Electronic Colloquium on Computational Complexity, Revision 1 of Report No. 173 (2016)

One-sided error communication complexity of Gap Hamming Distance

Egor Klenin [*1] and Alexander Kozachinskiy [†1, 2]

[1]Lomonosov Moscow State University
[2]National Research University Higher School of Economics

Abstract

Assume that Alice has a binary string x and Bob a binary string y, both of length n. Their goal is to output 0, if x and y are at least L-close in Hamming distance, and output 1, if x and y are at least U-far in Hamming distance, where $L < U$ are some integer parameters known to both parties. If the Hamming distance between x and y lies in the interval (L, U), they are allowed to output anything. This problem is called the Gap Hamming Distance. In this paper we study public-coin one-sided error communication complexity of this problem. The error with probability at most $1/2$ is allowed only for pairs at Hamming distance at least U. In this paper we establish lower bound $\Omega(L^2/U)$ for this complexity. The main result of the paper is the simultaneous protocol communicating $O((L^2/U)\log L)$ bits. Its complexity differs from the lower bound only by a $O(\log L)$ factor.

1 Communication complexity of GHD

Given two strings $x = x_1 \dots x_n \in \{0,1\}^n$, $y = y_1 \dots y_n \in \{0,1\}^n$, Hamming distance between x and y is defined as the number of positions, where x and y differ:

$$d(x,y) = |\{i \in \{1,\dots,n\} \mid x_i \neq y_i\}|.$$

Let $L < U \leqslant n$ be integer numbers. In this paper we consider the following communication problem $\mathsf{GHD}_{L,U}$, called the Gap Hamming Distance problem:

[*]yegorklenin@gmail.com
[†]kozlach@mail.ru

ISSN 1433-8092

Definition 1. *Let Alice receive an n-bit string x and Bob an n-bit string y such that either $d(x,y) \leqslant L$, or $d(x,y) \geqslant U$. They have to output 0, if the first inequality holds, and 1, if the second inequality holds. If the promise is not fulfilled, they may output anything.*

1.1 Prior work

1.1.1 Two-sided error upper bounds

Let $R(\mathsf{GHD}_{L,U})$ denote randomized two-sided error public coin communication complexity of $\mathsf{GHD}_{L,U}$. From [11] we know that $R(\mathsf{GHD}_{L,U})) = O(L^2/(U-L)^2)$ (assuming the constant error probability less than $1/2$).

The paper [9] established the upper bound $R(\mathsf{GHD}_{L,U}) = O(L \log L)$ in the case $U = L+1$, that is, there is no gap. This bound is much better than $O(L^2/(U-L)^2)$, which is $O(L^2)$ in this case.

It turns out that the protocols attaining these two upper bounds are *simultaneous*. That is, in these protocols Alice and Bob do not communicate at all, but rather send messages to the third party, Charlie, who then computes the output of the protocol. Charlie doesn't see inputs of Alice and Bob but sees public coins. The corresponding model, called simultaneous message passing (SMP) model, is even more restricted than one-way public-coin communication: every simultaneous protocol can be converted into one-way protocol with the same communication cost.

1.1.2 One-sided error public coin communication protocols

The one-sided error public coin communication complexity will be denoted by R^0. The superscript 0 means that the protocol is allowed to err only for input pairs at distance at least U. The maximal probability of error[1] is $1/2$ The superscript 1 will mean the opposite: protocols are allowed to err only for input pairs at distance at most L.

Let us first note that for all x, y we have

$$\mathsf{GHD}_{L,U}(x,y) = \neg\mathsf{GHD}_{n-U,n-L}(\neg x, y).$$

(Alice flips all bits of her input string.) Thus $\mathsf{GHD}_{L,U}(x,y)$ reduces to $\mathsf{GHD}_{n-U,n-L}$ and the other way around. This reduction maps 0-instances to 1-instances and vice verse. This observation implies that

$$R^1(\mathsf{GHD}_{L,U}) = R^0(\mathsf{GHD}_{n-U,n-L}).$$

[1] by the standard amplification argument we could have any constant between 0 and 1 instead of $1/2$.

Thus it suffices to study only one of these quantities and we will stick to R^0 (the error is allowed when the distance is at least U) .

The paper [8] noticed that for all $U > L$ it holds that $R^0(\mathsf{GHD}_{L,U}) = O(L \log n)$. Once again, there is a public coin SMP protocol attaining this bound (which is just a simple modification of the standard protocol for EQUALITY).

1.1.3 GHD and the lower bounds in data streams and property testing

Several works used GHD to obtain lower bounds for streaming algorithms and for property testing problems. As was discovered in [14] by Woodruff, there is a reduction from GHD to a number of fundamental data stream problems, including the problem of estimating frequency moments. More specifically, if there is a $\Omega(n)$ lower bound against any r-round two-sided error communication protocol for $\mathsf{GHD}_{n/2-\Theta(\sqrt{n}),\ n/2+\Theta(\sqrt{n})}$, then there is a $\Omega(1/(r\varepsilon^2))$ lower bound on the space complexity of any r-pass streaming algorithm estimating the frequency moments in a data stream within a factor of $(1+\varepsilon)$.

In [14] Woodruff proved $\Omega(n)$ lower bound against 1-round protocols (see also [10] for more direct and simple proof). In a subsequent works ([2, 3]) this lower bound was extended to $O(1)$-round protocols. Finally, $\Omega(n)$ lower bound in the most general setting, when there is no restriction on the number of rounds, was obtained in [6, 13, 12].

As it turns out, lower bounds on the one-sided error version of GHD are also useful. In [5] Buhrman, Cleve and Wigderson proved that for any constant $c < 1$ it holds that $R^1(\mathsf{GHD}_{0,\,cn}) = \Omega(n)$. Moreover, they showed that $\Omega(n)$ lower bound holds also for a weaker version of $\mathsf{GHD}_{0,\,cn}$ problem, in which Hamming distance between the inputs is either 0 or *exactly* cn (provided that cn is the even integer).

Blais, Brody and Matulef in [1] used this result to obtain lower bounds on testing decision trees and signed majorities with one-sided error.

Further, Brody and Woodruff ([4]) used the lower bound on the one-sided error GHD from [5] to obtain lower bounds for streaming algorithms with *one-sided approximation*, i.e., for algorithms which either always return an overestimate or always return an underestimate on the objective function. Their results include lower bounds for the problem of over(under)-estimating the number of non-zero rows in a matrix and the Earth Mover Distance between two multisets.

1.2 This work

In this paper we study public-coin one-sided error communication complexity R^0 of $\mathsf{GHD}_{L,U}$. Once again, the error is allowed only for pairs at Hamming distance at least U.

1.2.1 The upper bound

Our main result is a one-sided error public-coin simultaneous protocol for $\mathsf{GHD}_{L,U}$ with communication complexity $O((L^2/U)\log L)$. It is constructed in the following 4 steps (description of the protocol in this section is a bit informal, and the precise bounds can be found below in the paper).

Step 1. On this step we construct our main novel protocol, called *the Triangle Inequality Protocol*. It communicates $O((L^2/U)\log n)$ bits (which is a little bit more than required, since $\log L$ is replaced by $\log n$) and solves the $\mathsf{GHD}_{L,U}$ problem when the ratio U/L is larger than a certain constant.

The protocol works as follows. It randomly splits x and y in $b = O(L^2/U)$ blocks $x^1, \ldots, x^b$ and $y^1, \ldots, y^b$. The ith bit x_i of x goes in the block x^j where j is chosen at random with uniform probability distribution over $\{1, \ldots, b\}$, and decisions for different i's are independent. Each bit y_i of y goes in the block y^j with the same index as x_i goes in. This partition is made using the shared random source (so that the parties have the same partition). Both parties also read random strings $r^1, \ldots, r^b$ from the shared random source and Alice communicates $d(x^j, r^j)$ to Charlie for all $j = 1, \ldots, b$. Bob does the same with $d(y^j, r^j)$. Thus the communication is $b \log n = O((L^2/U)\log n)$. Charlie outputs 0 if the sum

$$\sum_{j=1}^{b} |d(x^j, r^j) - d(y^j, r^j)|$$

is at most L and 1 otherwise. By the triangle inequality each term in this sum is at most $d(x^j, y^j)$ and thus the sum is at most $d(x, y)$. Therefore this protocol does not err if $d(x, y) \leqslant L$.

On the other hand, if $d(x, y) \geqslant U \geqslant C'L$ for a certain constant C', then for any fixed j the average value of $d(x^j, y^j)$ is at least 2. From the properties of binomial distributions it follows that we have $d(x^j, y^j) \geqslant d(x, y)/10b$ with probability at least $1/3$. The value $d(x^j, r^j) - d(y^j, r^j)$ is distributed as the distance from the origin in a random walk with $d(x^j, y^j)$ steps along a line (each step has length 1 and is directed to the left or to the right with equal probabilities). From the properties of random walks it follows that for every j we have $|d(x^j, r^j) - d(y^j, r^j)| > \sqrt{d(x^j, y^j)}$ with constant positive

probability. These two facts imply that with constant probability the sum $\sum_{j=1}^{b} |d(x^j, r^j) - d(y^j, r^j)|$ is $\Omega(b\sqrt{d(x,y)/10b}) = \Omega(\sqrt{bd(x,y)})$. Recall that $b = O(L^2/U)$ and we assume that $d(x,y) \geqslant U$. If the constant hidden in O-notation is large enough then the lower bound $\Omega(\sqrt{bd(x,y)})$ for the sum $\sum_{j=1}^{b} |d(x^j, r^j) - d(y^j, r^j)|$ is larger than L.

Step 2. In [8] it was noticed that for all $L < U$ there is one-sided error public-coin simultaneous protocol for $\mathsf{GHD}_{L,U}$ with communication $O(L \log n)$. This protocol is just a modification of the standard protocol for equality and it never errs for inputs at distance at most L.

Our second protocol runs the Triangle Inequality Protocol if $U > C'L$ and the protocol from [8] otherwise. Notice that in the latter case $L = O(L^2/U)$, and thus we obtain protocols with communication $O((L^2/U) \log n)$ for all L, U.

Step 3. On this step we use the techniques from [9] to replace the $\log n$ factor by a $\log L$ factor. More specifically, we run the protocol from the second step for the strings u, v of length $O(L^{16})$ obtained from the original strings x, y by the following transformation. As in the Triangle Inequality Protocol, we split x, y into $b = O(L^{16})$ blocks and then replace each block by the parity of its bits. Obviously, $d(u,v) \leqslant d(x,y)$. We then show that $d(u,v) = d(x,y)$ with constant probability provided $d(x,y) \leqslant L^8$. Therefore this protocol has constant one-sided error probability for all input pairs with $d(x,y) \leqslant L^8$. By construction this protocol communicates $O((L^2/U) \log L)$ bits.

Step 4. Finally, to handle the case $d(x,y) > L^8$, we consider the following protocol. We run the protocol from step 3 and then a simplified version of the Triangle Inequality Protocol. If any of these two protocols output 1, we output 1 and otherwise 0. The simplified version of the Triangle Inequality Protocol works as follows. Alice and Bob read a random n-bit string r from the shared random source. They compute distance from their inputs to r. Observe that due to triangle inequality $|d(x,r) - d(y,r)| \leqslant d(x,y)$. Hence $d(x,y) \leqslant L$ implies $|d(x,r) - d(y,r)| \leqslant L$. On the other hand, if $d(x,y) > L^8$, then due to the properties of random walks with constant positive probability it holds that $|d(x,r) - d(y,r)| > L^4$.

Thus step 4 is reduced to the following communication problem. Alice holds a number $a \in \{0, 1, \ldots, n\}$, Bob holds a number $b \in \{0, 1, \ldots, n\}$ and it is known that either $|a - b| \leqslant L$ or $|a - b| > L^4$. The goal is to find out whether the first or the second inequality is true. We construct a public-coin simultaneous protocol with communication $O(\log L)$ which always outputs 0 when $|a - b| \leqslant L$ and which with some constant positive probability outputs 1 when $|a - b| > L^4$.

There is a simple SMP protocol communicating $O(\log L + \log\log n)$ bits to solve even a gap-less (L vs $L+1$) version of this problem. Let $p_1, \ldots, p_k$ be the first $k = (4L+2)\cdot \log_2(n+L)$ primes. Parties read a random $p \in \{p_1, \ldots, p_k\}$ from the shared random source. Alice and Bob communicate $a \pmod p$ and $b \pmod p$ to Charlie. He checks, whether there is $i \in [-L, L]$ such that $a + i \pmod p = b \pmod p$. If there is such i, he outputs 0, otherwise 1.

Clearly the protocol has one-sided error: if $|a-b| \leqslant L$, then $a + i \pmod p = b \pmod p$ is true for $i = b - a$ and the protocol always outputs 0. On the other hand, if $|a-b| > L$, then b is different from $a+i$ for every $i \in [-L, L]$ and hence there is at most $\log_2(|b-a-i|) \leqslant \log_2(n+L)$ primes, dividing $b-(a+i)$. By union bound the probability that the protocol outputs 0 in this case is at most $(2L+1)/(4L+2) = 0.5$

The problem with this protocol is that a and b range from 0 to n. To get rid of $O(\log\log n)$ term we do some additional hashing. Let $Z_0, \ldots, Z_n$ be $(n+1)$ independent symmetric Bernoulli random variables. Define

$$h : \{0, 1, \ldots, n\} \to \{0, 1, \ldots, m-1\}, \qquad h(s) = Z_0 + Z_1 + \ldots + Z_s \pmod m,$$

where $m = \mathsf{poly}(L)$. Clearly $|h(a) - h(b)| \leqslant |a-b|$. It turns out that $|a-b| > L^4$ implies that with some constant positive probability $|h(a)-h(b)| > L$. It remains to apply the protocol with communication cost $O(\log L + \log\log n)$, constructed earlier, to $(h(a), h(b))$ (but now instead of $\log\log n$ we will have $\log\log m = \log\log L$).

1.2.2 Lower bounds

As we mentioned, for every constant $c < 1$ for all $U < cn$ the one-sided error communication complexity (the error is allowed when the distance is 0) of $\mathsf{GHD}_{0,U}$ is $\Omega(n)$ ([5]). However, from the argument in [5] it is not clear how the constant hidden in $\Omega(n)$ depends on c. In this paper we answer this question by proving that the constant is $\Omega((1-c)^2)$. In other words, we show that the one-sided error complexity R^1 (the error is allowed when the distance is 0) of $\mathsf{GHD}_{0,U}$ is $\Omega((n-U)^2/n)$.

As a corollary we obtain the lower bound $\Omega(L^2/U)$ for one-sided error complexity R^0 of $\mathsf{GHD}_{L,U}$ (the error is allowed when the distance is at least U). As we explained earlier, R^1 of $\mathsf{GHD}_{0,U-L}$ on U-bit strings equals R^0 of $\mathsf{GHD}_{L,U}$ on U-bit strings. As the former is $\Omega((U-(U-L))^2/U)$ we obtain the lower bound $\Omega(L^2/U)$ for the latter. On the other hand, the problem $\mathsf{GHD}_{L,U}$ on U-bit strings reduces to the problem $\mathsf{GHD}_{L,U}$ on n-bit strings

(Alice and Bob append $n - U$ zeros to their strings) hence the one-sided complexity R^0 of the latter is also $\Omega(L^2/U)$.

1.2.3 The summary

Let us summarize our results.

Theorem 1. *Assume that $L > 1$. The one-sided error public-coin communication complexity R^0 of $\mathsf{GHD}_{L,U}$ on n-bit strings is at most*

$$O\left(\left(\frac{L^2}{U} + 1\right) \log L\right)$$

(The error is allowed only when the distance is at least U.) There is a public-coin simultaneous protocol attaining this bound.

Theorem 2. *The one-sided error public-coin communication complexity R^1 of $\mathsf{GHD}_{0,U}$ on n-bit strings is at least*

$$\Omega\left(\frac{(n-U)^2}{n} + 1\right).$$

(The error is allowed only when the distance is 0.)

Corollary 3. *The one-sided error public-coin communication complexity R^0 of $\mathsf{GHD}_{L,U}$ on n-bit strings is at least*

$$\Omega\left(\frac{L^2}{U} + 1\right).$$

(The error is allowed only when the distance is at least U.)

Thus our results determine the one-sided public-coin communication complexity of $\mathsf{GHD}_{L,U}$ (up to a factor $O(\log L)$) in the case when the parties are allowed to err only for input pairs at distance at least U. If the parties are allowed to err only for input pairs at distance at most L, then the one-sided public-coin communication complexity of $\mathsf{GHD}_{L,U}$ is $(n-U)^2/(n-L)$ up to a factor of $O(\log(n-U))$.

2 Preliminaries

Let $f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ be a Boolean function.

Definition 2. *A deterministic communication protocol is a rooted binary tree, in which each non-leaf vertex is associated either with Alice or with Bob and each leaf is labeled by* 0 *or* 1*. Each non-leaf vertex v, associated with Alice, is assigned a function $f_v : \mathcal{X} \to \{0,1\}$ and each non-leaf vertex u, associated with Bob, is assigned a function $g_u : \mathcal{Y} \to \{0,1\}$. For each non-leaf vertex one of its out-going edges is labeled by* 0 *and other one is labeled by* 1*.*

A computation according to a deterministic protocol runs as follows. Alice is given $x \in \mathcal{X}$, Bob is given $y \in \mathcal{Y}$. They start at the root of tree. If they are in a non-leaf vertex v, associated with Alice, Alice sends $f_v(x)$ to Bob and they move to the son of v by the edge labeled by $f_v(x)$. If they are in a non-leaf vertex, associated with Bob, they act in a similar same way, however this time it is Bob who sends a bit to Alice. When they reach a leaf, they output the bit which labels this leaf.

Definition 3. *Communication complexity of a deterministic protocol π, denoted by $CC(\pi)$, is defined as the depth of the corresponding binary tree.*

Randomized protocols with shared randomness (aka public-coin protocols) can be defined as follows:

Definition 4. *A public-coin communication protocol is a probability distribution over deterministic protocols. Communication complexity of a public-coin protocol τ, denoted by $CC(\tau)$, is defined as $\max_{\pi} CC(\pi)$, where π is taken over the deterministic protocols from the support of τ (recall that τ is the distribution).*

Given a public-coin protocol τ, Alice and Bob choose the deterministic protocol to be executed according to the distribution, defined by τ.

Definition 5. *We say that a public-coin protocol computes a partial function f with error probability ε, if for every pair of inputs (x, y) in the domain of f with probability at least $1 - \varepsilon$ that protocol outputs $f(x, y)$. Randomized communication complexity of f is defined as*

$$R_\varepsilon(f) = \min_{\pi} CC(\pi),$$

where minimum is over all protocols that compute f with error probability ε.

A deterministic *simultaneous* protocol τ is a triple $\langle \phi, \psi, \theta \rangle$ where

$$\phi : \mathcal{X} \to \{0,1\}, \qquad \psi : \mathcal{Y} \to \{0,1\}^m,$$

$$\theta : \{0,1\}^n \times \{0,1\}^m \to \{0,1\}.$$

The communication cost of τ is $n+m$. A public-coin simultaneous protocol π is a probability distribution over deterministic simultaneous protocols. Communication cost of π is the maximal possible communication cost of τ, where τ is a deterministic simultaneous protocol taken from the support of π.

Assume that Alice is given $x \in \mathcal{X}$ and Bob is given $y \in \mathcal{Y}$. The output of a public-coin simultaneous protocol π on (x, y) is a random variable defined as follows. Sample a deterministic simultaneous protocol $\tau = \langle \phi, \psi, \theta \rangle$ according to π. Output $\theta(\phi(x), \psi(y))$.

If for $i \in \{0,1\}$ we require that the protocol never errs on inputs from $f^{-1}(i)$, then the corresponding notion is called "randomized one-sided error communication complexity" and is denoted by $R^i_\varepsilon(f)$.

The Gap Hamming Distance problem is the problem of computing the following partial function:

$$\mathsf{GHD}^n_{L,U}(x,y) = \begin{cases} 0 & d(x,y) \leqslant L, \\ 1 & d(x,y) \geqslant U, \\ \text{undefined} & L < d(x,y) < U, \end{cases} \qquad \text{for } x, y \in \{0,1\}^n.$$

3 The upper bound

The protocol for Theorem 1 is a combination of three different protocols. The most important of them solves $\mathsf{GHD}_{L,U}$ with one sided error in the case when U/L exceeds some constant. Its communication length is $O((L^2/U + 1)\log n)$. We call that protocol the "Triangle Inequality Protocol", because it uses the triangle inequality for Hamming distance.

3.1 The Triangle Inequality Protocol

The following Lemma is the standard fact of Probability Theory:

Lemma 4. *There exists a positive constant $\alpha > 0$ such that for every m it holds that*

$$\Pr[S_m \geqslant \sqrt{m}] \geqslant \alpha,$$

where S_m denotes one-dimensional random walk with m steps, i.e, S_m is equal to the sum of m independent random variables, each taking the values 1 and -1 with probabilities 1/2.

Everywhere below α stands for the constant from Lemma 4.

Let $x, y \in \{0,1\}^n$ denote Alice's and Bob's input strings, respectively. The parties set $b = \lceil CL^2/U + 1 \rceil$, where $C = 360/\alpha^2$. Then they use public coins to sample a function $\chi : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, b\}$ uniformly at random. They use χ to divide x and y into b blocks

$$x^1, \ldots, x^b, \qquad y^1, \ldots, y^b.$$

The block x^j consists of all bits x_i of x such that $\chi(i) = j$. Similarly, y^j consists of all bits y_i with $\chi(i) = j$. The order in which bits of jth block are arranged is not important, the parties care only that they use the same order.

Then they use public coins to sample b random strings $r^1, \ldots, r^b$ of the same lengths, as $x^1, \ldots, x^b$ and $y^1, \ldots, y^b$. Alice then sends b numbers to Charlie:

$$d(x^1, r^1), \ldots, d(x^b, r^b).$$

In turn, Bob sends

$$d(y^1, r^1), \ldots, d(y^b, r^b).$$

Then Charlie computes the sum

$$T = \sum_{j=1}^{b} \left| d(x^j, r^j) - d(y^j, r^j) \right|.$$

If $T \leqslant L$, Charlie outputs 0. Otherwise he outputs 1.

If $d(x, y) \leqslant L$, then the protocol always outputs 0. Indeed, since Hamming distance satisfies the triangle inequality, we have that

$$T = \sum_{j=1}^{b} \left| d(x^j, r^j) - d(y^j, r^j) \right| \leqslant \sum_{j=1}^{b} d(x^j, y^j) = d(x, y) \leqslant L.$$

Thus this protocol has a one-sided error: it can err only if $d(x, y) \geqslant U$. Now we will estimate the probability of error in the case when $d(x, y) \geqslant U$.

Lemma 5. *Assume that $U \geqslant 2b$. Then the protocol for the input pair x, y at distance at least U outputs 1 with some positive constant probability (more specifically, with probability at least $\alpha/6$).*

Proof. Assume that U, L, x, y satisfy the assumption of the lemma. Fix $j = 1, \ldots, b$. First we have to understand what is the distribution of the

random variable $|d(x^j, r^j) - d(y^j, r^j)|$. By construction Alice and Bob choose a random function χ that governs the partition of x, y into blocks. For each i such that $x_i \neq y_i$ the probability that x_i, y_i land into the block with number j is $1/b$. Hence the random variable $d(x^j, y^j)$ has binomial distribution $B(d(x, y), 1/b)$ with parameters $d(x, y)$ and $1/b$, i. e., the probability of the event $d(x^j, y^j) = k$ equals

$$\binom{d(x,y)}{k}(1/b)^k(1-1/b)^{d(x,y)-k}.$$

The average value of $d(x^j, y^j)$ is thus equal to $d(x, y)/b$.

Once x^j, y^j are determined, Alice and Bob sample r^j. The value $d(x^j, r^j) - d(y^j, r^j)$ can be represented as the sum of $|x^j| = |y^j|$ terms where each term corresponds to a number i with $\chi(i) = j$. If $x_i = y_i$ then the term is 0. Otherwise it is either -1 or 1 depending on whether the respective bit of r^j is equal to x_i or to y_i. Thus for every fixed partition into blocks the value $|d(x^j, r^j) - d(y^j, r^j)|$ is distributed as the distance from origin in the random walk along the line with $d(x^j, y^j)$ independent steps where each step is 1 with probability $1/2$ and -1 with the same probability.

To finish the proof we will use the following fact about binomial distribution, which is proven in the Appendix.

Lemma 6. *If X is distributed according to the binomial distribution $B(n, p)$ and $pn \geqslant 2$, then*

$$\Pr\left[X > \frac{pn}{10}\right] \geqslant \frac{1}{3}.$$

Recall that the random variable $d(x^j, y^j)$ has binomial distribution $B(d(x, y), 1/b)$ and we assume that $d(x, y)/b \geqslant U/b \geqslant 2$. Hence by Lemma 6 with probability at least $1/3$ we have $d(x^j, y^j) \geqslant d(x, y)/10b$.

Fix any partition into blocks such that $d(x^j, y^j) \geqslant d(x, y)/10b$. By Lemma 4 with probability at least α we have

$$|d(x^j, r^j) - d(y^j, r^j)| \geqslant \sqrt{d(x^j, y^j)} \geqslant \sqrt{d(x, y)/10b}.$$

We have proved that for every fixed j with probability at least $\alpha/3$ we have $|d(x^j, r^j) - d(y^j, r^j)| \geqslant \sqrt{d(x, y)/10b}$. A simple averaging argument shows that with probability at least $\alpha/6$ the fraction of j that satisfy this inequality is bigger than $\alpha/6$. Indeed, let the random variable θ denote the fraction of j that satisfy this inequality. Its average is at least $\alpha/3$. On the other hand, we can upperbound its average by the sum

$$\Pr[\theta > \alpha/6] \cdot 1 + \Pr[\theta \leqslant \alpha/6] \cdot (\alpha/6) \leqslant \Pr[\theta > \alpha/6] + \alpha/6.$$

Thus with probability $\alpha/6$ we have

$$\sum_{j=1}^{b} |d(x^j, r^j) - d(y^j, r^j)| > (\alpha/6)b\sqrt{d(x,y)/10b} = (\alpha/6)\sqrt{b \cdot d(x,y)/10}.$$

Recall that $b = \lceil CL^2/U + 1 \rceil$, where $C = 360/\alpha^2$, and $d(x,y) \geqslant U$. So the right hand side of the last displayed inequality is strictly larger than L. $\square$

If the ratio U/L is larger than a certain constant then the protocol solves $\mathsf{GHD}_{L,U}$ with constant one-sided error-probability. One can verify that the assumption $U \geqslant 2b$ of Lemma 5 is met for all $U \geqslant 2CL + 2$.

Recall that the communication length of the protocol is $O((L^2/U + 1)\log n)$. Now we need a protocol with the same communication length for L, U such that $U \leqslant 2CL + 1$. Notice that in this case $L \leqslant 2CL^2/(U-1)$ and the upper bound $O((L^2/U)\log n)$ for communication boils down to $O(L \log n)$. A protocol with such performance was constructed in [8].

3.2 The protocol of [8]

For the reader's convenience and to stress that the protocol from [8] has one-sided error we give here its full description.

Here $\oplus$ stands for the bit-wise XOR over n-bit vectors and $\langle \cdot, \cdot \rangle : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ denotes the inner product over $\mathbb{F}_2$:

$$\langle a, b \rangle = \sum_{i=1}^{n} a_i b_i \pmod 2.$$

Let $x, y \in \{0,1\}^n$ denote Alice's and Bob's input strings, respectively. They use public coins to sample N vectors

$$R_1, \ldots, R_N \in \{0,1\}^n$$

independently uniformly at random. Alice sends $\langle x, R_1 \rangle, \ldots, \langle x, R_N \rangle$ to Charlie. Bob does the same with y. If there is $f \in \{0,1\}^n$ of Hamming weight at most L such that:

$$\langle x \oplus f, R_1 \rangle = \langle y, R_1 \rangle, \ldots, \langle x \oplus f, R_N \rangle = \langle y, R_N \rangle, \tag{1}$$

then Charlie outputs 0. Otherwise Charlie outputs 1.

Such protocol costs $O(N)$ bits. If $d(x,y) \leqslant L$, then the protocol outputs 0 with probability 1. Indeed, $f = x \oplus y$ (which is of Hamming weight at most L in this case) satisfies (1).

Now assume that $d(x,y) > L$. Then any $f \in \{0,1\}^n$ of Hamming weight at most L satisfies (1) only with probability at most 2^{-N} (because $x + f \neq y$). Hence the error probability of the protocol is at most $V_2(n,L) \cdot 2^{-N}$ in this case. Here $V_2(n,L)$ is the size of Hamming ball of radius L. As $V_2(n,L) \leqslant (n+1)^L$, it is enough to take $N = O(L \log n)$.

3.3 The simplified version of the Triangle Inequality Protocol

Thus for all L, U we have a public-coin simultaneous protocol with communication length $O((L^2/U+1)\log n)$ to solve $\mathsf{GHD}_{L,U}$ with constant one-sided error probability. To replace $\log n$ factor by $\log L$ factor we will need the following public-coin simultaneous protocol with communication length $O(\log L)$ to solve $\mathsf{GHD}_{L,(4L+2+m_1)^8}$ with constant one-sided error probability (here m_1 is the large enough constant to be defined later). Notice that $O((L^2/U+1)\log L)$ becomes $O(\log L)$ for $U = (4L+2+m_1)^8$.

The protocol. The parties use public coins to sample a vector $r \in \{0,1\}^n$ uniformly at random. Alice and Bob compute the distance from r to their input strings. If $d(x,y) \leqslant L$, then by Triangle Inequality we have $|d(x,r) - d(y,r)| \leqslant d(x,y) \leqslant L$. On the other hand, assume that $d(x,y) \geqslant (4L + 2 + m_1)^8$. From Lemma 4 it follows that in this case with constant positive probability we have $|d(x,r) - d(y,r)| \geqslant \sqrt{d(x,y)} > (4L+2)^4 + m_1^4$.

Consider the following auxiliary problem. Alice holds a number $a \in \{0,1,\ldots,n\}$, Bob holds a number $b \in \{0,1,\ldots n\}$ and it is promised that either $|a-b| \leqslant L$ or $|a-b| > (4L+2)^4 + m_1^4$. They want to know whether the first or the second inequality is true. As the previous paragraph shows, if there is a public-coin SMP protocol with communication length $O(\log L)$, which always outputs 0 when $|a-b| \leqslant L$ and which with constant positive probability outputs 1 when $|a-b| > (4L+2)^4 + m_1^4$, then we are done.

We will need the following lemma:

Lemma 7. *There is a positive integer m_0 and a positive real c such that the following holds. Assume that $m \geqslant m_0, N \geqslant m^4$. Consider N independent random variables $Z_1, \ldots, Z_N$, where each variable takes the values 0 and 1 with probabilities 1/2. Then for every $i \in \{0,1,\ldots,m-1\}$ it holds that:*

$$\Pr[Z_1 + \ldots + Z_N = i \pmod m] \geqslant \frac{c}{m}.$$

The proof of this Lemma can be found in Appendix.

Set $m_1 = m_0$. Take $m = \max\{m_1, (4L+2)\}$. Use public coins to sample $n+1$ independent random variables

$$Z_0, Z_1, Z_2, \ldots, Z_n,$$

where each variable takes the values 0 and 1 with probabilities 1/2.

Define

$$h(x) = Z_0 + Z_1 + \ldots + Z_x \pmod{m} \in \{0, 1, \ldots, m-1\}.$$

Let $p_1, \ldots, p_k$ be the first k primes, where $k = \lceil \log_2(m+L) \rceil \cdot (4L+2)$. Parties sample $p \in \{p_1, \ldots, p_k\}$ uniformly at random. Alice sends $h(a) \pmod{p}$ to Charlie, Bob does the same with b. Charlie checks, whether there is $i \in [-L, L]$ such that $h(a) + i = h(b) \pmod{p}$. If there is one, then Charlie outputs 0. Otherwise Charlie outputs 1.

The protocol communicates $O(\log p_k) = O(\log k) = O(\log L)$ bits. Further, it is easy to see that the protocol has one-sided error; this follows from the fact that the absolute value of $h(b) - h(a)$ is always bounded by $|a-b|$. That is, the protocol always outputs 0 when $|a-b| \leqslant L$.

Now, let's consider the case when $|a-b| > (4L+2)^4 + m_1^4$. Since the protocol is symmetric with respect to Alice and Bob we may assume that $a < b$. Then

$$h(b) - h(a) = Z_{a+1} + \ldots + Z_b \pmod{m}.$$

Observe that $N = |b-a|$ is at least m^4 and $m \geqslant m_1 = m_0$. Thus it is possible to apply Lemma 7 to $h(b) - h(a)$. Namely, Lemma 7 implies that with constant positive probability $h(b) - h(a)$ is different from any $i \in [-L, L]$. Indeed, if $h(b) - h(a) = i$ for some $i \in [-L, L]$, then it is also true that

$$h(b) - h(a) = Z_{a+1} + \ldots + Z_b = i \pmod{m} \tag{2}$$

Let E be the event that $h(b) - h(a)$ avoids (2) for every $i \in [-L, L]$. By Lemma (7) the probability of E is at least $c \cdot \frac{m-(2L+1)}{m} \geqslant \frac{c}{2}$.

Once again, E implies that $h(b)$ is different from $h(a) + i$ for all $i \in [-L, L]$. In particular, E implies that there are at most $\log_2 |h(b) - (h(a)+i)| \leqslant \log_2(m+L)$ different primes, dividing $h(b) - (h(a)+i)$. Thus, conditioned on E, for any fixed $i \in [-L, L]$ the equality $h(a) + i = h(b) \pmod{p}$ is true only with probability at most $1/(4L+2)$. Overall, by union bound the protocol outputs 1 with probability at least

$$\Pr[E] \left(1 - \frac{2L+1}{4L+2}\right) \geqslant c/4,$$

and the latter is the positive constant.

3.4 The final protocol for Theorem 1

The protocol. Step 1. Alice and Bob first run the Simplified Triangle Inequality Protocol from the previous section. If that protocol outputs 1 they output 1 and halt. Otherwise they proceed to Step 2.

Step 2. They divide x and y into $w = 2(4L+2+m_1)^{16}$ blocks randomly (as in the construction of the Triangle Inequality Protocol). Let

$$x^1, \ldots, x^w, \qquad y^1, \ldots, y^w$$

denote the resulting blocks. Let u_i be the XOR of all bits from x^i and let v_i be the XOR of all bits from y^i. Alice privately computes $u_1, \ldots, u_w$ and sets $u = u_1 \ldots u_w$. Bob privately computes $v_1, \ldots, v_w$ and sets $v = v_1 \ldots v_w$.

Recall that we have a protocol with communication length $O((L^2/U + 1)\log w) = O((L^2/U + 1)\log L)$ to solve $\mathsf{GHD}_{L,U}$ on w-bit strings with constant positive one-sided error probability.

Alice and Bob run this protocol for input pair (u, v) (and not (x, y)). They output the result of this run.

The communication length of the constructed protocol is $O((L^2/U + 1)\log L)$. We have to show that it has one-sided constant error probability.

If $d(x, y) \leqslant L$ then the run of the Simplified Triangle Inequality Protocol will output 0 with probability 1. Thus they proceed to Step 2. The distance between u and v does not exceed the distance between x and y and hence is at most L. Thus the run of the second protocol also outputs 0 with probability 1.

Assume that $d(x, y) \geqslant U$. If $d(x, y) \geqslant (4L+2+m_1)^8$, then the Simplified Triangle Inequality Protocol outputs 1 with positive constant probability, they output 1 and halt.

Assume that $U \leqslant d(x, y) < (4L + 2 + m_1)^8$. We claim that in this case with constant positive probability we have $d(u, v) = d(x, y)$. Indeed, consider any two positions in which x and y differ. Those positions land into the same block with probability $\frac{1}{w}$. By union bound, with probability at least

$$1 - \frac{d(x, y)^2}{w} \geqslant 1 - \frac{(4L + 2 + m_1)^{16}}{2(4L + 2 + m_1)^{16}} = 0.5$$

all the positions in which x and y differ land in different blocks. The latter means that for all i the blocks x^i and y^i differ in at most 1 position and hence $d(u, v) = d(x, y)$. Thus with probability at least $1/2$ we have $d(u, v) \geqslant U$ and Alice and Bob output 1 with positive constant probability on the second step.

4 The lower bound

In this section we prove Theorem 2.

Proof of Theorem 2. Let τ be a protocol witnessing $R^1_{\frac{1}{2}}(\mathsf{GHD}_{0,U})$. Then the following hold:

- for each $x \in \{0,1\}^n$ the protocol τ for input (x,x) outputs 0 with probability at least $\frac{1}{2}$;
- for all $x, y \in \{0,1\}^n$ with $d(x,y) \geqslant U$ the protocol τ always outputs 1.

By the standard averaging argument due to von Neumann there is a deterministic protocol π such that

- the communication complexity of π is at most $R^1_{\frac{1}{2}}(\mathsf{GHD}_{0,U})$;
- π outputs 0 for at least half of diagonal input pairs (x,x);
- π outputs 1 for all inputs pairs at Hamming distance at least U.

Consider any 0-leaf of π and the corresponding rectangle $R = A \times B \subset \{0,1\}^n \times \{0,1\}^n$. The number of diagonal pairs from R is equal to $|A \cap B|$. Diameter of $A \cap B$ must be less than U. Indeed, if there are $x, y \in A \cap B$ such that $d(x,y) \geqslant U$, then π outputs 0 for input pair (x,y).

It turns out that the largest set of diameter $2r < n$ is the Hamming ball of radius r and the diameter of the latter is at most $2^{n(1-c(1-2r/n)^2)}$ for some positive constant c (Lemma 11 in the Appendix).

Let $r = \lfloor U/2 \rfloor$. For $U = n$ the lower bound in Theorem 2 is constant and thus the statement is obvious. Therefore we may assume that $U < n$ and hence $r < n/2$. The diameter of $A \cap B$ is at most $2r$ (recall that the diameter of $A \cap B$ is strictly less than U). By Lemma 11 we have

$$|A \cap B| \leqslant 2^{n(1-c(1-2r/n)^2)} \leqslant 2^{n(1-c(1-U/n)^2)}.$$

We have shown that if R is the rectangle corresponding to a 0-leaf of π, then R covers at most $2^{n(1-c(1-U/n)^2)}$ diagonal pairs. As the total number of diagonal pairs covered by 0-leaves of π is at least 2^{n-1}, the number of 0-leaves in π is at least $2^{cn(1-U/n)^2-1}$. Thus we have

$$R^1_{\frac{1}{2}}(\mathsf{GHD}_{0,U}) \geqslant c \cdot \frac{(n-U)^2}{n} - 1. \tag{3}$$

Obviously we also have

$$R^1_{\frac{1}{2}}(\mathsf{GHD}_{0,U}) \geqslant 1. \tag{4}$$

From inequalities (3) and (4) we can easily deduce that

$$R^1_{\frac{1}{2}}(\mathsf{GHD}_{0,U}) \geqslant \Omega\left(\frac{(n-U)^2}{n} + 1\right)$$

(for example, we can add these inequalities with appropriate positive weights). □

References

[1] Blais, E., Brody, J., and Matulef, K. Property testing lower bounds via communication complexity. *Computational Complexity 21*, 2 (2012), 311–358.

[2] Brody, J., and Chakrabarti, A. A multi-round communication lower bound for gap hamming and some consequences. In *Computational Complexity, 2009. CCC'09. 24th Annual IEEE Conference on* (2009), IEEE, pp. 358–368.

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

[4] Brody, J., and Woodruff, D. P. Streaming algorithms with one-sided estimation. In *APPROX-RANDOM* (2011), Springer, pp. 436–447.

[5] Buhrman, H., Cleve, R., and Wigderson, A. Quantum vs. classical communication and computation. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing* (1998), ACM, pp. 63–68.

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

[7] Cohen, G., Honkala, I., Litsyn, S., and Lobstein, A. *Covering codes*, vol. 54. Elsevier, 1997.

[8] Gavinsky, D., Kempe, J., and de Wolf, R. Quantum communication cannot simulate a public coin. *arXiv preprint quant-ph/0411051* (2004).

[9] Huang, W., Shi, Y., Zhang, S., and Zhu, Y. The communication complexity of the hamming distance problem. *Information Processing Letters 99*, 4 (2006), 149–153.

[10] Jayram, T. S., Kumar, R., and Sivakumar, D. The one-way communication complexity of hamming distance. *Theory of Computing 4*, 1 (2008), 129–135.

[11] Kushilevitz, E., Ostrovsky, R., and Rabani, Y. Efficient search for approximate nearest neighbor in high dimensional spaces. *SIAM Journal on Computing 30*, 2 (2000), 457–474.

[12] Sherstov, A. A. The communication complexity of gap hamming distance. *Theory of Computing 8*, 1 (2012), 197–208.

[13] Vidick, T. A concentration inequality for the overlap of a vector on a large set, with application to the communication complexity of the gap-hamming-distance problem. *Chicago Journal of Theoretical Computer Science 1* (2012).

[14] Woodruff, D. Optimal space lower bounds for all frequency moments. In *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms* (2004), Society for Industrial and Applied Mathematics, pp. 167–175.

A Auxiliary results

A.1 Hamming Space

Definition 6. *The function*

$$h(x) = x \log_2 \frac{1}{x} + (1-x) \log_2 \frac{1}{1-x}$$

is called the Shannon function.

We need the following lemma about Shannon function.

Lemma 8. *There exists a constant $c > 0$ such that for all $x \in \left[0, \frac{1}{2}\right]$:*

$$1 - h(x) \geqslant c\,(1-2x)^2\,.$$

Proof. Let $f(x) = 1 - h(x)$. Simple calculations show that $f'\left(\frac{1}{2}\right) = 0, f''\left(\frac{1}{2}\right) > 0$. Hence there exists $\delta > 0$ such that for any $x \in \left[\frac{1}{2} - \delta, \frac{1}{2}\right]$ it holds that

$$1 - h(x) = f(x) \geqslant \frac{f''\left(\frac{1}{2}\right)}{4} \cdot \left(\frac{1}{2} - x\right)^2. \tag{5}$$

For the remaining x's we have

$$1 - h(x) \geqslant 1 - h\left(\frac{1}{2} - \delta\right) \geqslant \left(1 - h\left(\frac{1}{2} - \delta\right)\right) \cdot (1 - 2x)^2.$$

Hence we can set

$$c = \min\left\{\frac{f''\left(\frac{1}{2}\right)}{16}, 1 - h\left(\frac{1}{2} - \delta\right)\right\}.$$

□

For any $B \subset \{0,1\}^n$ define

$$\mathsf{diam}(B) = \max_{x,y \in B} d(x,y).$$

Let $V_2(n,r)$ denote the size of Hamming ball of radius r, that is

$$V_2(n,r) = \binom{n}{0} + \ldots + \binom{n}{r}.$$

We will use the following well-known facts about the size of Hamming balls.

Proposition 9 ([7])**.** *If $r \leqslant \frac{n}{2}$, then $V_2(n,r) \leqslant 2^{h\left(\frac{r}{n}\right)n}$.*

Proposition 10 ([7])**.** *If $B \subset \{0,1\}^n$, r is natural, $\mathsf{diam}(B) \leqslant 2r$ and $n \geqslant 2r+1$, then*

$$|B| \leqslant V_2(n,r).$$

Propositions 10, 9 and Lemma 8 easily imply the following

Lemma 11. *Assume that $r < n/2$. Then the cardinality of every set $B \subset \{0,1\}^n$ with $\mathsf{diam}(B) \leqslant 2r$ is at most $2^{n(1-c(1-(2r/n))^2)}$ for some absolute positive constant c.*

A.2 Probability Theory

Definition 7 (Probability distributions)**.** *Let $B(n,p)$ denote the binomial distribution with parameters $n \in \mathbb{N}$ and $p \in (0,1)$. For every natural n let S_n denote the one-dimensional random walk with n steps. More specifically, let S_n be equal to*

$$S_n = X_1 + \ldots + X_n,$$

where $X_1, \ldots, X_n$ are independent random variables taking values in $\{-1,1\}$, such that for each i the following holds: $\Pr[X_i = 1] = \Pr[X_i = -1] = \frac{1}{2}$.

Proof of Lemma 6. Let X be a random variable distributed according to the binomial distribution $B(n,p)$.The expectation and variation of X are given by:

$$\mathrm{E}X = pn, \qquad \mathrm{Var}X = p(1-p)n \leqslant pn.$$

Hence by Chebyshev inequality we get

$$\Pr\left[X \leqslant \frac{pn}{10}\right] \leqslant \frac{\mathrm{Var}X}{\left(pn \cdot \left(1 - \frac{1}{10}\right)\right)^2} \leqslant \frac{\frac{100}{81}}{pn} \leqslant \frac{100}{162} \leqslant \frac{2}{3}.$$

□

Proof of Lemma 7. Take m_0 to be the first natural satisfying the following two conditions:

- There exists $d > 0$ such that for all $N \geqslant m_0^4$ and for every k between $N/2 - \sqrt{N}$ and $N/2 + \sqrt{N}$ the following holds:

$$\Pr[Z_1 + \ldots + Z_N = k] = \binom{N}{k} 2^{-N} \geqslant \frac{d}{\sqrt{N}}.$$

- for all $m \geqslant m_0$, $N \geqslant m^4$ and $i \in \{0, 1, \ldots, m-1\}$ the number of k between $N/2 - \sqrt{N}$ and $N/2 + \sqrt{N}$ such that $k \equiv i \pmod{p}$ is at at least $\frac{\sqrt{N}}{m}$.

Set $c = d$ and observe that for all m, N such that $m \geqslant m_0$ and $N \geqslant m^4$ and for every $i \in \{0, 1, \ldots, m-1\}$ it holds that

$$\Pr[Z_1 + \ldots + Z_N \equiv i \pmod{m}] \geqslant \frac{\sqrt{N}}{m} \cdot \frac{d}{\sqrt{N}} = \frac{c}{m}.$$

□

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il