Making the protocol Las Vegas

In this section we consider protocols that always output a correct answer and prove Theorem 1.5. Note that when our original protocol outputs "disjoint" it is always correct and we mainly have to make sure that there is no error in the case when we output "not disjoint." Let us first establish that we cannot do this maintaining the complexity at $O(k)$ for small values of k.

Lemma 5.1. *For each Las Vegas protocol for the case* $k = 1$ *we have a communication complexity of at least* $\Omega(\log n)$.

Proof. For each singleton set $\{i\}$ and fixed random string fix the shortest accepting conversation when both players hold this set. If two of these conversations are equal then we can create a possible communication pattern with an incorrect output. Thus we have at least n different communication patterns and hence we use $\Omega(\log n)$ bits of communication on the average. □

We proceed to prove Theorem 1.5. We modify our original protocol to make sure that it always produces a certificate for its answer. Let us first give some intuition for the modifications we make to the protocol.

There are two places in the original protocol in which the players halt without a certificate. The first is not very interesting and happens if the index a of the first set Z_a containing S is greater than 2^{2s}. If we measure the expected communication, we can afford always to send a as it is expected to be small.

The more interesting reason for halting is when $t' > 3t/4$. It is easy to see that in this case it is quite likely that the sets have an intersection of size $\Omega(t)$ and thus if Bob chooses a random element from his set and sends it to Alice, we have a constant probability of having found an element in the intersection of the two sets and the protocol can safely terminate.

If the sets are disjoint, the case $t' > 3t/4$ happens with probability $\exp(-t)$ and since we need $\log n$ bits to specify an element, we get a contribution $\exp(-t)\log n$ to the expected communication complexity. For $t \gg \log\log n$ the geometric decay of k_i will ensure that the total contribution of these terms is $O(1)$. For smaller k_i we have to be more careful, and let Bob send an element only with a probability that would still make the expected cost small.

When the sets are not disjoint, this probability introduces "delay" in sending an element. However we will see that this does not affect the asymptotic complexity, since here we can afford a communication cost $\Omega(\log n)$ bits anyway.

Let us now describe the protocol.

Each player holds sets S and T which are updated every round, with the sizes s of S and t of T known to both players. We describe a round of the protocol assuming that $s \le t$. If $s > t$ the roles of Alice and Bob are interchanged.

As before, the random tape is interpreted as a long sequence of random subsets $Z_1, Z_2, \ldots$ of $[n]$.

1. Alice finds the first index, a, such that $S \subseteq Z_a$.

2. Bob sets $t' = |T \cap Z_a|$ and sends it to Alice. If $t' = 0$ they halt with the output "disjoint."

3. If $t' \le 3t/4$ they both update S, T accordingly and proceed to the next round.

4. If $t' > 3t/4$ then Bob flips a coin, whose probability of Heads is $\min(1, t/\log n)$. If it comes up Tails, Bob tells Alice, and they repeat the round again with the same S, T. If it comes up Heads, Bob picks a random element j of T, and sends it to Alice. If $j \in S$ Alice outputs "Not disjoint." Otherwise, they repeat the round with the same S, T.

Let us analyze the complexity of this protocol. Let m denote the size of the intersection of the initial inputs. The same analysis as in the previous section shows that in any round with $t > 16m$, the event $t' > 3t/4$ happens with probability at most $\exp(-t)$. Thus we expect again the k_i to decrease geometrically till that point.

Let us first analyze the communication used till the first round where $t < 16m$. Alice's message is in expectation $O(k_i)$. Bob's message is of length $O(\log k_i)$ for sending t', and with probability $t\exp(-t)/\log n$ it increases by an additional $\log n$ bits for the item j. The latter part is at most $t\exp(-t) \le O(1)$ bits in expectation and thus the total expected cost of each round is $O(k_i)$. The expected geometric decrease of k_i guarantees that in expectation the communication is $O(k)$ up to the point that $t < 16m$. In particular, if $m = 0$, this means that the total communication is $O(k)$ in expectation.

Now let us analyze the cost after t gets below $16m$ (which happens only if the inputs intersect). Note that now, there can be at most a constant number of rounds before which we have $t' > 3t/4$. Moreover, when this happens, we have a constant probability that the random $j \in T$ also satisfies $j \in S$, and the protocol halts. Note that Bob chooses to pick such j with probability at least $\min(1, m/\log n)$. So we expect $O(1 + (\log n)/m)$ repetitions of this round, before Bob chooses j. But each repetition costs only $O(m)$ bits, so in expectation, this part, as well as the cost of sending j, amount to a total of $O(m + \log n) \le O(k + \log n)$ bits, as promised.

Finally, we note that the protocol always halts with a certificate for the answer given.

6 Private randomness

As stated in the Introduction, the general transformation of Newman [7] gives a protocol for DISJ_k^n in the private coins model with complexity $O(k + \log\log(n^k)) = O(k + \log\log n)$. Let us show that the additive term is needed in the case when $k = 1$. This turns out to follow from a general lower bounds of Yao [13]. Let us assume that for any x, x' such that $x \ne x'$ there is a y such that $f(x,y) \ne f(x',y)$ and a similar property holds for y. We call such a function *non-redundant*. This is a natural assumption since if for $x \ne x'$ there is no such y we can consider x and x' to be the same input and reduce the set of possible inputs. Now we have the following lower bound [13, Theorem 5].

Theorem 6.1 ([13]). *For every non-redundant communication problem $f : X \times Y \to \{0,1\}$, the probabilistic communication complexity of f in the private coins model requires $\Omega(\log\log|X| + \log\log|Y|)$ bits.*

This proves that Theorem 1.4 is optimal up to constant factors as DISJ_1^n is the identity function on $[n]$. As Yao's paper does not contain a proof of this theorem, we give here a sketch of the proof (which probably exists somewhere in the literature).

Proof. (Sketch) Assume that at most d bits are exchanged in a probabilistic protocol P for f and, increasing the complexity by at most a factor of two, we assume that Alice and Bob each send every other bit. For every $x \in X$, let $v(x)$ denote the real vector of length $t \le 2^{d+1}$ whose entries are labeled by Boolean strings σ of even length at most d, such that $v(x)_\sigma$ is the probability that Alice sends 0 when holding input x given that σ describes the communication so far. It is not difficult to see that for every two inputs x, x' for Alice, and every input y for Bob, the probability that P accepts (x,y) and the probability that it accepts (x',y) differ at most by the L_1 distance of $v(x)$ and $v(x')$. But given that $f(x,y)$ and $f(x',y)$ differ for at least some y, the vectors $v(x)$ for all $x \in X$ must be at least 1/3 apart in L_1-distance. A standard volume argument shows that in dimension t there are at most $\exp(t)$ such vectors. It follows that $d = \Omega(\log\log|X|)$.

□

References

[1] * L. BABAI, P. FRANKL, AND J. SIMON: Complexity classes in communication complexity theory. In *Proc. 27th FOCS*, pp. 337–347. IEEE Computer Society, 1986. 1

[2] * L. BABAI AND P. KIMMEL: Randomized simultaneous messages: solution of a problem of Yao in communication complexity. In *Proc. 12th IEEE Symp. on Computational Complexity*, pp. 239–246. IEEE Computer Society, 1997. [CCC:10.1109/CCC.1997.612319]. 3

[3] * Z. BAR-YOSSEF, T. S. JAYRAM, R. KUMAR, AND D. SIVAKUMAR: Information statistics approach to data stream and communication complexity. *J. of Computer and System Sciences*, 68:702–732, 2004. [JCSS:10.1016/j.jcss.2003.11.006]. 1, 1.2

[4] * S. JUKNA: *Extremal Combinatorics*. Springer Verlag, 2001. 1

[5] * B. KALYANASUNDARAM AND G. SCHNITGER: The probabilistic communication complexity of set intersection. *SIAM J. on Discrete Mathematics*, 5:545–557, 1992. [SIDMA:10.1137/0405044]. 1, 1.2

[6] * E. KUSHILEVITZ AND N. NISAN: *Communication Complexity*. Cambridge University Press, 1997. 1

[7] * I. NEWMAN: Private vs. common random bits in communication complexity. *Information Processing Letters*, 39:67–71, 1991. [IPL:10.1016/0020-0190(91)90157-D]. 1, 1, 6

[8] * I. NEWMAN AND M. SZEGEDY: Public vs. private coins flips in one round communication games. In *Proc. 28th STOC*, pp. 561–570. ACM Press, 1996. [STOC:237814.238004]. 3

[9] * N. NISAN AND I. SEGAL: The communication requirements of efficient allocations and supporting lindhal prices. 2004. 1

[10] * I. PARNAFES, R. RAZ, AND A. WIGDERSON: Direct product results and the GCD problem in old and new communication models. In *Proc. 29th STOC*, pp. 363–372. ACM Press, 1997. [STOC:258533.258620]. 1

[11] * R. RAZ AND A. WIGDERSON: Monotone circuits for matching require linear depth. *Journal of the ACM*, 39:736–744, 1992. [JACM:146637.146684]. 1

[12] * A. A. RAZBOROV: The distributional complexity of disjointness. *Theoretical Computer Science*, 106:385–390, 1992. [TCS:10.1016/0304-3975(92)90260-M]. 1, 1.2

[13] * A. C.-C. YAO: Some complexity questions related to distributive computing. In *Proc. 11th STOC*, pp. 209–213. ACM Press, 1979. [STOC:800135.804414]. 1, 3, 6, 6.1

AUTHORS

Johan Håstad
Professor
Royal Institute of Technology, Stockholm, Sweden
johanh@kth.se
http://www.csc.kth.se/~johanh/

Avi Wigderson
Professor
Institute for Advanced Study
avi@ias.edu
http://www.math.ias.edu/~avi/

ABOUT THE AUTHORS

JOHAN HÅSTAD graduated from M.I.T. in 1986. His advisor was Shafi Goldwasser. His CS interests include cryptography, complexity theory and approximability of NP-hard optimization problems. He also enjoys table tennis.

AVI WIGDERSON graduated from Princeton University in 1983 under the supervision of Richard Lipton. He is interested in all aspects of theoretical computer science and their interactions with mathematics and the sciences. He likes collaborating (in general and) with Johan.