Electronic Colloquium on Computational Complexity, Report No. 64 (2024)

No Complete Problem for Constant-Cost Randomized Communication

Yuting Fang
Ohio State University, USA
fang.564@osu.edu

Lianna Hambardzumyan*
Hebrew University of Jerusalem, Israel
lianna.hambardzumyan@mail.huji.ac.il

Nathaniel Harms†
EPFL, Switzerland
nathaniel.harms@epfl.ch

Pooya Hatami‡
Ohio State University, USA
hatami.2@osu.edu

Abstract

We prove that the class of communication problems with public-coin randomized constant-cost protocols, called BPP^0, does not contain a complete problem. In other words, there is no randomized constant-cost problem $Q \in \mathsf{BPP}^0$, such that all other problems $P \in \mathsf{BPP}^0$ can be computed by a constant-cost *deterministic* protocol with access to an oracle for Q. We also show that the k-Hamming Distance problems form an infinite hierarchy within BPP^0. Previously, it was known only that Equality is not complete for BPP^0. We introduce a new technique, using Ramsey theory, that can prove lower bounds against arbitrary oracles in BPP^0, and more generally, we show that k-Hamming Distance matrices cannot be expressed as a Boolean combination of any constant number of matrices which forbid large Greater-Than subproblems.

*Research partially supported by ISF grants 921/22 and 2635/19.

†Supported by an NSERC postdoctoral fellowship and the Swiss State Secretariat for Education, Research, and Innovation (SERI) under contract number MB22.00026.

‡Supported by NSF grant CCF-1947546.

ISSN 1433-8092

1 Introduction

One of the main goals in communication complexity is to understand the power of randomized communication. The standard example is the EQUALITY problem, where two parties Alice and Bob are given strings $x, y \in \{0,1\}^n$, respectively, and must decide if $x = y$. Given a shared source of randomness, Alice and Bob can solve this problem with probability $3/4$ using only 2 bits of communication, regardless of input size, whereas a deterministic protocol requires n bits of communication. The EQUALITY problem is therefore one of the most extreme possible examples of the power of randomized communication, and to understand the power of randomness it is important to understand such extremes. For this purpose we define the class BPP^0 of communication problems that, like EQUALITY, have constant-cost randomized public-coin protocols (hereafter called merely *constant-cost protocols*, see Definition 1.6). The focused study of BPP^0 was initiated by [HHH22b, HWZ22], because:

- There are many connections to other areas, including operator theory and Fourier analysis [HHH22b], learning [LS09, FX15, HHP+22, HHM23, HZ24], graph sparsity [HWZ22, EHK22], and implicit graph representations [Har20, HWZ22, EHK22, EHZ23, HH22, NP24, HZ24].
- Communication complexity is often applied to find lower bounds for other problems, and constant vs. non-constant is the most basic lower bound question that one can ask, and yet it is often challenging – several surprisingly non-trivial and natural communication problems are in BPP^0 (e.g. computing small distances in planar graphs [Har20, HWZ22, EHK22], deciding incidence of certain low-dimensional point-halfspace arrangements [HWZ22, HZ24], etc.). There are many lower bound techniques in the literature, but they often do not help answer questions about constant-cost communication, so we must develop new techniques (as in this paper).
- Constant-cost communication is a more "fine-grained" approach to understanding randomized communication, which makes distinctions between different uses of randomness (e.g. public vs. private, or EQUALITY vs. GREATER-THAN) that are usually not differentiated, and allows for better understanding of "dimension-free" relations between matrix parameters [HHH22b].
- If we wish to identify the structure of problems which allow for efficient randomized communication, then we expect this structure to be most evident in the constant-cost problems. Some open problems about randomized communication remain open even when restricted to their BPP^0 versions, including the size of monochromatic rectangles [CLV19, HHH22b], the role of one- vs. two-sided error and the existence of a complete problem. Answering these questions for BPP^0 is a first step towards the more general answers. Constant-cost communication may be restrictive enough that one might even hope to find a complete characterization of the problems in this class to answer these questions.

See also the recent survey [HH24] for more details. Relevant to all of these motivations is the idea that BPP^0 might contain a *complete* problem, i.e. one "truly randomized" constant-cost protocol P, such that all other constant-cost protocols can be rewritten as deterministic protocols using P as a subroutine (see Section 1.1.1 for formal definitions). Identifying a complete problem for BPP^0 would answer almost all questions about BPP^0 and provide a nearly complete understanding of the most extreme examples of the power of randomized communication. Earlier work [HWZ22, HHH22b] proved that EQUALITY is *not* complete for BPP^0. We prove that there is *no* complete problem:

Theorem 1.1. *There is no complete problem for* BPP^0.

We also prove the following hierarchy of k-HAMMING DISTANCE problems in BPP^0. The k-

Hamming Distance problem asks two players to decide whether the Hamming distance between $x, y \in \{0,1\}^n$ is k; for constant k, this is known to be in BPP^0 (see Section 1.1.2).

Theorem 1.2. *There are infinitely many constants k such that k-Hamming Distance cannot be reduced to $(k-1)$-Hamming Distance.*

In particular, a simple application of our argument shows that this is true for $k \in \{1, 2\}$, recovering the result of [HHH22b, HWZ22] that 1-Hamming Distance does not reduce to Equality (i.e. 0-Hamming Distance), while also separating 1- and 2-Hamming Distance, whereas previously it was not known whether 1-Hamming Distance is complete for BPP^0.

Theorem 1.3. 2-Hamming Distance *cannot be reduced to* 1-Hamming Distance.

This hierarchy echoes a similar hierarchy of Integer Inner Product functions within BPP, established in [CLV19], though our proof is necessarily very different. Those functions are denoted IIP_d for constant $d \in \mathbb{N}$ and are defined on dn-bit integer vectors $x, y \in \mathbb{Z}^d$ with $\mathsf{IIP}_d^n(x, y) = 0$ if and only if $\langle x, y \rangle = 0$ (Definition 3.10). They are in the communication complexity class BPP but are conjectured to have non-constant cost (see e.g. [CHHS23]). We use IIP_d as an example to show that, even if a problem might have non-constant cost, and may therefore be "more complex" than any k-Hamming Distance problem, we can still easily separate k-Hamming Distance from them using our technique:

Theorem 1.4. *For any constant d, there exists a constant k such that k-Hamming Distance cannot be reduced to IIP_d.*

To prove the theorems above, especially Theorem 1.1, it is necessary to prove lower bounds on communication protocols with access to an oracle computing an arbitrary problem in BPP^0. There are many techniques in the literature for oracle lower bounds in communication, including two lower bound techniques against the Equality oracle in BPP^0 [HHH22b, HWZ22, HZ24] and several techniques for lower bounds against Equality in other communication complexity classes (e.g. [CLV19, PSW21, CHHS23, PSS23]). However, none of these techniques have succeeded in proving separations within BPP^0 against oracles other than Equality – not even against the 1-Hamming Distance oracle. An added challenge for proving lower bounds against arbitrary oracles in BPP^0 is that little is known about the structure of problems in BPP^0, including the basic question of whether they have large monochromatic rectangles [HHH22b]. We introduce a new Ramsey-theoretic lower bound technique that gives separations against *arbitrary* oracles in BPP^0. We give a proof overview and comparison to prior work in Section 1.2, after introducing definitions in Section 1.1. The proofs of the theorems above are in Section 3, and follow from a general lemma proved in Section 2 about the structure of oracle protocols for computing k-Hamming Distance.

Our final result, proved in Section 4, is of a different type and deals with BPP reductions. An *unbounded-size* BPP reduction, for a problem with n-bit inputs, allows $\operatorname{poly}\log n$-many oracle queries of arbitrary query size. It was proved in [CLV19] (see also [CHHS23]) that Equality is not complete for BPP, because IIP_d^n (for $d \geq 3$) is irreducible to Equality, and more generally the IIP_d problems form an infinite hierarchy under unbounded-size BPP reductions. This leaves open the question of how the IIP_d hierarchy interacts with k-Hamming Distance; in particular, whether IIP_d^n reduces to k-Hamming Distance for $k = \operatorname{poly}\log n$ under unbounded-size reductions. We prove this is not so, by showing certain "dimension-free" relations: the Equality oracle complexity and γ_2-norm of arbitrary $N \times N$ submatrices of k-Hamming Distance depends on N but *not* on the underlying dimension. This implies:

Theorem 1.5 (Informal). *For every $d \geq 3$, IIP_d^n does not reduce to k-Hamming Distance, under unbounded-size BPP reductions, for any $k = k(n) \leq n/(\log n)^{\omega(1)}$.*

A careful examination of [CLV19] gives the incomparable statement that, for every constant k, there exists an unspecified constant $d = d(k) \geq 6$ such that IIP_d^n requires $\Theta(n)$ k-Hamming Distance oracle queries. Our lower bound is $\Omega(n/k \log n)$ queries, but it applies to $d = 3$ and non-constant k.

We conclude the paper with a discussion and open problems in Section 5.

1.1 Preliminaries: Reductions, k-Hamming Distance, and Greater-Than

Let us now formalize the notions of reductions and completeness within BPP^0, and state some required facts about the k-Hamming Distance and Greater-Than problems.

1.1.1 Constant-Cost Reductions

A *communication problem* $\mathcal{P}$ is a sequence $\mathcal{P} = (P_N)_{N \in \mathbb{N}}$ where $P_N \in \{0,1\}^{N \times N}$ is an $N \times N$ Boolean matrix. We will use N to denote the size of the matrix and, when it is natural, we may define another parameter (e.g. sometimes writing n for the number of bits in the input). For any Boolean matrix M, we write $\mathsf{R}(M)$ for the minimum cost of a two-way public-coin randomized communication protocol computing M with success probability at least $2/3$, and we refer to standard texts [KN96, RY20] for an introduction to randomized communication. For a communication problem $\mathcal{P}$, we write $\mathsf{R}(\mathcal{P})$ for the function $N \mapsto \mathsf{R}(P_N)$.

Definition 1.6 (BPP^0). A problem $\mathcal{P}$ has *constant cost* if there exists a constant c such that $\mathsf{R}(\mathcal{P}) \leq c$, i.e. for all $N \in \mathbb{N}$, it holds that $\mathsf{R}(P_N) \leq c$. We define BPP^0 as the set of all problems $\mathcal{P}$ which have constant cost.

Our notation follows in spirit the notation for complexity classes like AC^i and NC^i: For any i, we think of BPP^i as being the class of communication problems $\mathcal{P}$ with $\mathsf{R}(\mathcal{P}) = O(\log^i \log N)$, i.e. $O(\log^i n)$ where $n = \lceil \log N \rceil$ is the number of bits required to represent the inputs. Therefore the standard communication complexity class BPP is $\mathsf{BPP} = \bigcup_{i=0}^{\infty} \mathsf{BPP}^i$.

Remark 1.7. Unlike BPP, the class BPP^0 remains unchanged regardless of whether it is defined in terms of two-way, one-way, or simultaneous communication protocols [HWZ22], but it is *not* equivalent to replace public randomness with private randomness.

We must now define communication protocols with oracle queries. It is common to study communication with oracles (see e.g. [BFS86, GPW18, CLV19, PSW21, CHHS23, PSS23]), but for BPP^0 the most natural definition of oracle queries is different from the standard one. Usually, the size of the input to the oracle query is restricted – on n-bit inputs, the oracle should be queried only on $\mathrm{poly}(n)$-bit inputs, because making $\mathrm{poly} \log n$ queries to problems with complexity $\mathrm{poly} \log m$ on query inputs of $m = \mathrm{poly}(n)$ bits will produce a protocol of complexity $\mathrm{poly} \log n$, preserving the usual notion of "efficiency". For BPP^0, the oracles should allow queries of *arbitrary* size, for the same reason that this preserves our notion of efficiency. This leads to the following definition which is implicit in prior works [HWZ22, EHK22, HHH22b] and explicit in [HZ24]:

For any set $\mathcal{M}$ of Boolean matrices, we will define the *query set* $\mathsf{QS}(\mathcal{M})$ of $\mathcal{M}$ as the set of all matrices Q obtained from matrices in $\mathcal{M}$ by the following operations:

1. Arbitrary row and column permutations;
2. Taking arbitrary submatrices; and

3. Duplicating arbitrary rows or columns.

The difference between our reductions and the standard reductions for communication complexity comes from Item (2). Problems in BPP^0 are hereditary, in a way that problems in standard BPP are not: if $\mathcal{P}$ has a randomized protocol with constant cost c, then any matrix $P \in \mathsf{QS}(\mathcal{P})$ also has a randomized protocol with cost c (because P is obtained by choosing a problem $P' \in \mathcal{P}$, taking a submatrix P'' of P', which cannot increase the communication cost, and then duplicating rows and columns and permuting them, which does not change the communication cost). On the other hand, a problem $\mathcal{P}$ in the standard BPP class with complexity $\mathrm{poly}(\log\log N)$ can have matrices $P \in \mathsf{QS}(\mathcal{P})$ with cost $\Theta(\log N)$, as in the following example:

Example 1.8. The $\log(n)$-Hamming Distance problem, defined on binary strings $\{0,1\}^n$, has cost $\Theta(\log(n)\log\log(n)) = \Theta(\log\log(N)\log\log\log(N))$ on matrices of size $N \times N = 2^n \times 2^n$ (which follows from optimal bounds on the communication complexity of k-Hamming Distance for non-constant $k = \log n$ [HSZZ06, Sağ18]). But since the VC dimension of the k-Hamming Distance matrices depends on k (see Definition 3.3 and Proposition 3.5), the set QS contains *all* matrices.

Definition 1.9 (Communication with Oracle Queries)**.** Let $\mathcal{M}$ be any set of Boolean matrices and let $P \in \{0,1\}^{N\times N}$. Then we write $\mathsf{D}^{\mathcal{M}}(P)$ for the minimum cost of a deterministic communication protocol with access to an oracle for $\mathcal{M}$, defined as follows. A *communication protocol* with oracle access to $\mathcal{M}$ is a binary tree T with inner nodes V and leaves L. Each inner node $v \in V$ is associated with an $N \times N$ matrix $Q_v \in \mathsf{QS}(\mathcal{M})$ and each leaf $\ell \in L$ is associated with an output bit $b_\ell \in \{0,1\}$. An input $x, y \in [N]$ then naturally defines a path from the root to a leaf $\ell \in L$, wherein at each node the (x, y) entry of the corresponding query matrix is used to decide whether to travel left or right. The output of the protocol is then the label of the reached leaf, b_ℓ.

For simplicity of the definition, we force every round of communication to be via an oracle query – the players cannot send messages directly to each other. As observed by [CLV19], a standard round of communication can be simulated by an oracle query as long as the set $\mathcal{M}$ is non-trivial, i.e. it contains at least one of the matrices $\left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]$, $\left[\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right]$, or their permutations or Boolean negations.

The following proposition is easy to prove using standard error-boosting techniques:

Proposition 1.10. *Let $\mathcal{P}, \mathcal{Q}$ be communication problems such that $\mathcal{Q} \in \mathsf{BPP}^0$ and $\mathsf{D}^{\mathcal{Q}}(\mathcal{P}) = O(1)$. Then $\mathcal{P} \in \mathsf{BPP}^0$.*

Then the following is the most natural definition of reductions within BPP^0.

Definition 1.11 (Constant-Cost Reductions)**.** Let $\mathcal{P}, \mathcal{Q}$ be communication problems. Then we say that $\mathcal{P}$ has a *constant-cost reduction* to $\mathcal{Q}$ (or simply it *reduces to* $\mathcal{Q}$) if $\mathsf{D}^{\mathcal{Q}}(\mathcal{P}) = O(1)$.

We now have the natural notion of completeness for BPP^0:

Definition 1.12 (Completeness in BPP^0)**.** A communication problem $\mathcal{Q}$ is *complete* for BPP^0 if $\mathcal{Q} \in \mathsf{BPP}^0$ and, for all $\mathcal{P} \in \mathsf{BPP}^0$, $\mathcal{P}$ is constant-cost reducible to $\mathcal{Q}$.

We will use an equivalent and often more convenient definition of constant-cost reductions in terms of Boolean combinations of query matrices. We refer the reader to [HZ24] for a simple proof[1].

Proposition 1.13. *Let $\mathcal{M}$ be a set of Boolean matrices and let $\mathcal{P}$ be a communication problem. Then $\mathsf{D}^{\mathcal{M}}(\mathcal{P}) = O(1)$ if and only if there exists a constant $c \in \mathbb{N}$ and a function $f : \{0,1\}^c \to \{0,1\}$*

[1]The idea is that $f(Q_1(x,y), \ldots, Q_c(x,y))$ is the function that simulates the communication protocol using the answers to all of the queries.

such that, for all $N \in \mathbb{N}$, there exist $Q_1, \ldots, Q_c \in \mathsf{QS}(\mathcal{M})$ such that

$$\forall x, y \in [N]: \qquad P_N(x,y) = f(Q_1(x,y), Q_2(x,y), \ldots, Q_c(x,y)) \,. \tag{1}$$

Remark 1.14. The above proposition claims that we have a "uniform" function $f : \{0,1\}^c \to \{0,1\}$, that works for *every* problem size N. Swapping the quantifiers to allow a different function $f_N : \{0,1\}^c \to \{0,1\}$ for each N does not increase the power of the protocol, because there is only a constant number of functions $\{0,1\}^c \to \{0,1\}$, so the choice of f_N can be encoded[2] in f.

Constant-cost reductions are also natural and useful in the study of *implicit graph representations*; see [Har20, HWZ22, EHK22, NP24, HZ24] for more on this connection, and [Cha23] for reductions motivated directly from implicit graph representations.

1.1.2 k-Hamming Distance

Let $\mathsf{dist}(x,y)$ denote the Hamming distance between two strings. There are two variations of the k-Hamming Distance problem. The Exact k-Hamming Distance problem is defined as $\mathsf{EHD}_k = (\mathsf{EHD}_k^n)_{n\in\mathbb{N}}$ where

$$\forall x, y \in \{0,1\}^n: \qquad \mathsf{EHD}_k^n(x,y) = 1 \text{ if and only if } \mathsf{dist}(x,y) = k \,.$$

The Threshold k-Hamming Distance problem is defined as $\mathsf{THD}_k = (\mathsf{THD}_k^n)_{n\in\mathbb{N}}$ where

$$\forall x, y \in \{0,1\}^n: \qquad \mathsf{THD}_k^n(x,y) = 1 \text{ if and only if } \mathsf{dist}(x,y) \le k \,.$$

Observe that $\mathsf{THD}_0 \equiv \mathsf{EHD}_0$ is the Equality problem. For other constant values of $k \ge 1$, the problems are also equivalent under constant-cost reductions: it is easy to show that $\mathsf{D}^{\mathsf{THD}_k}(\mathsf{EHD}_k) \le 2$ (since $\mathsf{EHD}_k \equiv \mathsf{THD}_k \wedge \neg\mathsf{THD}_{k-1}$, and THD_{k-1} can be computed by one query to THD_k by padding the input), and $\mathsf{D}^{\mathsf{EHD}_k}(\mathsf{THD}_k) \le k$ (since $\mathsf{THD}_k \equiv \bigvee_{t=0}^{k} \mathsf{EHD}_t$ and EHD_t can be computed by one query to EHD_k by padding the input). The two-way public-coin randomized communication cost of these problems is $O(k \log k)$ [Yao03, HSZZ06] (with a matching lower bound when $k < \sqrt{n}$ [Sağ18]) so for every constant k, EHD_k and THD_k are in BPP^0.

1.1.3 Greater-Than and Stability

The Greater-Than problem is $\mathsf{GT} = (\mathsf{GT}_t)_{t\in\mathbb{N}}$, where the matrix GT_t is defined on $i, j \in [t]$ as

$$\mathsf{GT}_t(i,j) = 1 \text{ if and only if } i \le j \,.$$

Following the terminology of [HWZ22], we say a problem $\mathcal{P}$ is *stable* if the largest Greater-Than subproblem within $\mathcal{P}$ has constant size. Formally:

Definition 1.15 (Stability)**.** A set $\mathcal{M}$ of matrices is *stable* if there exists a constant t such that $\mathsf{GT}_t \notin \mathsf{QS}(\mathcal{M})$. Equivalently, $\mathcal{M}$ is stable if there exists a constant t such that, for any matrix $M \in \mathcal{M}$, and any set of rows $x_1, \ldots, x_m$ and columns $y_1, \ldots, y_m$ of M which satisfy $M(x_i, y_j) = 1$ iff $i \le j$, it holds that $m \le t$.

It is equivalent to require $\neg\mathsf{GT}_t \notin \mathcal{M}$ instead of $\mathsf{GT}_t \notin \mathcal{M}$, where $\neg\mathsf{GT}_t$ denotes Boolean negation, because $\neg\mathsf{GT}_t$ is a submatrix of GT_{t+1}. We will require the following observation, which follows from the known fact that $\mathsf{R}(\mathsf{GT}_t) = \Theta(\log\log t)$ [Nis93, Vio15, RS15, BW16] and therefore $\mathsf{GT} \notin \mathsf{BPP}^0$.

[2] The first r queries can be used to select a function f_N out of a space of 2^r functions

Observation 1.16. *Every problem $\mathcal{Q} \in \mathsf{BPP}^0$ is stable.*

Remark 1.17. Stability is a necessary but not sufficient condition for a problem $\mathcal{Q}$ to belong to BPP^0. For example, any problem $\mathcal{Q} \in \mathsf{BPP}^0$ must contain at most $2^{O(N \log N)}$ $N \times N$ matrices [HWZ22], but the family of all $K_{2,2}$-free matrices (i.e. matrices with no 2×2 rectangle of 1s), which is stable, contains more matrices than this [LZ15]. Even restricting to problems $\mathcal{Q}$ that are both stable and have at most $2^{O(N \log N)}$ $N \times N$ matrices is insufficient to guarantee membership in BPP^0 [HHH22a].

Remark 1.18. The size of the largest Greater-Than inside a matrix M is also called the Littlestone dimension, which characterizes the number of mistakes made by an online learning algorithm [Lit88, ALMM19]. Any stable set of matrices describes a hypothesis class that is learnable in a bounded number of mistakes, while BPP^0 is the family of hypothesis classes that are learnable in a bounded number of mistakes with the *perceptron* algorithm, due to the bounded-margin embedding of [LS09].

1.2 Proof Overview and Comparison to Prior Work

1.2.1 Prior Techniques

We wish to prove lower bounds for the EHD_k problem, against arbitrary oracles. By Proposition 1.13, if we assume EHD_k reduces to $\mathcal{Q}$, we can write, for all $x, y \in \{0,1\}^n$,

$$\mathsf{EHD}_k^n(x,y) = f(Q_1(x,y), Q_2(x,y), \ldots, Q_c(x,y)), \tag{2}$$

where $Q_1, \ldots, Q_c \in \mathsf{QS}(\mathcal{Q})$. A natural approach to show that EHD_k cannot be reduced to $\mathcal{Q}$ is to define a complexity measure κ which is bounded on $\mathcal{Q}$, such that, say, $\kappa(f(Q_1, \ldots, Q_c)) \leq g(\kappa(Q_1), \ldots, \kappa(Q_c))$, for some function g independent of n, while $\kappa(\mathsf{EHD}_k^n) = \omega(1)$. This is the approach taken in several prior works [HHH22b, CLV19, CHHS23, PSS23] to prove lower bounds against the Equality oracle. The γ_2 norm, used in [HHH22b, CHHS23], cannot separate EHD_k from EHD_1, for any constant $k > 1$. Another measure, *η-area*, was introduced in [CLV19] to show separations within BPP. It is unclear whether this could be used to show separations within BPP^0, which would require a technical analysis of the monochromatic rectangles within all submatrices of EHD_k, and indeed all problems in BPP^0, whereas the rectangle analyses in [CLV19] fail for EHD_1. It is not even known whether every problem in BPP^0 has large (linear-size) monochromatic rectangles [HHH22b], which would be the first step in proving Theorem 1.1 using the technique of [CLV19].

A more structural (but non-quantitative) approach was taken in [HWZ22, HZ24], which depended fundamentally on the fact that the Equality oracle partitions the inputs into very simple monochromatic rectangles. Every step of that proof fails when Equality is replaced with EHD_1. The challenge with a structural approach is that it is difficult to understand the structure of an arbitrary Boolean combination of matrices $Q_1, \ldots, Q_c$, even if the structure of $Q_1, \ldots, Q_c$ are themselves well-understood, and the structure of matrices belonging to problems in BPP^0 is *not* well-understood.

1.2.2 Proof Overview

We overcome these challenges in a way that is conceptually simpler than the previous bounds against only the Equality oracle, and, unlike the prior work, does not involve any argument about monochromatic rectangles. Observe that EHD_k is permutation-invariant in the following way. For every pair of inputs $x, y \in \{0,1\}^n$, we think of (x, y) as defining a sequence of "dominoes",

i.e. pairs $ab \in \{0,1\}^2$, where x_1y_1 is the first domino, x_2y_2 is the second, and so on:

$$\begin{matrix} x \\ y \end{matrix} = \boxed{\begin{matrix} x_1 \\ y_1 \end{matrix}}\boxed{\begin{matrix} x_2 \\ y_2 \end{matrix}}\cdots\boxed{\begin{matrix} x_n \\ y_n \end{matrix}}$$

Then the output of EHD_k is invariant under permutations on these dominoes.

Our next ingredient is the basic observation that every problem $\mathcal{Q}$ in BPP^0 has a fixed constant t such that no Greater-Than communication problem of size larger than $t \times t$ exists in $\mathsf{QS}(\mathcal{Q})$, i.e. problems $\mathcal{Q} \in \mathsf{BPP}^0$ are "stable" (Observation 1.16).

The function $\mathsf{EHD}_k(x,y) = f(Q_1(x,y),\ldots,Q_c(x,y))$ is, as a whole, invariant under "domino permutations" on the input, but *a priori* we have no similar guarantee on the queries Q_i. Our goal is to show that each query Q_i must *also* be permutation-invariant. We accomplish this (in Section 2.2) by thinking of the query responses as a coloring of a hypergraph whose vertices are the coordinates $[n]$ of the input. Using only the permutation invariance of EHD_k and stability of the queries, we apply the hypergraph Ramsey theorem in stages, in each stage increasing the number of permutations under which the *queries* Q_i are invariant, until we achieve our goal.

From here, we see that, if there are two classes A and B of inputs (x,y), where A and B are equivalence classes under domino permutations, and furthermore the output of EHD_k is different on inputs A than on inputs B, then there must be a query $Q \in \mathsf{QS}(\mathcal{Q})$ that distinguishes *all* inputs in A from *all* inputs in B. In this way, we transform the task of finding a lower bound for EHD_k against any constant number of arbitrary queries in $\mathsf{QS}(\mathcal{Q})$, into the task of finding a lower bound for a partial subproblem of EHD_k against a *single* query from $\mathsf{QS}(\mathcal{Q})$, which is done in Section 3.

Our main idea, forcing the algorithm to behave a certain way using Ramsey theory, was unexpectedly inspired by unrelated works in *property testing*, which use a more direct application of Ramsey theory to force testing algorithms to process random samples in a certain way [Fis04, DKN15].

2 Permutation Invariance of k-Hamming Distance Protocols

We now prove the main lemma, which shows that any constant-cost reduction from EHD_k to an arbitrary stable set of matrices $\mathcal{M}$ (Definition 1.15) can be forced to use oracle queries that are invariant under permutations on the input.

Definition 2.1 (Permutation Invariance)**.** For a matrix $Q : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$, we say that Q is *permutation-invariant* if for every $x, y \in \{0,1\}^n$ and every permutation $\pi : [n] \to [n]$,

$$Q(x,y) = Q(x_\pi, y_\pi)\,,$$

where $x_\pi := (x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)})$, and y_π is defined similarly.

Lemma 2.2. *Let $k \in \mathbb{N}$ and let $\mathcal{Q}$ be any stable set of Boolean matrices. Suppose $\mathsf{D}^{\mathcal{Q}}(\mathsf{EHD}_k) = O(1)$. Then there exists a constant c and a function $f : \{0,1\}^c \to \{0,1\}$, such that, for all $n \in \mathbb{N}$, there are c* permutation-invariant *queries $Q_1,\ldots,Q_c \in \mathsf{QS}(\mathcal{Q})$ satisfying*

$$\forall x, y \in \{0,1\}^n \quad \mathsf{EHD}_k^n(x,y) = f(Q_1(x,y),\ldots,Q_c(x,y)).$$

We state some notation and the Ramsey theorem in Section 2.1 and prove the lemma in Section 2.2.

2.1 Setup: Dominoes and the Hypergraph Ramsey Theorem

Definition 2.3. (Domino) We call a pair $ab \in \{0,1\}^2$ a *domino* and we denote it as $\boxed{\frac{a}{b}}$. For any $n \in \mathbb{N}$ and for any pair $(x,y) \in \{0,1\}^n \times \{0,1\}^n$, the *dominoes* of (x,y) is the sequence

$$\boxed{\frac{x}{y}} := \left(\boxed{\frac{x_1}{y_1}}, \boxed{\frac{x_2}{y_2}}, \ldots, \boxed{\frac{x_n}{y_n}}\right).$$

For a set of dominoes $\Delta \subseteq \{0,1\}^2$, we denote the complement of Δ by $\overline{\Delta} = \{0,1\}^2 \setminus \Delta$.

Definition 2.4 (Type)**.** Let $\Delta \subseteq \{0,1\}^2$ be a set of dominoes. A Δ-*type* is a tuple $\langle \Gamma_\Delta, \tau \rangle$ containing a Δ-*signature* $\Gamma_\Delta \in \Delta^*$, which is an ordered sequence of dominoes in Δ, and a *tally* $\tau = [\tau_{ab}]_{a,b\in\{0,1\}}$, which is a sequence with $\tau_{ab} \in \mathbb{Z}$. For a pair $(x,y) \in \{0,1\}^n \times \{0,1\}^n$ and a set of dominoes $\Delta \subseteq \{0,1\}^2$, the Δ-*type* of (x,y), denoted $\chi_\Delta(x,y)$, is the tuple $\langle \Gamma_\Delta(x,y), \tau(x,y) \rangle$, where

- $\Gamma_\Delta(x,y)$ is the Δ-signature of (x,y): the subsequence of dominoes of (x,y) that belong to Δ,

$$\Gamma_\Delta(x,y) = \left(\boxed{\frac{x_i}{y_i}} \mid \text{ for all } i \in [n],\ \boxed{\frac{x_i}{y_i}} \in \Delta\right);$$

- $\tau(x,y) = [\tau_{ab}(x,y)]_{a,b\in\{0,1\}}$ denotes the *tally* of the dominoes of (x,y), where $\tau_{ab}(x,y)$ is the number of times $\boxed{\frac{a}{b}}$ occurs in the dominoes of (x,y).

For example, with $\Delta = \left\{\boxed{\frac{0}{1}}, \boxed{\frac{1}{0}}\right\}$:

$$\chi_\Delta\left(\boxed{\frac{0110000}{0101001}}\right) = \left\langle \Gamma_\Delta = \left(\boxed{\frac{1}{0}}, \boxed{\frac{0}{1}}, \boxed{\frac{0}{1}}\right), [\tau_{00}=3, \tau_{01}=2, \tau_{10}=1, \tau_{11}=1] \right\rangle.$$

Definition 2.5 (Shuffle Invariance)**.** For any matrix $Q : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ and a set of dominoes $\Delta \subseteq \{0,1\}^2$, we say that Q is Δ-*shuffle invariant* if

$$\forall x,y,u,v \in \{0,1\}^n : \qquad \chi_{\overline{\Delta}}(x,y) = \chi_{\overline{\Delta}}(u,v) \implies Q(x,y) = Q(u,v).$$

In other words, Q is Δ-*shuffle invariant* if its value on (x,y) is invariant under any swap of consecutive dominoes in the sequence $\boxed{\frac{x_1}{y_1}}\boxed{\frac{x_2}{y_2}}\cdots\boxed{\frac{x_n}{y_n}}$ which involve a domino in Δ; any permutation of the dominoes achieved by a sequence of such swaps preserves the relative order of dominoes *outside* Δ, and therefore preserves the $\overline{\Delta}$-type. Q is permutation-invariant if it is Δ-shuffle invariant for the full set of dominoes $\Delta = \{0,1\}^2$.

We require the well-known hypergraph Ramsey theorem. For any set T and any $0 \le t \le |T|$, write $\binom{T}{t}$ for the family of subsets of T of cardinality t.

Theorem 2.6 (Hypergraph Ramsey theorem)**.** *For any $\alpha, \beta \in \mathbb{N}$ and $\sigma \ge \alpha$, there exists $R = R(\alpha,\beta,\sigma)$ such that for any coloring $\kappa : \binom{[R]}{\alpha} \to [\beta]$, there exists a subset $T \subseteq [R]$ of size σ such that κ is constant on $\binom{T}{\alpha}$.*

We use an easy corollary of this theorem and provide a proof for the sake of completeness. For any set T and any $0 \le t \le |T|$, write $\binom{T}{\le t}$ for the set of subsets of T of cardinality at most t.

Corollary 2.7. *For any $\alpha, \beta \in \mathbb{N}$ and $\sigma \ge \alpha$, there exists $N = N(\alpha,\beta,\sigma)$ such that for any coloring $\kappa : \binom{[N]}{\le \alpha} \to [\beta]$, there exists $T \subseteq [N]$ of size σ such that κ is constant on $\binom{T}{\alpha'}$ for every $\alpha' \le \alpha$.*

Proof. For any $\alpha', \beta', \sigma' \in \mathbb{N}$, write $R(\alpha', b', \sigma')$ for the number obtained from Theorem 2.6.

We prove the statement by induction on α. For $\alpha = 1$ the conclusion is easy to obtain. Now assume $\alpha > 1$ and write $M := N(\alpha - 1, \beta, \sigma)$ for the number obtained by induction for parameters $\alpha - 1, \beta, \sigma$. We define $N := N(\alpha, \beta, \sigma) := R(\alpha, \beta, N(\alpha - 1, \beta, \sigma))$.

Let $\mathsf{col} : \binom{[N]}{\leq \alpha} \to [\beta]$. Let $\mathsf{col}_\alpha : \binom{[N]}{\alpha} \to [\beta]$ be the function col restricted to domain $\binom{[N]}{\alpha}$. Then by Theorem 2.6, there exists a set $T \subseteq [N]$ of size M such that col_α is constant on domain $\binom{T}{\alpha}$. Relabel the elements of T as $[M]$ and define the function $\mathsf{col}' : \binom{[M]}{\leq \alpha - 1} \to [\beta]$ as the function col restricted to the domain $\binom{T}{\leq \alpha - 1}$ with the elements of T relabeled as $[M]$. T has cardinality $M = N(\alpha - 1, \beta, \sigma)$, so by induction col' is constant on $\binom{[M]}{\alpha'}$ for each $\alpha' \leq \alpha - 1$. Then col is constant on $\binom{T}{\alpha'}$ for $\alpha' \leq \alpha - 1$, and since col_α is constant on $\binom{T}{\alpha}$, this implies the conclusion. □

2.2 Proof of Lemma 2.2

We now prove the permutation invariance lemma. Let $\mathcal{Q}$ be any stable set of matrices. For any set $\Delta \subseteq \{0,1\}^2$ of dominoes, consider the following statement:

Δ-Shuffle Property: *There exist a constant c and a function $f : \{0,1\}^c \to \{0,1\}$ such that for every $n \in \mathbb{N}$ there are $Q_1, \ldots, Q_c \in \mathsf{QS}(\mathcal{Q})$ such that*

$$\forall x, y \in \{0,1\}^n : \qquad \mathsf{EHD}_k^n(x, y) = f(Q_1(x, y), \ldots, Q_c(x, y)) \,, \tag{3}$$

and each Q_i is Δ-shuffle invariant.

Proposition 1.13 guarantees that the Δ-shuffle property holds for $\Delta = \emptyset$. Our goal is to show that it holds for $\Delta = \{0,1\}^2$.

Claim 2.8. *Let $\Delta \subseteq \{0,1\}^2$ be any set of dominoes, and suppose the Δ-shuffle property holds. Then for any $a \in \{0,1\}$, the $\left(\Delta \cup \left\{\left[\begin{smallmatrix} a \\ a \end{smallmatrix}\right]\right\}\right)$-shuffle property also holds.*

Proof of claim. Let $\Delta' = \Delta \cup \left\{\left[\begin{smallmatrix} a \\ a \end{smallmatrix}\right]\right\}$, and let $\overline{a}$ denote the negation of the bit a. Then, let $D = \left\{\left[\begin{smallmatrix} \overline{a} \\ \overline{a} \end{smallmatrix}\right], \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]\right\}$. For $n \in \mathbb{N}$, let $N = N(n, 2^b, n)$ be the number obtained from Corollary 2.7, where $b = c \cdot 3^n$, which will be justified below. We will embed $\{0,1\}^n$ into the larger domain of $\{0,1\}^N$ such that the embedding preserves the Hamming distance and the D-type of any two strings, and allows to show the Δ'-shuffle invariance of queries.

The first two properties are easy to satisfy. Take any subset of coordinates $T \subset [N]$ of size $|T| = n$, and let $\phi_T : \{0,1\}^n \to \{0,1\}^N$ be the map that writes $x \in \{0,1\}^n$ into the coordinates of T in the order-preserving way and sets the coordinates outside of T to be a. Observe that, for all $x, y \in \{0,1\}^n$,

$$\mathsf{dist}(x, y) = \mathsf{dist}(\phi_T(x), \phi_T(y)) \text{ and } \Gamma_D(x, y) = \Gamma_D(\phi_T(x), \phi_T(y)) \,.$$

Next, we choose a set T that helps us to show the Δ'-shuffle invariance of the queries. By assumption, there exists $f : \{0,1\}^c \to \{0,1\}$ and $Q'_1, \ldots, Q'_c : \{0,1\}^N \times \{0,1\}^N \to \{0,1\}$ with each $Q'_i \in \mathsf{QS}(\mathcal{Q})$ being Δ-shuffle invariant, such that

$$\forall X, Y \in \{0,1\}^N : \qquad \mathsf{EHD}_k^N(X, Y) = f(Q'_1(X, Y), \ldots, Q'_c(X, Y)).$$

where we write $X, Y \in \{0,1\}^N$ to distinguish them from lower-dimensional $x, y \in \{0,1\}^n$.

To each $S \subseteq [N]$ with $|S| \leq n$ we assign a color, which is a binary string $\mathsf{col}(S) \in \{0,1\}^b$, as follows. Let $s = |S|$ and define the color $q(d)$ of a domino vector $d \in D^s$ to be the sequence of c bits $q(d) = (Q'_1(U,V), \ldots, Q'_c(U,V))$, where $U, V \in \{0,1\}^N$ is the unique pair whose D-signature is d and the dominoes of d are written in the coordinates S of $[N]$. Now set $\mathsf{col}(S)$ to be $(q(d))_{d \in D^s}$, the concatenation of the colors $q(d)$ of all possible signature vectors $d \in D^s$, in lexicographic order of d. The total number of bits in $\mathsf{col}(S)$ is at most $c \cdot 3^s \leq c \cdot 3^n = b$, so there are at most 2^b colors.

By Corollary 2.7, there exists a set $T \subseteq [N]$ of size $|T| = n$ such that for every $s \leq n$, the subsets $S \subseteq T$ of cardinality $|S| = s$ each have the same color $\mathsf{col}(S)$. Let $\phi := \phi_T$ be the map defined above, which preserves the Hamming distance of $x, y \in \{0,1\}^n$ and their D-signature. For each $i \in [c]$, we now define the matrix $Q_i : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ by $Q_i(x,y) := Q'_i(\phi(x), \phi(y))$. Observe that Q_i is a submatrix of Q'_i, so $Q_i \in \mathsf{QS}(\mathcal{Q})$ since $\mathsf{QS}(\mathcal{Q})$ is closed under taking submatrices. Now,

$$\begin{aligned} \forall x, y \in \{0,1\}^n : \qquad f(Q_1(x,y), \ldots, Q_c(x,y)) &= f(Q'_1(\phi(x),\phi(y)), \ldots, Q'_c(\phi(x),\phi(y))) \\ &= \mathsf{EHD}_k^N(\phi(x), \phi(y)) \\ &= \mathsf{EHD}_k^n(x,y)\,. \end{aligned}$$

It remains to show that each Q_i is Δ'-shuffle invariant. Observe that each Q_i is Δ-shuffle invariant because it is a submatrix of Q'_i and Q'_i is Δ-shuffle invariant. Indeed, for any $x, y, u, v \in \{0,1\}^n$ and any $i \in [c]$,

$$\begin{aligned} \chi_{\overline{\Delta}}(x,y) = \chi_{\overline{\Delta}}(u,v) &\implies \chi_{\overline{\Delta}}(\phi(x),\phi(y)) = \chi_{\overline{\Delta}}(\phi(u),\phi(v)) \\ &\implies Q'_i(\phi(x),\phi(y)) = Q_i(\phi(u),\phi(v)) \\ &\implies Q_i(x,y) = Q_i(u,v)\,. \end{aligned}$$

Now let $x, y, u, v \in \{0,1\}^n$ such that $\chi_{\overline{\Delta'}}(x,y) = \chi_{\overline{\Delta'}}(u,v)$. We must show that $Q_i(x,y) = Q_i(u,v)$. First, assume that (x,y) and (u,v) have the same D-signature, so that the dominoes of (u,v) are obtained from those of (u,v) by a sequence of swaps of consecutive dominoes, where each swap involves an $\boxed{\frac{a}{a}}$ domino, (so that their subsequences of non-$\boxed{\frac{a}{a}}$ dominoes are the same). Then the sets

$$S_1 := \{i \in [N] \mid \phi(x)_i = \overline{a} \text{ or } \phi(y)_i = \overline{a}\} \qquad \text{and} \qquad S_2 := \{i \in [N] \mid \phi(u)_i = \overline{a} \text{ or } \phi(v)_i = \overline{a}\}$$

have the same size, thus also the same color $\mathsf{col}(S_1) = \mathsf{col}(S_2)$ because $S_1, S_2 \subseteq T$, and T was chosen so that all of its subsets of the same size have the same color. From the definition, there is some index $j \in [b]$ such that

$$\mathsf{col}(S_1)_j = Q'_i(\phi(x),\phi(y)) = Q_i(x,y) \qquad \text{and} \qquad \mathsf{col}(S_2)_j = Q'_i(\phi(u),\phi(v)) = Q_i(u,v),$$

and thus $Q_i(x,y) = Q_i(u,v)$ as desired. Finally, suppose (x,y) and (u,v) do not have the same D-signature, though they must still have the same $\overline{\Delta'}$-type, by assumption (meaning in particular $\Delta \neq \emptyset$). Consider the pairs (x', y') defined as

$$\boxed{\frac{x'}{y'}} = \boxed{\frac{x_{i_1}}{y_{i_2}}}\boxed{\frac{x_{i_2}}{y_{i_2}}} \cdots \boxed{\frac{x_{i_k}}{y_{i_k}}}\boxed{\frac{a}{a}}\boxed{\frac{a}{a}} \cdots \boxed{\frac{a}{a}}$$

$$\boxed{\frac{u'}{v'}} = \boxed{\frac{u_{j_1}}{v_{j_2}}}\boxed{\frac{u_{j_2}}{u_{j_2}}} \cdots \boxed{\frac{v_{j_k}}{v_{j_k}}}\boxed{\frac{a}{a}}\boxed{\frac{a}{a}} \cdots \boxed{\frac{a}{a}}$$

where $i_1 < i_2 < \cdots < i_k$ and $j_1 < j_2 < \cdots < j_k$ are the indices of the non-$\left[\frac{a}{a}\right]$ dominoes of (x, y) and (u, v) respectively. Observe that (x', y') has the same D-type and $\overline{\Delta'}$-type as (x, y), and (u', v') has the same D-type and $\overline{\Delta'}$-type as (u, v). This is because the order of non-$\left[\frac{a}{a}\right]$ dominoes is preserved, and the number of non-$\left[\frac{a}{a}\right]$ dominoes in (x, y) and (u, v) is the same since they have the same $\overline{\Delta'}$-type. By the argument above, we have $Q_i(x, y) = Q_i(x', y')$ and $Q_i(u, v) = Q_i(u', v')$ for each query Q_i. Finally, observe that (x', y') and (u', v') have the same $\overline{\Delta}$-type, since by assumption they have the same ($\overline{\Delta'} = \overline{\Delta \cup \left[\frac{a}{a}\right]}$)-type (meaning the order of non-$(\Delta \cup \left[\frac{a}{a}\right])$ dominoes is the same). Then (x', y') is obtained from (u', v') by swaps of consecutive dominoes involving dominoes in Δ. Since each query Q_i is Δ-shuffle invariant, we have $Q_i(x', y') = Q_i(u', v')$, and therefore

$$Q_i(x, y) = Q_i(x', y') = Q_i(u', v') = Q_i(u, v)\,,$$

as desired. □

Applying Claim 2.8 twice, with $\Delta = \emptyset$ and $a = 0$, and then with $\Delta = \left\{\left[\frac{0}{0}\right]\right\}$ and $a = 1$, we achieve the $\left\{\left[\frac{0}{0}\right], \left[\frac{1}{1}\right]\right\}$-shuffle property. We conclude with the following claim:

Claim 2.9. *Suppose that the $\left\{\left[\frac{0}{0}\right], \left[\frac{1}{1}\right]\right\}$-shuffle property holds. Then the $\left\{\left[\frac{0}{0}\right], \left[\frac{1}{1}\right], \left[\frac{0}{1}\right], \left[\frac{1}{0}\right]\right\}$-shuffle property also holds.*

Proof of claim. Since $\mathcal{Q}$ is stable, there exists a constant t such that neither GT_t or its complement $\neg\mathsf{GT}_t$ belong to $\mathsf{QS}(\mathcal{Q})$.

Take $\Delta = \left\{\left[\frac{0}{1}\right], \left[\frac{1}{0}\right]\right\}$. Take $N \in \mathbb{N}$ such that $N - n > 2t$, and embed $\{0,1\}^n$ into $\{0,1\}^N$ by the map $\phi : x \mapsto x00\cdots0$, that pads $N-n$ many 0's at the end of the input string. Let $Q'_1, \ldots, Q'_c \in \mathcal{Q}$ be $\overline{\Delta}$-shuffle invariant query matrices such that $\mathsf{EHD}^N_k(X, Y) = f(Q'_1(X, Y), \ldots, Q'_c(X, Y))$ for all $X, Y \in \{0,1\}^N$. Now for each $i \in [c]$, define $Q_i : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ as $Q_i(x, y) = Q'_i(\phi(x), \phi(y))$. Note that $\Gamma_\Delta(x, y) = \Gamma_\Delta(\phi(x), \phi(y))$.

Since each Q'_i is $\overline{\Delta}$-shuffle invariant, $Q'_i(X, Y)$ depends only on the Δ-type $\chi_\Delta(X, Y)$, and $Q_i(x, y)$ depends only on the Δ-type $\chi_\Delta(x, y)$. Therefore, for any Δ-type A, we write $Q'_i(A)$ for the value taken by Q'_i on all (X, Y) with $\chi_\Delta(X, Y) = A$.

Assume for the sake of contradiction that there exists $i \in [c]$ such that Q_i is not permutation-invariant. Then there exist $x, y, u, v \in \{0,1\}^n$ such that $Q'_i(\phi(x), \phi(y)) \neq Q'_i(\phi(u), \phi(v))$, and, for the two Δ-types $A = (\Gamma^A, \tau^A) = \chi_\Delta(\phi(x), \phi(y))$ and $B = (\Gamma^B, \tau^B) = \chi_\Delta(\phi(x), \phi(y))$:

1. $Q'_i(A) \neq Q'_i(B)$; and
2. The Δ-signatures Γ^A, Γ^B (subsequences of $\left[\frac{1}{0}\right]$ and $\left[\frac{0}{1}\right]$ dominoes) are permutations of each other, the tallies $\tau^A = \tau^B =: \tau$ are the same, and $\tau_{00} \geq N - n$ due to the padding in ϕ.

It suffices to consider Δ-types with tally τ and signatures Γ^A and Γ^B, where Γ^B is obtained by swapping a single consecutive pair of dominoes in Γ^A; if it holds that $Q'_i(A) = Q'_i(B)$ for any two Δ-types with tally τ and Δ-signatures Γ^A, Γ^B which differ only by swapping a consecutive pair of

dominoes, then it holds that $Q'_i(A) = Q'_i(B)$ for all Δ-types A, B with tally τ, since A may be transformed into B by a sequence of swaps of consecutive dominoes.

Thus, assume Γ^A and Γ^B differ by only one swap of consecutive dominoes. Then we may choose domino sequences Σ^A, Σ^B with the Δ-types A and B as follows.

For some domino subsequences $\left[\frac{C^\circ}{C_\bullet}\right] \in \Delta^{d_1}$ and $\left[\frac{D^\circ}{D_\bullet}\right] \in \Delta^{d_2}$, where $d_1, d_2 \in \mathbb{N}$ satisfy $d_1 + 2 + d_2 + \tau_{00} + \tau_{11} = N$, the following $\Sigma^A, \Sigma^B \in (\{0,1\}^2)^N$ have Δ-types A and B respectively:

$$\Sigma^A = \left(\left[\frac{C^\circ}{C_\bullet}\right]\left(\left[\frac{1}{0}\right]\left[\frac{0}{1}\right]\right)\left[\frac{D^\circ}{D_\bullet}\right]\left(\left[\frac{0}{0}\right]\right)^{\tau_{00}}\left(\left[\frac{1}{1}\right]\right)^{\tau_{11}}\right),$$
$$\Sigma^B = \left(\left[\frac{C^\circ}{C_\bullet}\right]\left(\left[\frac{0}{1}\right]\left[\frac{1}{0}\right]\right)\left[\frac{D^\circ}{D_\bullet}\right]\left(\left[\frac{0}{0}\right]\right)^{\tau_{00}}\left(\left[\frac{1}{1}\right]\right)^{\tau_{11}}\right).$$

To achieve a contradiction, we construct an impossibly large GT submatrix within Q'_i as follows. Let $e_i \in \{0,1\}^{\tau_{00}+2}$ be the string that has 1 in the ith coordinate and is 0 everywhere else. For $i \in [\tau_{00} + 2]$, define $h_i \in \{0,1\}^N$ as

$$h_i = \begin{cases} C^\circ \mid e_i \mid D^\circ \mid \underbrace{1 \ldots 1}_{\tau_{11}} & \text{if } i \text{ is even,} \\ C_\bullet \mid e_i \mid D_\bullet \mid \underbrace{1 \ldots 1}_{\tau_{11}} & \text{otherwise.} \end{cases}$$

Now consider the $\lfloor \tau_{00}/2 \rfloor \times \lfloor \tau_{00}/2 \rfloor$ submatrix M of Q'_i on rows h_{2i} and on columns h_{2j+1} for $i, j \in [\lfloor \tau_{00}/2 \rfloor]$. Note that if $i \leq j$, the pair (h_{2i}, h_{2j+1}) has Δ-signature Γ^A, and if $i > j$, the pair has the Δ-signature Γ^B. Since Q'_i has different values on the input pairs that have Δ-signatures Γ^A and Γ^B, respectively, the submatrix M is exactly $\mathsf{GT}_{\lfloor \tau_{00}/2 \rfloor}$ or its complement. Since $\tau_{00} \geq N - n > 2t$, this contradicts the fact that $\mathsf{QS}(\mathcal{Q})$ does not have GT_T for any $T > t$. □

3 Main Results: Separations within BPP^0

We now apply Lemma 2.2 to prove our main results. The essence of our technique is that it transforms the task of proving a lower bound for a (total) communication problem $\mathcal{P}$ against an arbitrary constant number of oracle queries, into a lower bound for a certain type of *partial* subproblem of $\mathcal{P}$ against a *single* oracle query. This type of partial problem is defined below.

3.1 Reduction to One Query

Recall that the *tally* $\tau(x, y)$ of two $x, y \in \{0,1\}^n$ counts the number of times each domino $\left[\frac{a}{b}\right]$ appears in the dominoes of (x, y).

Definition 3.1 (Two-Tally Matrix)**.** We say a partial matrix $M \in \{0, 1, *\}^{t \times t}$ is a *two-tally matrix* of EHD_k if there exist $n \in \mathbb{N}$ and $x^{(1)}, \ldots, x^{(t)}, y^{(1)}, \ldots, y^{(t)} \in \{0,1\}^n$ which satisfy the following:

1. The submatrix of EHD^n_k on rows $x^{(i)}$ and columns $y^{(j)}$ is a completion of M;
2. If $M(x^{(i)}, y^{(j)}) = M(x^{(i')}, y^{(j')}) \neq *$ have the same non-$*$ value in M, then $(x^{(i)}, y^{(j)})$ and $(x^{(i')}, y^{(j')})$ have the same tally $\tau(x^{(i)}, y^{(j)}) = \tau(x^{(i')}, y^{(j')})$, meaning that $(x^{(i)}, y^{(j)})$ and $(x^{(i')}, y^{(j')})$ are permutations of each other.

Our main results will follow by the application of the next lemma.

Lemma 3.2. *Let $\mathcal{Q}$ be any stable set of matrices, let k be any constant, and let M be a two-tally matrix of EHD_k. If $\mathsf{D}^{\mathcal{Q}}(\mathsf{EHD}_k) = O(1)$ then there is $L \in \mathsf{QS}(\mathcal{Q})$ that is a completion either of M, or its Boolean negation $\neg M$.*

Proof. Suppose $\mathsf{D}^{\mathcal{Q}}(\mathsf{EHD}_k) = O(1)$. By Lemma 2.2, there is a constant c and a function f such that for every n, there exist permutation-invariant matrices $Q_1, \ldots, Q_c \in \mathcal{Q}$ such that

$$\forall x, y \in \{0,1\}^n : \qquad \mathsf{EHD}_k^n(x,y) = f(Q_1(x,y), \ldots, Q_c(x,y)). \tag{4}$$

Consider now the two-tally matrix M of EHD_k defined by the vectors $x^{(1)}, \ldots, x^{(t)}, y^{(1)}, \ldots, y^{(t)} \in \{0,1\}^n$, and let i_0, j_0, i_1, j_1 be such that $M(x^{(i_b)}, y^{(j_b)}) = b$ for $b \in \{0,1\}$ (we may assume such pairs exist as otherwise the claim is trivial). By (4), there must exist ℓ such that $Q_\ell(x^{(i_0)}, y^{(j_0)}) \neq Q_\ell(x^{(i_1)}, y^{(j_1)})$, which by the permutation invariance of Q_ℓ implies that Q_ℓ distinguishes between 0s and 1s of M. Then the submatrix L of Q_ℓ, on rows $\{x^{(1)}, \ldots, x^{(t)}\}$ and columns $\{y^{(1)}, \ldots, y^{(t)}\}$, is a completion of either M or $\neg M$. □

3.2 No Complete Problem for BPP^0

We now prove a lower bound for EHD_k against queries of a general form, which will imply our main Theorem 1.1. For convenience, we will state the general result in terms of the VC dimension.

Definition 3.3 (VC Dimension)**.** The VC dimension of a matrix $M \in \{0,1\}^{N \times N}$ is the largest number d such that there are d columns $y^{(1)}, \ldots, y^{(d)}$ that are *shattered*, meaning that for every $S \subseteq [d]$ there is a row x^S such that $M(x^S, y^{(i)}) = 1$ if and only if $i \in S$.

Remark 3.4. For every problem $\mathcal{P} \in \mathsf{BPP}^0$ there is a constant d such that the VC dimension of any $P \in \mathcal{P}$ is at most d. If $\mathcal{M}$ is a set of matrices with unbounded VC dimension (for example, if $\mathcal{M}$ is the Set-Disjointness communication problem), then $\mathsf{QS}(\mathcal{M})$ is the set of all matrices, meaning in particular that $\mathsf{D}^{\mathcal{M}}(\mathcal{P}) = 1$ for all communication problems $\mathcal{P}$.

A simple argument shows that *any* fixed total matrix M appears as a two-tally matrix of EHD_k, for sufficiently large constant k. (This is not the same thing as saying EHD_k has unbounded VC dimension – for each constant k, the VC dimension of EHD_k remains bounded.)

Proposition 3.5. *For every constant k, there is a $2^k \times k$ matrix M of VC dimension k that is a two-tally submatrix of EHD_{k-1}.*

Proof. Let $n > 2k$, and let $y^{(1)}, \ldots, y^{(k)} \in \{0,1\}^n$ be the first k standard basis vectors. Now for every set $S \subseteq [k]$ let $x^S \in \{0,1\}^n$ be the string where the last $k - |S|$ bits are set to 1, and the bits $i \in S$ are set to 1, and the remaining bits are 0. Fix i and S and consider two cases:

- Suppose $i \in S$ and consider the domino sequence of $x^S, y^{(i)}$; we see that it contains 1 of $\left[\frac{1}{1}\right]$, $k-1$ of $\left[\frac{1}{0}\right]$, and $n-k$ of $\left[\frac{0}{0}\right]$, and $\mathsf{dist}(x^S, y^{(i)}) = k-1$.
- Now suppose $i \notin S$ and consider the domino sequence of $x^S, y^{(i)}$; we see that it contains 1 of $\left[\frac{0}{1}\right]$, k of $\left[\frac{1}{0}\right]$, and $n-k-1$ of $\left[\frac{0}{0}\right]$, and $\mathsf{dist}(x^S, y^{(i)}) = k+1$.

We see that for every set $S \subseteq [k]$, $\mathsf{EHD}^n_{k-1}(x^S, y^{(i)}) = 1$ iff $i \in S$ and therefore this $k \times 2^k$ submatrix has VC dimension k. From the above observations, this submatrix satisfies the conditions to be a two-tally submatrix of EHD_{k-1}. □

As a result, we get a general separation of EHD_k against oracle queries belonging to any stable set of matrices. Note that any stable set of matrices has constant VC dimension, because a forbidden GT_t subproblem implies a bound of t on the VC dimension.

Theorem 3.6. *Let $\mathcal{Q}$ be any stable set of matrices with VC dimension d. Then for any $k \geq d$, $\mathsf{D}^{\mathcal{Q}}(\mathsf{EHD}_k) = \omega(1)$.*

Proof. Assume $\mathsf{D}^{\mathcal{Q}}(\mathsf{EHD}_k) = O(1)$. By Proposition 3.5 there is a $2^{k+1} \times (k+1)$ two-tally matrix M of EHD_k, with VC dimension $k+1 > d$. Note that M and $\neg M$ are permutations of each other (meaning $\neg M$ is obtained from M by permuting its rows and columns), so $M, \neg M \notin \mathsf{QS}(\mathcal{Q})$. This contradicts Lemma 3.2. □

Our main Theorem 1.1 now follows as a corollary, since any problem $\mathcal{Q} \in \mathsf{BPP}^0$ must be stable (Observation 1.16).

Corollary 3.7 (Theorem 1.1)**.** *For every problem $\mathcal{Q} \in \mathsf{BPP}^0$, there exists a constant k such that $\mathsf{D}^{\mathcal{Q}}(\mathsf{EHD}_k) = \omega(1)$.*

This statement also holds for THD_k in place of EHD_k, since they are equivalent under constant-cost reductions. THD_k has a one-sided error protocol (unlike EHD_k) which also means there is no complete problem for the class of constant-cost problems with one-sided error.

3.3 The k-Hamming Distance Hierarchy

The VC dimension of Equality is 1, since $\left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ cannot occur as a submatrix. Therefore

$$\mathsf{D}^{\mathrm{EQ}}(\mathsf{EHD}_1) = \omega(1)\,,$$

by Theorem 3.6, recovering (qualitatively) the results of [HWZ22, HHH22b]. An immediate consequence of Theorem 3.6 is an infinite number of such separations, forming an infinite hierarchy within BPP^0.

Corollary 3.8 (Theorem 1.2)**.** *There are infinitely many $k \in \mathbb{N}$ such that $\mathsf{D}^{\mathsf{EHD}_k}(\mathsf{EHD}^n_{k+1}) = \omega(1)$.*

Proof. Fix any t. Then EHD_t is stable, so by Theorem 3.6, there is some constant $t' > t$ such that $\mathsf{D}^{\mathsf{EHD}_t}(\mathsf{EHD}_{t'}) = \omega(1)$. Then there must exist $t \leq k < t'$ such that $\mathsf{D}^{\mathsf{EHD}_k}(\mathsf{EHD}_{k+1}) = \omega(1)$, because, if $\mathsf{D}^{\mathsf{EHD}_k}(\mathsf{EHD}_{k+1}) = O(1)$ for every $t \leq k < t'$ then we would have $\mathsf{D}^{\mathsf{EHD}_t}(\mathsf{EHD}_{t'}) = O(1)$. □

In the above proof, it suffices to take $t' = 2^{\Theta(t \log t)}$. First observe that any $M \in \mathsf{QS}(\mathsf{EHD}_t)$ has $\mathsf{R}(M) \leq C \cdot t \log t$ for some constant C. On the other hand, we may choose a $t' \times t'$ two-tally matrix $M \in \mathsf{QS}(\mathsf{EHD}_{t'})$ with maximum randomized communication cost $\mathsf{R}(M) = \Theta(\log t')$, since Proposition 3.5 guarantees that every $t' \times t'$ matrix with unique columns exists as a two-tally submatrix of $\mathsf{EHD}_{t'}$. Therefore, if we choose $t' = 2^{\Theta(t \log t)}$ with a sufficiently large constant in the exponent, then we have a two-tally submatrix M of $\mathsf{EHD}_{t'}$ with $\mathsf{R}(M) = \Theta(\log t') > C \cdot t \log t$, so $M \notin \mathsf{QS}(\mathsf{EHD}_t)$ (and the same holds for $\neg M$).

However, the VC dimension argument does not suffice to separate EHD_1 from EHD_2, because the VC dimension of EHD_1 is 3 (the first 3 standard basis vectors are shattered), which only proves $\mathsf{D}^{\mathsf{EHD}_1}(\mathsf{EHD}_3) = \omega(1)$. We tighten this separation by choosing a different two-tally matrix M.

Theorem 3.9 (Restatement of Theorem 1.3)**.** $\mathsf{D}^{\mathsf{EHD}_1}(\mathsf{EHD}_2) = \omega(1)$.

Proof. First consider the matrix M defined as the submatrix of EHD_2^7 on rows $\mathcal{X}$ and columns $\mathcal{Y} = \mathcal{Y}_0 \cup \mathcal{Y}_1$, where

$$\begin{aligned}\mathcal{X} &:= \{0011000, 1100000\}\\ \mathcal{Y}_0 &:= \{0000011, 0000101, 0000110\}\\ \mathcal{Y}_1 &:= \{1010000, 1001000, 010100\}\,.\end{aligned}$$

Observe that for $\beta \in \{0,1\}$, $\mathcal{X} \times \mathcal{Y}_\beta$ is a β-monochromatic rectangle of EHD_2^7, and that the distance between any two $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is either 2 or 4. Now, observe that for any two distinct strings $a, b \in \{0,1\}^7$ with Hamming weight 2, there exists a $\delta = \delta(a,b) \in \{0,1\}^7$ with Hamming weight 2 such that $\mathsf{dist}(a, \delta) = 2$ and $\mathsf{dist}(b, \delta) = 4$. We now extend M to a partial matrix M' by adding the columns $\{\delta(a,b) \,|\, a, b \in \mathcal{X}, a \neq b\}$ and the rows $\{\delta(a,b) \,|\, a, b \in \mathcal{Y}, a \neq b\}$, and taking the entries

$$M'(x,y) := \begin{cases} 1 & \text{if } \mathsf{dist}(x,y) = 2\\ 0 & \text{if } \mathsf{dist}(x,y) = 4\\ * & \text{otherwise.}\end{cases}$$

This matrix agrees with $\mathsf{EHD}_2^7(x,y)$ whenever $\mathsf{dist}(x,y) \in \{2,4\}$, and every row and every column is distinct. Since the weight of every string is the same and all non-$*$ entries have distance 2 or 4, it must hold that all 1-valued entries x, y of M' are domino permutations of each other, and the same for all 0-valued entries, making M' a two-tally matrix of EHD_2.

Assume for the sake of contradiction that $\mathsf{D}^{\mathsf{EHD}_1}(\mathsf{EHD}_2) = O(1)$. Then by Lemma 3.2, there is $L \in \mathsf{QS}(\mathsf{EHD}_1)$ that is a completion of M' or $\neg M'$. However, M' and $\neg M'$ both contain the submatrix $K_{2,3}$ (the 2×3 all-1s matrix), and it is known that EHD_1 does not contain $K_{2,3}$. Since M' has distinct rows and columns, it cannot be obtained from a submatrix of EHD_1 by copying rows and columns, and therefore any completion L of M' or $\neg M'$ cannot belong to $\mathsf{QS}(\mathsf{EHD}_1)$, a contradiction. □

3.4 Separating k-Hamming Distance and Integer Inner Product

Our final application separates k-Hamming Distance from Integer Inner Product:

Definition 3.10 (Integer Inner Product)**.** For any fixed constant d, the Integer Inner Product problem in dimension d is defined as $\mathsf{IIP}_d = (\mathsf{IIP}_d^n)_{n\in\mathbb{N}}$ where $\mathsf{IIP}_d^n : \{0,1\}^{dn} \times \{0,1\}^{dn} \to \{0,1\}$ is the function defined on $x, y \in \{0,1\}^{dn}$, interpreted as the binary representation of integer vectors $x = (x_1, x_2, \ldots, x_d)$ and $y = (y_1, y_2, \ldots, y_d)$ in domain $[-2^{n-1}, 2^{n-1}]^d$, where

$$\mathsf{IIP}_d^n(x,y) := \begin{cases} 0 & \text{if } \langle x, y\rangle = 0\\ 1 & \text{otherwise.}\end{cases}$$

It is known that $\mathsf{R}(\mathsf{IIP}_d^n) = O(d \cdot \log n)$ [CLV19], so $\mathsf{IIP}_d \in \mathsf{BPP}$ for every constant d, but it is conjectured that $\mathsf{IIP}_d \notin \mathsf{BPP}^0$ (see e.g. [CHHS23]). It was shown in [CLV19] that there is an

infinite sequence $d_1 < d_2 < \cdots$ of constants such that $\mathsf{D}^{\mathsf{IIP}_{d_i}}(\mathsf{IIP}_{d_{i+1}}) = \Theta(n)$; in other words, they form an infinite hierarchy within BPP. We will show that oracles to these functions, which each have much larger randomized communication complexity than any EHD_k, nevertheless cannot simulate EHD_k in BPP^0.

We require the following lemma.

Lemma 3.11. *For any constant d, IIP_d is stable and has VC dimension at most d.*

Proof. It suffices to prove the following claim.

Claim 3.12. *Let $t \geq 1$, and let $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^t$ be any finite sets of points with $\vec{0} \notin \mathcal{Y}$, and consider any sequences of points $x_1, \ldots, x_m \in \mathcal{X}$ and $y_1, \ldots, y_m \in \mathcal{Y}$ such that $\forall i, j \in [m]$, $\langle x_i, y_j \rangle = 0$ if and only if $i \leq j$. Then $m \leq t$.*

Proof of claim. We prove the claim by induction on t. One may easily check that the claim is true in the base case $t = 1$ where $\langle x_i, y_j \rangle = 0$ iff $x_i = 0$. Now assume $t \geq 2$. Let $x_1, \ldots, x_m \in \mathcal{X}$ and $y_1, \ldots, y_m \in \mathcal{Y}$ be sequences satisfying the condition $\langle x_i, y_j \rangle = 0$ iff $i \leq j$. Since $y_m \neq \vec{0}$ and $\langle x_i, y_m \rangle = 0$ for all $i \in [m]$, it defines a perpendicular subspace $\mathcal{W}$ of dimension $t - 1$, $\mathcal{W} := \{x \in \mathbb{R}^t : \langle x, y_m \rangle = 0\}$, such that $\{x_1, \ldots, x_m\} \subseteq \mathcal{W}$. Let $y'_1, \ldots, y'_{m-1}$ denote the projections of $y_1, \ldots, y_{m-1}$ into $\mathcal{W}$, and observe for each $j \in [m-1]$ that $\langle x_m, y_j \rangle \neq 0$ by definition, so $y'_j \neq \vec{0}$ since $x_m \in \mathcal{W}$. Finally note that for all $i, j \in [m-1]$, it holds that $\langle x_i, y'_j \rangle = 0$ iff $\langle x_i, y_j \rangle = 0$, and therefore we may apply the induction hypothesis to $x_1, \ldots, x_{m-1}$ and $y'_1, \ldots, y'_{m-1}$ to conclude that $m - 1 \leq t - 1$. □

To conclude the proof of the lemma, observe that taking finite sets $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^d$ that may include $\vec{0}$ cannot increase the size m of the ordered sequences $x_1, \ldots, x_m$ and $y_1, \ldots, y_m$ in the claim by more than 1. The bound on the VC dimension is known in the literature, and also follows from the above proof, since this exhibits a $(d+1) \times (d+1)$ matrix with unique rows and columns that cannot be a submatrix of IIP_d. □

We now apply Theorem 3.6 to separate EHD_k from IIP_d and obtain the formal statement of Theorem 1.4.

Theorem 3.13. *For any constant d and any $k \geq d$, $\mathsf{D}^{\mathsf{IIP}_d}(\mathsf{EHD}_k) = \omega(1)$.*

There are two conjectures in the literature which would imply a stronger version of this theorem for any constant d and $k = 1$; see Section 5. Note that $\mathsf{D}^{\mathsf{IIP}_2}(\mathsf{EHD}_0) = O(1)$ since EHD_0 is Equality.

4 The Complexity of Submatrices of k-Hamming Distance

Problems in BPP^0 satisfy the hereditary property that the randomized communication cost $\mathsf{R}(\cdot)$ remains bounded by taking submatrices. As discussed in Section 1.1.1, it is not true in general (outside BPP^0) that $\mathsf{R}(\cdot)$ is preserved, as a function of the matrix size, by taking arbitrary submatrices. It is helpful, for proving lower bounds, to understand when hereditary properties hold for other complexity measures as well. We will show in this section that $\mathsf{D}^{\mathrm{EQ}}(\mathsf{EHD}_k)$, and the γ_2-norm of EHD_k, are also preserved when taking submatrices of EHD_k, and we use these hereditary properties to prove new lower bounds against EHD_k oracles within BPP.

The γ_2-norm is an important norm in communication complexity. For a matrix $M \in \{0,1\}^{N\times N}$ it is defined as

$$\|M\|_{\gamma_2} = \min_{A,B:\, M=AB} \|A\|_{\text{row}} \|B\|_{\text{col}}\,,$$

where $\|A\|_{\text{row}}, \|B\|_{\text{col}}$ denote the largest ℓ_2 norm of any row and column, respectively. See [LS09, HHH22b, CHHS23] for a discussion of this quantity and its relation to communication complexity.

We prove the following proposition in Section 4.1.

Proposition 4.1. *For any k and any $N \times N$ query matrix $Q \in \mathsf{QS}(\mathsf{THD}_k)$, we have*

1. $\mathsf{D}^{\mathrm{EQ}}(Q) = O(k \log\log N)$, *and*
2. $\|Q\|_{\gamma_2} = (\log N)^{O(k)}$.

Item (2) above follows from Item (1) by the result of [HHH22b] that for any Boolean matrix M,

$$\mathsf{D}^{\mathrm{EQ}}(M) \geq \frac{1}{2} \log \|M\|_{\gamma_2}. \tag{5}$$

It was proved later in [CHHS23] that for any $d \geq 3$,

$$\|\mathsf{IIP}^n_d\|_{\gamma_2} = 2^{\Omega(n)}. \tag{6}$$

This combined with (5) recovers the lower-bound of [CLV19], showing that for every $d \geq 3$,

$$\mathsf{D}^{\mathrm{EQ}}(\mathsf{IIP}^n_d) = \Omega(n). \tag{7}$$

Combining this with Proposition 4.1 shows that even IIP^n_3 does not reduce to the k-Hamming Distance under BPP reductions allowing queries of unbounded size, for any $k \leq n/(\log n)^{\omega(1)}$:

Theorem 4.2 (Restatement of Theorem 1.5)**.** *For any $d \geq 3$ and $k \leq n/(\log n)^{\omega(1)}$, we have $\mathsf{D}^{\mathsf{THD}_k}(\mathsf{IIP}^n_d) = (\log n)^{\omega(1)}$. In particular, when k is a constant, $\mathsf{D}^{\mathsf{THD}_k}(\mathsf{IIP}^n_d) = \Omega\left(n/\log n\right)$.*

Proof. Consider a $\mathsf{D}^{\mathsf{THD}_k}$ protocol for IIP^n_d with q queries. By Proposition 4.1, each of the queries is a $2^{dn} \times 2^{dn}$ matrix $Q \in \mathsf{QS}(\mathsf{THD}_k)$ that can be simulated with $O(k \log n)$ Equality oracle queries. This gives a protocol for IIP^n_d with $O(qk \log n)$ Equality queries. Using (7), we have $q = \Omega(n/k \log n)$. Therefore, $q = (\log n)^{\omega(1)}$, as long as $k \leq n/(\log n)^{\omega(1)}$, and $q = \Omega(n/\log n)$ when k is a constant. □

4.1 Replacing k-Hamming Distance Queries with Equality Queries

We now show that any $N \times N$ matrix $Q \in \mathsf{QS}(\mathsf{THD}_k)$ can be reduced to $O(k \log\log N)$ Equality oracle calls. When $Q = \mathsf{THD}^n_k$, i.e. when the matrix has size $2^n \times 2^n$ and the inputs are $x, y \in \{0,1\}^n$, this can be done easily using binary search to find the first differing bit, and removing it and repeating up to $k+1$ times. But this simple protocol becomes inefficient when Q is a submatrix of THD^d_k for $d \gg \log N$, i.e. the inputs x, y are chosen from subsets $X, Y \subseteq \{0,1\}^d$ with $|X|, |Y| \ll 2^d$, so the number of coordinates is very large and naïve binary search gives $O(k \log d)$ instead of $O(k \log\log N)$. Such a d-dependent bound is not useful for our purposes, as our protocols are allowed to query oracles of arbitrary dimension.

We will need the following simple lemma. For a binary string $x \in \{0,1\}^d$ and a set $A \subseteq [d]$, write $x_A \in \{0,1\}^{|A|}$ for the substring of x on indices A.

Lemma 4.3. *Let $Z \subseteq \{0,1\}^d$ be a set of N binary strings. Then there exists a partition $[d] = A \cup B$ such that, for all $x, y \in Z$:*

1. *If $x_A = y_A$, then x_B, y_B differ on at most $3 \log N$ bits.*
2. *If $x_B = y_B$, then x_A, y_A differ on at most $3 \log N$ bits.*

Proof. Choose a partition $[d] = A \cup B$ uniformly at random. Fix an arbitrary pair $x, y \in Z$. If x and y differ on at most $3 \log N$ bits, then regardless of A and B, the properties hold for x, y. Assume otherwise, and note that in this case, the probability that $x_A = y_A$ or $x_B = y_B$ is at most $2 \cdot 2^{-3 \log N} < 1/\binom{N}{2}$. Thus, by a union bound over all the $\binom{N}{2}$ choices of x, y, there exists a choice of A and B that satisfies the claim. □

We also require the next proposition, which is well-known and is achieved by performing binary search on the bits in the binary representation of the inputs $i, j \in [N]$.

Proposition 4.4. $\mathsf{D}^{\mathrm{EQ}}(\mathsf{GT}_N) = O(\log \log N)$.

Now we show that any $N \times N$ submatrix of THD_k, with arbitrary dimension, can be computed efficiently using EQ oracles.

Proof of Proposition 4.1. Let $Q \in \mathsf{QS}(\mathsf{THD}_k)$, and let $X, Y \subseteq \{0,1\}^d$ be the sets of rows and columns of Q respectively, for some dimension d. Write $Z = X \cup Y$ for the set of all relevant binary strings. In the statement of Proposition 4.1 we have defined $N = |X| = |Y|$ but in the proof here we write $N = |Z|$ for convenience, which does not affect the conclusion. We first define the following procedure Bounded Diameter Threshold Distance, which uses Equality oracle queries to compute the Hamming distance (up to threshold k), on inputs that are promised to belong to a set of N inputs of diameter at most $3 \log N$. This protocol works by transforming a low-diameter set into a low-Hamming-weight set. This subroutine will be used in Threshold Distance protocol determining whether $\mathsf{dist}(x, y) > k$.

Bounded Diameter Threshold Distance(Z, x, y, k)

Requires $x, y \in Z$, $\forall u, v \in Z, \mathsf{dist}(u, v) \leq 3 \log N$. ▷ *Write* $N = |Z|$.
Alice and Bob, without communication, determine the following:
  Pick the lexicographically first $z \in Z$.
  $I \leftarrow \{i \mid \exists w \in Z, (w \oplus z)_i = 1\}$. ▷ *Note that* $|I| = O(N \log N)$.
Alice lets $S \leftarrow \{i \mid (x \oplus z)_i = 1\}$.
Bob lets $T \leftarrow \{j \mid (y \oplus z)_j = 1\}$. ▷ *Note that* $S, T \subseteq I$ *and* $|S|, |T| \leq 3 \log N$.
Initialize $c \leftarrow 0$. ▷ *The number of differing bits that are confirmed so far.*
while $c < k$ and $\textsc{Eq}(S, T) = 0$ **do**
  $c \leftarrow c + 1$.
  Initiate $S' \leftarrow S$ and $T' \leftarrow T$, and $K = \lceil 3 \log N \rceil$.
  while $K > 1$ **do**
    ▷ *Here we determine the smallest element in the symmetric difference of* S' *and* T'.
    $K \leftarrow \lceil K/2 \rceil$.
    $S_1 \leftarrow$ the first $\min\{|S'|, K\}$ elements of S'. $S_2 \leftarrow S' \setminus S_1$.
    $T_1 \leftarrow$ the first $\min\{|T'|, K\}$ elements of T'. $T_2 \leftarrow T' \setminus T_1$.
    if $\textsc{Eq}(S_1, T_1) = 1$ **then** $S' \leftarrow S_1$, $T' \leftarrow T_1$.
    else $S' \leftarrow S_2, T' \leftarrow T_2$.
  if $|S'| = 0$ **then**
    Let j be such that $T' = \{j\}$. Bob lets $T \leftarrow T - \{j\}$.
  else if $|T'| = 0$ **then**
    Let i be such that $S' = \{i\}$. Alice lets $S \leftarrow S - \{i\}$.
  else
    Let i, j be such that $S' = \{i\}$ and $T' = \{j\}$ ▷ *Alice knows* i *and Bob knows* j.
    Alice and Bob determine whether $i < j$ using Proposition 4.4 on domain I.
    if $i < j$ **then** Alice lets $S \leftarrow S - \{i\}$.
    else Bob lets $T \leftarrow T - \{j\}$.
if $S = T$ **then return** c.
else return $\perp$.

Claim 4.5. *Let* $k \in \mathbb{N}$ *and* $Z \subseteq \{0,1\}^d$ *be shared inputs to both parties, where* Z *satisfies* $|Z| = N$ *and* $\mathsf{dist}(u, v) \leq 3 \log N$ *for all* $u, v \in Z$. *Then on inputs* $x, y \in Z$, *the protocol* Bounded Diameter Threshold Distance(Z, x, y, k) *uses at most* $O(k \log \log N)$ *calls to the* Equality *oracle and outputs the following:*

- *If* $\mathsf{dist}(x, y) \leq k$, *the protocol outputs* $\mathsf{dist}(x, y)$.
- *Otherwise the protocol outputs* $\perp$.

The analysis of the number of Equality oracle calls is elementary. The correctness of the Bounded Diameter Threshold Distance protocol follows from the fact that the inner loop satisfies the following invariant. Let $i \in \mathbb{N}$ be the smallest element in the symmetric difference of S', T'. Then

- If $S_1 \neq T_1$ then $i \in S_1 \cup T_1$, and it is the smallest element in the symmetric difference of those two sets.
- If $S_1 = T_1$ then $i \in S_2 \cup T_2$, and it is the smallest element in the symmetric difference of those two sets.

We now use the bounded-diameter search sub-protocol to construct the full search protocol.

Threshold Distance(x, y, k)

if $\textsc{Eq}(x, y) = 1$ **then return** 0.
else if $k = 0$ **then return** $\perp$.
else if $k > 0$ **then**
  Partition $[d] = A \cup B$ according to Lemma 4.3 applied with $Z = X \cup Y$, where $N \leq 2^{n+1}$.
  if $\textsc{Eq}(x_A, y_A) = 1$ and $\textsc{Eq}(x_B, y_B) = 0$ **then**
    $Z_B \leftarrow \{z_B \mid z \in Z \wedge\ z_A = x_A = y_A\}$.
    return Bounded Diameter Threshold Distance(Z_B, x_B, y_B, k).
      ▷ *Returns the correct value due to Lemma 4.3 and Claim 4.6.*
  else if $\textsc{Eq}(x_B, y_B) = 1$ and $\textsc{Eq}(x_A, y_A) = 0$ **then**
    $Z_A \leftarrow \{z_A \mid z \in Z \wedge\ z_B = x_B = y_B\}$.
    return Bounded Diameter Threshold Distance(Z_A, x_A, y_A, k).
      ▷ *Returns the correct value due to Lemma 4.3 and Claim 4.6.*
  else ▷ *In this case* $\mathsf{dist}(x_A, y_A), \mathsf{dist}(x_B, y_B) \geq 1$.
    $t \leftarrow$ Threshold Distance$(x_A, y_A, k-1)$.
      ▷ *Returns* $t = \mathsf{dist}(x_A, y_A)$ *if* $\mathsf{dist}(x_A, y_A) \leq k-1$.
    if $t = \perp$ **then return** $\perp$. ▷ $\mathsf{dist}(x_A, y_A) + \mathsf{dist}(x_B, y_B) > (k-1)+1$.
    $r \leftarrow$ Threshold Distance$(x_B, y_B, k-t)$.
      ▷ *Returns* $r = \mathsf{dist}(x_B, y_B)$ *if* $\mathsf{dist}(x_B, y_B) \leq k - \mathsf{dist}(x_A, y_A)$.
    if $r = \perp$ **then return** $\perp$.
    else return $t + r$.

The proposition follows immediately from the next claim.

Claim 4.6. *Let $k \in \mathbb{N}$ be a shared input to both parties. Then on inputs $x \in X, y \in Y$, the protocol* Threshold Distance(x, y, k) *uses at most $O(k \log\log N)$ calls to the* Equality *oracle and outputs the following:*

- *If* $\mathsf{dist}(x, y) \leq k$*, the protocol outputs* $\mathsf{dist}(x, y)$.
- *If* $\mathsf{dist}(x, y) > k$*, the protocol outputs* $\perp$.

The correctness of the Threshold Distance protocol follows from the claims in the comments, and the observation that the number of Equality oracle queries is $O(k \log\log N)$, which can be computed by an elementary recurrence. □

5 Discussion and Open Problems

Our Theorem 1.2 shows that there is an infinite hierarchy of k-Hamming Distance problems within BPP^0 that are irreducible to lower levels of the hierarchy. We expect it to be the case that $\mathsf{D}^{\mathsf{EHD}_k}(\mathsf{EHD}_{k+1}) = \omega(1)$ for every constant k. Indeed, it seems natural to expect that, for inputs on n bits, $\mathsf{D}^{\mathsf{EHD}_k}(\mathsf{EHD}_{k+1}) = \Omega(\log n)$, which matches an easy binary search based upper bound of $O(\log n)$. This was proved for $k = 0$ in [HHH22b]. It is possible that the question could be answered using the technique of [CLV19] combined with an analysis of monochromatic rectangles in EHD_k.

Question 5.1. *Is it the case that* $\mathsf{D}^{\mathsf{EHD}_k}(\mathsf{EHD}^n_{k+1}) = \Theta(\log n)$*?*

One question that arose in the course of this work is whether a certain *dimension reduction* result holds for k-Hamming Distance. Constant-cost reductions (Definition 1.11) to the k-Hamming Distance problem EHD_k allow queries of arbitrarily large dimension. The question is whether any such query can be replaced with a constant number of queries to EHD_k with dimension $O(\log N)$ where N is the original domain size. Formally:

Question 5.2. *Let* $M \in \{0,1\}^{N \times N}$ *be an arbitrary submatrix of* EHD^n_k *where* n *is arbitrarily large. Is there an absolute constant* c *and a function* $f : \{0,1\}^c \to \{0,1\}$ *such that*

$$\forall i, j \in [N] : \qquad M(i,j) = f(H_1(i,j), H_2(i,j), \ldots, H_c(i,j)),$$

where each H_i *is an* $N \times N$ *submatrix of* EHD^d_k *with* $d = O(\log N)$*?*

If this question has a positive answer, it may be helpful in the future for lower bounds against Hamming Distance oracles; it is one strategy that we tried in pursuit of Theorem 1.2.

An important question left open by this paper is whether the k-Hamming Distance captures the entirety of BPP^0, up to reductions. In other words, for every problem $\mathcal{P} \in \mathsf{BPP}^0$, there exists a constant k such that $\mathsf{D}^{\mathsf{EHD}_k}(\mathcal{P}) = O(1)$. We do not believe this to be the case, but there is not any example of a problem $\mathcal{P} \in \mathsf{BPP}^0$ in the literature that might serve as a candidate counterexample.

Finally, we point out two conjectures in the literature that would imply a stronger form of Theorem 3.13 that holds for any constant d and $k = 1$. The first conjecture is that, if $\mathcal{Q}$ and $\mathcal{P}$ are any problems where $\mathsf{D}^{\mathcal{Q}}(\mathcal{P}) = O(1)$ and $\mathcal{Q}$ has bounded sign-rank (which holds in particular for IIP_d [CHHS23]), $\mathcal{P}$ also has bounded sign-rank [HHP+22]. The second conjecture is that EHD_1 has unbounded sign-rank [HHP+22].

Acknowledgments

Thanks to Viktor Zamaraev for many helpful discussions and for helpful comments on the presentation of the main proof, and thanks to the anonymous reviewers for their comments which helped improve the presentation of this article.

References

[ALMM19] Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private PAC learning implies finite Littlestone dimension. In *STOC'19—Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 852–860. ACM, New York, 2019.

[BFS86] Laszlo Babai, Peter Frankl, and Janos Simon. Complexity classes in communication complexity theory. In *Proceedings of the 27th Annual Symposium on Foundations of Computer Science*, SFCS '86, page 337–347, USA, 1986. IEEE Computer Society.

[BW16] Mark Braverman and Omri Weinstein. A discrepancy lower bound for information complexity. *Algorithmica*, 76:846–864, 2016.

[Cha23] Maurice Chandoo. Logical labeling schemes. *Discrete Mathematics*, 346(10):113565, 2023.

[CHHS23] Tsun-Ming Cheung, Hamed Hatami, Kaave Hosseini, and Morgan Shirley. Separation of the Factorization Norm and Randomized Communication Complexity. In Amnon

Ta-Shma, editor, *38th Computational Complexity Conference (CCC 2023)*, volume 264 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 1:1–1:16, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[CLV19] Arkadev Chattopadhyay, Shachar Lovett, and Marc Vinyals. Equality Alone Does not Simulate Randomness. In Amir Shpilka, editor, *34th Computational Complexity Conference (CCC 2019)*, volume 137 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 14:1–14:11, Dagstuhl, Germany, 2019. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[DKN15] I. Diakonikolas, D. M. Kane, and V. Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1183–1202, Los Alamitos, CA, USA, oct 2015. IEEE Computer Society.

[EHK22] Louis Esperet, Nathaniel Harms, and Andrey Kupavskii. Sketching Distances in Monotone Graph Classes. In Amit Chakrabarti and Chaitanya Swamy, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2022)*, volume 245 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 18:1–18:23, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[EHZ23] Louis Esperet, Nathaniel Harms, and Viktor Zamaraev. Optimal Adjacency Labels for Subgraphs of Cartesian Products. In Kousha Etessami, Uriel Feige, and Gabriele Puppis, editors, *50th International Colloquium on Automata, Languages, and Programming (ICALP 2023)*, volume 261 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 57:1–57:11, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[Fis04] Eldar Fischer. On the strength of comparisons in property testing. *Information and Computation*, 189(1):107–116, 2004.

[FX15] Vitaly Feldman and David Xiao. Sample complexity bounds on differentially private learning via communication complexity. volume 44, pages 1740–1764, 2015.

[GPW18] Mika Göös, Toniann Pitassi, and Thomas Watson. The landscape of communication complexity classes. *Computational Complexity*, 27(2):245–304, 2018.

[Har20] Nathaniel Harms. Universal Communication, Universal Graphs, and Graph Labeling. In Thomas Vidick, editor, *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*, volume 151 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 33:1–33:27, Dagstuhl, Germany, 2020. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[HH22] H. Hatami and P. Hatami. The implicit graph conjecture is false. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1134–1137, Los Alamitos, CA, USA, nov 2022. IEEE Computer Society.

[HH24] Hamed Hatami and Pooya Hatami. Structure in communication complexity and constant-cost complexity classes. *arXiv preprint arXiv:2401.14623*, 2024.

[HHH22a] Lianna Hambardzumyan, Hamed Hatami, and Pooya Hatami. A counter-example to the probabilistic universal graph conjecture via randomized communication complexity. *Discrete Applied Mathematics*, 322:117–122, 2022.

[HHH22b] Lianna Hambardzumyan, Hamed Hatami, and Pooya Hatami. Dimension-free bounds and structural results in communication complexity. *Israel Journal of Mathematics*, 253(2):555–616, 2022.

[HHM23] Hamed Hatami, Kaave Hosseini, and Xiang Meng. A Borsuk-Ulam lower bound for sign-rank and its applications. In *STOC'23—Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 463–471. ACM, New York, [2023] ©2023.

[HHP+22] Hamed Hatami, Pooya Hatami, William Pires, Ran Tao, and Rosie Zhao. Lower Bound Methods for Sign-Rank and Their Limitations. In Amit Chakrabarti and Chaitanya Swamy, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2022)*, volume 245 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 22:1–22:24, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[HSZZ06] Wei Huang, Yaoyun Shi, Shengyu Zhang, and Yufan Zhu. The communication complexity of the hamming distance problem. *Information Processing Letters*, 99(4):149–153, 2006.

[HWZ22] Nathaniel Harms, Sebastian Wild, and Viktor Zamaraev. Randomized communication and implicit graph representations. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2022, page 1220–1233, New York, NY, USA, 2022. Association for Computing Machinery.

[HZ24] Nathaniel Harms and Viktor Zamaraev. Randomized communication and implicit representations for matrices and graphs of small sign-rank. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1810–1833, 2024.

[KN96] Eyal Kushilevitz and Noam Nisan. *Communication Complexity*. Cambridge University Press, 1996.

[Lit88] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2:285–318, 1988.

[LS09] Nati Linial and Adi Shraibman. Learning complexity vs. communication complexity. *Combin. Probab. Comput.*, 18(1-2):227–245, 2009.

[LZ15] Vadim V Lozin and Victor Zamaraev. Boundary properties of factorial classes of graphs. *Journal of Graph Theory*, 78(3):207–218, 2015.

[Nis93] Noam Nisan. The communication complexity of threshold gates. *Combinatorics, Paul Erdős is Eighty*, 1:301–315, 1993.

[NP24] Moni Naor and Eugene Pekel. Adjacency sketches in adversarial environments. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1067–1098. SIAM, Philadelphia, PA, 2024.

[PSS23] Toniann Pitassi, Morgan Shirley, and Adi Shraibman. The strength of equality oracles in communication. In *14th Innovations in Theoretical Computer Science Confer-*

ence, volume 251 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages Art. No. 89, 19. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2023.

[PSW21] Toniann Pitassi, Morgan Shirley, and Thomas Watson. Nondeterministic and randomized boolean hierarchies in communication complexity. *Computational Complexity*, 30:1–48, 2021.

[RS15] Sivaramakrishnan Natarajan Ramamoorthy and Makrand Sinha. On the communication complexity of greater-than. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 442–444, 2015.

[RY20] Anup Rao and Amir Yehudayoff. *Communication Complexity and Applications*. Cambridge University Press, 2020.

[Sağ18] Mert Sağlam. Near log-convexity of measured heat in (discrete) time and consequences. In *59th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2018*, pages 967–978. IEEE Computer Soc., Los Alamitos, CA, 2018.

[Vio15] Emanuele Viola. The communication complexity of addition. *Combinatorica*, 35:703–747, 2015.

[Yao03] Andrew Chi-Chih Yao. On the power of quantum fingerprinting. In *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing*, pages 77–81. ACM, New York, 2003.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il