Electronic Colloquium on Computational Complexity, Report No. 351 (2026)

Communication Complexity of Disambiguations of Gap Hamming Distance

Yogesh Dahiya Mika Göös Nathaniel Harms Hamed Hatami
Pooya Hatami Shachar Lovett Weiqiang Yuan

Abstract

We consider the Gap Hamming Distance problem in which Alice and Bob receive n-bit strings and must decide whether their Hamming distance is at most $n/3$ or at least $2n/3$. This promise problem has public-coin randomized communication complexity $O(1)$ since the players can estimate the Hamming distance by comparing their strings on a few shared random coordinates. The required output is undefined, or "ambiguous", on inputs with Hamming distance between $n/3$ and $2n/3$.

We prove that any disambiguation of this problem to a total function has randomized communication complexity $\Omega(n^{2/3})$. As a corollary, we obtain an $O(1)$ versus $\Omega(n^{2/3})$ separation between randomized and pseudo-deterministic communication complexity. The proof was first found by an AI model and then reworked by the authors.

1 Introduction

Denote the Hamming distance between two strings $x, y \in \{\pm 1\}^n$ by $\mathrm{dist}(x, y)$, and let $\gamma \in (0, 1/2)$ be a fixed parameter. The *γ-Gap Hamming Distance problem* is the promise problem in which Alice and Bob receive x and y, respectively, and must distinguish between the cases

$$\mathrm{dist}(x, y) \le \gamma n \qquad \text{and} \qquad \mathrm{dist}(x, y) \ge (1 - \gamma)n.$$

The randomized (public-coin) communication complexity of this problem is $O_\gamma(1)$, since the players can estimate $\mathrm{dist}(x, y)$ to sufficient accuracy by comparing their bits on a constant-size set of shared random coordinates.

A *disambiguation* of a promise problem is an extension of it to a total Boolean function, that is, a total function that agrees with it on all inputs that satisfy the promise. We prove that every disambiguation of the γ-Gap Hamming Distance problem has large randomized complexity.

Theorem 1.1. *Let $\gamma \in (0, 1/2)$ be fixed. Every total function $F\colon \{\pm 1\}^n \times \{\pm 1\}^n \to \{0, 1\}$ satisfying*

$$F(x, y) = \begin{cases} 0 & \text{if } \mathrm{dist}(x, y) \le \gamma n, \\ 1 & \text{if } \mathrm{dist}(x, y) \ge (1 - \gamma)n \end{cases}$$

has randomized communication complexity $\Omega_\gamma(n^{2/3})$.

Previously, the best known lower bound for this problem was $\Omega(\log \log n)$, due to [BHH+26]. A linear lower bound $\Omega_\gamma(n)$ is conjectured in [GHR+26].

ISSN 1433-8092

Pseudo-deterministic communication. Our result is partly motivated by the study of pseudo-determinism in communication complexity. When solving a search problem, a randomized algorithm that is only required to be correct with high probability may nevertheless output different (but still correct) outputs between different independent executions. This may be undesirable behaviour, so we may demand a randomized algorithm whose outputs are consistent (with high probability) between different executions. Such an algorithm is called *pseudo-deterministic*. Pseudo-determinism was introduced by Gat and Goldwasser [GG11] (but also independently by Huynh and Nordström [HN12] in the context of proof complexity) and has since been studied across a wide range of computational models, including learning algorithms [OS18, ILPS22], sequential and parallel algorithms [GGR13, GG17], average-case and approximation algorithms [Hol17, DPV18], interactive proofs [GGH18], low-space algorithms [GL19], and streaming algorithms [GGMW20]. The central question in this line of research is to what extent pseudo-determinism can be separated from general randomized computation.

Formally, a communication *search problem* is a relation $\mathcal{R} \subseteq \{\pm 1\}^n \times \{\pm 1\}^n \times \mathcal{Z}$: Alice and Bob receive x and y, respectively, and must output some z with $(x, y, z) \in \mathcal{R}$. A *pseudo-deterministic* protocol Π for $\mathcal{R}$ is a randomized protocol such that, for every (x, y), there is a fixed value $h(x, y)$ with $(x, y, h(x, y)) \in \mathcal{R}$ such that $\Pr[\Pi(x, y) = h(x, y)] \geq \frac{2}{3}$. We call $h(x, y)$ the *canonical output*.

Although we aim to compare pseudodeterministic protocols to randomized protocols, there is no standard definition of what it means for a randomized protocol to solve a search problem. For *functions*, the standard definition is that the algorithm is required to output the unique correct answer with probability at least $2/3$, and this choice of constant $2/3$ is essentially arbitrary because we can efficiently amplify the success probability to any other constant. We cannot always amplify the success probability for search problems (because the algorithm may output different correct answers on different executions), so we cannot simply choose a fixed constant probability of success. The paper [GHR+26] discusses these nuances in more detail and surveys various definitions in the literature. However, for this paper, we can say that a randomized protocol Π solves $\mathcal{R}$ if $\Pr[(x, y, \Pi(x, y)) \in \mathcal{R}] \geq \frac{2}{3}$ for every (x, y). This is justified by the fact that the problems we study do allow efficient amplification of the success probability.

The problem of separating randomized and pseudo-deterministic communication complexity was posed in [GIPS21]. The first separations, $O(1)$ versus $\Omega(\log \log n)$ [BHH+26] and $O(\log n)$ versus $\Omega(\sqrt{n})$ [GHR+26, Wat26], were obtained only recently. Theorem 1.1 improves on both of these results, giving a separation of $O(1)$ versus $\Omega(n^{2/3})$. This is because the γ-Gap Hamming Distance problem is a relation (where the relation $\mathcal{R}$ allows either 0 or 1 outputs on distances between γn and $(1-\gamma)n$), and a pseudodeterministic algorithm for this problem is equivalent to a randomized algorithm that computes a disambiguation.

Furthermore, as an application of our main result, we obtain an $O(1)$ versus $\Omega(n^{2/3})$ separation for Hamming-distance estimation: for a fixed $\varepsilon \in (0, 1/2)$, the players must output a number t satisfying

$$|\operatorname{dist}(x, y) - t| \leq \varepsilon n. \tag{1}$$

Estimating the Hamming distance from $O_\varepsilon(1)$ shared random coordinates solves this problem with $O_\varepsilon(1)$ communication. Now suppose that Π is a pseudo-deterministic protocol with canonical output $h(x, y)$, and consider the total function

$$F(x, y) \coloneqq \mathbf{1}[h(x, y) \geq n/2].$$

Note that F satisfies the conditions of Theorem 1.1 with $\gamma = \frac{1/2-\varepsilon}{2}$. Moreover, running Π and

outputting 1 if and only if its output is at least $n/2$ computes F with error at most $1/3$. We therefore obtain the following.

Corollary 1.2. *For every fixed $\varepsilon \in (0, 1/2)$, the search problem (1) has randomized communication complexity $O_\varepsilon(1)$ and pseudo-deterministic communication complexity $\Omega_\varepsilon(n^{2/3})$.*

Parity decision trees. For $z \in \{\pm 1\}^n$, let $|z| \coloneqq |\{i : z_i = -1\}|$. For $\gamma \in (0, 1/2)$, the γ-Gap Majority problem is the promise problem $f\colon \{\pm 1\}^n \to \{0, 1, *\}$ defined by

$$f(z) = \begin{cases} 0 & \text{if } |z| \le \gamma n, \\ 1 & \text{if } |z| \ge (1-\gamma)n, \\ * & \text{otherwise.} \end{cases} \tag{2}$$

The γ-Gap Hamming Distance problem is its XOR-lift $F(x, y) = f(x \odot y)$, where $\odot$ denotes coordinatewise multiplication, the $\{\pm 1\}$ representation of XOR.

Each parity query to $x \odot y$ can be simulated with two bits of communication by exchanging the parities of x and y on the queried set. Thus a randomized parity decision tree for any disambiguation of γ-Gap Majority yields a randomized communication protocol for a disambiguation of γ-Gap Hamming Distance, with at most twice the cost. By Theorem 1.1, we obtain the following.

Corollary 1.3. *For every fixed $\gamma \in (0, 1/2)$, every disambiguation of the γ-Gap Majority problem has randomized parity decision tree complexity $\Omega_\gamma(n^{2/3})$.*

This improves Gavinsky's lower bound of $\Omega_\gamma(\sqrt{n})$ for the same problem [Gav25]. In fact, our proof of Theorem 1.1 follows his iterative strategy.

AI methodology. An AI model first produced a proof of an $\Omega(\sqrt{n})$ lower bound, and after further generic prompts, it improved the bound to $\Omega(n^{2/3})$. The authors extracted the key ideas from these AI-generated proofs and used them to write an argument in the form they found most natural and intuitive. The exposition in this note was drafted by the authors, with AI tools used for editing, typesetting, and spell-checking.

Notation. We write $\mathrm{R}_\varepsilon(F)$ for the public-coin randomized communication complexity of F with error at most ε on every input. We write $\mathcal{U}_m$ for the uniform distribution on $\{\pm 1\}^m$, and $\mathrm{Ber}(\beta)$ for the distribution on $\{\pm 1\}$ that assigns probability β to 1 and $1 - \beta$ to -1.

We use H to denote the usual Shannon entropy. The *min-entropy* of a distribution μ on a finite set Ω and its *deficit* are defined respectively as (all logarithms are to base 2)

$$H_\infty(\mu) \coloneqq -\log \|\mu\|_\infty \qquad \text{and} \qquad D_\infty(\mu) \coloneqq \log(|\Omega|\, \|\mu\|_\infty).$$

For $\rho \in [-1, 1]$, a random variable (X, Y) taking values in $\{\pm 1\}^m \times \{\pm 1\}^m$ is *ρ-correlated* if each marginal is uniform, the pairs (X_i, Y_i) are independent, and $\mathbb{E}[X_i Y_i] = \rho$ for every i. For $x \in \{\pm 1\}^m$, we write $N_\rho(x)$ for the conditional distribution of Y given $X = x$: independently in each coordinate, $Y_i = x_i$ with probability $(1 + \rho)/2$ and $Y_i = -x_i$ with probability $(1 - \rho)/2$.

2 Proof overview

We prove Theorem 1.1 by contradiction. Suppose that F has a randomized protocol with small communication cost. Replacing $F(x, y)$ by $1 - F(x, -y)$ if necessary, we may assume that

$$\mathop{\mathbb{E}}_{\mathcal{U}_n \times \mathcal{U}_n} F \geq \frac{1}{2}.$$

We follow the iterative strategy of Gavinsky [Gav25]. By exploiting the low-cost protocol for F, we progressively grow a set K of coordinates and force Alice's and Bob's inputs to agree on most of the coordinates in K, while maintaining $\mathbb{E}\, F \geq 1/2$. Eventually, every pair of inputs consistent with the forced coordinates K has Hamming distance $< \gamma n$, so F is guaranteed to be 0, which contradicts the maintained property $\mathbb{E}\, F \geq 1/2$.

More precisely, we maintain a product distribution $\mu \times \nu$ on the inputs, a set K of processed coordinates, and a reference string $a \in \{\pm 1\}^K$, with two properties:

(i) $\mathbb{E}_{\mu \times \nu}\, F \geq 1/2$;

(ii) for every (x, y) in the support of $\mu \times \nu$, both $\mathrm{dist}(x_K, a)$ and $\mathrm{dist}(y_K, a)$ are small (and therefore $\mathrm{dist}(x_K, y_K)$ is small).

We start with uniform inputs and K empty. Each round adds new coordinates to K, in two steps:

1. *Planting a mostly agreeing block (Lemma 4.1).* We choose a large block I outside K and a reference string w on it, and restrict both inputs to be close to w on I, while keeping $\mathbb{E}\, F > 1/4$. We then add I to K and extend a by w.

2. *Restoring the expectation (Lemma 4.2).* We use the low-cost protocol for F to find a rectangle (leaf of the protocol) of sufficiently large probability on which $\mathbb{E}\, F \geq 1/2$. Conditioning on this rectangle restores the expectation.

Eventually, every pair of inputs in the support differs in at most γn coordinates. Then $F = 0$ on the entire support, contradicting (i).

To see the idea, suppose first that X and Y are uniform. We replace them with a ρ-correlated pair, obtained by choosing a random set $I \subseteq [n]$, including each coordinate independently with probability ρ, and sampling X, Y uniformly subject to $X_I = Y_I$. The following theorem shows that, for $\rho = \Theta(d^{-1/2})$, this changes the acceptance probability of the protocol by only a small constant.

Theorem 2.1 (Girish–Sinha–Tal–Wu, [GSTW23, Theorem 1.5])**.** *For every randomized protocol Π with output in $\{0, 1\}$ that communicates at most d bits, and every $\rho \in [-1, 1]$,*

$$\left| \Pr_{X,Y \sim \mathcal{U}_m} [\Pi(X, Y) = 1] - \Pr_{X \sim \mathcal{U}_m,\, Y \sim N_\rho(X)} [\Pi(X, Y) = 1] \right| \leq O\big(|\rho| \sqrt{d}\big).$$

Since the protocol has small error, the expectation of F also changes by only a small constant. By averaging, we can then fix a block I of size $\Omega(\rho n)$ and a common value $w = X_I = Y_I$ while maintaining $\mathbb{E}\, F > 1/4$.

After we restrict to a rectangle to restore the expectation, the inputs are no longer uniform, so Theorem 2.1 does not apply directly. To handle this, we use a coupling (Lemma 3.1): any random

string X of high min-entropy can be coupled with a uniform random string U so that X and U are close in expected Hamming distance. Applying this separately to Alice's and Bob's inputs gives uniform proxy strings U and V, and we can apply the correlation argument above to (U, V) instead of inputs. Since each input is close to its proxy string, both inputs are close to $w = U_I = V_I$ on the block I.

The coupling also guarantees that, after fixing (I, w) and discarding a small exceptional set, the inputs on the remaining coordinates still have high min-entropy. Moreover, conditioned on (I, w), Alice's and Bob's inputs are independent.

3 Coupling with the uniform distribution

The following lemma couples a distribution μ on $\{\pm 1\}^m$ of high min-entropy with the uniform distribution on $\{\pm 1\}^m$, using a sequential maximal coupling due to Marton [Mar86]. The resulting pair (X, U), with $X \sim \mu$ and U uniform, has two properties: (a) X and U are close in expected Hamming distance, and (b) for every set I of coordinates, revealing U_I does not significantly increase the probabilities of individual values of $X_{\bar{I}}$, except on a small exceptional set.

Lemma 3.1 (Uniform coupling)**.** *Let μ be a distribution on $\{\pm 1\}^m$ with $D_\infty(\mu) \le D$. There is a coupling (X, U) of μ and $\mathcal{U}_m$ with the following properties:*

(a) $\mathbb{E}\big[\mathrm{dist}(X, U)\big] \le \sqrt{mD}$.

(b) For every $I \subseteq [m]$ and $s > 0$, the following holds. For $u \in \{\pm 1\}^I$, let ν_u be the distribution of $(X_{\bar{I}} \mid U_I = u)$. Then

$$\Pr_{(X,U)}\Big[\nu_{U_I}(X_{\bar{I}}) > 2^{-(m-|I|)+D+s}\Big] \le 2^{-s}. \tag{3}$$

Proof. For each $i \in [m]$, let

$$a_i(x_{<i}) \coloneqq \mathbb{E}_{X\sim\mu}\big[X_i \;\big|\; X_{<i} = x_{<i}\big].$$

Take independent $T_1, \dots, T_m$, each uniform on $[-1, 1]$, and for $i = 1, \dots, m$ in turn, set

$$X_i = 1 \iff T_i \le a_i(X_{<i}), \qquad U_i = 1 \iff T_i \le 0.$$

Since $\Pr[T_i \le a] = (1 + a)/2$, we have $X \sim \mu$ and $U \sim \mathcal{U}_m$. Moreover,

$$\Pr[X_i \ne U_i \mid X_{<i}] = \frac{|a_i(X_{<i})|}{2}.$$

Part (a). Conditioned on $X_{<i} = x_{<i}$, the bit X_i has distribution $\mathrm{Ber}\big(\frac{1+a_i(x_{<i})}{2}\big)$. By the chain rule for entropy and the elementary bound $H\big(\mathrm{Ber}(\frac{1+a}{2})\big) \le 1 - a^2/4$ for $a \in [-1, 1]$,

$$H(X) = \sum_{i=1}^m H(X_i \mid X_{<i}) \le m - \frac{1}{4}\sum_{i=1}^m \mathbb{E}\big[a_i(X_{<i})^2\big].$$

Rearranging, and using $H(X) \ge H_\infty(X) = m - D_\infty(\mu) \ge m - D$, we obtain

$$\frac{1}{4}\sum_{i=1}^m \mathbb{E}\big[a_i(X_{<i})^2\big] \le m - H(X) \le D. \tag{4}$$

Hence, by Cauchy–Schwarz and (4),

$$\mathbb{E}\big[\mathrm{dist}(X,U)\big] = \frac{1}{2}\sum_{i=1}^{m}\mathbb{E}\big|a_i(X_{<i})\big| \le \sqrt{mD}.$$

Part (b). Fix $I \subseteq [m]$ and let $Z \coloneqq (U_I, X_{\bar{I}})$. Since U_I is uniform,

$$\mathbb{E}\big[2^{m-|I|}\,\nu_{U_I}(X_{\bar{I}})\big] = 2^{m-|I|}\sum_{u,y}\Pr[Z=(u,y)]\,\frac{\Pr[Z=(u,y)]}{\Pr[U_I=u]} = 2^m\sum_{z}\Pr[Z=z]^2.$$

As $2^m\,\|\mu\|_\infty \le 2^D$, Markov's inequality reduces (3) to showing that the collision probability of Z satisfies

$$\mathrm{col}(Z) \coloneqq \sum_{z}\Pr[Z=z]^2 \le \|\mu\|_\infty . \tag{5}$$

We prove (5) by induction on m. For $a \in \mathrm{supp}(X_1)$, let $p_a \coloneqq \Pr[X_1 = a]$ and let μ_a be the distribution of $(X_{>1} \mid X_1 = a)$. Then $\|\mu_a\|_\infty \le \|\mu\|_\infty / p_a$.

Note that, conditioned on any $(X_1, U_1) = (a, b)$ of positive probability, the pair $(X_{>1}, U_{>1})$ is the coupling above for μ_a. So by induction, the conditional distribution of $Z_{>1}$ has collision probability at most $\|\mu\|_\infty / p_a$.

In the case $1 \notin I$, we have $Z_1 = X_1$, so

$$\mathrm{col}(Z) = \sum_a p_a^2\,\mathrm{col}(Z_{>1} \mid X_1 = a) \le \sum_a p_a^2 \cdot \frac{\|\mu\|_\infty}{p_a} = \|\mu\|_\infty .$$

If $1 \in I$, then $Z_1 = U_1$ is uniform. By convexity of collision probability and the induction hypothesis,

$$\begin{aligned}\mathrm{col}(Z) &= \frac{1}{2}\,\mathbb{E}_{U_1}\big[\mathrm{col}(Z_{>1} \mid U_1)\big] \le \frac{1}{2}\,\mathbb{E}_{X_1,U_1}\big[\mathrm{col}(Z_{>1} \mid X_1, U_1)\big] \\ &\le \frac{\|\mu\|_\infty}{2}\,\mathbb{E}_{X_1}\Big[\frac{1}{p_{X_1}}\Big] \le \|\mu\|_\infty ,\end{aligned}$$

where the last inequality follows from $\mathbb{E}[1/p_{X_1}] = |\,\mathrm{supp}(X_1)| \le 2$. This completes the proof of (5), and hence of part (b). □

4 Proof of Theorem 1.1

Throughout this section, $d \coloneqq \mathrm{R}_{1/100}(F)$, and $\rho \coloneqq c_0/\sqrt{d}$, where $c_0 > 0$ is an absolute constant small enough that the bound in Theorem 2.1 is at most $1/100$. For a distribution μ on $\{\pm1\}^n$ and a set $J \subseteq [n]$, we write μ_J for the marginal of μ on J.

Lemma 4.1 (Planting a mostly agreeing block)**.** *Let $\mu \times \nu$ be a product distribution on $\{\pm1\}^n \times \{\pm1\}^n$ with $\mathbb{E}_{\mu\times\nu} F \ge 1/2$. Let $J \subseteq [n]$ be a set of $m \ge 100/\rho$ coordinates such that μ_J and ν_J have deficit at most D, and set $b \coloneqq 20\rho\sqrt{mD}$.*

Then there exist a set $I \subseteq J$ of size at least $\rho m/2$, a string $w \in \{\pm1\}^I$, and a product distribution $\widetilde{\mu} \times \widetilde{\nu}$, supported within the support of $\mu \times \nu$, such that:

(i) every $(x, y) \in \mathrm{supp}(\widetilde{\mu} \times \widetilde{\nu})$ satisfies $\mathrm{dist}(x_I, w) \le b$ and $\mathrm{dist}(y_I, w) \le b$;

(ii) $\mathbb{E}_{\widetilde{\mu}\times\widetilde{\nu}} F > 1/4$;

(iii) $\widetilde{\mu}_{J\setminus I}$ *and* $\widetilde{\nu}_{J\setminus I}$ *have deficit at most* $D+7$.

Proof. Apply Lemma 3.1 to μ_J and ν_J to obtain couplings (X_J, U_J) and (Y_J, V_J), where U_J and V_J are uniform on $\{\pm 1\}^J$. Consider the following protocol Π of cost d, in which Alice and Bob receive $u, v \in \{\pm 1\}^J$, respectively:

- Alice samples X_J from the coupling conditioned on $U_J = u$, and then samples $X_{[n]\setminus J}$ from μ conditioned on X_J.
- Bob does the same with $V_J = v$ to obtain Y, using independent randomness.
- They run the protocol for F of cost d and error at most $1/100$ on (X, Y).

We run Π on inputs (u, v) drawn from two different distributions. First, let (U'_J, V'_J) be independent and uniform, and let X' and Y' be the strings that Alice and Bob sample on input (U'_J, V'_J); then $(X', Y') \sim \mu \times \nu$. Second, let (U_J, V_J) be ρ-correlated, and let X and Y be the strings sampled on input (U_J, V_J). Note that X and Y are not independent in this case.

By Theorem 2.1, the acceptance probabilities of Π in the two runs differ by at most $1/100$. Since the original protocol for F errs with probability at most $1/100$ on every input, we obtain

$$\mathbb{E}\, F(X,Y) \geq \mathbb{E}\, \Pi(U_J, V_J) - \frac{1}{100} \geq \mathbb{E}\, \Pi(U'_J, V'_J) - \frac{2}{100} \geq \mathbb{E}\, F(X', Y') - \frac{3}{100} \geq \frac{47}{100}. \tag{6}$$

A ρ-correlated pair (U_J, V_J) can be sampled by first choosing a random set $I \subseteq J$ that includes each coordinate independently with probability ρ, then sampling $V_I = U_I$ uniformly, and finally sampling $U_{J\setminus I}$ and $V_{J\setminus I}$ independently and uniformly.

Write $R := (I, U_I)$, let p_R and q_R be the conditional distributions of $X_{J\setminus I}$ and $Y_{J\setminus I}$ given R, and put $M_I := 2^{-(m-|I|)+D+5}$. Consider the events

$$\mathcal{A} := \big\{\operatorname{dist}(X_I, U_I) \leq b,\ p_R(X_{J\setminus I}) \leq M_I\big\},$$
$$\mathcal{B} := \big\{\operatorname{dist}(Y_I, V_I) \leq b,\ q_R(Y_{J\setminus I}) \leq M_I\big\}.$$

We call a value (I_0, w) of R *good* if

$$|I_0| \geq \rho m/2 \qquad \text{and} \qquad \mathbb{E}\big[F(X,Y)\,\mathbf{1}_{\mathcal{A}\cap\mathcal{B}} \;\big|\; R = (I_0, w)\big] > \frac{1}{4}. \tag{7}$$

We first show that a good value exists, and then use it to construct $\widetilde{\mu}$ and $\widetilde{\nu}$.

A good R exists. We bound the probability that $\mathcal{A}$ or $\mathcal{B}$ fails. Since (X_J, U_J) is independent of I and each coordinate belongs to I with probability ρ, by Lemma 3.1 (a),

$$\mathbb{E}\big[\operatorname{dist}(X_I, U_I)\big] = \rho\, \mathbb{E}\big[\operatorname{dist}(X_J, U_J)\big] \leq \rho\sqrt{mD},$$

so, since $b = 20\rho\sqrt{mD}$, Markov's inequality gives $\Pr[\operatorname{dist}(X_I, U_I) > b] \leq 1/20$ (where we may assume $D > 0$). Applying Lemma 3.1 (b) with $s = 5$ for each fixed value of I gives $\Pr[p_R(X_{J\setminus I}) > M_I] \leq 1/32$. By the union bound, and the same argument applied to (Y_J, V_J), we obtain

$$\Pr[\mathcal{A}^c] \leq \frac{1}{20} + \frac{1}{32} \qquad \text{and} \qquad \Pr[\mathcal{B}^c] \leq \frac{1}{20} + \frac{1}{32}.$$

Finally, by the Chernoff bound, $\Pr[|I| < \rho m/2] \le e^{-\rho m/8} < 1/100$.

Let $\mathcal{G} \coloneqq \{|I| \ge \rho m/2\} \cap \mathcal{A} \cap \mathcal{B}$. By the union bound, $\Pr[\mathcal{G}^c] < \frac{1}{5}$, so by (6),

$$\mathbb{E}\big[F(X,Y)\,\mathbf{1}_{\mathcal{G}}\big] \ge \frac{47}{100} - \frac{1}{5} > \frac{1}{4}.$$

Averaging over R, some value (I_0, w) of R satisfies $\mathbb{E}\big[F(X,Y)\,\mathbf{1}_{\mathcal{G}} \;\big|\; R = (I_0,w)\big] > \frac{1}{4}$, and such a value satisfies the goodness condition (7).

Completing the proof. Fix a good value (I_0, w) of R. Conditioned on $R = (I_0, w)$, the inputs X and Y are independent. Let

$$\alpha \coloneqq \Pr\big[\mathcal{A} \;\big|\; R = (I_0,w)\big], \qquad \beta \coloneqq \Pr\big[\mathcal{B} \;\big|\; R = (I_0,w)\big],$$

and let $\widetilde{\mu}$ and $\widetilde{\nu}$ be the distributions of X given $R = (I_0, w)$ and $\mathcal{A}$, and of Y given $R = (I_0, w)$ and $\mathcal{B}$, respectively.

Property (i) holds by the definition of $\mathcal{A}$ and $\mathcal{B}$, since $U_{I_0} = V_{I_0} = w$. For (ii), note that

$$\frac{1}{4} < \mathbb{E}\big[F(X,Y)\,\mathbf{1}_{\mathcal{A}\cap\mathcal{B}} \;\big|\; R = (I_0,w)\big] = \alpha\beta\;\mathbb{E}_{\widetilde{\mu}\times\widetilde{\nu}}\,F \le \min\Big\{\alpha,\ \beta,\ \mathbb{E}_{\widetilde{\mu}\times\widetilde{\nu}}\,F\Big\},$$

which also gives $\alpha, \beta > 1/4$.

For (iii), let p be the distribution of $X_{J\setminus I_0}$ given $R = (I_0, w)$. Then every $z \in \{\pm 1\}^{J\setminus I_0}$ satisfies

$$\widetilde{\mu}_{J\setminus I_0}(z) \le \frac{p(z)\,\mathbf{1}_{\{p(z)\le M_{I_0}\}}}{\alpha} < 4M_{I_0} = 2^{-(m-|I_0|)+D+7},$$

and the same holds for $\widetilde{\nu}$. □

Lemma 4.2 (Restoring the expectation)**.** *Let ξ be any distribution with $\mathbb{E}_\xi F \ge 1/4$. Then there is a rectangle $A \times B$ such that*

$$\xi(A\times B) \ge \frac{2^{-d}}{5} \qquad \textit{and} \qquad \mathbb{E}_\xi[F \mid A\times B] \ge \frac{1}{2}.$$

Proof. Fix the randomness of a protocol for F of cost d and error at most $1/100$ to obtain a deterministic protocol with probability of error at most $1/100$ under ξ. Let $\mathcal{L}$ be the set of leaves on which it outputs 1; recall that each leaf is a rectangle $A \times B$. Without loss of generality, we may assume that $\mathbb{E}_\xi[F \mid L] \ge 1/2$ for all $L \in \mathcal{L}$ (since otherwise we may flip the output of the protocol on L while reducing the probability of error). Since the probability of error is at most $1/100$ and $\mathbb{E}_\xi F \ge 1/4$, we have $\sum_{L\in\mathcal{L}} \xi(L) \ge \frac{1}{4} - \frac{1}{100} \ge \frac{1}{5}$. Since $|\mathcal{L}| \le 2^d$, there is a leaf L with $\xi(L) \ge 2^{-d}/5$. □

Remark 4.3. Applying Lemma 4.2 with a product distribution $\mu\times\nu$ and conditioning on $A\times B$ gives the product distribution $\mu(\cdot \mid A) \times \nu(\cdot \mid B)$. Since $\mu(A), \nu(B) \ge 2^{-d}/5$, conditioning multiplies every probability by at most $5 \cdot 2^d$. Hence, for every $J \subseteq [n]$, the deficits of the marginals on J increase by at most $d + \log 5 < d + 3$.

Proof of Theorem 1.1. As in the overview, assume $\mathbb{E}_{\mathcal{U}_n \times \mathcal{U}_n} F \geq 1/2$. Note that $d \geq 1$, and suppose, towards a contradiction, that $d \leq c_\gamma n^{2/3}$ for a sufficiently small constant $c_\gamma > 0$.

For each step $t \in \{0, \dots, T\}$ (with T determined below), we maintain a product distribution $\mu_t \times \nu_t$ with $\mathbb{E}_{\mu_t \times \nu_t} F \geq 1/2$, a set $K_t \subseteq [n]$ of planted coordinates with a reference string $a_t \in \{\pm 1\}^{K_t}$, and numbers Δ_t and D_t such that every (x, y) in the support satisfies $\mathrm{dist}(x_{K_t}, a_t), \mathrm{dist}(y_{K_t}, a_t) \leq \Delta_t$, and the marginals of μ_t and ν_t on $J_t := [n] \setminus K_t$ have deficit at most D_t. Write $m_t := |J_t|$. Initially, $K_0 = \varnothing$, $\mu_0 = \nu_0 = \mathcal{U}_n$, and $\Delta_0 = D_0 = 0$.

As long as $m_t > \gamma n/2$, we apply Lemma 4.1 with $J = J_t$, followed by Lemma 4.2, and add the new block I to K_t to obtain K_{t+1}, and we extend a_t by w to obtain a_{t+1}. Since both lemmas only shrink the support, the distance bounds on K_t are preserved, and by Lemma 4.1 (iii) and Remark 4.3,

$$m_{t+1} \leq (1 - \rho/2)\, m_t, \qquad D_{t+1} \leq D_t + d + 10, \qquad \Delta_{t+1} \leq \Delta_t + 20\rho\sqrt{m_t D_t}.$$

Consequently, the process stops after $T \leq \lceil (2/\rho) \ln(2/\gamma) \rceil = O_\gamma(1/\rho)$ rounds. Moreover, $D_t \leq t(d + 10)$ and $m_t \leq n$, so

$$\Delta_T \leq 20\rho\sqrt{n(d+10)}\, T^{3/2} = O_\gamma\Big(\sqrt{n(d+10)/\rho}\Big) = O_\gamma\big(\sqrt{n}\, d^{3/4}\big).$$

Since $\rho m_t > c_0 \gamma n/(2\sqrt{d}) \geq c_0 \gamma n^{2/3}/(2\sqrt{c_\gamma})$ before stopping, the condition $\rho m_t \geq 100$ of Lemma 4.1 holds for sufficiently large n.

By choosing c_γ sufficiently small, we obtain $\Delta_T \leq \gamma n/4$. Thus every pair (x, y) in the final support satisfies

$$\mathrm{dist}(x, y) \leq 2\Delta_T + m_T \leq \gamma n,$$

forcing $F(x, y) = 0$. This contradicts $\mathbb{E}_{\mu_T \times \nu_T} F \geq 1/2$. Hence $\mathrm{R}_{1/100}(F) = \Omega_\gamma(n^{2/3})$. □

Acknowledgements

H.H. wishes to thank Amir Yehudayoff and Arkadev Chattopadhyay for earlier discussions on attempts to solve this problem.

M.G. and W.Y. are supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number MB22.00026. H.H. and N.H. are supported by NSERC Discovery Grants. Y.D. and S.L. are supported by Simons Investigator Award #929894. S.L. is also supported by NSF Award CCF-2425349.

References

[BHH+26] Ari Blondal, Hamed Hatami, Pooya Hatami, Chavdar Lalov, and Sivan Tretiak. Borsuk–Ulam and replicable learning of large-margin halfspaces. In *Proceedings of the 58th Annual ACM Symposium on Theory of Computing (STOC)*, pages 529–540, 2026. doi:10.1145/3798129.3800771.

[DPV18] Peter Dixon, A. Pavan, and N. V. Vinodchandran. On pseudodeterministic approximation algorithms. In *43rd International Symposium on Mathematical Foundations of Computer Science (MFCS 2018)*, volume 117 of *LIPIcs*, pages 61:1–61:11, 2018. doi:10.4230/LIPIcs.MFCS.2018.61.

[Gav25] Dmytro Gavinsky. Unambiguous parity-query complexity. *Random Structures & Algorithms*, 66(3):e70010, 2025. doi:10.1002/rsa.70010.

[GG11] Eran Gat and Shafi Goldwasser. Probabilistic search algorithms with unique answers and their cryptographic applications. Technical Report TR11-136, Electronic Colloquium on Computational Complexity, 2011. URL: https://eccc.weizmann.ac.il/report/2011/136/.

[GG17] Shafi Goldwasser and Ofer Grossman. Bipartite perfect matching in pseudo-deterministic NC. In *44th International Colloquium on Automata, Languages, and Programming (ICALP 2017)*, volume 80 of *LIPIcs*, pages 87:1–87:13, 2017. doi:10.4230/LIPIcs.ICALP.2017.87.

[GGH18] Shafi Goldwasser, Ofer Grossman, and Dhiraj Holden. Pseudo-deterministic proofs. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, volume 94 of *LIPIcs*, pages 17:1–17:18, 2018. doi:10.4230/LIPIcs.ITCS.2018.17.

[GGMW20] Shafi Goldwasser, Ofer Grossman, Sidhanth Mohanty, and David P. Woodruff. Pseudo-deterministic streaming. In *11th Innovations in Theoretical Computer Science Conference (ITCS)*, volume 151 of *LIPIcs*, pages 79:1–79:25, 2020. doi:10.4230/LIPIcs.ITCS.2020.79.

[GGR13] Oded Goldreich, Shafi Goldwasser, and Dana Ron. On the possibilities and limitations of pseudodeterministic algorithms. In *Proceedings of the 4th Innovations in Theoretical Computer Science Conference (ITCS 2013)*, pages 127–138, 2013. doi:10.1145/2422436.2422453.

[GHR+26] Mika Göös, Nathaniel Harms, Artur Riazanov, Anastasia Sofronova, Dmitry Sokolov, and Weiqiang Yuan. Pseudodeterministic communication complexity. In *Proceedings of the 58th Annual ACM Symposium on Theory of Computing (STOC)*, pages 2030–2039, 2026. doi:10.1145/3798129.3800907.

[GIPS21] Shafi Goldwasser, Russell Impagliazzo, Toniann Pitassi, and Rahul Santhanam. On the pseudo-deterministic query complexity of NP search problems. In *36th Computational Complexity Conference (CCC)*, volume 200 of *LIPIcs*, pages 36:1–36:22, 2021. doi:10.4230/LIPIcs.CCC.2021.36.

[GL19] Ofer Grossman and Yang P. Liu. Reproducibility and pseudo-determinism in log-space. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2019)*, pages 606–620, 2019. doi:10.1137/1.9781611975482.38.

[GSTW23] Uma Girish, Makrand Sinha, Avishay Tal, and Kewen Wu. Fourier growth of communication protocols for XOR functions. In *Proceedings of the 64th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 721–732, 2023. doi:10.1109/FOCS57990.2023.00047.

[HN12] Trinh Huynh and Jakob Nordström. On the virtue of succinct proofs: Amplifying communication complexity hardness to time-space trade-offs in proof complexity. In *Proceedings of the 44th Symposium on Theory of Computing (STOC)*, pages 233–248, New York, NY, USA, 2012. ACM. doi:10.1145/2213977.2214000.

[Hol17] Dhiraj Holden. A note on unconditional subexponential-time pseudo-deterministic algorithms for BPP search problems, 2017. URL: https://arxiv.org/abs/1707.05808, arXiv:1707.05808.

[ILPS22] Russell Impagliazzo, Rex Lei, Toniann Pitassi, and Jessica Sorrell. Reproducibility in learning. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC 2022)*, pages 818–831, 2022. doi:10.1145/3519935.3519973.

[Mar86] Katalin Marton. A simple proof of the blowing-up lemma. *IEEE Transactions on Information Theory*, 32(3):445–446, 1986. doi:10.1109/tit.1986.1057176.

[OS18] Igor Carboni Oliveira and Rahul Santhanam. Pseudo-derandomizing learning and approximation. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018)*, volume 116 of *LIPIcs*, pages 55:1–55:19, 2018. doi:10.4230/LIPIcs.APPROX-RANDOM.2018.55.

[Wat26] Thomas Watson. Pseudodeterminism and MA $\neq$ NP$^{\text{BPP}}$ in communication complexity. Technical report, arXiv, 2026. doi:10.48550/ARXIV.2608.26425.

Yogesh Dahiya
University of California San Diego, La Jolla, California, USA
E-mail address: ydahiya@ucsd.edu

Mika Göös
EPFL, Lausanne, Switzerland
E-mail address: mika.goos@epfl.ch

Nathaniel Harms
University of British Columbia, Vancouver, Canada
E-mail address: nharms@cs.ubc.ca

Hamed Hatami
McGill University, Montréal, Canada
E-mail address: hatami@cs.mcgill.ca

Pooya Hatami
The Ohio State University, Columbus, Ohio, USA
E-mail address: hatami.2@osu.edu

Shachar Lovett
University of California San Diego, La Jolla, California, USA
E-mail address: slovett@ucsd.edu

Weiqiang Yuan
EPFL, Lausanne, Switzerland, and University of Copenhagen, Copenhagen, Denmark
E-mail address: weiqiang.yuan@di.ku.dk

ECCC ISSN 1433-8092
https://eccc.weizmann.ac.il