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

# One-sided error communication complexity of Gap Hamming Distance

Egor Klenin, Alexander Kozachinskiy

November 6, 2016

**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 the upper bound $O((L^2/U)\log L)$ and the lower bound $\Omega(L^2/U)$ for this complexity. These bounds differ only by a $O(\log L)$ factor.

The best upper bounds for communication complexity of GHD known before are the following. The upper bounds $O(L \log n)$ for one-sided error complexity from [5] and $O(L \log L)$ for two-sided error complexity from [6], which do not depend on $U$ and hold for all $U > L$. Our communication protocol outperforms those from [5] and [6] in the case when the ration $U/L$ is not bounded by a constant. The other known upper bound $O(L^2/(U-L)^2)$ holds for two-sided error complexity of GHD [7]. If $U$ is greater than $L+\sqrt{L}$, then the protocol from [7] outperforms ours, however it has two-sided error. It is worth to note that all mentioned protocols run in one round.

From technical viewpoint, our achievement is a new protocol to prove that $x, y$ are far on the basis of a large difference between distances from $x$ and $y$ to a randomly chosen string.

Our lower bound $\Omega(L^2/U)$ (for the one-sided error communication complexity of GHD) generalizes the lower bound $\Omega(U)$ established in [1], [2] for $U = O(L)$.

## 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 $y$



ISSN 1433-8092

differ:

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

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:

**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 public coin communication protocols

Randomized two-sided error public coin communication complexity of $\mathsf{GHD}_{L,U}$ has been extensively studied. From [7] we know that it is at most $O(L^2/(U-L)^2)$ (assuming the constant error probability). The long standing open problem was to obtain the matching lower bound (it would have imply some matching lower bounds for the data stream algorithms). Finally it was solved ([3], [10], [8]) for $L = n/2 - \Theta(\sqrt{n})$ and $U = n/2 + \Theta(\sqrt{n})$: it was shown that two-sided error communication complexity of $\mathsf{GHD}_{L,U}$ is $\Omega(n)$ for such $L, U$. This lower bound matches the upper bound $O(L^2/(U-L)^2)$, which is $O(n)$ in this case.

The paper [6] established the upper bound $O(L \log L)$ for two-sided error protocols 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.

### 1.1.2 One-sided error public coin communication protocols

The one-sided error public coin communication complexity will be denoted by $R^0_\varepsilon$. The superscript 0 means that the protocol is allowed to err only for input pairs at distance at least $U$. The parameter $\varepsilon$ is the maximal probability of error. The superscript 1 will mean the opposite: protocols are allowed to err only for input pairs at distance at most $L$. The error probability $\varepsilon$ is assumed to be a constant less than 1. By amplification the error probability can be made arbitrarily small. Therefore in the sequel we will drop the subscript $\varepsilon$ in the notations $R^0_\varepsilon$ and $R^1_\varepsilon$.

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

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

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 [5] establishes the upper bound $O(L \log n)$ for one-sided error complexity $R^0$ of $\mathsf{GHD}_{L,U}$ in the case $U = L + 1$, that is, there is no gap.

In the papers [1], [2] the case $L = 0$ was studied (under the name *Gap-Equality problem*). It was shown that if $U < (1 - \Omega(1))\, n$, then the one-sided error complexity $R^1$ (the error is allowed when the distance is 0) of $\mathsf{GHD}_{0,\,U}$ is $\Omega(n)$ (moreover, if $U$ is even, $\Omega(n)$ lower bound holds also for a weaker version of $\mathsf{GHD}_{0,\,U}$ problem, in which Hamming distance between the inputs is either 0 or *exactly* $U$).

## 1.2 This work

In this paper we study public-coin one-sided error communication complexity $R^0$ of $\mathsf{GHD}_{L,U}$ when the error probability is constant. 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 protocol for $\mathsf{GHD}_{L,U}$ with communication complexity $O((L^2/U) \log L)$. It is constructed in the following 4 steps.

*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 in one round. 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 $i$th 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 Alice and Bob have the same partition). Both parties also read random strings $r^1, \ldots, r^j$ from the shared random source and Alice communicates to Bob $d(x^j, r^j)$ for all $j = 1, \ldots, b$. Thus the communication is $b \log n = O((L^2/U) \log n)$. Bob computes $d(y^j, r^j)$ for all $j = 1, \ldots, b$ and 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 $L$.

On the remaining three steps we use the known techniques and protocols.

*Step 2.* In [5], for all $L < U$ a one-sided error protocol for $\mathsf{GHD}_{L,U}$ with communication $O(L \log n)$ was constructed. In that protocol Alice sends hash value of length $O(L \log n)$ of $x$ to Bob, then Bob looks for a string with the same hash value at distance at most $L$ from his string and outputs 0 iff there is such string. The required bound for probability of error follows from the bound $(n + 1)^L$ for the volume of the Hamming ball of radius $L$.

Our second protocol runs the Triangle Inequality Protocol if $U > C'L$ and the protocol from [5] 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 [6] 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 $O(L^4)$ obtained from the original strings $x, y$ by the following transformation. As in the Triangle Inequality Protocol, we split $x, y$ into $b = O(L^4)$ 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 $d(x, y) \leqslant L^2$. Therefore this protocol has constant one-sided error probability for all input pairs with $d(x, y) \leqslant L^2$. By construction this protocol communicates $O((L^2/U) \log L)$ bits.

*Step 4.* Finally, to handle the case $d(x, y) > L^2$, 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 hash function $h$ from the shared random source with $4L + 2$ values and a random $n$-bit string $r$. Alice sends $h(d(x, r))$ to Bob. Bob then outputs 0 iff there is an integer $d'$ with $|d' - d(y, r)| \leqslant L$ and $h(d') = h(d(x, r))$. If $d(x, y) \leqslant L$ then triangle inequality guarantees that $|d(x, r) - d(y, r)| \leqslant d(x, y) \leqslant L$ and hence the protocol always outputs 0. Assume now that $d(x, y) > L^2$. The properties of random walks imply that $|d(x, r) - d(y, r)| > L$ with constant probability. If this happens, then every $d'$ with $|d' - d(y, r)| \leqslant L$ differs from $d(x, r)$. For each $d' \neq d(x, r)$ the probability of event $[h(d') = d(x, r)]$ is at most $1/(4L+2)$. By union bound, conditional to the event $[|d(x, r) - d(y, r)| > L]$, the protocol outputs 0 with probability at most $(2L + 1)/(4L + 2) = 1/2$.

### 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 $\Omega(n)$ ([1]). However, from the argument in [1] 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.** *The one-sided error public-coin communication complexity $R^0$ of $\mathsf{GHD}_{L,U}$ on $n$-bit strings is at most*

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

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

**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 $\pi$, denoted by $CC(\pi)$, is defined as the depth of the corresponding binary tree.*

Randomizes 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 $\tau$, denoted by $CC(\tau)$, is defined as $\max_{\pi} CC(\pi)$, where $\pi$ is taken over the deterministic protocols from the support of $\tau$ (recall that $\tau$ is the distribution).*

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

**Definition 5.** *We say that a public-coin protocol computes a partial function $f$ with error probability $\varepsilon$, 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 $\varepsilon$.*

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:

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

# 3 The lower bound

In this section we prove Theorem 2.

*Proof of Theorem 2.* Let $\tau$ 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 $\tau$ 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 $\tau$ always outputs 1.

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

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

Consider any 0-leaf of $\pi$ and the corresponding rectangle $R = A \times B \subset \{0,1\}^n \times \{0,1\}^n$. The set of all diagonal pairs from $R$ equals $A \cap B$. Its diameter must be less than $U$. Indeed, if there are $x, y \in A \cap B$ such that $d(x,y) \geqslant U$, then $\pi$ 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 10 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 10 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 $\pi$, 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 $\pi$ is at least $2^{n-1}$, the number of 0-leaves in $\pi$ is at least $2^{cn(1-U/n)^2-1}$. Thus we have

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

Obviously we also have

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

From inequalities (1) and (2) 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). □

# 4 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.

## 4.1 The Triangle Inequality Protocol

Let $x, y \in \{0,1\}^n$ denote Alice's and Bob's input strings, respectively. Alice and Bob set $b = \lceil CL^2/U \rceil$, where $C$ is a constant to be defined later. Then they use public coins to sample a function $\chi : \{1,2,\ldots,n\} \to \{1,2,\ldots,b\}$ uniformly at random. They use $\chi$ 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 $j$th block are arranged is not important, Alice and Bob 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 Bob:

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

Then Bob computes the sum

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

If $T \leqslant L$, Bob 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 4.** *There is a positive constants $C$ such that the following holds. Assume that $b = \lceil CL^2/U \rceil$ and $U \geqslant 2b$. Then the protocol for the input pair $x, y$ at distance at least $U$ outputs 1 with some positive constant probability.*

*Proof.* Assume that $C, U, L, b, x, y$ 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 $\chi$ 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 blocks 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 $1/b$: 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 facts about binomial distributions and random walks, which are proven in the Appendix.

**Lemma 5.** *If $X$ is distributed according to the binomial distribution $B(n,p)$ and $pn \geqslant 2$, then*

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

**Lemma 6.** *Let $S_n$ denote the sum of $n$ independent random variables where each variable takes the values 1 and $-1$ with probabilities 1/2. Then for all $n$ $|S_n| \geqslant \sqrt{n}$ with some positive constant probability.*

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 5 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 6 with some positive constant probability 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 some positive constant probability $\alpha$ 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 $\alpha/2$ the fraction of $j$ that satisfy this inequality is at least $\alpha/2$. Indeed, let the random variable $\theta$ denote the fraction of $j$ that satisfy this inequality. Its average is at least $\alpha$. On the other hand, we can upperbound its average by the sum

$$\Pr[\theta > \alpha/2] \cdot 1 + \Pr[\theta \leqslant \alpha/2] \cdot (\alpha/2) \leqslant \Pr[\theta > \alpha/2] + \alpha/2.$$

Thus with probability $\alpha/2$ we have

$$\sum_{j=1}^{b} |d(x^j, r^j) - d(y^j, r^j)| \geqslant (\alpha/2)b\sqrt{d(x,y)/10b} = (\alpha/2)\sqrt{b \cdot d(x,y)/10}.$$

Recall that $b = \lceil CL^2/U \rceil$ and $d(x,y) \geqslant U$. If $C$ is a large enough constant then the right hand side of the last displayed inequality is larger than $L$. $\square$

Let $C$ be an integer constant satisfying Lemma 4. 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 4 is met for all $U \geqslant 2CL + 1$.

Recall that the communication length of the protocol is $O(L^2 \log n/U)$. Now we need a protocol with the same communication length for $L, U$ such that $U \leqslant 2CL$. Notice that in this case the upper bound $O(L^2 \log n/U)$ for communication boils down to $O(L \log n)$. A protocol with such performance was constructed in [5].

## 4.2 The protocol of [5]

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

$$h : \{0,1\}^n \to \{1, 2, \ldots, 2 \cdot V_2(n, L)\},$$

where $V_2(n, L)$ stands for the cardinality of the Hamming ball of radius $L$. Alice sends $h(x)$ to Bob. If there exists $z \in \{0,1\}^n$ such that $d(z,y) \leqslant L$ and $h(z) = h(x)$, then Bob outputs 0. Otherwise Bob outputs 1.

The protocol communicates $O\left(\log(2 \cdot V_2(n, L)\right) = O(L \log n + 1)$ bits.

If $d(x,y) \leqslant L$ then the protocol outputs 0 with probability 1. If $d(x,y) > L$ then the protocol can err. An error may occur, if there is a $z \in \{0,1\}^n$ such that $d(z,y) \leqslant L$ and $h(z) = h(x)$. Any such $z$ must be different from $x$. Hence for every fixed $z$ the probability of error is at most $1/2V_2(n, L)$. By union bound the protocol errs with probability at most $V_2(n, L)(1/2V_2(n, L)) = 1/2$.

## 4.3 The simplified version of the Triangle Inequality Protocol

Thus for all $L, U$ we have a 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 protocol with communication length $O(\log L)$ to solve $\mathsf{GHD}_{L,L^2+1}$ with constant one-sided error probability. (Notice that $O((L^2/U + 1) \log L)$ becomes $O(\log L)$ for $U = L^2 + 1$.)

*The protocol.* Alice and Bob use public coins to sample a vector $r \in \{0,1\}^n$ uniformly at random and a function $h : \{0, 1, \ldots, n\} \to \{1, 2, \ldots, 4L + 2\}$ uniformly at random. Alice sends $h(d(x,r))$ to Bob. If there is an integer $i$ such that $|i - d(y,r)| \leqslant L$ and $h(i) = h(d(x,r))$, then Bob outputs 0. Otherwise Bob outputs 1.

The protocol communicates $O(\log(4L + 2))$ bits. If $d(x,y) \leqslant L$, then the protocol outputs 0 with probability 1. Indeed, for $i = d(x,r)$ we have $|i - d(y,r)| \leqslant d(x,y) \leqslant L$ by the triangle inequality.

Assume that $d(x,y) \geqslant L^2+1$. In this case by Lemma 6 we have $|d(x,r) - d(y,r)| \geqslant \sqrt{d(x,y)} > L$ with positive constant probability. If this happens, every $i$ with $|i - d(y,r)| \leqslant L$ differs from $d(x,r)$. There are $2L+1$ such $i$'s and for each of them the probability of event $[h(i) = d(x,r)]$ is at most $1/(4L+2)$. By union bound, conditional to the event $[|d(x,r) - d(y,r)| > L]$, the protocol outputs 0 with probability at most $(2L+1)/(4L+2) = 1/2$. Hence the protocol outputs 1 with positive constant probability.

## 4.4 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 $2L^4$ blocks randomly (as in the construction of the Triangle Inequality Protocol). Let

$$x^1, \ldots, x^{2L^4}, \qquad y^1, \ldots, y^{2L^4}$$

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 $u_1, \ldots, u_{2L^4}$ and sets $u = u_1 \ldots u_{2L^4}$. Bob privately computes $v_1, \ldots, v_{2L^4}$ and sets $v = v_1 \ldots v_{2L^4}$.

Then they run the protocol with communication length $O((L^2/U+1)\log n)$ constructed earlier for input pair $(u,v)$ (and not $(x,y)$). In that protocol they use parameters $L$ and $b = \lceil CL^2/U \rceil$ for the number of blocks, where $C$ is the constant from Lemma 4. They output the result of this run.

The communication length of the constructed protocol is $O((L^2/U + 1)\log(L+2))$. 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 $d(x,y) > L^2$, then the Simplified Triangle Inequality Protocol outputs 1 with probability 1/2, they output 1 and halt.

Assume that $U \leqslant d(x,y) \leqslant L^2$. 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 $\frac{1}{2L^4}$. By union bound, with probability at least $1 - \frac{d(x,y)^2}{2L^4} \geqslant 1 - \frac{L^4}{2L^4} = 0.5$ 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.

# References


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

[2] 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.

[3] 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.

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

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

[6] 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.

[7] 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.

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

[9] Shevtsova, I. On the absolute constants in the berry-esseen type inequalities for identically distributed summands. *arXiv preprint arXiv:1111.6554* (2011).

[10] 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).


# 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 7.** *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

$$1 - h(x) = f(x) \geqslant \frac{f''\left(\frac{1}{2}\right)}{4} \cdot \left(\frac{1}{2} - x\right)^2. \tag{3}$$

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 8** ([4])**.** *If $r \leqslant \frac{n}{2}$, then $V_2(n,r) \leqslant 2^{h\left(\frac{r}{n}\right)n}$.*

**Proposition 9** ([4])**.** *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 9, 8 and Lemma 7 easily imply the following

**Lemma 10.** *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 $2^{n(1-c(1-(2r/n))^2)}$.*

## A.2 Probability Theory

**Definition 7** (Probability distributions)**.** *Let $\mathcal{N}(0,1)$ denote the standard normal random variable. 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:* $\Pr[X_i = 1] = \Pr[X_i = -1] = \frac{1}{2}$.

*The proof of Lemma 5.* 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

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

□

For the proof of Lemma 6 we will need the following

**Proposition 11** (Berry–Esseen inequality, [9])**.** *For every real $x$ and natural $n$ the following holds:*

$$\left|\Pr\left[\frac{S_n}{\sqrt{n}} < x\right] - \Pr[\mathcal{N}(0,1) < x]\right| \leqslant \frac{1}{2\sqrt{n}}.$$

*The proof of Lemma 6.* Let $n_0$ be the first natural such that $\frac{1}{\sqrt{n_0}} < \Pr[\mathcal{N}(0,1) > 1]$. From the Berry-Esseen inequality it follows that for any $n \geqslant n_0$:

$$\begin{aligned}\Pr[S_n < \sqrt{n}] = \Pr\left[\frac{S_n}{\sqrt{n}} < 1\right] &\leqslant \Pr[\mathcal{N}(0,1) < 1] + \frac{1}{2\sqrt{n}} \\ &\leqslant \Pr[\mathcal{N}(0,1) < 1] + \frac{\Pr[\mathcal{N}(0,1) \geqslant 1]}{2} \\ &= 1 - \frac{\Pr[\mathcal{N}(0,1) \geqslant 1]}{2},\end{aligned}$$

and thus:

$$\Pr[|S_n| \geqslant \sqrt{n}] \geqslant \Pr[S_n \geqslant \sqrt{n}] \geqslant \frac{\Pr[\mathcal{N}(0,1) > 1]}{2}.$$

For any $n < n_0$ we have that

$$\Pr[|S_n| \geqslant \sqrt{n}] \geqslant \Pr[S_n = n] = 2^{-n} > 2^{-n_0}.$$

□



ECCC ISSN 1433-8092

http://eccc.hpi-web.de