Electronic Colloquium on Computational Complexity, Report No. 15 (2009)

A Multi-Round Communication Lower Bound for Gap Hamming and Some Consequences*

Joshua Brody
Department of Computer Science
Dartmouth College
Hanover, NH 03755, USA
jbrody@cs.dartmouth.edu

Amit Chakrabarti
Department of Computer Science
Dartmouth College
Hanover, NH 03755, USA
ac@cs.dartmouth.edu

February 19, 2009

Abstract

The Gap-Hamming-Distance problem arose in the context of proving space lower bounds for a number of key problems in the data stream model. In this problem, Alice and Bob have to decide whether the Hamming distance between their n-bit input strings is large (i.e., at least $n/2 + \sqrt{n}$) or small (i.e., at most $n/2 - \sqrt{n}$); they do not care if it is neither large nor small. This $\Theta(\sqrt{n})$ gap in the problem specification is crucial for capturing the approximation allowed to a data stream algorithm.

Thus far, for randomized communication, an $\Omega(n)$ lower bound on this problem was known only in the one-way setting. We prove an $\Omega(n)$ lower bound for randomized protocols that use any constant number of rounds.

As a consequence we conclude, for instance, that ε-approximately counting the number of distinct elements in a data stream requires $\Omega(1/\varepsilon^2)$ space, even with multiple (a constant number of) passes over the input stream. This extends earlier one-pass lower bounds, answering a long-standing open question. We obtain similar results for approximating the frequency moments and for approximating the empirical entropy of a data stream.

In the process, we also obtain tight $n - \Theta(\sqrt{n}\log n)$ lower and upper bounds on the one-way deterministic communication complexity of the problem. Finally, we give a simple combinatorial proof of an $\Omega(n)$ lower bound on the one-way randomized communication complexity.

1 Introduction

This paper concerns communication complexity, which is a heavily-studied basic computational model, and is a powerful abstraction useful for obtaining results in a variety of settings not necessarily involving communication. To cite but two examples, communication complexity has been applied to prove lower bounds on circuit depth (see, e.g., [KW90]) and on query times for static data structures (see, e.g., [MNSW98, Păt08]). The basic setup involves two players, Alice and Bob, each of whom receives an input string. Their goal is to compute some function of the two strings, using a protocol that involves exchanging a *small* number of bits. When communication complexity is applied as a lower bound technique — as it often is — one seeks to

*Work supported in part by an NSF CAREER Award CCF-0448277 and NSF grant EIA-98-02068.

ISSN 1433-8092

prove that there does not exist a nontrivial protocol, i.e., one that communicates only a sublinear number of bits, for computing the function of interest. Naturally, such a proof is more challenging when the protocol is allowed to be *randomized* and err with some small probability on each input.

The textbook by Kushilevitz and Nisan [KN97] provides detailed coverage of the basics of communication complexity, and of a number of applications, including the two mentioned above. In this paper, we only recap the most basic notions, in Section 2.

Our focus here is on a specific communication problem — the Gap-Hamming-Distance problem — that, to the best of our knowledge, was first formally studied by Indyk and Woodruff [IW03] in FOCS 2003. They studied the problem in the context of proving space lower bounds for the Distinct Elements problem in the data stream model. We shall discuss their application shortly, but let us first define our communication problem precisely.

The Problem. In the Gap-Hamming-Distance problem, Alice receives a Boolean string $x \in \{0,1\}^n$ and Bob receives $y \in \{0,1\}^n$. They wish to decide whether x and y are "close" or "far" in the Hamming sense. That is, they wish to output 0 if $\Delta(x,y) \leq n/2 - \sqrt{n}$ and 1 if $\Delta(x,y) \geq n/2 + \sqrt{n}$. They do not care about the output if neither of these conditions holds. Here, Δ denotes Hamming distance. In the sequel, we shall be interested in a parametrized version of the problem, where the thresholds are set at $n/2 \pm c\sqrt{n}$, for some parameter $c \in \mathbb{R}^+$.

Our Results. While we prove a number of results about the Gap-Hamming-Distance problem here, there is a clear "main theorem" that we wish to highlight. Technical terms appearing below are defined precisely in Section 2.

Theorem 1 (Main Theorem, Informal). *Suppose a randomized $\frac{1}{3}$-error protocol solves the Gap-Hamming-Distance problem using k rounds of communication. Then, at least one message must be $n/2^{O(k^2)}$ bits long. In particular, any protocol using a constant number of rounds must communicate $\Omega(n)$ bits in some round. In fact, these bounds apply to deterministic protocols with low distributional error under the uniform distribution.*

Notice that our lower bound applies to the *maximum* message length, not just the *total* length.

At the heart of our proof is a round elimination lemma that lets us "eliminate" the first round of communication, in a protocol for the Gap-Hamming-Distance problem, and thus derive a shorter protocol for an "easier" instance of the same problem. By repeatedly applying this lemma, we eventually eliminate all of the communication. We also make the problem instances progressively easier, but, if the original protocol was short enough, at the end we are still left with a nontrivial problem. The resulting contradiction lower bounds the length of the original protocol. We note that this underlying "round elimination philosophy" is behind a number of key results in communication complexity [MNSW98, Sen03, CR04, ADHP06, Cha07, VW07, CJP08].

Besides the above theorem, we also prove tight lower *and upper* bounds of $n - \Theta(\sqrt{n}\log n)$ on the one-way deterministic communication complexity of Gap-Hamming-Distance. Only $\Omega(n)$ lower bounds were known before. We also prove an $\Omega(n)$ one-way randomized communication lower bound. This matches earlier results, but our proof has the advantage of being purely combinatorial. (We recently learned that Woodruff [Woo09] had independently discovered a similar combinatorial proof. We present our proof nevertheless, for pedagogical value, as it can be seen as a generalization of our deterministic lower bound proof.)

Motivation and Relation to Prior Work. We now describe the original motivation for studying the Gap-Hamming-Distance problem. Later, we discuss the consequences of our Theorem 1. In the data stream

model, one wishes to compute a real-valued function of a massively long input sequence (the data stream) using very limited space, hopefully sublinear in the input length. To get interesting results, one almost always needs to allow randomized approximate algorithms. A key problem in this model, that has seen much research [FM85, AMS99, BJK$^+$04, IW03, Woo09], is the Distinct Elements problem: the goal is to estimate the number of distinct elements in a stream of m elements (for simplicity, assume that the elements are drawn from the universe $[m] := \{1, 2, \ldots, m\}$).

An interesting solution to this problem would give an nontrivial tradeoff between the quality of approximation desired as the space required to achieve it. The best such result [BJK$^+$04] achieved a multiplicative $(1+\varepsilon)$-approximation using space $\widetilde{O}(1/\varepsilon^2)$, where the $\widetilde{O}$-notation suppresses $\log m$ and $\log(1/\varepsilon)$ factors. It also processed the input stream in a single pass, a very desirable property. Soon afterwards, Indyk and Woodruff [IW03] gave a matching $\Omega(1/\varepsilon^2)$ space lower bound for one-pass algorithms for this problem, by a reduction from the Gap-Hamming-Distance communication problem. In SODA 2004, Woodruff [Woo04] improved the bound, extending it to the full possible range of subconstant ε, and also applied it to the more general problem of estimating frequency moments $F_p := \sum_{i=1}^{n} f_i^p$, where f_i is the frequency of element i in the input stream. A number of other natural data stream problems have similar space lower bounds via reductions from Gap-Hamming, a more recent example being the computation of the empirical entropy of a stream [CCM07].

The idea behind the reduction is quite simple: Alice and Bob can convert their Gap-Hamming inputs into suitable streams of integers, and then simulate a one-pass streaming algorithm using a single round of communication in which Alice sends Bob the memory contents of the algorithm after processing her stream. In this way, an $\Omega(n)$ one-way communication lower bound translates into an $\Omega(1/\varepsilon^2)$ one-pass space lower bound. Much less simple was the proof of the communication lower bound itself. Woodruff's proof [Woo04] required intricate combinatorial arguments and a fair amount of complex calculations. Jayram et al. [JKS07] later provided a rather different proof, based on a simple geometric argument, coupled with a clever reduction from the INDEX problem. A version of this proof is given in Woodruff's Ph.D. thesis [Woo07]. In Section 5, we provide a still simpler direct combinatorial proof, essentially from first principles.

All of this left open the tantalizing possibility that a second pass over the input stream could drastically reduce the space required to approximate the number of distinct elements — or, more generally, the frequency moments F_p. Perhaps $\widetilde{O}(1/\varepsilon)$ space was possible? This was a long-standing open problem [Kum06] in data streams. Yet, some thought about the underlying Gap-Hamming communication problem suggested that the linear lower bound ought to hold for general communication protocols, not just for one-way communication. This prompted the following natural conjecture.

Conjecture 2. *A $\frac{1}{3}$-error randomized communication protocol for the Gap-Hamming-Distance problem must communicate $\Omega(n)$ bits in total, irrespective of the number of rounds of communication.*

An immediate consequence of the above conjecture is that a second pass does *not* help beat the $\Omega(1/\varepsilon^2)$ space lower bound for the aforementioned streaming problems; in fact, no constant number of passes helps. Our Theorem 1 does *not* resolve Conjecture 2. However, it *does* imply the $\Omega(1/\varepsilon^2)$ space lower bound with a constant number of passes. This is because we *do* obtain a linear communication lower bound with a constant number of rounds.

Finer Points. To better understand our contribution here, it is worth considering some finer points of previously known lower bounds on Gap-Hamming-Distance, including some "folklore" results. The earlier one-way $\Omega(n)$ bounds were *inherently* one-way, because the INDEX problem has a trivial two-round protocol. Also, the nature of the reduction implied a distributional error lower bound for Gap-Hamming only under a somewhat artificial input distribution. Our bounds here, including our one-way randomized bound,

overcome this problem, as does the recent one-way bound of Woodruff [Woo09]: they apply to the uniform distribution. As noted by Woodruff [Woo09], this has the desirable consequence of implying space lower bounds for the Distinct Elements problem under weaker assumptions about the input stream: it could be random, rather than adversarial.

Intuitively, the uniform distribution is the hard case for the Gap-Hamming problem. The Hamming distance between two uniformly distributed n-bit strings is likely to be just around the $n/2 \pm \Theta(\sqrt{n})$ thresholds, which means that a protocol will have to work hard to determine which threshold the input is at. Indeed, this line of thinking suggests an $\Omega(n)$ lower bound for distributional complexity — under the uniform distribution — on the *gapless* version of the problem. Our proofs here confirm this intuition, at least for a constant number of rounds.

It is relatively easy to obtain an $\Omega(n)$ lower bound on the *deterministic* multi-round communication complexity of the problem. One can directly demonstrate that the communication matrix contains no large monochromatic rectangles (see, e.g. [Woo07]). Indeed, the argument goes through even with gaps of the form $n/2 \pm \Theta(n)$, rather than $n/2 \pm \Theta(\sqrt{n})$. It is also easy to obtain an $\Omega(n)$ bound on the randomized complexity of the gapless problem, via a reduction from DISJOINTNESS. Unfortunately, the known hard distributions for DISJOINTNESS are far from uniform, and DISJOINTNESS is actually very easy under a uniform input distribution. So, this reduction does not give us the results we want.

Furthermore, straightforward rectangle-based methods (discrepancy/corruption) fail to effectively lower bound the randomized communication complexity of our problem. This is because there *do* exist very large near-monochromatic rectangles in its communication matrix. This can be seen, e.g., by considering all inputs (x, y) with $x_i = y_i = 0$ for $i \in [n/100]$.

Connection to Decision Trees and Quantum Communication. We would like to bring up two other illuminating observations. Consider the following query complexity problem: the input is a string $x \in \{0,1\}^n$ and the desired output is 1 if $|x| \geq n/2 + \sqrt{n}$ and 0 if $|x| \leq n/2 - \sqrt{n}$. Here, $|x|$ denotes the Hamming weight of x. The model is a randomized decision tree whose nodes query individual bits of x, and whose leaves give outputs in $\{0,1\}$. It is not hard to show that $\Omega(n)$ queries are needed to solve this problem with $\frac{1}{3}$ error. Essentially, one can do no better than sampling bits of x at random, and then $\Omega(1/\varepsilon^2)$ samples are necessary to distinguish a biased coin that shows heads with probability $\frac{1}{2} + \varepsilon$ from one that shows heads with probability $\frac{1}{2} - \varepsilon$.

The Gap-Hamming-Distance problem can be seen as a generalization of this problem to the communication setting. Certainly, any efficient decision tree for the query problem implies a correspondingly efficient communication protocol, with Alice acting as the querier and Bob acting as the responder (say). Conjecture 2 says that no better communication protocols are possible for this problem.

This query complexity connection brings up another crucial point. The *quantum* query complexity of the above problem can be shown to be $O(\sqrt{n})$, by the results of Nayak and Wu [NW99]. This in turn implies an $O(\sqrt{n}\log n)$ quantum communication protocol for Gap-Hamming, essentially by carefully "implementing" the quantum query algorithm, as in Razborov [Raz02]. Therefore, any technique that seeks to prove an $\Omega(n)$ lower bound for Gap-Hamming (under classical communication) must necessarily fail for quantum protocols. This rules out several recently-developed methods, such as the factorization norms method of Linial and Shraibman [LS07] and the pattern matrix method of Sherstov [She08].

Connections to Recent Work. Our multi-round $\Omega(n)$ bound turns out to also have applications [ABC09] to the communication complexity of several distributed "functional monitoring" problems, studied recently by Cormode et al. [CMY08] in SODA 2008. Also, our lower bound approach here uses and extends a subspace-finding technique recently developed by Brody [Bro09] to prove lower bounds on multiparty

pointer jumping.

2 Basic Definitions, Notation and Preliminaries

We begin with definitions of our central problem of interest, and quickly recall some standard definitions from communication complexity. Along the way, we also introduce some notation that we use in the rest of the paper.

Definition 1. For strings $x, y \in \{0,1\}^n$, the Hamming distance between x and y, denoted $\Delta(x,y)$, is defined as the number of coordinates $i \in [n]$ such that $x_i \neq y_i$.

Definition 2 (Gap-Hamming-Distance problem). Suppose $n \in \mathbb{N}$ and $c \in \mathbb{R}^+$. The c-Gap-Hamming-Distance partial function, on n-bit inputs, is denoted $\mathrm{GHD}_{c,n}$ and is defined as follows.

$$\mathrm{GHD}_{c,n}(x,y) \;=\; \begin{cases} 1\,, & \text{if } \Delta(x,y) \geq n/2 + c\sqrt{n}\,, \\ 0\,, & \text{if } \Delta(x,y) \leq n/2 - c\sqrt{n}\,, \\ \star\,, & \text{otherwise.} \end{cases}$$

We also use $\mathrm{GHD}_{c,n}$ to denote the corresponding communication problem where Alice holds $x \in \{0,1\}^n$, Bob holds $y \in \{0,1\}^n$, and the goal is for them to communicate and agree on an output bit that matches $\mathrm{GHD}_{c,n}(x,y)$. By convention, $\star$ matches both 0 and 1.

Protocols. Consider a communication problem $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1,\star\}^n$ and a protocol $\mathcal{P}$ that attempts to solve f. We write $\mathcal{P}(x,y)$ to denote the output of $\mathcal{P}$ on input (x,y): note that this may be a random variable, dependent on the internal coin tosses of $\mathcal{P}$, if $\mathcal{P}$ is a randomized protocol. A deterministic protocol $\mathcal{P}$ is said to be correct for f if $\forall\,(x,y) : \mathcal{P}(x,y) = f(x,y)$ (the "=" is to be read as "matches"). It is said to have *distributional error* ε under an input distribution ρ if $\Pr_{(x,y)\sim\rho}[\mathcal{P}(x,y) \neq f(x,y)] \leq \varepsilon$. A *randomized protocol* $\mathcal{P}$, using a public random string r, is said to be have error ε if $\forall\,(x,y) : \Pr_r[\mathcal{P}(x,y) \neq f(x,y)] \leq \varepsilon$. A protocol $\mathcal{P}$ is said to be a *k-round protocol* if it involves exactly k messages, with Alice and Bob taking turns to send the messages; by convention, we usually assume that Alice sends the first message and the recipient of the last message announces the output. A 1-round protocol is also called a *one-way protocol*, since the entire communication happens in the Alice $\to$ Bob direction.

Communication Complexity. The deterministic communication complexity $D(f)$ of a communication problem f is defined to be the minimum, over deterministic protocols $\mathcal{P}$ for f, of the number of bits exchanged by $\mathcal{P}$ for a worst-case input (x,y). By suitably varying the class of protocols over which the minimum is taken, we obtain, e.g., the ε-error randomized, one-way deterministic, ε-error one-way randomized, and ε-error ρ-distributional deterministic communication complexities of f, denoted $R_\varepsilon(f)$, $D^{\to}(f)$, $R^{\to}_\varepsilon(f)$, and $D_{\rho,\varepsilon}(f)$, respectively. When the error parameter ε is dropped, it is tacitly assumed to be $\frac{1}{3}$; as is well-known, the precise value of this constant is immaterial for asymptotic bounds.

Definition 3 (Near-Orthogonality). We say that strings $x, y \in \{0,1\}^n$ are c-near-orthogonal, and write $x \perp_c y$, if $|\Delta(x,y) - n/2| < c\sqrt{n}$. Here, c is a positive real quantity, possibly dependent on n. Notice that $\mathrm{GHD}_{c,n}(x,y) = \star \;\Leftrightarrow\; x \perp_c y$.

The distribution of the Hamming distance between two uniform random n-bit strings — equivalently, the distribution of the Hamming weight of a uniform random n-bit string — is just an unbiased binomial distribution $\mathrm{Binom}(n, \frac{1}{2})$. We shall use the following (fairly loose) bounds on the tail of this distribution (see, e.g., Feller [Fel68]).

Fact 3. *Let* $T_n(c) = \Pr_x[x \not\perp_c 0^n]$*, where* x *is distributed uniformly at random in* $\{0,1\}^n$*. Let* $T(c) = \lim_{n\to\infty} T_n(c)$*. Then*

$$2^{-3c^2-2} \;\le\; T(c) \;\approx\; \frac{e^{-2c^2}}{c\sqrt{2\pi}} \;\le\; 2^{-c^2}.$$

There are two very natural input distributions for $\mathrm{GHD}_{c,n}$: the uniform distribution on $\{0,1\}^n \times \{0,1\}^n$, and the (non-product) distribution that is uniform over all inputs for which the output is precisely defined. We call this latter distribution $\mu_{c,n}$.

Definition 4 (Distributions). For $n \in \mathbb{N}$, $c \in \mathbb{R}^+$, let $\mu_{c,n}$ denote the uniform distribution on the set $\{(x,y) \in \{0,1\}^n \times \{0,1\}^n : x \not\perp_c y\}$. Also, let $\mathcal{U}_n$ denote the uniform distribution on $\{0,1\}^n$.

Using Fact 3, we can show that for a constant c and suitably small ε, the distributional complexities $D_{\mathcal{U}_n \times \mathcal{U}_n,\varepsilon}(\mathrm{GHD}_{c,n})$ and $D_{\mu_{c,n},\varepsilon}(\mathrm{GHD}_{c,n})$ are within constant factors of each other. This lets us work with the latter and draw conclusions about the former. The latter has the advantage that it is meaningful for any $\varepsilon < \frac{1}{2}$, whereas the former is only meaningful if $\varepsilon < \frac{1}{2}T(c)$.

Let $\mathcal{B}(x,r)$ denote the Hamming ball of radius r centered at x. We need use the following bounds on the volume (i.e., size) of a Hamming ball. Here, $H : [0,1] \to [0,1]$ is the binary entropy function.

Fact 4. *If* $r = c\sqrt{n}$*, then* $(\sqrt{n}/c)^r < |\mathcal{B}(x,r)| < n^r$.

Fact 5. *If* $r = \alpha n$ *for some constant* $0 < \alpha < 1$*, then* $|\mathcal{B}(x,r)| \le 2^{nH(\alpha)}$.

3 Main Theorem: Multi-Round Lower Bound

3.1 Some Basics

In order to prove our multi-round lower bound, we need a simple — yet, powerful — combinatorial lemma, known as Sauer's Lemma [Sau72]. For this, we recall the concept of Vapnik-Chervonenkis dimension. Let $S \subseteq \{0,1\}^n$ and $I \subseteq [n]$. We say that S shatters I if the set obtained by restricting the vectors in S to the coordinates in I has the maximum possible size, $2^{|I|}$. We define VC-dim(S) to be the maximum $|I|$ such that S shatters I.

Lemma 6 (Sauer's Lemma). *Suppose* $S \subseteq \{0,1\}^n$ *has* VC-dim$(S) < d$*. Then*

$$|S| \;\le\; \sum_{i=0}^{d} \binom{n}{d}.$$

When $d = \alpha n$ for some constant α, then the above sum can be upper bounded by $2^{nH(\alpha)}$. This yields the following corollary.

Corollary 7. *If* $|S| \ge 2^{nH(\alpha)}$*, for a constant* α*, then* VC-dim$(S) \ge \alpha n$.

We now turn to the proof proper. It is based on a round elimination lemma that serves to eliminate the first round of communication of a GHD protocol, yielding a shorter protocol, but for GHD instances with weakened parameters. To keep track of all relevant parameters, we introduce the following notation.

Definition 5. A $[k,n,s,c,\varepsilon]$-protocol is a deterministic k-round protocol for $\mathrm{GHD}_{c,n}$ that errs on at most an ε fraction of inputs, under the input distribution $\mu_{c,n}$, and in which each message is s bits long.

The next lemma gives us the "end point" of our round elimination argument.

Lemma 8. *There exists no* $[0, n, s, c, \varepsilon]$*-protocol with* $n > 1$*,* $c = o(\sqrt{n})$*, and* $\varepsilon < \frac{1}{2}$*.*

Proof. With these parameters, $\mu_{c,n}$ has nonempty support. This implies $\Pr_{\mu_{c,n}}[\mathrm{GHD}_{c,n}(x, y) = 0] = \Pr_{\mu_{c,n}}[\mathrm{GHD}_{c,n}(x, y) = 1] = \frac{1}{2}$. Thus, a 0-round deterministic protocol, which must have constant output, cannot achieve error less than $\frac{1}{2}$. □

3.2 The Round Elimination Lemma

The next lemma is the heart of our proof. To set up its parameters, we set $t_0 = (48 \ln 2) \cdot 2^{11k}$, $t = 2^{15k}$, and $b = T^{-1}(1/8)$, and we define a sequence $\langle (n_i, s_i, c_i, \varepsilon_i) \rangle_{i=0}^{k}$ as follows:

$$\left.\begin{array}{lll} n_0 = n\,, & & n_{i+1} = n_i/3\,, \\ s_0 = t_0 s\,, & & s_{i+1} = t s_i\,, \\ c_0 = 10\,, & & c_{i+1} = 2c_i\,, \\ \varepsilon_0 = 2^{-2^{11k}}\,, & & \varepsilon_{i+1} = \varepsilon_i / T(c_{i+1})\,. \end{array}\right\} \text{ for } 0 \le i < k\,. \tag{1}$$

Lemma 9 (Round Elimination for GHD). *Suppose* $0 \le i < k$ *and* $s_i \le n_i/20$*. Suppose there exists a* $[k - i, n_i, s_i, c_i, \varepsilon_i]$*-protocol. Then there exists a* $[k - i - 1, n_{i+1}, s_{i+1}, c_{i+1}, \varepsilon_{i+1}]$*-protocol.*

Proof. Let $(n, s, c, \varepsilon) = (n_i, s_i, c_i, \varepsilon_i)$ and $(n', s', c', \varepsilon') = (n_{i+1}, s_{i+1}, c_{i+1}, \varepsilon_{i+1})$. Also, let $\mu = \mu_{c,n}$, $\mu' = \mu_{c',n'}$, $\mathrm{GHD} = \mathrm{GHD}_{c,n}$ and $\mathrm{GHD}' = \mathrm{GHD}_{c',n'}$. Let $\mathcal{P}$ be a $[k - i, n, s, c, \varepsilon]$-protocol. Assume, WLOG, that Alice sends the first message in $\mathcal{P}$.

Call a string $x_0 \in \{0,1\}^n$ "good" if

$$\Pr_{(x,y)\sim\mu}[\mathcal{P}(x, y) \ne \mathrm{GHD}(x, y) \mid x = x_0] \le 2\varepsilon\,. \tag{2}$$

By the error guarantee of $\mathcal{P}$ and Markov's inequality, the number of good strings is at least 2^{n-1}. There are $2^s \le 2^{n/20}$ different choices for Alice's first message. Therefore, there is a set $M \subseteq \{0,1\}^n$ of good strings such that Alice sends the same first message $\mathfrak{m}$ on every input $x \in M$, with $|M| \ge 2^{n-1-n/20} \ge 2^{nH(1/3)}$. By Corollary 7, VC-dim$(M) \ge n/3$. Therefore, there exists a set $I \subseteq [n]$, with $|I| = n/3 = n'$, that is shattered by M. For strings $x' \in \{0,1\}^{n'}$ and $x'' \in \{0,1\}^{n-n'}$, we write $x' \circ x''$ to denote the string in $\{0,1\}^n$ formed by plugging in the bits of x' and x'' (in order) into the coordinates in I and $[n] \setminus I$, respectively.

We now give a suitable $(k - i - 1)$-round protocol $\mathcal{Q}$ for GHD', in which Bob sends the first message. Consider an input $(x', y') \in \{0,1\}^{n'} \times \{0,1\}^{n'}$, with Alice holding x' and Bob holding y'. By definition of shattering, there exists an $x'' \in \{0,1\}^{n-n'}$ such that $x := x' \circ x'' \in M$. Alice and Bob agree beforehand on a suitable x for each possible x'. Suppose Bob were to pick a uniform random $y'' \in \{0,1\}^{n-n'}$ and form the string $y := y' \circ y''$. Then, Alice and Bob could simulate $\mathcal{P}$ on input (x, y) using only $k - i - 1$ rounds of communication, with Bob starting, because Alice's first message in $\mathcal{P}$ would always be $\mathfrak{m}$. Call this randomized protocol $\mathcal{Q}_1$. We define $\mathcal{Q}$ to be the protocol obtained by running t instances of $\mathcal{Q}_1$ in parallel, using independent random choices of y'', and outputting the majority answer. Note that the length of each message in Q is $ts = s'$. We shall now analyze the error.

Suppose $x'' \perp_b y''$. Let $d_1 = \Delta(x, y) - n/2$, $d_2 = \Delta(x', y') - n'/2$ and $d_3 = \Delta(x'', y'') - (n - n')/2$. Clearly, $d_1 = d_2 + d_3$. Also,

$$|d_1| \ge |d_3| - |d_2| \ge c'\sqrt{n'} - b\sqrt{n - n'} \ge \frac{(c' - b\sqrt{2})\sqrt{n}}{\sqrt{3}} \ge c\sqrt{n}\,,$$

where we used (1) and our choice of b. Thus, $x \not\perp_c y$. The same calculation also shows that d_1 and d_3 have the same sign, as $|d_3| > |d_2|$. Therefore $\text{GHD}(x, y) = \text{GHD}'(x', y')$.

For the rest of the calculations in this proof, fix an input x' for Alice, and hence, x'' and x as well. For a fixed y', let $\mathcal{E}(y')$ denote the event that $\mathcal{P}(x, y) \neq \text{GHD}(x, y)$: note that y'' remains random. Using the above observation (at step (3) below), we can bound the probability that $\mathcal{Q}_1$ errs on input (x', y') as follows.

$$\begin{aligned}
\Pr_y\left[Q_1(x', y') \neq \text{GHD}'(x', y') \mid y'\right] &\leq \Pr_y\left[\mathcal{P}(x, y) \neq \text{GHD}(x, y) \vee \text{GHD}(x, y) \neq \text{GHD}'(x', y') \mid y'\right] \\
&\leq \Pr_{y''}\left[\mathcal{E}(y')\right] + \Pr_y\left[\text{GHD}(x, y) \neq \text{GHD}'(x', y') \mid y'\right] \\
&\leq \Pr_{y''}\left[\mathcal{E}(y')\right] + \Pr_{y''}\left[x'' \not\perp_b y''\right] \qquad (3) \\
&\leq \Pr_{y''}\left[\mathcal{E}(y')\right] + T(b) \\
&= \Pr_{y''}\left[\mathcal{E}(y')\right] + 1/8\,, \qquad (4)
\end{aligned}$$

where step (4) follows from our choice of b. To analyze $\mathcal{Q}$, notice that during the t-fold parallel repetition of $\mathcal{Q}_1$, y' remains fixed while y'' varies. Thus, it suffices to understand how the repetition drives down the sum on the right side of (4). Unfortunately, for some values of y', the sum may exceed $\frac{1}{2}$, in which case it will be driven *up*, not down, by the repetition. To account for this, we shall bound the *expectation* of the first term of that sum, for a random y'.

To do so, let $z \sim \mu \mid x$ be a random string independent of y. Notice that z is uniformly distributed on a subset of $\{0,1\}^n$ of size $2^n T(c)$, whereas y is uniformly distributed on a subset of $\{0,1\}^n$ of size $2^n T(c')$. (We are now thinking of x as being fixed and both y' and y'' as being random.) Therefore,

$$\begin{aligned}
\mathbb{E}_{y'}\left[\Pr_{y''}\left[\mathcal{E}(y')\right]\right] = \Pr_y\left[\mathcal{E}(y')\right] &= \Pr_y[\mathcal{P}(x, y) \neq \text{GHD}(x, y)] \\
&\leq \Pr_z[\mathcal{P}(x, z) \neq \text{GHD}(x, z)] \cdot T(c)/T(c') \\
&\leq 2\varepsilon T(c)/T(c')\,, \qquad (5)
\end{aligned}$$

where (5) holds because x, being good, satisfies (2). Thus, by Markov's inequality,

$$\Pr_{y'}\left[\Pr_{y''}\left[\mathcal{E}(y')\right] \geq \frac{1}{8}\right] \leq 16\varepsilon T(c)/T(c')\,. \qquad (6)$$

If, for a particular y', the *bad event* $\Pr_{y''}[\mathcal{E}(y')] \geq \frac{1}{8}$ does *not* occur, then the right side of (4) is at most $1/8 + 1/8 = 1/4$. In other words, $\mathcal{Q}_1$ errs with probability at most $1/4$ for this y'. By standard Chernoff bounds, the t-fold repetition in $\mathcal{Q}$ drives this error down to $(e/4)^{t/4} \leq 2^{-t/10} \leq \varepsilon_0 \leq \varepsilon$. Combining this with (6), which bounds the probability of the bad event, we get

$$\Pr_{y',r}\left[\mathcal{Q}(x', y') \neq \text{GHD}'(x', y')\right] \leq 16\varepsilon T(c)/T(c') + \varepsilon \leq \varepsilon/T(c') = \varepsilon'\,,$$

where r denotes the internal random string of $\mathcal{Q}$ (i.e., the collection of y''s used).

Note that this error bound holds for *every* fixed x', and thus, when $(x', y') \sim \mu'$. Therefore, we can fix Bob's random coin tosses in $\mathcal{Q}$ to get the desired $[k - i - 1, n', s', c', \varepsilon']$-protocol. □

3.3 The Lower Bound

Having established our round elimination lemma, we obtain our lower bound in a straightforward fashion.

Theorem 10 (Multi-round Lower Bound). *Let $\mathcal{P}$ be a k-round $\frac{1}{3}$-error randomized communication protocol for $\mathrm{GHD}_{c,n}$, with $c = O(1)$, in which each message is s bits long. Then*

$$s \;\geq\; \frac{n}{2^{O(k^2)}}\,.$$

Remark. This is a formal restatement of Theorem 1.

Proof. For simplicity, assume $c \leq c_0 = 10$. Our proof easily applies to a general $c = O(1)$ by a suitable modification of the parameters in (1). Also, assume $n \geq 2^{4k^2}$, for otherwise there is nothing to prove.

By repeating $\mathcal{P}$ $(48 \ln 2) \cdot 2^{11k} = t_0$ times, in parallel, and outputting the majority of the answers, we can reduce the error to $2^{-2^{11k}} = \varepsilon_0$. The size of each message is now $t_0 s = s_0$. Fixing the random coins of the resulting protocol gives us a $[k, n_0, s_0, c_0, \varepsilon_0]$-protocol $\mathcal{P}_0$.

Suppose $s_i \leq n_i/20$ for all i, with $0 \leq i < k$. We then repeatedly apply Lemma 9 k times, starting with $\mathcal{P}_0$. Eventually, we end up with a $[0, n_k, s_k, c_k, \varepsilon_k]$-protocol. Examining (1), we see that $n_k = n/3^k$, $s_k = 2^{15k^2} s_0 = (48 \ln 2) 2^{15k^2+11k} s$, and $c_k = 10 \cdot 2^k$. Notice that $n_k \geq 2^{4k^2}/3^k > 1$ and $c_k = o(\sqrt{n_k})$. We also see that $\langle c_i \rangle_{i=1}^{k}$ is an increasing sequence, whence $\varepsilon_{i+1}/\varepsilon_i = 1/T(c_{i+1}) \leq 1/T(c_k) \leq 2^{3{c_k}^2+2}$, where the final step uses Fact 3. Thus,

$$\varepsilon_k \;\leq\; \varepsilon_0 \big(2^{3c_k^2+2}\big)^k \;=\; 2^{-2^{11k}} \cdot 2^{(3(10 \cdot 2^k)^2+2) \cdot k} \;=\; 2^{-2^{11k}+300k \cdot 2^{2k}+2k} \;<\; \frac{1}{2}\,.$$

In other words, we have a $[0, n_k, s_k, c_k, \varepsilon_k]$-protocol with $n_k > 1$, $c_k = o(\sqrt{n_k})$ and $\varepsilon_k < \frac{1}{2}$. This contradicts Lemma 8.

Therefore, there must exist an i such that $s_i \geq n_i/20$. Since $\langle s_i \rangle_{i=1}^{k}$ is increasing and $\langle n_i \rangle_{i=1}^{k}$ is decreasing, $s_k \geq n_k/20$. By the above calculations, $(48 \ln 2) 2^{15k^2+11k} s \geq n/(20 \cdot 3^k)$, which implies $s \geq n/2^{O(k^2)}$, as claimed. □

Notice that, for constant k, the argument in the above proof in fact implies a lower bound for deterministic protocols with small enough constant distributional error under $\mu_{c,n}$. This, in turn, extends to distributional error under the uniform distribution, as remarked earlier.

4 Tight Deterministic One-Way Bounds

The main result of this section is the following.

Theorem 11. $D^{\rightarrow}(\mathrm{GHD}_{c,n}) = n - \Theta(\sqrt{n} \log n)$ *for all constant c.*

Definition 6. Let $x_1, x_2, y \in \{0,1\}^n$. We say that y *witnesses* x_1 and x_2 or that y is a witness for (x_1, x_2) if $x_1 \not\perp_c y$, $x_2 \not\perp_c y$, and $\mathrm{GHD}_{c,n}(x_1, y) \neq \mathrm{GHD}_{c,n}(x_2, y)$.

Intuitively, if (x_1, x_2) have a witness, then they cannot be in the same message set. For if Alice sent the same message on x_1 and x_2 and Bob's input y was a witness for (x_1, x_2) then whatever Bob were to output, the protocol would err on either (x_1, y) or (x_2, y). The next lemma characterizes which (x_1, x_2) pairs have witnesses.

Lemma 12. *For all $x_1, x_2 \in \{0,1\}^n$, there exists y that witnesses (x_1, x_2) if and only if $\Delta(x_1, x_2) \geq 2c\sqrt{n}$.*

Proof. On the one hand, suppose y witnesses (x_1, x_2). Then assume WLOG that $\Delta(x_1, y) \leq n/2 - c\sqrt{n}$ and $\Delta(x_2, y) \geq n/2 + c\sqrt{n}$. By the triangle inequality, $\Delta(x_1, x_2) \geq \Delta(x_2, y) - \Delta(x_1, y) = 2c\sqrt{n}$. Conversely, suppose $\Delta(x_1, x_2) \geq 2c\sqrt{n}$. Let $L = \{i : x_1[i] = x_2[i]\}$, and let $R = \{i : x_1[i] \neq x_2[i]\}$. Suppose y agrees with x_1 on all coordinates from R and half the coordinates from L. Then, $\Delta(x_1, y) = |L|/2 = (n - \Delta(x_1, x_2))/2 \leq n/2 - c\sqrt{n}$. Furthermore, y agrees with x_2 on *no* coordinates from R and half the coordinates from L, so $\Delta(x_1, y) = |L|/2 + |R| \geq n/2 + c\sqrt{n}$. □

We show that it is both necessary and sufficient for Alice to send different messages on x_1 and x_2 whenever $\Delta(x_1, x_2)$ is "large". To prove this, we need the following theorem, due to Bezrukov [Bez87] and a claim that is easily proved using the probabilistic method (a full proof of the claim appears in the appendix).

Theorem 13. *Call a subset $A \subseteq \{0,1\}^n$ d-maximal if it is largest, subject to the constraint that $\Delta(x, y) \leq d$ for all $x, y \in A$.*

1. *If $d = 2t$ then $\mathcal{B}(x, t)$ is d-maximal for any $x \in \{0,1\}^n$.*
2. *If $d = 2t + 1$ then $\mathcal{B}(x, t) \cup \mathcal{B}(y, t)$ is d-maximal for any $x, y \in \{0,1\}^n$ such that $\Delta(x, y) = 1$.* □

Claim 14. *It is possible to cover $\{0,1\}^n$ with at most $2^{n - O(\sqrt{n}\log n)}$ Hamming balls, each of radius $c\sqrt{n}$.* □

Proof of Theorem 11. For the lower bound, suppose for the sake of contradiction that there is a protocol where Alice sends only $n - c\sqrt{n}\log n$ bits. By the pigeonhole principle, there exists a set $M \subseteq \{0,1\}^n$ of inputs of size $|M| \geq 2^n / 2^{n - c\sqrt{n}\log n} = 2^{c\sqrt{n}\log n} = n^{c\sqrt{n}}$ upon which Alice sends the same message. By Theorem 13, the Hamming ball $\mathcal{B}(x, c\sqrt{n})$ is $2c\sqrt{n}$-maximal, and by Fact 4, $|\mathcal{B}(x, c\sqrt{n})| < |M|$. Therefore, there must be $x_1, x_2 \in M$ with $\Delta(x_1, x_2) > 2c\sqrt{n}$. By Lemma 12, there exists a y that witnesses (x_1, x_2). No matter what Bob outputs, the protocol errs on either (x_1, y) or on (x_2, y).

For a matching upper bound, Alice and Bob fix a covering $\mathcal{C} = \{\mathcal{B}(x_0, r)\}$ of $\{0,1\}^n$ by Hamming balls of radius $r = c\sqrt{n}$. On input x, Alice sends Bob the Hamming ball $\mathcal{B}(x_0, r)$ containing x. Bob selects some $x' \in \mathcal{B}(x_0, r)$ such that $x' \not\perp_c y$ and outputs $\text{GHD}(x', y)$. The correctness of this protocol follows from Lemma 12, as $\Delta(x, x') \leq 2c\sqrt{n}$ since they are both in $\mathcal{B}(x_0, c\sqrt{n})$. The cost of the protocol is given by Claim 14, which shows that it suffices for Alice to send $\log\left(2^{n - O(\sqrt{n}\log n)}\right) = n - O(\sqrt{n}\log n)$ bits to describe each Hamming ball. □

5 One Round Randomized Lower Bound

Next, we develop a one-way lower bound for randomized protocols. Note that our lower bound applies to the uniform distribution, which, as mentioned in Section 1, implies space lower bounds for the Distinct Elements problem under weaker assumptions about the input stream. Woodruff [Woo09] recently proved similar results, also for the uniform distribution. We include our lower bound as a natural extension of the deterministic bound.

Theorem 15. $R_\varepsilon^{\rightarrow}(\text{GHD}_{c,n}) = \Omega(n)$.

Proof. For the sake of clarity, fix $c = 2$ and $\varepsilon = 1/10$, and suppose $\mathcal{P}$ is a one-round, ε-error, $o(n)$-bit protocol for $\text{GHD}_{c,n}$.

Definition 7. For $x \in \{0,1\}^n$, let $Y_x := \{y : x \not\perp_2 y\}$. Say that x is *good* if $\Pr_{y \in Y_x}[\mathcal{P}(x,y) = \text{GHD}(x,y)] \leq 2\varepsilon$. Otherwise, call x *bad*.

By Markov's inequality, at most a $1/2$-fraction of x are *bad*. Next, fix Alice's message $\mathfrak{m}$ to maximize the number of *good* x, and let $M = \{x \in \{0,1\}^n : x \text{ is good and Alice sends } \mathfrak{m} \text{ on input } x\}$. It follows that

$$|M| \geq 2^{n-1}/2^{o(n)} > 2^{n(1-o(1))}.$$

Our goal is to show that since $|M|$ is large, we must err on a $> 2\varepsilon$-fraction of $y \in Y_x$ for some $x \in M$, contradicting the goodness of x. Note that it suffices to show that a 4ε fraction of $y \in Y_{x_1}$ witness x_1 and x_2.

$|M| \geq 2^{n(1-o(1))}$, so by Fact 5 and Theorem 13, There exist x_1, x_2 with $\Delta(x_1,x_2) \geq 1 - o(1)$. Next, we'd like to determine the probability that a random $y \in Y_{x_1}$ witnesses (x_1,x_2). Without loss of generality, let $x_1 = 0^n$. Let $w(x) := \Pr_{y \in Y_{x_1}}[\text{GHD}(x,y) \neq \text{GHD}(x_1,y)]$. The following lemma shows that $w(x)$ is an increasing function of $|x|$. We leave the proof until the appendix.

Lemma 16. *For all $x, x' \in \{0,1\}^n$, $w(x) \geq w(x') \Leftrightarrow |x| \geq |x'|$, with equality if and only if $|x| = |x'|$.*

We compute $w(x)$ by conditioning on $|y|$:

$$w(x) \;=\; \sum_{n_1 \leq n/2 - c\sqrt{n}} \Pr\left[\Delta(x,y) \geq n/2 + c\sqrt{n} \mid |y| = n_1\right] \cdot \Pr[|y| = n_1]\,.$$

Fix $|x| =: m$, pick a random y with $|y| = n_1$, and suppose there are k coordinates i such that $x_i = y_i$. Then, $\Delta(x,y) = (m-k) + (n_1 - k) = m + n_1 - 2k$. Hence,

$$\Delta(x,y) \geq n/2 + c\sqrt{n} \iff k \leq \frac{m+n_1}{2} - \frac{n}{4} - \frac{c}{2}\sqrt{n}\,.$$

Note that given a random y with weight $|y| = n_1$, the probability that exactly k of m coordinates have $x_i = y_i = 1$ follows the hypergeometric distribution $\text{Hyp}(k; n, m, n_1)$. Therefore, we can express the probability $\Pr_{|y|=n_1}[\Delta(x,y) \geq n/2 + c\sqrt{n}]$ as

$$\Pr_{|y|=n_1}\left[\Delta(x,y) \geq n/2 + c\sqrt{n}\right] \;=\; \sum_{k \leq \frac{m+n_1}{2} - \frac{n}{4} - \frac{c}{2}\sqrt{n}} \text{Hyp}(k; n, m, n_1)\,.$$

Finally, we show that $w(x) > 4\varepsilon$ for a suitably large constant $|x|$ with the following claims, whose proofs are left to the appendix.

Claim 17. *Conditioned on $|y| \leq n/2 - 2\sqrt{n}$, we have $\Pr[|y| \geq n/2 - 2.1\sqrt{n}] \leq \frac{1}{3}$.*

Claim 18. *For all $d < n/2 - 2.1\sqrt{n}$, we have $\Pr[\Delta(x_2,y) \geq n/2 + d\sqrt{n}] \geq 0.95$.*

Its easy to see from the previous two claims that $w(x) > 0.95 \cdot (2/3) > 4\varepsilon$. □

6 Concluding Remarks

Our most important contribution here was to prove a multi-round lower bound on a fundamental problem in communication complexity, the Gap-Hamming Distance problem. As a consequence, we extended several known $\Omega(1/\varepsilon^2)$-type space bounds for various data stream problems, such as the Distinct Elements problem, to multi-pass algorithms. These resolve long-standing open questions.

The most immediate open problem suggested by our work is to resolve Conjecture 2. It appears that proving the conjecture true is going to require a technique other than round elimination, or else, an *extremely* powerful round elimination lemma that does not lose a constant fraction of the input length at each step. On the other hand, proving the conjecture false is also of great interest, and such a proof might extend to nontrivial data stream algorithms, albeit with a super-constant number of passes.

Acknowledgements

We would like to thank Anna Gal, T. S. Jayram and David Woodruff for stimulating discussions about the problem at various points of time.

References

[ABC09] Chrisil J. Arackaparambil, Joshua Brody, and Amit Chakrabarti. Functional monitoring without monotonicity. Manuscript, 2009.

[ADHP06] Micah Adler, Erik D. Demaine, Nicholas J. A. Harvey, and Mihai Pătraşcu. Lower bounds for asymmetric communication channels and distributed source coding. In *Proc. 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 251–260, 2006.

[AMS99] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999. Preliminary version in *Proc. 28th Annu. ACM Symp. Theory Comput.*, pages 20–29, 1996.

[Bez87] Sergei Bezrukov. Specification of the maximal sized subsets of the unit cube with respect to given diameter. *Problems of Information Transmission*, 1:106–109, 1987.

[BJK+04] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, D. Sivakumar, and Luca Trevisan. Counting distinct elements in a data stream. In *Proc. 6th International Workshop on Randomization and Approximation Techniques in Computer Science*, pages 128–137, 2004.

[Bro09] Joshua Brody. The maximum communication complexity of multi-party pointer jumping. Manuscript, 2009.

[CCM07] Amit Chakrabarti, Graham Cormode, and Andrew McGregor. A near-optimal algorithm for computing the entropy of a stream. In *Proc. 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 328–335, 2007.

[Cha07] Amit Chakrabarti. Lower bounds for multi-player pointer jumping. In *Proc. 22nd Annual IEEE Conference on Computational Complexity*, pages 33–45, 2007.

[CJP08] Amit Chakrabarti, T. S. Jayram, and Mihai Pătraşcu. Tight lower bounds for selection in randomly ordered streams. In *Proc. 19th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 720–729, 2008.

[CMY08] Graham Cormode, S. Muthukrishnan, and Ke Yi. Algorithms for distributed functional monitoring. In *Proc. 19th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1076–1085, 2008.

[CR04] Amit Chakrabarti and Oded Regev. An optimal randomised cell probe lower bound for approximate nearest neighbour searching. In *Proc. 45th Annual IEEE Symposium on Foundations of Computer Science*, pages 473–482, 2004.

[Fel68] William Feller. *An Introduction to Probability Theory and its Applications*. John Wiley, New York, NY, 1968.

[FM85] Philippe Flajolet and G. Nigel Martin. Probabilistic counting algorithms for data base applications. *J. Comput. Syst. Sci.*, 31(2):182–209, 1985.

[HS05] Don Hush and Clint Scovel. Concentration of the hypergeometric distribution. *Statistics and Probability Letters*, 75(2):127–132, 2005.

[IW03] Piotr Indyk and David Woodruff. Tight lower bounds for the distinct elements problem. In *Proc. 45th Annual IEEE Symposium on Foundations of Computer Science*, pages 283–289, 2003.

[JKS07] T. S. Jayram, Ravi Kumar, and D. Sivakumar. The one-way communication complexity of gap hamming distance. Manuscript, 2007.

[KN97] Eyal Kushilevitz and Noam Nisan. *Communication Complexity*. Cambridge University Press, Cambridge, 1997.

[Kum06] Ravi Kumar. Story of distinct elements, 2006. talk at IITK Workshop on Algorithms for Data Structures.

[KW90] Mauricio Karchmer and Avi Wigderson. Monotone circuits for connectivity require super-logarithmic depth. *SIAM J. Disc. Math.*, 3(2):255–265, 1990. Preliminary version in *Proc. 20th Annual ACM Symposium on the Theory of Computing*, pages 539–550, 1988.

[LS07] Nati Linial and Adi Shraibman. Lower bounds in communication complexity based on factorization norms. In *Proc. 39th Annual ACM Symposium on the Theory of Computing*, pages 699–708, 2007.

[MNSW98] Peter Bro Miltersen, Noam Nisan, Shmuel Safra, and Avi Wigderson. On data structures and asymmetric communication complexity. *J. Comput. Syst. Sci.*, 57(1):37–49, 1998. Preliminary version in *Proc. 27th Annual ACM Symposium on the Theory of Computing*, pages 103–111, 1995.

[NW99] Ashwin Nayak and Felix Wu. The quantum query complexity of approximating the median and related statistics. In *Proc. 31st Annual ACM Symposium on the Theory of Computing*, pages 384–393, 1999.

[Păt08] Mihai Pătraşcu. (data) structures. In *Proc. 49th Annual IEEE Symposium on Foundations of Computer Science*, 2008.

[Raz02] Alexander A. Razborov. Quantum communication complexity of symmetric predicates. *Izvestiya of the Russian Academy of Science, Mathematics*, 67:0204025, 2002.

[Sau72] N. Sauer. On the density of families of sets. *J. Combin. Theory Ser. A*, 13:145–147, 1972.

[Sen03] Pranab Sen. Lower bounds for predecessor searching in the cell probe model. In *Proc. 18th Annual IEEE Conference on Computational Complexity*, pages 73–83, 2003.

[She08] Alexander A. Sherstov. The pattern matrix method for lower bounds on quantum communication. In *STOC '08: Proceedings of the 40th annual ACM symposium on Theory of computing*, pages 85–94, New York, NY, USA, 2008. ACM.

[VW07] Emanuele Viola and Avi Wigderson. One-way multi-party communication lower bound for pointer jumping with applications. In *Proc. 48th Annual IEEE Symposium on Foundations of Computer Science*, pages 427–437, 2007.

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

[Woo07] David P. Woodruff. *Efficient and Private Distance Approximation in the Communication and Streaming Models*. PhD thesis, MIT, 2007.

[Woo09] David Woodruff. The average case complexity of counting distinct elements. In *Proc. 12th International Conference on Database Theory*, 2009.

APPENDIX

A Proofs of Technical Lemmas

We begin with a proof of Claim 14, which we state here for convenience.

Claim 19 (Restatement of Claim 14). *For any constant c, it is possible to cover $\{0,1\}^n$ with at most $2^{n-O(\sqrt{n}\log n)}$ Hamming balls, each with radius $r = c\sqrt{n}$.*

Proof. We use the probabilistic method. Let $r := c\sqrt{n}$. For $x \in \{0,1\}^n$, let $\mathcal{B}_x := \mathcal{B}(x,r)$ be the Hamming ball of radius r centered at x. For a t to be determined later, pick $x_1, \ldots, x_t$ independently and uniformly at random from $\{0,1\}^n$. We want to show that with nonzero probability, the universe $\{0,1\}^n$ is covered by these t Hamming balls $\mathcal{B}_{x_1}, \ldots, \mathcal{B}_{x_t}$.

Now, fix any $x \in \{0,1\}^n$ and any $1 \le i \le t$. Since x_i was picked uniformly at random, each x is equally likely to be in $\mathcal{B}_{x_i}$. Therefore,

$$\Pr[x \in \mathcal{B}_{x_i}] = \frac{|\mathcal{B}_{x_i}|}{2^n} \ge 2^{\theta(\sqrt{n}\log n)-n}$$

where inequality stems from Fact 4.

Let $BAD_x = \bigwedge_{1\le i\le t} x \notin \mathcal{B}_{x_i}$ be the event that x is not covered by any of the Hamming balls we picked at random, and let $BAD = \bigvee BAD_x$ be the event that *some* x is not covered by the Hamming balls. We want to limit $\Pr[BAD]$. BAD_x occurs when $x \notin \mathcal{B}_{x_i}$ for all x_i. Therefore, using $1 - x \le e^{-x}$ for all real x,

$$\Pr[BAD_x] = \left(1 - 2^{\theta(\sqrt{n}\log n)-n}\right)^t \le \mathrm{e}^{-t\cdot 2^{\theta(\sqrt{n}\log n)-n}}.$$

By the union bound,

$$\Pr[BAD] \le 2^n \Pr[BAD_x] = 2^{n-\frac{t}{\ln 2}2^{\theta(n\sqrt{n})-n}}.$$

Picking $t = \ln 2(n+1)2^{n-\theta(\sqrt{n}\log n)} = 2^{n-\theta(\sqrt{n}\log n)}$ ensures that $\Pr[BAD] < 1$. Therefore, there exists a set of $t = 2^{n-\theta(\sqrt{n}\log n)}$ Hamming balls of radius $c\sqrt{n}$ that cover $\{0,1\}^n$. □

Recall that $w(x) = \Pr_{y\in Y_0}[\mathrm{GHD}(x,y) \ne \mathrm{GHD}(\vec{0},y)]$.

Lemma 20 (Restatement of Lemma 16). *For all $x, x' \in \{0,1\}^n$, $w(x) \le w(x')$ if and only if $|x| \le |x'|$, with equality if and only if $|x| = |x'|$.*

Proof. If $|x| = |x'|$, then $w(x) = w(x')$ by symmetry. Further, note that $\mathrm{GHD}(x,y) = 0$ if and only if $\mathrm{GHD}(-x,y) = 1$. Therefore, it suffices to handle the case where $|y| \le n/2 - c\sqrt{n}$ and $\mathrm{GHD}(\vec{0},y) = 0$.

For the rest of the proof, we assume that $x_i = x'_i$, except for the nth coordinate, where $x_n = 0$ and $x'_n = 1$. Thus, $|x| = |x'| - 1$. We show that $w(x) < w(x')$; the rest of the lemma follows by induction.

Let Y be the set of strings with Hamming weight $|y| \le n/2 - c\sqrt{n}$. Partition Y into the following three sets:

- $A := \{y : |y| = n/2 + c\sqrt{n} \wedge y_n = 0\}$.
- $B := \{y : |y| < n/2 + c\sqrt{n} \wedge y_n = 0\}$.
- $C := \{y : y_n = 1\}$.

Note the one-to-one correspondence between strings in B and strings in C obtained by flipping the nth bit. Now, consider any $y \in B$ such that y witnesses $(\vec{0}, x')$ but not $(\vec{0}, x)$. Flipping the nth bit of y yields a string $y' \in C$ such that Y witnesses $(\vec{0}, x)$ but not $(\vec{0}, x')$. Hence among $y \in B \cup C$ there is an equal number of witnesses for x and x'. For any $y \in A$, $y_n = 0$, whence $|y - x'| = |y - x| + 1$. Therefore, any y that witnesses $(\vec{0}, x)$ must also witness $(\vec{0}, x')$, whence $w(x) \leq w(x')$. □

Many claims in this paper require tight upper and lower tail bounds for binomial and hypergeometric distributions. We use Chernoff bounds where they apply. For other bounds, we approximate using normal distributions. We use Feller [Fel68] as a reference.

Definition 8. For $x \in \mathbb{R}$, let $\phi(x) := e^{-x^2/2}/\sqrt{2\pi}$ and

$$N(x) := \int_x^\infty \phi(y)dy.$$

$N(x)$ is the cumulative distribution function of the normal distribution. We use it in Fact 3 to approximate $T(x)$. Here, we'll also use it to approximate tails of the binomial and hypergeometric distributions.

Lemma 21 (Feller, Chapter VII, Lemma 2.). *For all $x > 0$,*

$$\phi(x)\left(\frac{1}{x} - \frac{1}{x^3}\right) < N(x) < \phi(x)\frac{1}{x}.$$

Theorem 22 (Feller, CHapter VII, Theorem 2.). *For fixed z_1, z_2,*

$$Pr[n/2 + (z_1/2)\sqrt{n} \leq |y| \leq n/2 + (z_2/2)\sqrt{n}] \sim N(z_1) - N(z_2).$$

Theorem 23. *For any γ such that $\gamma = \omega(1)$ and $\gamma = o(n^{1/6})$, we have*

$$\sum_{k > n/2 + \gamma\sqrt{n}/2} \binom{n}{k} \sim N(\gamma).$$

Claim 24 (Restatement of Claim 17). *Conditioned on $|y| \leq n/2 - 2\sqrt{n}$,*

$$Pr[|y| \geq n/2 - 2.1\sqrt{n}] \leq 1/3.$$

Proof. By Theorem 22 and Lemma 21, we have

$$\begin{aligned} \Pr[n/2 - 2.1\sqrt{n} \leq |y| \leq n/2 - 2\sqrt{n}] &\sim N(4) - N(4.2) \\ &\leq \phi(4)/4 - \phi(4.2)(4.2^{-1} - 4.2^{-3}) \\ &\leq 2.0219 * 10^{-5} \end{aligned}$$

By Fact3, $\Pr[|y| \geq n/2 - 2\sqrt{n}] \leq 2^{-3 \cdot 2^2 - 2} = 2^{-14} = 6.1035 \cdot 10^{-5}$. Putting the two terms together, we get

$$\Pr[|y| \geq n/2 - 2.1\sqrt{n} || y| \leq n/2 - 2\sqrt{n}] \leq \frac{2.0219 \cdot 10^{-5}}{6.1035 \cdot 10^{-5}} \leq 1/3.$$

□

Claim 25 (Restatement of Claim 18). *For all $d < n/2 - 2.1\sqrt{n}$,*

$$\Pr[\Delta(x_2, y) \geq n/2 + 2\sqrt{n}] \geq 0.95.$$

Proof. The proof follows from the following claim, instantiated with $c = 2$ and $\alpha = 2.1$. □

Claim 26. *For all $\alpha > c$, $|x| = \gamma n$, and all $\gamma \geq 1 - (1 - c/\alpha)/4$,*

$$\Pr_{|y|=n/2-\alpha\sqrt{n}}[\Delta(x,y) \geq n/2 + c\sqrt{n}] \geq 1 - \exp\left(-\frac{2(\alpha-c)\alpha^2(1+o(1))}{3\alpha+c}\right).$$

Proof. Let $m := |x| = \gamma n$ and let $n_1 = n/2 - \alpha\sqrt{n}$. Then, the probability that a random y with $|y| = n_2$ can be expressed using the hypergeometric distribution $\mathrm{Hyp}(k; n, m, n_1)$. Let the m set bits of x be the defects. The probability of k of the n_1 bits of y are defective is $\mathrm{Hyp}(k; n, m, n_1)$. Note that $\Delta(x,y) = (m-k) + (n_1 - k) = m + n_1 - 2k$. Therefore,

$$\Delta(x,y) \geq n/2 + c\sqrt{n} \Leftrightarrow k \leq \frac{m+n_1}{2} - \frac{n}{4} - \frac{c}{2}\sqrt{n}. = \frac{\gamma n}{2} - \frac{\alpha+c}{2}\sqrt{n}$$

We express the probability $\Pr_{|y|=n_1}[\Delta(x,y) \geq n/2 + c\sqrt{n}]$ as

$$\Pr_{|y|=n_1}[\Delta(x,y) \geq n/2 + c\sqrt{n}] = \Pr_{K\sim\mathrm{Hyp}(k;n,m,n_1)}[K \leq \frac{\gamma n}{2} - \frac{\alpha+c}{2}\sqrt{n}].$$

Next, we use a concentration of measure result due to Hush and Scovel [HS05]. Here, we present a simplified version.

Theorem 27 (Hush, Scovel). *Let $m = \gamma n > n_1 = n/2 - \alpha\sqrt{n}$, and let $\beta = n/m(n-m)$.*

$$\Pr[K - E[K] > \eta] < \exp(-2\beta\eta^2(1+o(1))).$$

The expected value of a random variable K distributed according to $\mathrm{Hyp}(K; n, m, n_1)$ is

$$E[K] = \frac{mn_1}{n} = \frac{\gamma n}{n}\left(\frac{n}{2} - \alpha\sqrt{n}\right) = \frac{\gamma n}{2} - \gamma\alpha\sqrt{n}.$$

Set $\eta := (\alpha - c)\sqrt{n}/4$. Note that

$$E[K] + \eta = \frac{\gamma n}{2} - \gamma\alpha\sqrt{n} + \frac{\alpha-c}{4}\sqrt{n} \leq \frac{\gamma n}{2} - \frac{\alpha+c}{2}\sqrt{n} = \frac{m+n_1}{2} - \frac{n}{4} - \frac{c}{2}\sqrt{n}$$

where the inequality holds because $\gamma \geq 1 - (1 - c/\alpha)/4$. Note also that $(1 - c/\alpha)/4 = (\alpha - c)/4\alpha$, so $1 - (1 - c/\alpha)/4 = (3\alpha + c)/4\alpha$. By Theorem 27

$$\begin{aligned}
\Pr[K > \frac{\gamma n}{2} - \frac{\alpha+c}{2}\sqrt{n}] &= \Pr[K - E[K] > \eta] \\
&< \exp\left(-\frac{2n\eta^2(1+o(1))}{m(n-m)}\right) \\
&= \exp\left(-\frac{2(\alpha-c)^2(1+o(1))}{16\gamma(1-\gamma)}\right) \\
&\leq \exp\left(-\frac{2(\alpha-c)^2(4\alpha)^2(1+o(1))}{16(\alpha-c)(3\alpha+c)}\right) \\
&= \exp\left(-\frac{2(\alpha-c)\alpha^2(1+o(1))}{3\alpha+c}\right)
\end{aligned}$$

It follows that $\Pr[K \leq \frac{\gamma n}{2} - \frac{\alpha+c}{2}\sqrt{n}] \geq 1 - \exp\left(-\frac{2(\alpha-c)\alpha^2(1+o(1))}{3\alpha+c}\right)$. □

Claim 28. *For any $x_L \in \{0,1\}^{n_L}$, $\mathrm{GHD}(x_L, y_L)$ is defined for at least a $\geq e^{-2(c')^2}/5c'$-fraction of $y_L \in \{0,1\}^{n_L}$.*

Proof. Without loss of generality, assume $x_L = \vec{0}$. Then, $\mathrm{GHD}(x_L, y_L)$ is defined for all y such that $|y| \leq n_L/2 - c'\sqrt{n_L}$ or $|y| \geq n_L/2 + c'\sqrt{n_L}$. Note that for any constant $x > c'$,

$$\begin{aligned}
\Pr_y[|y| \leq \frac{n_L}{2} - c'\sqrt{n_L}] &\geq \Pr[\frac{n_L}{2} - x\sqrt{n_L} \leq |y| \leq \frac{n_L}{2} - c'\sqrt{n_L}] \\
&\geq N(2c') - N(2x) \\
&\geq \phi(2c')\left(\frac{1}{2c'} - \frac{1}{(2c')^3}\right) - \frac{\phi(2x)}{2x} \\
&= \frac{e^{-(2c')^2/2}}{\sqrt{2\pi}}(\left(\frac{1}{2c'} - \frac{1}{(2c')^3}\right) - \frac{e^{-2x^2}}{2x\sqrt{2\pi}} \\
&\geq \frac{e^{-2(c')^2}}{10c'}
\end{aligned}$$

$\Pr[|y| \geq n_L/2 + c'\sqrt{n_L}]$ is bounded in the same fashion. □

ECCC ISSN 1433-8092

http://eccc.hpi-web.de/