The One-Way Communication Complexity of Gap Hamming Distance

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

Abstract

Consider the gap Hamming distance problem (GapHD) for vectors of length n with the promise that the distance is either at least $\frac{n}{2} + \sqrt{n}$ or at most $\frac{n}{2} - \sqrt{n}$. The goal is to find out which of these two cases occurs. Woodruff [Woo04] gave a linear lower bound for the one-way communication complexity of GapHD. In this note we give a simple proof of this result. Our proof uses a simple reduction and avoids the VC-dimension arguments used in the previous paper. As shown in [Woo04], this implies an $\Omega(1/\epsilon^2)$-space lower bound for approximating frequency moments to within a factor $1 + \epsilon$ in the data stream model.

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 Hamming distance problem for vectors x and y of length n each with the promise that either $H(x, y) \leq \frac{n}{2} - \sqrt{n}$ or $H(x, y) \geq \frac{n}{2} + \sqrt{n}$. The goal is to find out which of these two cases occurs. In the *one-way communication model* [KN97], Alice gets x, Bob gets y and Alice sends a single message to Bob using which Bob outputs the desired answer. 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 cheapest one-way protocol for GapHD.

Woodruff [Woo04] showed an $\Omega(n)$ lower bound for GapHD and used it to obtain an $\Omega(1/\epsilon^2)$-space lower bound for approximating frequency moments to within a factor $1+\epsilon$ 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 [Woo04]. We will present two different reductions: the first reduction uses Rademacher sums and the second reduction treats the indexing problem from a geometric viewpoint.

Recall 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 [BJKS02] for a sharp bound in terms of the error probability). The main result of this note is the following:

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

*IBM Almaden Research Center, 650 Harry Rd, San Jose, CA 95120. E-mail: jayram@almaden.ibm.com.

†Yahoo! Research, 701 First Ave., Sunnyvale, CA 94089. E-mail: ravikumar@yahoo-inc.com. Work done while the author was at the IBM Almaden Research Center.

‡Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043. E-mail: siva@google.com. Work done while the author was at the IBM Almaden Research Center.

Proof. 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 mapping $0 \mapsto +1$ and $1 \mapsto -1$. Let e_i denote the standard basis vector corresponding to Bob's input.

Alice and Bob will use public randomness to realize an instance $(x, y) \in \{-1, 0, +1\}^N$ 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. Note that $H(x, y) = |\{k : \mathrm{sgn}(\langle u, r^k\rangle) \neq \mathrm{sgn}(\langle e_i, r^k\rangle)\}|$.

We will show that for any $r \sim \mu$,

$$\Pr[\mathrm{sgn}(\langle u, r\rangle) \neq \mathrm{sgn}(\langle e_i, r\rangle)] \begin{cases} \geq \frac{1}{2} + \frac{c}{\sqrt{n}} & \text{if } u_i = -1, \\ \leq \frac{1}{2} - \frac{c}{\sqrt{n}} & \text{if } u_i = +1, \end{cases} \tag{1}$$

for some positive constant $c > 0$.

We will use the following version of Chernoff's bound (e.g., see [McD98]):

Chernoff's Bound. *Let $X_1, X_2, \ldots, X_N$ be N i.i.d. binary random variables and $X = \sum_{k=1}^N X_k$. Then,* $\Pr[X - \mathbb{E}[X] > \epsilon] \leq e^{-2\epsilon^2/N}$ *and* $\Pr[X - \mathbb{E}[X] < -\epsilon] \leq e^{-2\epsilon^2/N}$.

Set $N = 4n/c^2$ and $\epsilon = \sqrt{N}$. 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 $N = O(n)$, this proves the linear lower bound for GapHD.

We now establish (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 $\mathrm{sgn}(\langle e_i, r\rangle) = \mathrm{sgn}(r_i)$. Write $\langle u, r\rangle = u_i r_i + \sum_{j \neq i} u_j r_j = u_i r_i + w$, where $w \triangleq \sum_{j \neq i} u_j r_j$. Note that w is independent of r_i. Fix a value for w and there are 2 cases to consider:

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

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

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

$$\begin{aligned} \Pr[\mathrm{sgn}(\langle u, r\rangle) \neq \mathrm{sgn}(\langle e_i, r\rangle) \mid w] &= \Pr[\mathrm{sgn}(u_i r_i) \neq \mathrm{sgn}(r_i) \mid w] \\ &= \Pr[\mathrm{sgn}(u_i r_i) \neq \mathrm{sgn}(r_i)] \\ &= \begin{cases} 1 & \text{if } u_i = -1, \\ 0 & \text{if } u_i = +1 \end{cases} \end{aligned} \tag{3}$$

Now w is the sum of $n-1$ iid random variables each of which is distributed uniformly in $\{-1,+1\}$. Since n is odd, $\Pr[w = 0] = c/\sqrt{n}$ some constant $c > 0$. Combining this with (2) and (3), we conclude

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

Geometry: The key idea is to view u and e_i as vectors in Euclidean space and apply the inner product protocol given in [KNR99]. This protocol uses the technique of [GW95] 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. By rotational symmetry, it suffices to consider the 2-dimensional plane determined by u and e_i wherein the direction of r is uniform in that plane. If $\hat{u}$ denotes the unit vector in the direction of u, then it follows that

$$\Pr[\mathrm{sgn}(\langle u, r\rangle) \neq \mathrm{sgn}(\langle e_i, r\rangle)] = \frac{\arccos(\langle \hat{u}, e_i\rangle)}{\pi} = \frac{1}{\pi} \cdot \arccos\left(\frac{u_i}{\sqrt{n}}\right) \tag{4}$$

Now, for any $z \in [-1, 1]$, $\arccos(z) = \frac{\pi}{2} - \arcsin(z)$. Using a simple approximation of $\arcsin(z)$ for small z, it can be shown that there exists a constant c such that $\arcsin\left(\frac{+1}{\sqrt{n}}\right) \geq \frac{c}{\sqrt{n}}$ and $\arcsin\left(\frac{-1}{\sqrt{n}}\right) \leq -\frac{c}{\sqrt{n}}$. Substituting in (4), we conclude

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

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 [KNR99]).

References

[BJKS02] Z. Bar-Yossef, T.S. Jayram, R. Kumar, and D. Sivakumar. Information theory methods in communication complexity. In *Proceedings of the 17th Annual IEEE Conference on Computational Complexity*, pages 93–102, 2002.

[GW95] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6):1115–1145, 1995.

[KN97] E. Kushilevitz and N. Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[KNR99] Ilan Kremer, Noam Nisan, and Dana Ron. On randomized one-round communication complexity. *Computational Complexity*, 8(1):21–49, 1999.

[McD98] C. McDiarmid. *Probabilistic methods for algorithmic discrete mathematics*, volume 16 of *Algorithms and Combinatorics*. Springer-Verlag, Berlin, 1998.

[Woo04] David Woodruff. Optimal space lower bounds for all frequency moments. In *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete algorithms*, pages 167–175, 2004.