Electronic Colloquium on Computational Complexity, Report No. 63 (2011)

THE COMMUNICATION COMPLEXITY OF GAP HAMMING DISTANCE

ALEXANDER A. SHERSTOV*

ABSTRACT. In the *gap Hamming distance* problem, two parties must determine whether their respective strings $x, y \in \{0,1\}^n$ are at Hamming distance less than $n/2 - \sqrt{n}$ or greater than $n/2 + \sqrt{n}$. In a recent tour de force, Chakrabarti and Regev (STOC '11) proved the long-conjectured $\Omega(n)$ bound on the randomized communication complexity of this problem. In follow-up work several months ago, Vidick (2010; ECCC TR11-051) discovered a simpler proof. We contribute a new proof, which is simpler yet and a page-and-a-half long.

1. INTRODUCTION

The *gap Hamming distance* problem features two communicating parties, the first of which receives a vector $x \in \{-1,+1\}^n$ and the second a vector $y \in \{-1,+1\}^n$. The two vectors are chosen such that the Hamming distance between them is either noticeably smaller than $n/2$ or noticeably larger than $n/2$. The objective is to reliably determine which is the case by exchanging as few bits of communication as possible. Throughout this paper, communication is assumed to be randomized, and the communication protocol is to produce the correct answer with probability $2/3$. Formally, gap Hamming distance is the communication problem that corresponds to the following Boolean function on $\{-1,+1\}^n \times \{-1,+1\}^n$:

$$\mathrm{GHD}_n(x,y) = \begin{cases} -1 & \text{if } \langle x,y\rangle \leqslant -\sqrt{n}, \\ +1 & \text{if } \langle x,y\rangle \geqslant \sqrt{n}, \end{cases} \tag{1.1}$$

where $\langle x,y\rangle = \sum x_i y_i$ is the usual inner product. Note that the function is left undefined when $\langle x,y\rangle$ is small in absolute value. The choice of $\sqrt{n}$ for the gap is natural for reasons of measure and represents the most difficult case from the standpoint of proof. All other gap values reduce to this canonical case.

The gap Hamming distance problem, or GHD for short, was proposed by Indyk and Woodruff [8] as a means to understand the complexity of several streaming tasks. More specifically, lower bounds on the communication complexity of gap Hamming distance imply lower bounds on the memory requirements of estimating the number of distinct elements in a data stream. Depending on the communication model used (one-way, constant-round, or unbounded-round), one obtains a lower bound for the corresponding class of streaming algorithms (one pass or multiple passes). Applications of GHD have been discovered to several other streaming tasks, including the computation of frequency moments [18] and entropy [6].

The communication complexity of gap Hamming distance has been the subject of much study over the past few years. In what follows, we will summarize the detailed chronological account from [7]. The original paper by Indyk and Woodruff [8] proved a linear lower bound on the *one-way* communication complexity of a problem closely related to GHD.

[*] Microsoft Research, Cambridge, MA 02142. Email: sherstov@alumni.cs.utexas.edu.

ISSN 1433-8092

A linear lower bound on the one-way communication complexity of GHD itself was obtained by Woodruff [18], with shorter and more elementary proofs discovered subsequently by Jayram, Kumar, and Sivakumar [10], Woodruff [19], and Brody and Chakrabarti [4]. A few years ago, Brody and Chakrabarti [4] obtained a linear lower bound for *constant-round* protocols solving GHD. The dependence on the number of rounds was improved by Brody et al. [5].

The communication complexity of GHD in the canonical, unbounded-round model, which is more compelling for streaming applications, proved to be a challenge to analyze. A lower bound of $\Omega(\sqrt{n})$ is immediate by a reduction to the *disjointness* problem [11, 14]. Coincidentally, GHD has a quantum communication protocol with cost $O(\sqrt{n}\log n)$, so that the many quantum methods discovered to date were of no use in moving beyond the $\sqrt{n}$ barrier. Finally, in a recent tour de force, Chakrabarti and Regev [7] settled the problem definitively with a linear lower bound on the unbounded-round communication complexity of gap Hamming distance:

THEOREM 1.1 (Chakrabarti and Regev). *Any randomized communication protocol that solves* GHD_n *with probability* $2/3$ *on every input has communication complexity* $\Omega(n)$.

The proof in [7] is quite involved. In follow-up work several months ago, Vidick [17] discovered a simpler proof. This paper contributes a new proof of Theorem 1.1, which is a page-and-a-half in length (Secs. 3–4) and simpler than the proofs of Chakrabarti and Regev [7] and Vidick [17]. In what follows, we give a detailed overview of previous work and our approach.

1.1. Some Terminology. Let $f\colon X \times Y \to \{-1,+1\}$ be a given communication problem. A common starting point in proving lower bounds on randomized communication complexity is Yao's minimax theorem [20]: to rule out a randomized protocol for f with cost c and error probability at most ϵ, one defines a probability distribution μ on $X \times Y$ and argues that with respect to μ, every *deterministic* protocol with cost c errs on more than an ϵ fraction of the inputs. This approach is *complete* in that one can always prove a tight lower bound on randomized communication in this manner.

The challenge in Yao's program is establishing the hardness of a given distribution μ for deterministic communication protocols. By far the most common solution since the 1980s is given by the *corruption method*, pioneered by Yao himself [20]. In more detail, a deterministic communication protocol with cost c gives a partition $X \times Y = \bigcup_{i=1}^{2^c} R_i$, where each set R_i is the Cartesian product of a subset of X and a subset of Y. The sets R_i are called *rectangles*, and the output of the deterministic protocol is constant on each rectangle. To prove a lower bound on communication, one defines a probability measure μ on $X \times Y$ and argues that every rectangle R with nontrivial measure is ϵ-*corrupted* by elements of $f^{-1}(+1)$, in the sense that

$$\mu(R \cap f^{-1}(+1)) > \epsilon\mu(R \cap f^{-1}(-1)) \tag{1.2}$$

for some constant $\epsilon > 0$. Provided that $f^{-1}(-1)$ has reasonable measure, (1.2) bounds from below the total number of rectangles in any partition of $X \times Y$. By symmetry, the roles of $+1$ and -1 can be interchanged throughout this argument. Furthermore, the argument applies unchanged to *partial* functions f, whose domain is a proper subset of $X \times Y$.

Over the years, many approaches have been used to prove (1.2). For product measures μ, a particularly general method was discovered by Babai, Frankl, and Simon [2] almost thirty years ago. It plays a key role in several subsequent papers and this work. In detail, let

μ be the uniform measure on $X \times Y$. For the sake of contradiction, suppose that $R = A \times B$ is a large rectangle that is not ϵ-corrupt, $A \subseteq X$, $B \subseteq Y$. We may assume that no row of R is 2ϵ-corrupt because any offending rows can discarded without affecting the size of R much. The proof is completed in two steps.

STEP 1: IDENTIFYING A HARD CORE.
Using the hypothesis that A is large, one identifies elements $x_1, x_2, \ldots, x_k \in A$ that are "very dissimilar" and collectively "representative" of X. Naturally, what those words mean depends on context but one gets the right idea by thinking about k random elements of X. Typically k is tiny, exponentially smaller than $|A|$. We will call $\{x_1, x_2, \ldots, x_k\}$ a *hard core* of A because at an intuitive level, these few elements capture the full complexity of A.

STEP 2: CORRUPTION.
Using the hypothesis that B is large and $x_1, x_2, \ldots, x_k$ are representative, one shows that the rectangle $\{x_1, x_2, \ldots, x_k\} \times B$ is 2ϵ-corrupt, a contradiction to the fact that no row of $R = A \times B$ is 2ϵ-corrupt.

This program is successful in practice because it is *much* easier to analyze the corruption of a rectangle $\{x_1, x_2, \ldots, x_k\} \times B$ for a small and highly structured collection elements $x_1, x_2, \ldots, x_k$. Babai, Frankl, and Simon [2] used this approach to establish, with an exceedingly elegant and short proof, an $\Omega(\sqrt{n})$ lower bound on the communication complexity of set disjointness. In that work, X and Y both referred to the family of subsets of $\{1, 2, \ldots, n\}$ of cardinality $\sqrt{n}$, and the hard core used in Step 1 was a collection of $k = \epsilon\sqrt{n}$ subsets that are mostly disjoint.

1.2. Previous Work. We are now in a position to outline the proofs of Theorem 1.1 due to Chakrabarti and Regev [7] and Vidick [17]. Both works study a continuous version of gap Hamming distance, in which the parties receive inputs $x, y \in \mathbb{R}^n$ drawn according to Gaussian measure and need to determine whether their inner product is less than $-\sqrt{n}$ or greater than $\sqrt{n}$. It was shown earlier [5] that the discrete and continuous versions of gap Hamming distance are essentially equivalent from the standpoint of communication complexity. The proof in [7] has two steps. Let $R = A \times B$ be rectangle of nonnegligible Gaussian measure. For reasons of measure, we may assume that the vectors in A, B have Euclidean norm $\sqrt{n}$, up to a multiplicative factor $1 \pm \epsilon$.

STEP 1: IDENTIFYING A HARD CORE.
Using the hypothesis that A has nontrivial measure, one identifies $\Omega(n)$ vectors $x_1, x_2, \ldots, x_i, \ldots \in A$ that are almost orthogonal. Precisely, the Euclidean norm of the projection of x_i onto $\operatorname{span}\{x_1, \ldots, x_{i-1}\}$ is a small constant fraction of the norm of x_i.

In retrospect, a system of near-orthogonal vectors is a natural choice for a hard core because GHD is defined in terms of inner products. *That* such a system of vectors can always be chosen from A was proven by Raz [13], who used this fact to obtain a lower bound for another linear-algebraic communication problem (deciding subspace membership). In light of the program of Babai, Frankl, and Simon [2], it is tempting to proceed to Step 2 and argue that the rectangle $\{x_1, x_2, \ldots, x_i, \ldots, \} \times B$ is heavily corrupted. Unfortunately, gap Hamming distance *does* have rectangles that are large and almost uncorrupted, and one cannot apply the corruption method directly. Instead, Chakrabarti and Regev [7] prove the following.

STEP $2'$: ANTICONCENTRATION.
With probability $\Omega(1)$, a random pair $(x, y) \in \{x_1, x_2, \dots, x_i, \dots\} \times B$ has $|\langle x, y\rangle| = \Omega(\sqrt{n})$.

The two steps above immediately give the following statement: for any sets $A, B \subseteq \mathbb{R}^n$ of nonnegligible measure, random vectors $x \in A$, $y \in B$ obey $|\langle x, y\rangle| = \Omega(\sqrt{n})$ with constant probability. This *anticoncentration* result is the technical centerpiece of Chakrabarti and Regev's proof. The authors actually derive a much stronger statement, giving a detailed characterization of the distribution of $\langle x, y\rangle$. To complete the proof, they use a criterion for high communication complexity due to Jain and Klauck [9], known as the *smooth rectangle bound*. Specifically, Chakrabarti and Regev use their anticoncentration result to argue that in any partition of $\mathbb{R}^n \times \mathbb{R}^n$, only a small constant measure of inputs can be covered by large uncorrupted rectangles. Settling this claim requires the introduction of a second measure, call it λ, to account for covering by large rectangles. The smooth rectangle bound [9] was discovered very recently and overcomes limitations of Yao's corruption bound—at the expense of being more challenging to use.

In follow-up work, Vidick [17] discovered a simpler proof of the anticoncentration property for $\langle x, y\rangle$, by taking a *matrix-analytic* view of the problem as opposed to the purely measure- and information-theoretic treatment in [7]. Vidick first shows that for any $A \subseteq \mathbb{R}^n$ of nonnegligible Gaussian measure, the matrix $M = \mathbf{E}_{x\in A}[xx^{\mathsf{T}}]$ has a relatively spread out spectrum, with a constant fraction of singular values on the order of $\Omega(1)$. Since $\mathbf{E}_{x\in A, y\in B}[\langle x, y\rangle^2] = \mathbf{E}_{y\in B}[y^{\mathsf{T}} M y]$, the author of [17] is able to use this spectral property of M to prove anticoncentration for $\langle x, y\rangle$. Vidick's proof ingeniously exploits the rotation-invariance of Gaussian measure and requires just the Bernstein inequality and the Berry-Esseen theorem for independent Gaussian variables. With anticoncentration established, Vidick uses the Jain-Klauck criterion to prove the lower bound on communication complexity.

1.3. Our Proof.

This paper contributes a new proof of Theorem 1.1, which is a page-and-a-half in length (Secs. 3–4) and simpler than the proofs of Chakrabarti and Regev [7] and Vidick [17]. Our approach departs from previous work on two counts. First, we use Yao's original corruption method for proving communication lower bounds, rather than the recent and more involved criterion of Jain and Klauck. Second, the authors of [7, 17] work with an extension of the problem to Gaussian space, whereas we are able to give a direct argument for the hypercube. As we show, the discrete setting allows for a treatment that is much simpler both in formalism and in substance; contrast the proofs of Lemma 4.4 in [13] and Lemma 3.1 in this paper to get an idea.

Our main technical tool is *Talagrand's concentration inequality* [15, 1, 16]. It states that for any given subset $S \subset \{-1, +1\}^n$ of constant measure, nearly all the points of the hypercube lie at a short Euclidean distance from the convex hull of S. Talagrand's concentration inequality has yielded results whose range and depth are out of proportion to the inequality's easy proof [1, 16]. We use the following well-known consequence of Talagrand's inequality: the projection of a random vector $x \in \{-1, +1\}^n$ onto a given linear subspace $V \subseteq \mathbb{R}^n$ has Euclidean norm $\sqrt{\dim V} \pm O(1)$ almost surely.

We now give a more detailed description of the proof. What we actually obtain is an $\Omega(n)$ lower bound on the communication complexity of *gap orthogonality*, a problem in which the two parties receive vectors $x, y \in \{-1, +1\}^n$ and need to reliably tell whether they are nearly orthogonal or far from orthogonal. Formally, gap orthogonality is the partial

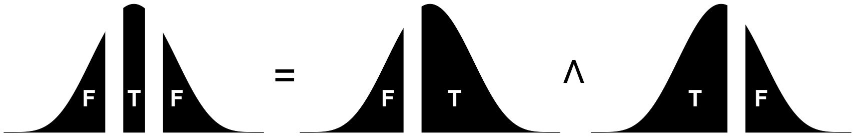

Figure 1: Reduction from gap orthogonality to gap Hamming distance (T = "true," F = "false").

Boolean function on $\{-1,+1\}^n \times \{-1,+1\}^n$ given by

$$\mathrm{ORT}_n(x,y) = \begin{cases} -1 & \text{if } |\langle x,y\rangle| \leqslant \sqrt{n}, \\ +1 & \text{if } |\langle x,y\rangle| \geqslant 2\sqrt{n}. \end{cases} \tag{1.3}$$

Gap orthogonality readily reduces to gap Hamming distance, as suggested pictorially in Figure 1. Hence, it suffices to prove an $\Omega(n)$ lower bound for gap orthogonality.

It seems at first that nothing of substance is gained by switching from gap Hamming distance to gap orthogonality. In actuality, the latter is preferable in that it allows the use of Yao's corruption method. Indeed, the corruption property for ORT_n is *equivalent* to the anticoncentration of $|\langle x,y\rangle|$. Thus, we just need to establish the anticoncentration: for some absolute constant $\epsilon > 0$ and any sets $A, B \subseteq \{-1,+1\}^n$ of uniform measure at least $2^{-\epsilon n}$, the inner product $\langle x,y\rangle$ for random $x \in A$, $y \in B$ cannot be too concentrated around zero. We give a short proof of this result, which combines selected ideas of [7] and [17] with some new elements.

STEP 1: IDENTIFYING A HARD CORE.
One can select a family of $\Omega(n)$ near-orthogonal vectors $x_1, x_2, \dots, x_i, \dots \in A$. Formally, the projection of x_i onto $\mathrm{span}\{x_1, x_2, \dots, x_{i-1}\}$ has Euclidean norm no greater than a *third* of the norm of x_i.

STEP 2: ANTICONCENTRATION & CORRUPTION.
Fix the vectors so constructed. Then with probability exponentially close to 1, a random $y \in \{-1,+1\}^n$ will have nonnegligible inner product (absolute value at least $\sqrt{n}/4$) with one or more of the vectors x_i.

Step 1 is a trivial consequence of Talagrand's concentration inequality; earlier works by Raz [13] and Chakrabarti and Regev [7] used an analogue of this claim for the sphere $\mathbb{S}^{n-1}$ with Haar measure, whose proof was more involved. To prove Step 2, we switch to the matrix-analytic view of Vidick [17] but give a simpler and more direct argument. Specifically, we consider the matrix M with rows $x_1, x_2, \dots, x_i, \dots$, which by construction is close in norm to an *orthogonal* matrix. It follows that a constant fraction of M's singular values are large, on the order of $\Omega(\sqrt{n})$. Applying Talagrand a second time, we get that a random vector $y \in \{-1,+1\}^n$ will have a constant fraction of its Euclidean norm in the linear subspace corresponding to the large singular values of M, except with probability exponentially small. This completes Step 2. The sought anticoncentration property falls out as a corollary, for purely combinatorial reasons. This proves corruption for gap orthogonality.

2. PRELIMINARIES

Notation. The symbol $[k]$ stands for the set $\{1, 2, \dots, k\}$. The inner product of $x, y \in \mathbb{R}^n$ is denoted $\langle x,y\rangle = \sum x_i y_i$. Likewise, $\langle A,B\rangle = \sum A_{ij} B_{ij}$ for matrices $A = [A_{ij}]$ and $B = [B_{ij}]$. The Boolean values "true" and "false" are represented in this paper by -1 and $+1$, respectively. In particular, Boolean functions take on values ± 1. A *partial* function on

X is a function whose domain of definition, dom f, is a proper subset of X. For a Boolean string x, the symbol x^k stands for the concatenation $xx\ldots x$ (k times).

Linear algebra. The Frobenius norm of a real matrix $M = [M_{ij}]$ is given by $\|M\|_{\mathrm{F}} = (\sum M_{ij}^2)^{1/2}$. We denote the singular values of M by $\sigma_1(M) \geqslant \sigma_2(M) \geqslant \cdots \geqslant 0$. Very precise estimates are known of the rth singular value, including the *Hoffman-Wielandt inequality*. For us, the following very crude bound is all that is needed.

FACT 2.1. *For all real matrices $M, \tilde{M}$ and all r,*

$$\sigma_{r+1}(M) \geqslant \frac{1}{\operatorname{rk} M - r}\left(\frac{\langle M, \tilde{M}\rangle}{\sigma_1(\tilde{M})} - \|M\|_{\mathrm{F}}\sqrt{r}\right).$$

Proof. Abbreviate $\sigma_i = \sigma_i(M)$. The r largest singular values of M sum to $\sigma_1 + \cdots + \sigma_r \leqslant (\sigma_1^2 + \cdots + \sigma_r^2)^{1/2}\sqrt{r} \leqslant \|M\|_{\mathrm{F}}\sqrt{r}$, and the remaining ones sum to at most $(\operatorname{rk} M - r)\sigma_{r+1}$. At the same time, $\sum \sigma_i \geqslant \langle M, \tilde{M}\rangle/\sigma_1(\tilde{M})$ by the singular value decomposition. □

The Euclidean norm of a vector is denoted $\|x\| = (\sum x_i^2)^{1/2}$. The dimension of a linear subspace V is denoted dim V. For a linear subspace $V \subseteq \mathbb{R}^n$ and a vector x, we let $\operatorname{proj}_V x$ denote the projection of x onto V. The following fact is immediate from Talagrand's concentration inequality [1, Thm. 7.6.1].

FACT 2.2 (Talagrand). *Let $V \subseteq \mathbb{R}^n$ be a linear subspace. Let $c > 1$ be a suitably large absolute constant. Then for a random vector $x \in \{-1, +1\}^n$ and all $t > 0$,*

$$\mathbf{P}[|\|\operatorname{proj}_V x\| - \sqrt{\dim V}| > t + c] < 4\exp\left(-\frac{t^2}{c}\right).$$

The short derivation of Fact 2.2 can be found in [16] or in Appendix A of this paper.

Communication complexity. Fix finite sets X, Y and let f be a (possibly partial) Boolean function on $X \times Y$. A randomized communication protocol is said to *compute f with error* ϵ if for all $(x, y) \in$ dom f, the output of the protocol on (x, y) is $f(x, y)$ with probability at least $1 - \epsilon$. The least communication cost of such a protocol is known as the *ϵ-error communication complexity of f*, denoted $R_\epsilon(f)$. For all constants $\epsilon \in (0, 1/2)$, one has $R_\epsilon(f) = \Theta(R_{1/3}(f))$.

A *rectangle* of $X \times Y$ is any set of the form $A \times B$, where $A \subseteq X$, $B \subseteq Y$. One of the earliest and best known criteria for high randomized communication complexity is Yao's *corruption bound* [20, 2, 12].

THEOREM 2.3 (Corruption bound). *Let f be a* (*possibly partial*) *Boolean function on $X \times Y$. Given $\epsilon, \delta > 0$, suppose that there is a distribution μ on $X \times Y$ such that*

$$\mu(R \cap f^{-1}(+1)) > \epsilon\mu(R \cap f^{-1}(-1))$$

for every rectangle $R \subseteq X \times Y$ with $\mu(R) > \delta$. Then

$$2^{R_\xi(f)} \geqslant \frac{1}{\delta}\left(\mu(f^{-1}(-1)) - \frac{\xi}{\epsilon}\right).$$

We gave an informal proof of Theorem 2.3 in the Introduction. For a rigorous treatment, see, e.g., [3, Lem. 3.5].

3. Corruption of Gap Orthogonality

We start by showing that any subset of $\{-1,+1\}^n$ of nontrivial size contains $n/10$ near-orthogonal vectors. See Raz [13, Lem. 4.4] for a similar result for the sphere $\mathbb{S}^{n-1}$.

Lemma 3.1. *Let $\alpha > 0$ be a sufficiently small constant. Fix $A \subseteq \{-1,+1\}^n$ with $|A| > 2^{(1-\alpha)n}$. Then for $k = \lfloor n/10 \rfloor$ there exist $x_1, x_2, \ldots, x_k \in A$ such that for each i,*

$$\| \operatorname{proj}_{\operatorname{span}\{x_1,x_2,\ldots,x_i\}} x_{i+1} \| \leqslant \frac{\sqrt{n}}{3}. \tag{3.1}$$

Proof. The proof is by induction. Having selected $x_1, x_2, \ldots, x_i \in A$, pick $x_{i+1} \in \{-1,+1\}^n$ uniformly at random. Then $\mathbf{P}[x_{i+1} \in A] > 2^{-\alpha n}$. On the other hand, Fact 2.2 implies that the projection of x_{i+1} onto $\operatorname{span}\{x_1, x_2, \ldots, x_i\}$ has Euclidean norm at most $\sqrt{n}/3$, except with probability $2^{-\alpha n}$. Thus, there exists $x_{i+1} \in A$ that obeys (3.1). □

The next lemma shows that given any family of near-orthogonal vectors in $\{-1,+1\}^n$, a random vector in $\{-1,+1\}^n$ will almost surely have a substantial inner product with some vector from the family. The proof uses the *lower* bound in Talagrand's theorem.

Lemma 3.2. *Fix vectors $x_1, x_2, \ldots, x_m \in \{-1,+1\}^n$ that obey (3.1) for all i. Then*

$$\mathop{\mathbf{P}}_{y\in\{-1,+1\}^n}\left[\max_{i=1,\ldots,m} |\langle y, x_i\rangle| \leqslant \frac{\sqrt{n}}{4}\right] \leqslant \mathrm{e}^{-\Omega(m)}. \tag{3.2}$$

Proof. Let $\tilde{x}_1, \ldots, \tilde{x}_m$ be orthogonal vectors obtained from $x_1, \ldots, x_m$ by the Gram-Schmidt process, i.e., $\tilde{x}_i = x_i - \operatorname{proj}_{\operatorname{span}\{x_1,\ldots,x_{i-1}\}} x_i$. Let M and $\tilde{M}$ be the $m \times n$ matrices with rows $x_1, \ldots, x_m$ and $\tilde{x}_1, \ldots, \tilde{x}_m$, respectively. Then $\langle M, \tilde{M}\rangle \geqslant \frac{8}{9}nm$ by (3.1). Also $\sigma_1(\tilde{M}) \leqslant \sqrt{n}$ since $\tilde{M}$ has orthogonal rows. Thus, Fact 2.1 gives $\sigma_{\lceil m/4 \rceil}(M) \geqslant 0.51\sqrt{n}$.

Let $M = \sum_{i=1}^m \sigma_i(M) u_i v_i^{\mathsf{T}}$ be the singular value decomposition of M. Define $V = \{v_i : \sigma_i(M) \geqslant 0.51\sqrt{n}\}$, so that $|V| \geqslant m/4$ by the previous paragraph. For all vectors y,

$$\|My\|^2 = \sum_{i=1}^m \sigma_i(M)^2 \langle y, v_i\rangle^2 \geqslant 0.26n \sum_{v\in V} \langle y, v\rangle^2 = 0.26n \| \operatorname{proj}_{\operatorname{span} V} y\|^2.$$

Since $\operatorname{span} V$ has dimension at least $m/4$, Fact 2.2 guarantees that $\|My\|^2 > mn/16$ with probability $1 - \exp(-\Omega(m))$ for random $y \in \{-1,+1\}^n$. This implies (3.2). □

The main result of this section is immediate from the previous two lemmas for basic combinatorial reasons; cf. the well-known proof of an $\Omega(\sqrt{n})$ lower bound on the communication complexity of disjointness due to Babai, Frankl, and Simon [2].

Theorem 3.3. *Let $\epsilon > 0$ be a small enough constant. Let $A, B \subseteq \{-1,+1\}^n$ be given such that $\mathbf{P}_{x\in A, y\in B}[x \star y] \leqslant \epsilon$, where $x \star y$ is shorthand for $|\langle x, y\rangle| > \sqrt{n}/4$. Then*

$$4^{-n}|A|\,|B| = \exp(-\Omega(n)).$$

Proof. Assume that $|A| > 2 \cdot 2^{(1-\alpha)n}$, where $\alpha > 0$ is the constant from Lemma 3.1. We will show that a random $y \in \{-1,+1\}^n$ occurs in B with probability at most $\exp(-\Omega(n))$.

The argument is a combinatorial accounting for what kinds of elements arise in B and is closely analogous to Theorem 8.3 in [2]. Define $A' = \{x \in A : \mathbf{P}_{y\in B}[x \star y] \leqslant 2\epsilon\}$, so that $|A'| \geqslant \frac{1}{2}|A|$. Fix $x_1, x_2, \ldots, x_k \in A'$ to obey (3.1) for all i, where $k = \lfloor n/10 \rfloor$. Define $B' = \{y \in B : \mathbf{P}_{i\in[k]}[x_i \star y] \leqslant 3\epsilon\}$, so that $|B'| \geqslant \frac{1}{3}|B|$. Then $2^{-n}|B'|$ is a lower bound on the probability that a random $y \in \{-1,+1\}^n$ has $|\langle y, x_i\rangle| \leqslant \sqrt{n}/4$ for at least $(1-3\epsilon)k$ indices i. By Lemma 3.2 and the union bound, this probability cannot exceed $\binom{k}{3\epsilon k}\mathrm{e}^{-\Omega(k)} \leqslant \mathrm{e}^{-\Omega(n)}$. ∎

4. Main Result

In this final section, we prove the sought $\Omega(n)$ bound on the communication complexity of gap Hamming distance and gap orthogonality, defined in (1.1) and (1.3).

Main Theorem. $R_{1/3}(\mathrm{ORT}_n) = \Omega(n)$.

Proof. Consider the partial Boolean function f_n on $\{-1,+1\}^n \times \{-1,+1\}^n$ defined as -1 when $|\langle x,y\rangle| \leqslant \sqrt{n}/8$ and $+1$ when $|\langle x,y\rangle| \geqslant \sqrt{n}/4$. With μ the uniform distribution on $\{-1,+1\}^n \times \{-1,+1\}^n$, Theorem 3.3 guarantees that $\mu(R \cap f_n^{-1}(+1)) > \epsilon\mu(R)$ for all rectangles R with $\mu(R) > 2^{-\epsilon n}$, where $\epsilon > 0$ is a small constant. Since $\mu({f_n}^{-1}(-1)) = \Theta(1)$, the corruption bound (Theorem 2.3) shows that $R_{1/3}(f_n) = \Omega(n)$. Finally, $f_n(x,y) = \mathrm{ORT}_{64n}(x^{64}, y^{64})$. ∎

Corollary. $R_{1/3}(\mathrm{GHD}_n) = \Omega(n)$.

Proof. Immediate from the reduction in Figure 1. Formally, for n a square,

$$\begin{aligned}\mathrm{ORT}_n(x,y) = \mathrm{GHD}_{10n+15\sqrt{n}}&\left(x^{10}(-1)^{15\sqrt{n}}, y^{10}(+1)^{15\sqrt{n}}\right)\\ &\wedge \neg\mathrm{GHD}_{10n+15\sqrt{n}}\left(x^{10}(+1)^{15\sqrt{n}}, y^{10}(+1)^{15\sqrt{n}}\right).\end{aligned}$$ ∎

Acknowledgments

The author is thankful to Amit Chakrabarti, Dmitry Gavinsky, Oded Regev, Thomas Vidick, Ronald de Wolf, and David Woodruff for their valuable feedback on an earlier version of this manuscript.

References

[1] N. Alon and J. Spencer. *The Probabilistic Method*. John Wiley & Sons, 3rd edition, 2008.

[2] L. Babai, P. Frankl, and J. Simon. Complexity classes in communication complexity theory. In *Proc. of the 27th Symposium on Foundations of Computer Science (FOCS)*, pages 337–347, 1986.

[3] P. Beame, T. Pitassi, N. Segerlind, and A. Wigderson. A strong direct product theorem for corruption and the multiparty communication complexity of disjointness. *Computational Complexity*, 15(4):391–432, 2006.

[4] J. Brody and A. Chakrabarti. A multi-round communication lower bound for gap Hamming and some consequences. In *Proc. of the 24th Conf. on Computational Complexity (CCC)*, pages 358–368, 2009.

[5] J. Brody, A. Chakrabarti, O. Regev, T. Vidick, and R. de Wolf. Better gap-Hamming lower bounds via better round elimination. In *Proc. of the 14th Intl. Workshop on Randomization and Computation (RANDOM)*, pages 476–489, 2010.

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

[7] A. Chakrabarti and O. Regev. An optimal lower bound on the communication complexity of gap-Hamming-distance. In *Proc. of the 43rd ACM Symposium on Theory of Computing (STOC)*, 2011. To appear.

[8] P. Indyk and D. P. Woodruff. Tight lower bounds for the distinct elements problem. In *Proc. of the 44th Symposium on Foundations of Computer Science (FOCS)*, pages 283–289, 2003.

[9] R. Jain and H. Klauck. The partition bound for classical communication complexity and query complexity. In *Proc. of the 25th Conf. on Computational Complexity (CCC)*, pages 247–258, 2010.

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

[11] B. Kalyanasundaram and G. Schnitger. The probabilistic communication complexity of set intersection. *SIAM J. Discrete Math.*, 5(4):545–557, 1992.

[12] E. Kushilevitz and N. Nisan. *Communication complexity*. Cambridge University Press, New York, 1997.

[13] R. Raz. Exponential separation of quantum and classical communication complexity. In *Proc. of the 31st Symposium on Theory of Computing (STOC)*, pages 358–367, 1999.

[14] A. A. Razborov. On the distributional complexity of disjointness. *Theor. Comput. Sci.*, 106(2):385–390, 1992.

[15] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de L'IHÉS*, 81(1):73–205, 1996.

[16] T. Tao. Talagrand's concentration inequality. Weblog entry, 2009. http://terrytao.wordpress.com/2009/06/09/talagrands-concentration-inequality/.

[17] T. Vidick. 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, 2010. Manuscript. Revised version in *Electronic Colloquium on Computational Complexity* (ECCC), Report TR11-051, 2011.

[18] D. P. Woodruff. Optimal space lower bounds for all frequency moments. In *Proc. of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 167–175, 2004.

[19] D. P. Woodruff. The average-case complexity of counting distinct elements. In *Proc. of the 12th International Conference, on Database Theory (ICDT)*, pages 284–295, 2009.

[20] A. C.-C. Yao. Lower bounds by probabilistic arguments. In *Proc. of the 24th Symposium on Foundations of Computer Science (FOCS)*, pages 420–428, 1983.

Appendix A. More on Talagrand's Concentration Inequality

For a subset $S \subseteq \mathbb{R}^n$ and a point $x \in \mathbb{R}^n$, let $\rho(x,S) = \inf_{y \in \operatorname{conv} S} \|x - y\|$ be the Euclidean distance from x to the convex hull of S. Talagrand's inequality [15, 1] states that for any reasonably large subset S of the hypercube, almost all the points of the hypercube lie at a short distance from the convex hull of S. In more detail:

Theorem A.1 (Talagrand). *For a fixed set $S \subseteq \{-1,+1\}^n$ and a random $x \in \{-1,+1\}^n$,*

$$\mathbf{P}[x \in S]\,\mathbf{P}[\rho(x,S) > t] \leqslant \mathrm{e}^{-t^2/16}.$$

In the terminology of Euclidean distances, Fact 2.2 states that for every linear subspace $V \subseteq \mathbb{R}^n$ and a random $x \in \{-1,+1\}^n$,

$$\mathbf{P}\left[\left|\rho(x,V) - \sqrt{n - \dim V}\right| > t + O(1)\right] \leqslant 4\mathrm{e}^{-\Omega(t^2)}. \tag{A.1}$$

To explain why this is a consequence of Talagrand's inequality, we will closely follow the treatment in a recent expository article by Tao [16]. Fix $a \geqslant 0$ and consider the set $S = \{x \in \{-1,+1\}^n : \rho(x,V) \leqslant a\}$. Then $\mathbf{P}[\rho(x,V) \leqslant a]\,\mathbf{P}[\rho(x,S) > t] \leqslant \exp(-t^2/16)$ by Talagrand. But by the triangle inequality, $\rho(x,V) > a + t$ implies $\rho(x,S) > t$, whence

$$\mathbf{P}[\rho(x,V) \leqslant a]\,\mathbf{P}[\rho(x,V) > a + t] \leqslant \mathrm{e}^{-t^2/16} \tag{A.2}$$

for any a. Let m be the median value of $\rho(x,V)$. The two tail bounds

$$\mathbf{P}[\rho(x,V) > m + t] \leqslant 2\mathrm{e}^{-t^2/16}, \tag{A.3}$$

$$\mathbf{P}[\rho(x,V) \leqslant m - t] \leqslant 2\mathrm{e}^{-t^2/16} \tag{A.4}$$

result from letting $a = m$ and $a = m - t$, respectively, in (A.2). Consequently, $\rho(x,V)$ is sharply concentrated around its median. The sharp concentration means among other things that the median is within an additive constant of $\mathbf{E}[\rho(x,V)^2]^{1/2} = \sqrt{n - \dim V}$. Along with (A.3) and (A.4), this settles (A.1).

ECCC ISSN 1433-8092

http://eccc.hpi-web.de