

The communication complexity of the Hamming distance problem

Wei Huang^{a,1}, Yaoyun Shi^{a,1}, Shengyu Zhang^{b,2}, Yufan Zhu^{a,*,1}

^a Department of Electrical Engineering and Computer Science, University of Michigan, 1301 Beal Avenue, Ann Arbor, MI 48109-2122, USA

^b Computer Science Department, Princeton University, NJ 08544, USA

Received 20 February 2005; received in revised form 28 September 2005; accepted 9 January 2006

Communicated by P.M.B. Vitányi

Abstract

We investigate the randomized and quantum communication complexity of the HAMMING DISTANCE problem, which is to determine if the Hamming distance between two n -bit strings is no less than a threshold d . We prove a quantum lower bound of $\Omega(d)$ qubits in the general interactive model with shared prior entanglement. We also construct a classical protocol of $O(d \log d)$ bits in the restricted *Simultaneous Message Passing* model with public random coins, improving previous protocols of $O(d^2)$ bits [A.C.-C. Yao, On the power of quantum fingerprinting, in: Proceedings of the 35th Annual ACM Symposium on Theory of Computing, 2003, pp. 77–81], and $O(d \log n)$ bits [D. Gavinsky, J. Kempe, R. de Wolf, Quantum communication cannot simulate a public coin, quant-ph/0411051, 2004].

© 2006 Published by Elsevier B.V.

Keywords: Computational complexity; Communication complexity; Hamming distance

1. Introduction

Communication complexity was introduced by Yao [17] and has been extensively studied afterward not only for its own intriguing problems, but also for its many applications ranging from circuit lower bounds to data streaming algorithms. We refer the reader to the monograph [12] for an excellent survey.

We recall some basic concepts below. Let n be an integer and $X = Y = \{0, 1\}^n$. Let $f: X \times Y \rightarrow \{0, 1\}$ be a Boolean function. Consider the scenario where two parties, Alice and Bob, who know only $x \in X$ and $y \in Y$, respectively, communicate interactively with each other to compute $f(x, y)$. The *deterministic communication complexity* of f , denoted by $D(f)$, is defined to be the minimum integer k such that there is a protocol for computing f using no more than k bits of communication on any pair of inputs. The *randomized communication complexity* of f , denoted by $R^{\text{pub}}(f)$, is similarly defined, with the exception that Alice and Bob can use publicly announced random bits and that they are required to compute $f(x, y)$ correctly with probability at least $2/3$. One of the central themes on the classical communication complexity studies is to understand how randomness helps in saving the communication

* Corresponding author.

E-mail addresses: weihuang@eecs.umich.edu (W. Huang), shiyy@eecs.umich.edu (Y. Shi), szhang@cs.princeton.edu (S. Zhang), yufanzhu@eecs.umich.edu (Y. Zhu).

¹ Supported in part by NSF grants 0347078 and 0323555.

² Supported in part by NSF grants CCR-0310466 and CCF-0426582.

cost. A basic finding of Yao [17] is that there are functions f such that $R(f) = O(\log D(f))$. One example is the EQUALITY problem, which simply checks whether $x = y$.

Later results show that different ways of using randomness result in quite subtle changes on communication complexity. A basic finding in this regard, due to Newman [13], is that public-coin protocols can save at most $O(\log n)$ bits over protocols in which Alice and Bob toss private (and independent) coins. The situation is, however, dramatically different in the *Simultaneous Message Passing* (SMP) model, also introduced by Yao [17], where Alice and Bob each send a message to a third person, who then outputs the outcome of the protocol. Apparently, this is a more restricted model and for any function, the communication complexity in this model is at least that in the general interactive communication model. Denote by $R^\parallel(f)$ and $R^{\parallel,\text{pub}}(f)$ the communication complexities in the SMP model with private and public random coins, respectively. It is interesting to note that $R^{\parallel,\text{pub}}(\text{EQUALITY}) = O(1)$ but $R^\parallel(\text{EQUALITY}) = \Theta(\sqrt{n})$ [2,14,5].

Yao also initiated the study of quantum communication complexity [18], where Alice and Bob are equipped with quantum computational power and exchange quantum bits. Allowing an error probability of no more than $1/3$ in the interactive model, the resulting communication complexity is the *quantum communication complexity* of f , denoted by $Q(f)$. If the two parties are allowed to share prior *quantum entanglement*, the quantum analogy of randomness, the communication complexity is denoted by $Q^*(f)$. Similarly, the quantum communication complexities in the SMP model are denoted by Q^\parallel and $Q^{\parallel,*}$, depending on whether prior entanglement is shared. The following relations among the measures are easy to observe.

$$Q^*(f) \leq \frac{R^{\text{pub}}(f)}{Q^{\parallel,*}(f)} \leq R^{\parallel,\text{pub}}(f). \quad (1)$$

Two very interesting problems in both communication models are the power of quantumness, i.e., determining the biggest gap between quantum and randomized communication complexities, and the power of shared entanglement, i.e., determining the biggest gap between quantum communication complexities with and without shared entanglement. An important result for the first problem by Buhrman et al. [7] is $Q^\parallel(\text{EQUALITY}) = O(\log n)$, an exponential saving compared to the randomized counterpart result $R^\parallel(\text{EQUALITY}) = \Theta(\sqrt{n})$ mentioned above. This

exponential separation is generalized by Yao [19], showing that $R^{\parallel,\text{pub}}(f) = \text{constant}$ implies $Q^\parallel(f) = O(\log n)$. As an application, Yao considered the HAMMING DISTANCE problem defined below. For any $x, y \in \{0, 1\}^n$, the Hamming weight of x , denoted by $|x|$, is the number of 1's in x , and the Hamming distance of x and y is $|x \oplus y|$, with “ \oplus ” being bit-wise XOR.

Definition 1.1. For $1 \leq d \leq n$, the d -HAMMING DISTANCE problem is to compute the following Boolean function $\text{HAM}_{n,d} : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\}$, with $\text{HAM}(x, y) = 1$ if and only if $|x \oplus y| > d$.

Lemma 1.2. (Yao [19].) $R^{\parallel,\text{pub}}(\text{HAM}_{n,d}) = O(d^2)$.

In a recent paper [10], Gavinsky et al. gave another classical protocol, which is an improvement over Yao's when $d \gg \log n$.

Lemma 1.3. (Gavinsky et al. [10].) $R^{\parallel,\text{pub}}(\text{HAM}_{n,d}) = O(d \log n)$.

In this paper, we observe a lower bound for $Q^*(\text{HAM}_{n,d})$, which is also a lower bound for $R^{\parallel,\text{pub}}(\text{HAM}_{n,d})$ according to Eq. (1).

Notice that $\text{HAM}(x, y) = n - \text{HAM}(x, \bar{y})$, where $\bar{y} \stackrel{\text{def}}{=} 11 \cdots 1 \oplus y$. Therefore

$$Q^*(\text{HAM}_{n,d}) = Q^*(\text{HAM}_{n,n-d}),$$

and we need only consider the case $d \leq n/2$.

Proposition 1.4. For any $d \leq n/2$, $Q^*(\text{HAM}_{n,d}) = \Omega(d)$.

We then construct a public-coin randomized SMP protocol that almost matches the lower bound and improves both of the above protocols.

Theorem 1.5. $R^{\parallel,\text{pub}}(\text{HAM}_{n,d}) = O(d \log d)$.

We shall prove the above two results in the following sections. Finally we discuss open problems and a plausible approach for closing the gap.

Other related work. Ambainis et al. [3] considered the *error-free* communication complexity, and proved that any *error-free* quantum protocol for the Hamming Distance problem requires at least $n - 2$ qubits of communication in the interactive model, for any $d \leq n - 1$. Feigenbaum et al. [9] started the secure multiparty approximate computation of the Hamming distance.

2. Lower bound of the quantum communication complexity of the Hamming distance problem

For proving the lower bound, we restrict $\text{HAM}_{n,d}$ on those pairs of inputs with equal Hamming distance. More specifically, for an integer k , $1 \leq k \leq n$, define $X_k = Y_k \stackrel{\text{def}}{=} \{x: x \in \{0, 1\}^n, |x| = k\}$. Let $\text{HAM}_{n,k,d}: X_k \times Y_k \rightarrow \{0, 1\}$ be the restriction of $\text{HAM}_{n,d}$ on $X_k \times Y_k$.

Before proving Proposition 1.4, we briefly introduce some related results. Let $x, y \in \{0, 1\}^n$. The DISJOINTNESS problem is to compute the following Boolean function $\text{DISJ}_n: \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\}$, $\text{DISJ}_n(x, y) = 1$ if and only if there exists an integer i , $1 \leq i \leq n$, so that $x_i = y_i = 1$. It is known that $R(\text{DISJ}_n) = \Theta(n)$ [11,15], and $Q^*(\text{DISJ}_n) = \Theta(\sqrt{n})$ [16,14].

We shall use an important lemma in Razborov [16], which is more general than his remarkable lower bound on quantum communication complexity of DISJOINTNESS. Here we may abuse the notation by viewing $x \in \{0, 1\}^n$ as the set $\{i \in [n]: x_i = 1\}$.

Lemma 2.1. (Razborov [16].) *Suppose $k \leq n/4$ and $l \leq k/4$. Let $D: [k] \rightarrow \{0, 1\}$ be any Boolean predicate such that $D(l) \neq D(l-1)$. Let $f_{n,k,D}: X_k \times Y_k \rightarrow \{0, 1\}$ be such that $f_{n,k,D}(x, y) \stackrel{\text{def}}{=} D(|x \cap y|)$. Then $Q^*(f_{n,k,D}) = \Omega(\sqrt{kl})$.*

Proof of Proposition 1.4. Consider D in Lemma 2.1 such that $D(t) = 1$ if and only if $t < l$. For any $x, y \in X_k$, we have $|x \cap y| = k - \text{HAM}(x, y)/2$. Let $l = k - d/2$, then $k - \text{HAM}(x, y)/2 < l$ if and only if $\text{HAM}(x, y) > d$. Therefore, $D(|x \cap y|) = 1$ if and only if $\text{HAM}(x, y) > d$. This implies that $f_{n,k,D}$ and $\text{HAM}_{n,k,d}$ are actually the same function, and thus $Q^*(f_{n,k,D}) = Q^*(\text{HAM}_{n,k,d})$.

To use Lemma 2.1, the following two constraints on k and l need to be satisfied: $k \leq n/4$ and $l \leq k/4$. When $d \leq 3n/8$, let $k = 2d/3 \leq n/4$, then $l = 2d/3 - d/2 = d/6 \leq n/16$. Both requirements for k and l are satisfied. So applying Lemma 2.1, we get $Q^*(\text{HAM}_{n,k,d}) = Q^*(f_{n,k,D}) = \Omega(\sqrt{kl}) = \Omega(d)$.

For $3n/8 < d \leq n/2$, it is reduced to the above case ($d \leq 3n/8$) rather than Lemma 2.1. Let $m = \lceil 8d/5 - 3n/5 \rceil$. Fix first m bits in x to be all 1's, and use x' to denote $x_{m+1} \dots x_n$. Similarly, fix first m bits of y to be all 0's, and use y' to denote $y_{m+1} \dots y_n$. Put $n' = n - m$, $k' = n'/4$, and $d' = d - m$. Then $\text{HAM}(x, y) = \text{HAM}(x', y') + m$ and $Q^*(\text{HAM}_{n,d})(x, y) \geq Q^*(\text{HAM}_{n',k',d'})(x', y')$. It is easy to verify that

$d' \leq 3n'/8$ and $d' = \Omega(d)$. Employing the result of the case that $d \leq 3n/8$, we have $Q^*(\text{HAM}_{n',k',d'}) = \Omega(d')$. Thus $Q^*(\text{HAM}_{n,d}) \geq Q^*(\text{HAM}_{n',k',d'}) = \Omega(d') = \Omega(d)$. \square

3. Upper bound of the classical communication complexity of the Hamming distance problem

To prove Theorem 1.5, we reduce the $\text{HAM}_{n,d}$ problem to $\text{HAM}_{16d^2,d}$ problem by the following lemma.

Lemma 3.1.

$$R^{\parallel, \text{pub}}(\text{HAM}_{n,d}) = O(R^{\parallel, \text{pub}}(\text{HAM}_{16d^2,d})).$$

Note that Theorem 1.5 immediately follows from Lemma 3.1 because by Lemma 1.3, $R^{\parallel, \text{pub}}(\text{HAM}_{n,d}) = O(d \log n)$, thus $R^{\parallel, \text{pub}}(\text{HAM}_{16d^2,d}) = O(d \log d^2) = O(d \log d)$. Now by Lemma 3.1, we have

$$R^{\parallel, \text{pub}}(\text{HAM}_{n,d}) = O(d \log d).$$

So in what follows, we shall prove Lemma 3.1. Define a partial function $\text{HAM}_{n,d|2d}(x, y)$ with domain $\{(x, y): x, y \in \{0, 1\}^n, |x \oplus y| \text{ is either less than } d \text{ or at least } 2d\}$ as follows:

$$\text{HAM}_{n,d|2d}(x, y) = \begin{cases} 0 & \text{if } \text{HAM}(x, y) \leq d, \\ 1 & \text{if } \text{HAM}(x, y) > 2d. \end{cases} \quad (2)$$

Then

Lemma 3.2.

$$R^{\parallel, \text{pub}}(\text{HAM}_{n,d|2d}) = O(1).$$

Proof. We revise Yao's protocol [19] to design an $O(1)$ protocol for $\text{HAM}_{n,d|2d}$. Assume the Hamming distance between x and y is k . Alice and Bob share some random public string, which consists of a sequence of γn (γ is some constant to be determined later) random bits, each of which is generated independently with probability $p = 1/(2d)$ of being 1. Denote this string by $z_1, z_2, \dots, z_{\gamma}$, each of length n . Party A sends the string $a = a_1 a_2 \dots a_{\gamma}$ to the referee, where $a_i = x \cdot z_i \pmod{2}$. Party B sends the string $b = b_1 b_2 \dots b_{\gamma}$ to the referee, where $b_i = y \cdot z_i \pmod{2}$. The referee announces $\text{HAM}_{n,d}(x, y) = 1$ if and only if the Hamming distance between a and b is more than $m = (1/2 - q)\gamma$ where $q = ((1 - 1/d)^d + (1 - 1/d)^{2d})/4$.

Now we prove the above protocol is correct with probability at least $49/50$. Let $c_i = a_i \oplus b_i$. Notice that the Hamming distance between a and b is the number of 1's in $c = c_1 c_2 \dots c_{\gamma}$. We need the following lemma by Yao [19]:

Lemma 3.3. Assume that the Hamming distance between x and y is k . Given c as defined above, each c_i is an independent random variable with probability α_k of being 1, where $\alpha_k = 1/2 - 1/2(1 - 1/d)^k$.

Since α_k is an increasing function over k , to separate $k \leq d$ from $k > 2d$, it would be sufficient to discriminate the two cases that $k = d$ and $k = 2d$. Let N_k be a random variable denoting the number of 1's in c , and $E(N_k)$ and $\sigma(N_k)$ denote corresponding expectation and standard deviation, respectively. Then we have $E(N_k) = \alpha_k \gamma$, and $\sigma(N_k) \leq (\alpha_k \gamma)^{1/2}$. Thus $E(N_{2d}) - E(N_d) = \gamma(\alpha_{2d} - \alpha_d) = \frac{1}{2}\gamma(1 - \frac{1}{d})^d(1 - (1 - \frac{1}{d})^d) \geq \frac{1}{8}\gamma$. Let $\gamma = 20000$, then $E(N_{2d}) - E(N_d) \geq 2500$, while $\sigma(N_d), \sigma(N_{2d}) < (\frac{1}{2}\gamma)^{1/2} = 100$. The cutoff point in the protocol is the middle of $E(N_d)$ and $E(N_{2d})$. By Chebyshev Inequility, with probability of at most $1/100$, $|N_d - E(N_d)| > 10\sigma(N_d) = 1000$. So does N_{2d} . Thus with probability of at least $49/50$, the number of 1's in c being more than cutoff point implies $k > 2d$ and vice versa. Therefore, $O(\gamma)$ communication is sufficient to discriminate the case $\text{HAM}(x, y) > 2d$ and $\text{HAM}(x, y) \leq d$ with error probability of at most $1/50$. \square

The following fact is also useful

Fact 1. If $2d$ balls are randomly thrown into $16d^2$ buckets, then with probability of at least $7/8$, each bucket has at most one ball.

Proof. There are $\binom{2d}{2}$ pairs of balls. The probability of one specific pair of balls falling into the same bucket is $\frac{1}{16d^2} \cdot \frac{1}{16d^2} \cdot 16d^2 = \frac{1}{16d^2}$. Thus the probability of having a pair of balls in the same bucket is upper bounded by $\frac{1}{16d^2} \cdot \binom{2d}{2} < 1/8$. Thus Fact 1 holds. \square

Now we are ready to prove Lemma 3.1.

Proof of Lemma 3.1. If $16d^2 \geq n$, the lemma is obviously true by appending 0's to x and y .

If $16d^2 < n$, suppose we already have a protocol P_1 of C communication to distinguish the cases $|x \oplus y| \leq d$ and $d < |x \oplus y| \leq 2d$ with error probability at most $1/8$. Then we can have a protocol of $C + O(1)$ communication for $\text{HAM}_{n,d}$ with error probability at most $1/4$. Actually, by repeating the protocol for $\text{HAM}_{n,d|2d}(x, y)$ several times, we can have a protocol P_2 of $O(1)$ communication to distinguish the cases $|x \oplus y| \leq d$ and $|x \oplus y| > 2d$ with error probability at most $1/8$. Now the whole protocol P is as follows. Alice sends the concatenation of $m_{A,1}$ and $m_{A,2}$, which are her messages

when she runs P_1 and P_2 , respectively. So does Bob send the concatenation of his two corresponding messages $m_{B,1}$ and $m_{B,2}$. The referee then runs protocol P_i on $(m_{A,i}, m_{B,i})$ and gets the results r_i . The referee now announces $|x \oplus y| \leq d$ if and only if both r_1 and r_2 say $|x \oplus y| \leq d$.

It is easy to see that the protocol is correct. If $|x \oplus y| \leq d$, then both protocols announces so with probability at least $7/8$, and thus P says so with probability at least $3/4$. If $|x \oplus y| > d$, then one of the protocols gets the correct range of $|x \oplus y|$ with probability at least $7/8$, and thus P announces $|x \oplus y| > d$ with probability at least $7/8$ too.

Now it remains to design a protocol of $O(R^{\text{pub}}(\text{HAM}_{16d^2,d}))$ communication to distinguish $|x \oplus y| \leq d$ and $d < |x \oplus y| \leq 2d$. First we assume that n is divisible by $16d^2$, otherwise we pad some 0's to the end of x and y . Using the public random bits, Alice divides x randomly into $16d^2$ parts evenly, Bob also divides y correspondingly. Let A_i, B_i ($1 \leq i \leq 16d^2$) denote corresponding parts of x, y . By Fact 1, with probability at least $7/8$, each pair A_i, B_i would contain at most one bit on which x and y differ. Therefore, the Hamming distance of A_i and B_i would be either 0 or 1, i.e., the Hamming distance of A_i and B_i equals the parity of $A_i \oplus B_i$, which is further equal to $\text{PARITY}(A_i) \oplus \text{PARITY}(B_i)$. Let a_i denote the parity of A_i , b_i denote the parity bit of B_i , and let $a = a_1 a_2 \dots a_{16d^2}$, $b = b_1 b_2 \dots b_{16d^2}$. Then $\text{HAM}_{16d^2,d}(a, b) = \text{HAM}_{n,d}(x, y)$ with probability at least $7/8$. So we run the best protocol for $\text{HAM}_{16d^2,d}$ on the input (a, b) , and use the answer to distinguish $|x \oplus y| \leq d$ and $d < |x \oplus y| \leq 2d$. \square

4. Discussion

We conjecture that our quantum lower bound in Lemma 1.4 is tight. It seems plausible to remove the $O(\log d)$ factor in our upper bound. Recently, Aaronson and Ambainis [1] sharpened the upper bound of the Set Disjointness problem from $O(\sqrt{n} \log n)$ to $O(\sqrt{n})$ using quantum local search instead of Grover's search. In their method, it takes only constant communication qubits to synchronize two parties and simulate each quantum query. From Yao's protocol [19], one can easily derive an $O(d \log d)$ two way interactive quantum communication protocol using quantum counting [6] and the connection between quantum query and communication [8]. Methods similar to [1] might help to remove the $O(\log d)$ factor in this upper bound.

Acknowledgement

We thank Alexei Kitaev for suggesting the relation of the HAMMING DISTANCE problem and the DISJOINTNESS problem, Ronald de Wolf for pointing out a mistake in our earlier draft, Jialin Zhang for suggesting to remove the $O(d \log d)$ term in Lemma 3.1 in our earlier draft and anonymous reviewers for helping us improve the presentation of this paper.

References

- [1] S. Aaronson, A. Ambainis, Quantum search of spatial regions, in: Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science, 2003, pp. 200–209.
- [2] A. Ambainis, Communication complexity in a 3-computer model, *Algorithmica* 16 (3) (1996) 298–301.
- [3] A. Ambainis, W. Gasarch, A. Srinivasan, A. Utis, Lower bounds on the deterministic and quantum communication complexity of hamming distance, *cs.CC/0411076*, 2004.
- [4] Z. Bar-Yossef, T.S. Jayram, Ravi Kumar, D. Sivakumar, Information statistics approach to data stream and communication complexity, *Journal of Computer and System Sciences* 68 (4) (2004) 702–732.
- [5] L. Babai, P.G. Kimmel, Randomized simultaneous messages: solution of a problem of Yao in communication complexity, in: Proceedings of the 12th Annual IEEE Conference on Computational Complexity, 1997, pp. 239–246.
- [6] G. Brassard, P. Høyer, A. Tapp, Quantum counting, in: Proceedings of 25th International Colloquium on Automata, Languages and Programming, 1998.
- [7] H. Buhrman, R. Cleve, J. Watrous, R. de Wolf, Quantum fingerprinting, *Physical Review Letters* 87 (16) (2001).
- [8] H. Buhrman, R. Cleve, A. Wigderson, Quantum vs. classical communication and computation, in: Proceedings of the 30th Annual ACM Symposium on Theory of Computing, 1998, pp. 63–68.
- [9] J. Feigenbaum, Y. Ishai, T. Malkin, K. Nissim, M.J. Strauss, R.N. Wright, Secure multiparty computation of approximations, in: Proceedings of 28th International Colloquium on Automata, Languages and Programming, 2001.
- [10] D. Gavinsky, J. Kempe, R. de Wolf, Quantum communication cannot simulate a public coin, *quant-ph/0411051*, 2004.
- [11] B. Kalyanasundaram, G. Schnitger, The probabilistic communication complexity of set intersection, *SIAM Journal on Discrete Mathematics* 5 (4) (1992) 545–557.
- [12] E. Kushilevitz, N. Nisan, *Communication Complexity*, Cambridge University Press, Cambridge, 1997.
- [13] I. Newman, Private vs. common random bits in communication complexity, *Information Processing Letters* 39 (2) (1991) 67–71.
- [14] I. Newman, M. Szegedy, Public vs. private coin flips in one round communication games, in: Proceedings of the 28th Annual ACM Symposium on Theory of Computing, 1996, pp. 561–570.
- [15] A.A. Razborov, Applications of matrix methods to the theory of lower bounds in computational complexity, *Combinatorica* 10 (1) (1990) 81–93.
- [16] A.A. Razborov, Quantum communication complexity of symmetric predicates, *Izvestiya Math.* 67 (1) (2003) 145–159 (English version); also in: *quant-ph/0204025*.
- [17] A.C.-C. Yao, Some complexity questions related to distributive computing, in: Proceedings of the 11th Annual ACM Symposium on Theory of Computing, 1979, pp. 209–213.
- [18] A.C.-C. Yao, Quantum circuit complexity, in: Proceedings of the 34th Annual IEEE Symposium on Foundations of Computer Science, 1993, pp. 352–361.
- [19] A.C.-C. Yao, On the power of quantum fingerprinting, in: Proceedings of the 35th Annual ACM Symposium on Theory of Computing, 2003, pp. 77–81.