Electronic Colloquium on Computational Complexity, Report No. 6 (2020)

THE COMMUNICATION COMPLEXITY OF THE EXACT GAP-HAMMING PROBLEM

ANUP RAO AND AMIR YEHUDAYOFF

ABSTRACT. We prove a sharp lower bound on the distributional communication complexity of the exact gap-hamming problem.

1. INTRODUCTION

The gap-hamming function $\mathsf{GH} = \mathsf{GH}_{n,k} : \{\pm 1\}^n \to \{0, 1, \star\}$ is defined by

$$\mathsf{GH}(x, y) = \begin{cases} 1 & \langle x, y\rangle \geq k, \\ 0 & \langle x, y\rangle \leq -k, \\ \star & \text{otherwise,} \end{cases}$$

where $\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i$ is the standard inner product (the Hamming distance between x and y is $\frac{n-\langle x,y\rangle}{2}$). This problem naturally fits into the framework of two-party communication complexity; for background and definitions, see the books [7, 9]. Alice gets x, Bob gets y, and their goal is to compute $\mathsf{GH}(x, y)$. It is a promise problem — the protocol is allowed to compute any value when the input corresponds to a $\star$, and it needs to be correct only on the remaining inputs. The standard choice for k is $\lceil\sqrt{n}\rceil$, so we write GH_n to denote $\mathsf{GH}_{n,\lceil\sqrt{n}\rceil}$.

The gap-hamming problem was introduced by Indyk and Woodruff in the context of streaming algorithms [5], and was subsequently studied and used in many works and in various contexts (see [6, 12, 1, 2, 3] and references within). Proving a sharp $\Omega(n)$ lower bound on its randomized communication complexity was a central open problem for almost ten years, until Chakrabarti and Regev [4] solved it. Later, Vidick [11], Sherstov [10], and [8] found simpler proofs. The difficulties in proving this lower bound are explained in [4, 10].

The exact gap-hamming function is defined by

$$\mathsf{EGH}_{n,k}(x, y) = \begin{cases} 1 & \langle x, y\rangle = k, \\ 0 & \langle x, y\rangle = -k, \\ \star & \text{otherwise.} \end{cases}$$

As before, we write EGH_n to denote $\mathsf{EGH}_{n,\lceil\sqrt{n}\rceil}$. The exact gap-hamming function is easier to compute than gap-hamming; the protocol only needs to worry about inputs

Part of this work was done while the authors were visiting the Simons Institute for the Theory of Computing. A.Y. is supported by ISF grant 1162/15.

ISSN 1433-8092

whose inner product has magnitude *exactly* k. Proving a sharp lower bound on the randomized communication complexity of EGH was left as an open problem.

One of the difficulties in proving a lower bound for EGH is the following somewhat surprising property: *for infinitely many values of* n*, the deterministic communication complexity of* EGH_n *is* 2. The reason is that there is a simple deterministic protocol of length 2 that computes $\langle X,Y\rangle \bmod 4$ for all n. The players announce the parities of their inputs $\frac{n-\sum_{j=1}^n X_j}{2} \bmod 2$ and $\frac{n-\sum_{j=1}^n Y_j}{2} \bmod 2$. Because $n = \langle X,Y\rangle \bmod 2$, this data determines $\langle X,Y\rangle \bmod 4$. For example, this deterministic protocol computes EGH_n when $\sqrt{n}$ is an odd integer, because then we have $-\sqrt{n} \neq \sqrt{n} \bmod 4$.

We overcome this difficulty and show that EGH is extraordinary in that although it is a natural problem with communication complexity $O(1)$ for infinitely many n's, the following holds.

Theorem. *The randomized communication complexity of* EGH_n *is at least* $\Omega(n)$ *for infinitely many values of* n*.*

There is a natural reduction between different parameters n,k, and from randomized protocols to distributional protocols. Denote by $U_{n,k}$ the uniform distribution over the set of pairs $(x,y) \in \{\pm 1\}^n \times \{\pm 1\}^n$ so that $\langle x,y\rangle \in \{\pm k\}$. For each integer t, given inputs $x,y \in \{\pm 1\}^n$, the players can use padding and public randomness (and no communication) to generate (X',Y') that is distributed according to $U_{tn,tk}$ for $k = \langle x,y\rangle$. In other words, from a protocol that solves $\mathsf{EGH}_{tn,tk}$ over the distribution $U_{tn,tk}$, we get a randomized protocol that solves $\mathsf{EGH}_{n,k}$. So, to prove the lower bound stated above, it suffices to prove the following distributional lower bound.

Theorem 1. *For every* $\beta > 0$*, there are constants* $n_0 > 0$ *and* $\alpha > 0$ *so that the following holds. Let* n,k *be positive even integers so that* $n > n_0$ *and* $k < \alpha\sqrt{n}$*. Any protocol that computes* $\mathsf{EGH}_{n,k}$ *over inputs from* $U_{n,k}$ *with success probability* $2/3$ *must have communication complexity at least* $(1-\beta)n$*.*

Theorem 1 is sharp in the following two senses. First, if $k \neq n \bmod 2$ then $\mathsf{EGH}_{n,k}$ is trivial, and if k is odd then the deterministic communication complexity of $\mathsf{EGH}_{n,k}$ is 2. Secondly, for every $\alpha > 0$, there is $\beta > 0$ so that if $k > \alpha\sqrt{n}$ then the randomized communication complexity of $\mathsf{EGH}_{n,k}$ is at most $(1-\beta)n$. In the randomized protocol, Alice gets x, Bob gets y and the public randomness is a sequence $I_1, I_2, \ldots, I_m$ of i.i.d. uniform elements in $[n]$ for $m \leq O(\frac{n}{\alpha^2})$. By a standard coupon collector argument, the number of (distinct) elements in the set $S = \{I_1,\ldots,I_m\}$ is at most $(1-\beta)n - 1$ with probability at least $\frac{5}{6}$. If $|S| > (1-\beta)n-1$, the parties "abort", and otherwise Alice sends to Bob the value of x_s for all $s \in S$. Bob uses this data to compute $z = \mathsf{sign}\big(\sum_{j=1}^m x_{I_j} y_{I_j}\big)$. Bob sends the output of the protocol z to Alice. Chernoff's bound says that if $\mathsf{EGH}_{n,k}(x,y) \neq \star$ then $\Pr[z = \mathsf{EGH}_{n,k}(x,y)] \geq \frac{5}{6}$. The union bound implies that the overall success probability is at least $\frac{2}{3}$.

The lower bounds [4, 11, 10, 8] for GH are based on anti-concentration. Roughly speaking, these works prove that $\Pr[\langle X,Y\rangle \in I] < p$ for all small intervals $I \subset \mathbb{R}$

and some small $p > 0$. The main ingredient for our lower bound on the complexity of EGH is the following "smoothness" result (which implies anti-concentration).

Theorem 2. *For every $\epsilon > 0$, there is $c_0 > 0$ so that the following holds. Let $A, B \subseteq \{\pm 1\}^n$ be of size $|A| \cdot |B| \geq 2^{(1+\epsilon)n}$. Let (X, Y) be uniformly distributed in $A \times B$. For every integer k,*

$$\big| \Pr[\langle X, Y\rangle = k] - \mathbb{P}[\langle X, Y\rangle = k+4\big| \leq \frac{c_0}{n}.$$

Here is a simple application of the smoothness theorem. Consider the function f defined by $f(k) = \Pr[\langle X, Y\rangle = k]$, where here X, Y are uniformly random in a large rectangle as in Theorem 2. The theorem shows that the "derivative" of f is bounded from above, so that if f takes a large value at a point then it takes large values on a large neighborhood of that point. For example, if $f(k_0) \geq \Omega(\frac{1}{\sqrt{n}})$ for some k_0 then $f(k) \geq \frac{9}{10} f(k_0)$ for all k so that $|k - k_0| \ll \sqrt{n}$ and $k = k_0 \,\mathsf{mod}\, 4$. In particular, $f(k_0) \leq O(\frac{1}{\sqrt{n}})$.

Theorem 2 is sharp in the following two senses. First, even for the case $A = B = \{\pm 1\}^n$, there is a k so that[1]

$$|\Pr[\langle X, Y\rangle = k] - \Pr[\langle X, Y\rangle = k+4| \geq \Omega(\tfrac{1}{n}).$$

So, $O(\frac{1}{n})$ is the best upper bound possible. Secondly, as the deterministic protocol described above shows, there are sets A, B of size $|A| = |B| = 2^{n-1}$ so that for all $j \in \{1, 2, 3\}$,

$$|\Pr[\langle X, Y\rangle = 0] - \Pr[\langle X, Y\rangle = j]| = \Pr[\langle X, Y\rangle = 0] = \Omega(\tfrac{1}{\sqrt{n}})$$

So, $+4$ is the minimum gap for which an $O(\frac{1}{n})$ upper bound holds.

2. Smoothness

To prove smoothness, we use the following theorem that was initially used to prove anti-concentration [8].

Theorem 3. *For every $\beta > 0$ and $\delta > 0$, there is $c > 0$ so that the following holds. Let $B \subseteq \{\pm 1\}^n$ be of size $2^{\beta n}$. For each $\theta \in [0, 1]$, for all but $2^{n(1-\beta+\delta)}$ vectors $x \in \{\pm 1\}^n$ it holds that*

$$\left| \mathop{\mathbb{E}}_{Y} \left[\exp(2\pi i \theta \langle x, Y\rangle)\right] \right| < 2\exp(-cn \sin^2(4\pi\theta)).$$

Surprisingly, the constant 4π on the r.h.s. on the theorem above plays a crucial role in our arguments.

[1]For an integer $k = \frac{n}{2} - \sqrt{n}$, we have $\binom{n}{k+1} - \binom{n}{k} = \binom{n}{k+1}\frac{n-2k-1}{n-k} \gtrsim \frac{2^n}{n}$.

Proof of Theorem 2. Let $\beta > 0$ be so that $|B| = 2^{\beta n}$ so that $|A| \geq 2^{(1-\beta+\epsilon)n}$. Theorem 3 with $\delta = \frac{\epsilon}{3}$ promises that for each $\theta \in [0,1]$, the size of

$$A_\theta = \left\{ x \in A : \left| \mathop{\mathbb{E}}_{Y} \left[\exp(2\pi i \theta \langle x, Y \rangle) \right] \right| > 2\exp(-cn\sin^2(4\pi\theta)) \right\}$$

is at most $2^{n(1-\beta+\delta)}$. For each $x \in A$, define $S_x = \{\theta \in [0,1] : x \in A_\theta\}$.

Fix x such that $|S_x| \leq 2^{-\delta n}$. Bound

$$\begin{aligned}
&\big| \Pr_Y[\langle x, Y\rangle = k] - \Pr_Y[\langle x, Y\rangle = k+4] \big| \\
&= \left| \mathop{\mathbb{E}}_{Y} \left[\int_0^1 \exp(2\pi i\theta(\langle x, Y\rangle - k)) - \exp(2\pi i \theta(\langle x, Y\rangle - k - 4))\, \mathrm{d}\theta \right] \right| \\
&\leq \int_0^1 |\exp(4\pi i\theta) - \exp(-4\pi i \theta)| \cdot \left| \mathop{\mathbb{E}}_{Y} \left[\exp(2\pi i\theta \langle x, Y\rangle) \right] \right| \mathrm{d}\theta \\
&\leq 2\int_0^1 |\sin(4\pi\theta)| \cdot \left| \mathop{\mathbb{E}}_{Y} \left[\exp(2\pi i\theta \langle x, Y\rangle) \right] \right| \mathrm{d}\theta.
\end{aligned}$$

Continue to bound

$$\begin{aligned}
&\int_0^1 |\sin(4\pi\theta)| \cdot \left| \mathop{\mathbb{E}}_{Y} \left[\exp(2\pi i\theta \langle x, Y\rangle) \right] \right| \mathrm{d}\theta \\
&\leq 2^{-\delta n} + \int_0^1 |\sin(4\pi\theta)| \cdot \exp(-cn\sin^2(4\pi\theta))\, \mathrm{d}\theta.
\end{aligned}$$

The integral goes around the circle twice, and it is identical in each quadrant. So,

$$\begin{aligned}
&\int_0^1 |\sin(4\pi\theta)| \cdot \exp(-cn\sin^2(4\pi\theta))\, \mathrm{d}\theta \\
&= 8\int_0^{1/8} \sin(4\pi\theta) \cdot \exp(-cn\sin^2(4\pi\theta))\, \mathrm{d}\theta \\
&\leq 32\pi \int_0^\infty \theta \cdot \exp(-16cn\theta^2)\, \mathrm{d}\theta \\
&\leq \tfrac{c_1}{n} \int_0^\infty \phi \cdot \exp(-\phi^2)\, \mathrm{d}\phi \leq \tfrac{c_2}{n},
\end{aligned}$$

where $c_1, c_2 > 0$ depend on ϵ, and we used $\frac{\eta}{\pi} \leq \sin(\eta) \leq \eta$ for $0 \leq \eta \leq \frac{\pi}{2}$.

Finally, because

$$\mathop{\mathbb{E}}_{x} |S_x| = \mathop{\mathbb{E}}_{\theta} \tfrac{|A_\theta|}{2^n} \leq 2^{n(-\beta+\delta)},$$

the number of $x \in A$ for which $|S_x| > 2^{-\delta n}$ is at most $2^{-\delta n}|A|$. Hence,

$$\big| \Pr_{X,Y}[\langle X, Y\rangle = k] - \Pr_{X,Y}[\langle X, Y\rangle = k+4] \big| \leq 2^{-\delta n} + 2\big(2^{-\delta n} + \tfrac{c_2}{n}\big) \leq \tfrac{c_0}{n}. \qquad \square$$

3. The lower bound

Proof of Theorem 1. Suppose the assertion of the theorem is false. The space of inputs can be partitioned into rectangles $R_1, \ldots, R_L$ with $L \leq 2^{(1-\beta)n}$, where the output of the protocol on each R_ℓ is fixed.

Let X, Y be i.i.d. uniformly at random in $\{\pm 1\}^n$. Let E denote the event that $|\langle X, Y\rangle| = k$. Define the collection of "typical" rectangles as

$$\mathbb{T} = \Big\{\ell \in [L] \ : \ \Pr_{X,Y}[E|R_\ell] \geq \tfrac{\Pr_{X,Y}[E]}{10} \quad \& \quad \Pr_{X,Y}[R_\ell] \geq 2^{-\left(1-\frac{\beta}{2}\right)n}\Big\}.$$

For $\alpha \leq 2$, because $k = n \,\mathsf{mod}\, 2$, we have $\Pr_{X,Y}[E] \geq \frac{p}{\sqrt{n}}$ for some universal constant $p > 0$. The contribution of non-typical rectangles is small:

$$\begin{aligned}\sum_{\ell \notin \mathbb{T}} \Pr_{X,Y}[R_\ell|E] &= \tfrac{1}{\Pr_{X,Y}[E]} \sum_{\ell \notin \mathbb{T}} \Pr_{X,Y}[R_\ell] \Pr_{X,Y}[E|R_\ell] \\ &< \tfrac{1}{\Pr_{X,Y}[E]} \left(L 2^{-\left(1-\frac{\beta}{2}\right)n} + \tfrac{\Pr_{X,Y}[E]}{10} \right) < \tfrac{1}{5},\end{aligned}$$

for n large enough. Because $k = -k \,\mathsf{mod}\, 4$ and $|k| < \alpha\sqrt{n}$, for each $\ell \in \mathbb{T}$, Theorem 2 with $\epsilon \geq \frac{\beta}{2}$ implies that

$$\begin{aligned}&|\Pr_{X,Y}[\langle X, Y\rangle = k | R_\ell \wedge E] - \Pr_{X,Y}[\langle X, Y\rangle = -k | R_\ell \wedge E]| \\ &= |\Pr_{X,Y}[\langle X, Y\rangle = k | R_j] - \Pr_{X,Y}[\langle X, Y\rangle = -k | R_j]| \cdot \tfrac{1}{\Pr_{X,Y}[E|R_j]} \\ &\leq \alpha\sqrt{n}\tfrac{c_0}{n} \cdot \tfrac{10\sqrt{n}}{p} < \tfrac{1}{6},\end{aligned}$$

for α small enough. So, the probability of error conditioned on R_ℓ for $\ell \in \mathbb{T}$ is at least $\frac{5}{12}$. The total probability of error is at least

$$\sum_{\ell \in \mathbb{T}} \Pr_{X,Y}[R_\ell|E] \cdot \tfrac{5}{12} > \tfrac{4}{5} \cdot \tfrac{5}{12} = \tfrac{1}{3}.$$

This contradicts the correctness of the protocol. □

Acknowledgement. We thank Oded Regev for helpful suggestions.

References

[1] J. Brody and A. Chakrabarti. A multi-round communication lower bound for gap hamming and some consequences. In *CCC*, pages 358–368, 2009.

[2] J. Brody, A. Chakrabarti, O. Regev, T. Vidick, and R. De Wolf. Better gap-hamming lower bounds via better round elimination. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 476–489. 2010.

[3] A. Chakrabarti, G. Cormode, and A. McGregor. A near-optimal algorithm for estimating the entropy of a stream. *ACM Transactions on Algorithms*, 6(3):51, 2010.

[4] A. Chakrabarti and O. Regev. An optimal lower bound on the communication complexity of gap-hamming-distance. *SIAM Journal on Computing*, 41(5):1299–1317, 2012.

[5] P. Indyk and D. Woodruff. Tight lower bounds for the distinct elements problem. In *FOCS*, pages 283–288, 2003.
[6] T. S. Jayram, R. Kumar, and D. Sivakumar. The one-way communication complexity of hamming distance. *Theory of Computing*, 4(1):129–135, 2008.
[7] E. Kushilevitz and N. Nisan. *Communication Complexity*. Cambridge University Press, 2006.
[8] A. Rao and A. Yehudayoff. Anti-concentration in most directions. *arXiv:1811.06510*, 2018.
[9] A. Rao and A. Yehudayoff. *Communication complexity and applications*. Cambridge University Press, 2020.
[10] A. A. Sherstov. The communication complexity of gap hamming distance. *Theory of Computing*, 8(1):197–208, 2012.
[11] T. Vidick. A concentration inequality for the overlap of a vector on a large set. *Chicago Journal of Theoretical Computer Science*, 1:1–12, 2012.
[12] D. P. Woodruff. The average-case complexity of counting distinct elements. In *ICDT*, pages 284–295, 2009.

School of Computer Science, University of Washington
E-mail address: anuprao@cs.washington.edu

Department of Mathematics, Technion-IIT
E-mail address: amir.yehudayoff@gmail.com

ECCC ISSN 1433-8092
https://eccc.weizmann.ac.il