THEORY OF COMPUTING, Volume 4 (2008), pp. 129–135
http://theoryofcomputing.org

The One-Way Communication Complexity of Hamming Distance

T. S. Jayram Ravi Kumar* D. Sivakumar*

Received: January 23, 2008; revised: September 12, 2008; published: October 11, 2008.

Abstract: Consider the following version of the Hamming distance problem for ± 1 vectors of length n: the promise is that the distance is either at least $\frac{n}{2}+\sqrt{n}$ or at most $\frac{n}{2}-\sqrt{n}$, and the goal is to find out which of these two cases occurs. Woodruff (*Proc. ACM-SIAM Symposium on Discrete Algorithms, 2004*) gave a linear lower bound for the randomized one-way communication complexity of this problem. In this note we give a simple proof of this result. Our proof uses a simple reduction from the indexing problem and avoids the VC-dimension arguments used in the previous paper. As shown by Woodruff (*loc. cit.*), this implies an $\Omega(1/\varepsilon^2)$-space lower bound for approximating frequency moments within a factor $1+\varepsilon$ in the data stream model.

ACM Classification: F.2.2

AMS Classification: 68Q25

Key words and phrases: One-way communication complexity, indexing, Hamming distance

1 Introduction

The *Hamming distance* $H(x,y)$ between two vectors x and y is defined to be the number of positions i such that $x_i \neq y_i$. Let GapHD denote the following promise problem: the input consists of ± 1 vectors x and y of length n, together with the promise that either $H(x,y) \leq \frac{n}{2}-\sqrt{n}$ or $H(x,y) \geq \frac{n}{2}+\sqrt{n}$, and the goal is to find out which of these two cases occurs. In the *one-way communication model* [6], Alice gets x, Bob gets y and Alice sends a single message to Bob using which Bob outputs the desired answer.

*Part of the work done while the author was at the IBM Almaden Research Center.

DOI: 10.4086/toc.2008.v004a006

We will also allow the protocols to be randomized in which case both Alice and Bob have access to a public random string, and the correct answer must be output with probability at least 2/3. The cost of such a protocol is the maximum number of bits communicated by Alice over all inputs. The randomized one-way communication complexity of GapHD is the cost of the minimum-cost one-way protocol for GapHD.

Woodruff [11] showed a tight $\Omega(n)$ lower bound for GapHD and used it to obtain an $\Omega(1/\varepsilon^2)$-space lower bound for approximating frequency moments within a factor $1+\varepsilon$ in the data stream model. In this note we show a simpler proof of the linear lower bound for GapHD; our proof uses an easy reduction from the *indexing* problem and avoids the VC-dimension arguments in [11]. We will present two different reductions: the first reduction uses Rademacher sums and the second reduction treats the indexing problem from a geometric viewpoint.

Subsequent to discussing our proof with Woodruff, he obtained [12] two alternative proofs of the $\Omega(n)$ lower bound for GapHD. The first proof [12, Section 4.3] is similar in spirit to ours, and presents a reduction from the indexing problem. The second proof [12, Section 4.4] establishes the $\Omega(n)$ lower bound for GapHD under the uniform product distribution of inputs, using combinatorial and information-theoretic arguments. For the more general version of the gap Hamming problem where the goal is to distinguish between $H(x,y) \le d$ and $H(x,y) \ge d(1+\varepsilon)$, an algorithm of Kushilevitz, Ostrovsky, and Rabani [7] yields an $O(1/\varepsilon^2)$ upper bound (independent of d). Finally, the multi-round randomized communication complexity of GapHD is still open; we conjecture an $\Omega(n)$ lower bound.

2 Main result

In this note we give a simple proof of the following result.

Theorem 1 (Woodruff)**.** *The randomized one-way communication complexity of* GapHD *is linear in the length of the input.*

Proof. We begin by recalling the indexing problem: Alice gets a set $T \subseteq [n]$, Bob gets an element $i \in [n]$, and the goal is to compute whether $i \in T$. We know that this has an $\Omega(n)$ lower bound in the one-way communication model (e. g., see [5] or [2] for a sharp bound in terms of the error probability; for completeness, we include a self-contained elementary proof of this result in Section 3).

Let Alice's input be $T \subseteq [n]$ and Bob's input be $i \in [n]$. Transform T to a vector $u \in \{-1,+1\}^n$ by setting $u_j = -1$ if $j \in T$ and $u_j = +1$ if $j \notin T$. Let e_j denote the standard 0-1 basis vector corresponding to Bob's input.

Alice and Bob will use public randomness to realize an instance $x, y \in \{-1,+1\}^N$ of GapHD, for some N to be specified later, as follows. Pick N i.i.d. vectors $r^1, r^2, \ldots, r^N$ in $\mathbb{R}^n$ where the distribution μ of each r^k will be specified later. Define $x_k \triangleq \mathrm{sgn}(\langle u, r^k\rangle)$ and $y_k \triangleq \mathrm{sgn}(\langle e_i, r^k\rangle)$ for all k. Since $r^k \in \{-1,+1\}^n$, we have $\mathrm{sgn}(\langle e_i, r^k\rangle) = r^k_i$. Also note that

$$H(x,y) = |\{k : \mathrm{sgn}(\langle u, r^k\rangle) \ne \mathrm{sgn}(\langle e_i, r^k\rangle)\}| = |\{k : \mathrm{sgn}(\langle u, r^k\rangle) \ne r^k_i\}|\,.$$

We will show that if $r \sim \mu$,

$$\Pr[\mathrm{sgn}(\langle u, r\rangle) \ne r_i] \begin{cases} \ge \frac{1}{2} + \frac{c}{\sqrt{n}} & \text{if } u_i = -1, \\ \le \frac{1}{2} - \frac{c}{\sqrt{n}} & \text{if } u_i = +1, \end{cases} \tag{2.1}$$

for some positive constant $c > 0$.

We will use the following version of Chernoff's bound (e. g., see [8]):

Chernoff's bound. *Let $X_1, X_2, \ldots, X_N$ be N i.i.d. 0-1 random variables and $X = \sum_{k=1}^{N} X_k$. Then*

$$\Pr[X - \mathbb{E}[X] > \varepsilon] \leq e^{-2\varepsilon^2/N} \quad \text{and} \quad \Pr[X - \mathbb{E}[X] < -\varepsilon] \leq e^{-2\varepsilon^2/N}.$$

Set $N = 4n/c^2$ and $\varepsilon = 1/\sqrt{N}$. By Chernoff's bound, with probability at least 2/3, we have that either $H(x,y) \geq \frac{N}{2} + \sqrt{N}$ if $u_i = -1$, or $H(x,y) \leq \frac{N}{2} - \sqrt{N}$ if $u_i = +1$. Therefore, given a protocol for GapHD, we have a protocol for the indexing problem. Since the indexing problem has a linear lower bound and $N = O(n)$, this proves the linear lower bound for GapHD.

We now establish (2.1) by giving two different proofs.

Rademacher sums: Assume that n is odd. Let μ be the uniform distribution over the vectors in $\{-1,+1\}^n$ and let $r \sim \mu$. Note that since both u and r are ± 1 vectors and n is odd, $\langle u, r \rangle \neq 0$. Write $\langle u, r \rangle = u_i r_i + w$, where $w \triangleq \sum_{j \neq i} u_j r_j$. Note that w is independent of r_i. Fix a value for w; there are two cases to consider:

- If $w \neq 0$, then $|w| \geq 2$ since w is a sum of an even number of ± 1 values. Therefore, $\operatorname{sgn}(\langle u, r \rangle) = \operatorname{sgn}(w)$, implying that

$$\Pr[\operatorname{sgn}(\langle u, r \rangle) \neq r_i \mid w \neq 0] = \Pr[\operatorname{sgn}(w) \neq r_i \mid w \neq 0] = \frac{1}{2}. \tag{2.2}$$

- If $w = 0$, then $\operatorname{sgn}(\langle u, r \rangle) = u_i r_i$. Using the independence of w and r_i, we obtain

$$\begin{aligned} \Pr[\operatorname{sgn}(\langle u, r \rangle) \neq r_i \mid w = 0] &= \Pr[u_i r_i \neq r_i \mid w = 0] \\ &= \Pr[u_i r_i \neq r_i] \\ &= \begin{cases} 1 & \text{if } u_i = -1, \\ 0 & \text{if } u_i = +1. \end{cases} \end{aligned} \tag{2.3}$$

Now w is the sum of $n-1$ i.i.d. random variables each of which is distributed uniformly in $\{-1,+1\}$. Since n is odd, assuming it is large enough, $\Pr[w = 0] \geq c'/\sqrt{n}$, for some constant $c' > 0$ (by Stirling's formula). Combining this with (2.2) and (2.3), and letting $c = c'/2$, we obtain (2.1).

Geometry: The key idea is to view u and e_i as vectors in Euclidean space and apply the inner product protocol given in [5]. This protocol uses the technique of [4], which arose in the context of rounding the solution of a semi-definite program. For the sake of completeness, we sketch this argument. Define μ such that $r \sim \mu$ is a uniformly chosen n-dimensional unit vector. Let r' be the projection of r in the plane spanned by u and e_i. Then by rotational symmetry, the direction of r' is uniform in this plane. If $\hat{u}$ denotes the unit vector in the direction of u, then it follows[1] that

$$\Pr[\operatorname{sgn}(\langle u, r \rangle) \neq r_i] = \frac{\arccos(\langle \hat{u}, e_i \rangle)}{\pi} = \frac{1}{\pi} \cdot \arccos\left(\frac{u_i}{\sqrt{n}}\right). \tag{2.4}$$

[1] Note that the events $\langle u, r \rangle = 0$ and $\langle e_i, r \rangle = 0$ have probability measure zero; if either happens, Alice and Bob will abandon the protocol; this adds a negligible quantity to the error probability.

Now, for any $z \in [-1,1]$, $\arccos(z) = \frac{\pi}{2} - \arcsin(z)$. Since $\sin(z) \le z$ for $z \ge 0$, we have $\arcsin(z) \ge z$ for $0 \le z \le 1$. Substituting in (2.4), we conclude

$$\Pr[\mathrm{sgn}(\langle u,r\rangle) \ne \mathrm{sgn}(\langle e_i,r\rangle)] \begin{cases} \ge \frac{1}{2} + \frac{1}{\pi\sqrt{n}} & \text{if } u_i = -1, \\ \le \frac{1}{2} - \frac{1}{\pi\sqrt{n}} & \text{if } u_i = +1, \end{cases}$$

as required. □

Remark. The geometric approach shown above uses an infinite amount of randomness which is not part of the standard model. However, the important point is that the space of inputs and messages are finite, therefore, the lower bounds for indexing and consequently for the GapHD will continue to hold. Alternatively, one can also prove the above bounds using finite amount of randomness by considering finite-precision versions of the random vectors (as was done in [5]; see also [9]).

3 Indexing lower bound

In this section, we present an elementary and self-contained proof of an $\Omega(n)$ lower bound on the randomized one-way communication complexity of the indexing problem. A more general result, involving the VC-dimension and shatter coefficients, appears in [5, 2]. The proof is intended to be elementary to present, and therefore, the constants are not chosen optimally. The proof borrows ideas from [3]; the use of error-correcting codes for data stream lower bounds also appears in [1].

The indexing problem may also be thought of as a problem on strings: Alice is given a string $x \in \{0,1\}^n$ and Bob is given an index $i \in [n]$, and the goal is for Bob to learn x_i, after receiving a single message from Alice. Alice and Bob are both allowed to be randomized. From the easy direction of Yao's min-max principle, it follows that if there is a randomized communication protocol that solves every instance of indexing with probability of error at most δ, then for any distribution μ on the inputs for indexing, there is a deterministic protocol that works correctly with probability at least $1-\delta$ when the inputs are chosen according to μ. Furthermore, if the randomized protocol is one-way, then so is the resulting deterministic protocol. The rest of this section presents an ensemble of distributions $\mu = \mu(\delta) = \{\mu_n\}$ (for suitably large $\delta > 0$) such that for some absolute constant $\varepsilon > 0$ and for sufficiently large n, in any deterministic communication protocol that works correctly on all but δ fraction of inputs from μ_n, Alice must communicate at least εn bits to Bob.

An (n,d)-error correcting code is a collection $\mathcal{C}$ of strings in $\{0,1\}^n$ such that for any distinct pair of strings $x,y \in \mathcal{C}$, $H(x,y) \ge d$, where H denotes the Hamming distance. It is well-known[2] that there are constants $\alpha > 0, \Delta > 0$ such that for all sufficiently large n, there is an $(n,\Delta n)$-error correcting code $\mathcal{C}_n$ of size $2^{\alpha n}$. The distribution μ_n we will use will pick a string x for Alice uniformly from the collection $\mathcal{C}_n$, and an index $i \in [n]$ for Bob uniformly at random. Let $0 < \delta \le \Delta/4$.

Let Π denote a deterministic one-way communication protocol that succeeds with probability at least $1-\delta$ when the inputs are drawn according to μ_n, and in which the number of bits sent by Alice

[2] For example, Justesen codes (see [10]) are a simple and explicit construction. It can be shown via the probabilistic method, using Chernoff bounds, that there exists α such that for any $\Delta < 1/2$ and sufficiently large n, there is an $(n,\Delta n)$-error correcting code of size $2^{\alpha n}$. For our proof, the explicitness of the code is not necessary.

is at most c. Since Π is a one-way protocol, there is a pair of functions $A : \{0,1\}^n \longrightarrow \{0,1\}^c$ and $B : \{0,1\}^c \times [n] \longrightarrow \{0,1\}$ that describe the actions of Alice and Bob in the protocol Π. First note that if $x \neq y$, since $H(x,y) \geq 4\delta n$, under the uniform distribution of $i \in [n]$, we have $x_i \neq y_i$ with probability at least 4δ. Now suppose $x,y \in \mathcal{C}_n$, $x \neq y$, and $A(x) = A(y) = z$. Then, for at least $4\delta n$ distinct values of i, $B(z,i)$ agrees with precisely one of x_i and y_i; equivalently, for at least $2\delta n$ distinct values of i, either $B(z,i) \neq x_i$ or $B(z,i) \neq y_i$. Since $\Pi(x,i) = B(z,i) = \Pi(y,i)$, we have $\Pr_i[\Pi(x,i) \neq x_i] \geq 2\delta$ or $\Pr_i[\Pi(y,i) \neq y_i] \geq 2\delta$. Since this is true for any pair of strings in $\mathcal{C}_n$ that agree under the map $A(\cdot)$, it follows that for any $z \in \{0,1\}^c$, for all but one x satisfying $A(x) = z$, we have $\Pr_i[\Pi(x,i) \neq x_i] \geq 2\delta$. If the total error of the protocol under the distribution μ_n is no more than δ, we have:

$$\frac{2^{\alpha n} - 2^c}{2^{\alpha n}} 2\delta < \delta\,,$$

whence it follows that $c > \alpha n - 1$.

Acknowledgments

We thank Laci Babai and the anonymous reviewers for many useful comments.

References

[1] * N. ALON, Y. MATIAS, AND M. SZEGEDY: The space complexity of approximating the frequency moments. *J. Comput. System Sci.*, 58(1):137–147, 1999. [JCSS:10.1006/jcss.1997.1545]. 3

[2] * Z. BAR-YOSSEF, T.S. JAYRAM, R. KUMAR, AND D. SIVAKUMAR: Information theory methods in communication complexity. In *Proc. 17th Annual IEEE Conf. Computational Complexity*, pp. 93–102. IEEE, 2002. [CCC:10.1109/CCC.2002.1004344]. 2, 3

[3] * Z. BAR-YOSSEF, R. KUMAR, AND D. SIVAKUMAR: Reductions in streaming algorithms, with an application to counting triangles in graphs. In *Proc. 13th ACM-SIAM Symp. Discrete Algorithms (SODA)*, pp. 623–632. SIAM, 2002. [SODA:545381.545464]. 3

[4] * M. X. GOEMANS AND D. P. WILLIAMSON: Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42(6):1115–1145, 1995. [JACM:10.1145/227683.227684]. 2

[5] * I. KREMER, N. NISAN, AND D. RON: On randomized one-round communication complexity. *Comput. Complexity*, 8(1):21–49, 1999. [Springer:w5pccfda9jbhpgdj]. 2, 2, 2, 3

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

[7] * E. KUSHILEVITZ, R. OSTROVSKY, AND Y. RABANI: Efficient search for approximate nearest neighbor in high dimensional spaces. *SIAM J. Comput.*, 30(2):457–474, 2000. [SICOMP:10.1137/S0097539798347177]. 1

[8] * C. MCDIARMID: *Probabilistic Methods for Algorithmic Discrete Mathematics*, volume 16 of *Algorithms and Combinatorics*. Springer-Verlag, Berlin, 1998. 2

[9] * I. NEWMAN: Private vs. common random bits in communication complexity. *Inform. Process. Lett.*, 39(2):67–71, 1991. [IPL:10.1016/0020-0190(91)90157-D]. 2

[10] * J. H. VAN LINT: *Introduction to Coding Theory*. Springer, 1999. 2

[11] * D. WOODRUFF: Optimal space lower bounds for all frequency moments. In *Proc. 15th Annual ACM-SIAM Symp. Discrete Algorithms (SODA)*, pp. 167–175. SIAM, 2004. [SODA:982792.982817]. 1

[12] * D. WOODRUFF: *Efficient and Private Distance Approximation in the Communication and Streaming Models*. PhD thesis, MIT, 2007. 1

AUTHORS

T. S. Jayram [About the author]
IBM Almaden Research Center
650 Harry Rd
San Jose, CA 95120
jayram@almaden.ibm.com
http://www.almaden.ibm.com/cs/people/jayram

Ravi Kumar [About the author]
Yahoo! Research
701 First Ave
Sunnyvale, CA 94089
ravikumar@yahoo-inc.com
http://research.yahoo.com/Ravi_Kumar

D. Sivakumar [About the author]
Google Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
siva@google.com
http://globofthoughts.blogspot.com

ABOUT THE AUTHORS

T. S. JAYRAM (aka JAYRAM THATHACHAR) has been at IBM Research since 1998. He obtained his Bachelors degree in Computer Science from IIT Madras, Chennai and his Masters and Ph. D. in Computer Science and Engineering from the University of Washington, Seattle, under the supervision of Paul Beame. He now manages the Algorithms and Computation group at IBM Almaden. His primary research interests are in massive data sets and computational complexity.

RAVI KUMAR has been at Yahoo! Research since July 2005. Prior to this, he was a research staff member at the IBM Almaden Research Center in the Computer Science Principles and Methodologies group. He obtained his Ph. D. in Computer Science from Cornell University under Ronitt Rubinfeld in December 1997. His primary interests are web algorithms, algorithms for large data sets, and theory of computation.

D. SIVAKUMAR (SIVA) is interested in theoretical computer science, organizing the world's information, and everything in between. He has been at Google Research since 2005, before which he was on the Research Staff at the IBM Almaden Research Center, San Jose, CA. Before finding his home in the pleasant environs of northern California, he spent a few years in the steam chamber of Houston, TX, as a member of the faculty of Computer Science at the University of Houston, and a few years in the freezer that is Buffalo, NY, where he received his Ph. D. in Computer Science from the State University of New York under the wise tutelage of Ken Regan, Jin-Yi Cai, and Alan Selman.