SIAM J. COMPUT.
Vol. 15, No. 4, November 1986

003

COMMUNICATION COMPLEXITY OF COMPUTING THE HAMMING DISTANCE*

KING F. PANG† AND ABBAS EL GAMAL‡

Abstract. Let $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$. Persons A and B are given $\mathbf{x}$ and $\mathbf{y}$ respectively. They communicate in order that both find the Hamming Distance $d_H^n(\mathbf{x},\mathbf{y})$. Three communication models, viz, deterministic, ε-error and ε-randomized, are considered. Let $C(d_H^n)$, $C_\varepsilon(d_H^n)$ and $D_\varepsilon(d_H^n)$ be the respective minimum number of bits that must be communicated under the three models. It is shown that

$$n+\log(n+1-\sqrt{n}) \leqq C(d_H^n) \leqq n+\lceil \log(n+1) \rceil.$$

It is also shown that both $C_\varepsilon(d_H^n)$ and $D_\varepsilon(d_H^n)$ are lower bounded by $\Omega(n)$, thus solving an open problem posed by Yao.

Key words. communication complexity, randomized protocol, Hamming distance, combinatorial extremal problem

AMS(MOS) subject classifications. 05, 68

1. Introduction. Let $\mathscr{X}$, $\mathscr{Y}$ and $\mathscr{Z}$ be three finite sets and $f: \mathscr{X} \times \mathscr{Y} \to \mathscr{Z}$. Person A is given $\mathbf{x} \in \mathscr{X}$ and person B is given $\mathbf{y} \in \mathscr{Y}$. They communicate according to an agreed-upon protocol, with the objective of computing $f(\mathbf{x},\mathbf{y})$. We consider three communication models which differ in the types of protocols employed and the level of correctness of the computation.

(i) *Deterministic model*: When A (or B) transmits, his message is a function of $\mathbf{x}$ (or $\mathbf{y}$) and all the previous messages. When the communication terminates, both A and B are required to know the correct value of $f(\mathbf{x},\mathbf{y})$, for all $(\mathbf{x},\mathbf{y}) \in \mathscr{X} \times \mathscr{Y}$. $C(f)$ is the minimum (over all deterministic protocols that satisfy the error-free requirement) number of bits communicated under the worst case input.

(ii) *ε-error model*: The ε-error model is deterministic in the sense of (i). However, when it terminates, both A and B are allowed to arrive at an incorrect value of $f(\mathbf{x},\mathbf{y})$, for as many as $\varepsilon \cdot \|\mathscr{X}\| \cdot \|\mathscr{Y}\|$ (arbitrary) pairs $(\mathbf{x},\mathbf{y}) \in \mathscr{X} \times \mathscr{Y}$. The ε-error communication complexity of f, $C_\varepsilon(f)$, is then similarly defined as $C(f)$, where the minimization is over all deterministic protocols satisfying the ε-error requirement. With a uniform density on $\mathscr{X} \times \mathscr{Y}$, the average case ε-error complexity $\bar{C}_\varepsilon(f)$ can also be defined.

(iii) *ε-randomized model*: When A (or B) transmits, he chooses randomly from a set of messages. The messages in this set and the probability density on it are specified by $\mathbf{x}$ (or $\mathbf{y}$) and the messages already transmitted. The error requirement is the following: averaged over all the possible sequences of messages sent during the communication, *for all* inputs $(\mathbf{x},\mathbf{y}) \in \mathscr{X} \times \mathscr{Y}$, the probability that the end result is different from $f(\mathbf{x},\mathbf{y})$ is no more than a constant $0 \leqq \varepsilon \leqq 1$. With a uniform density on $\mathscr{X} \times \mathscr{Y}$, let $D_P(f)$ be the average (over all random outcomes) number of bits communicated in protocol P averaged over all inputs. The ε-randomized communication complexity of f, $D_\varepsilon(f)$, is then defined as the minimum of $D_P(f)$, over all protocols that satisfy the ε-error condition.

* Received by the editors May 29, 1984, and in revised form July 2, 1985.

† Information Systems Laboratory, Electrical Engineering Department, Stanford University, Stanford, California 94305. The work of this author was partially supported by the National Science Foundation under contract 80-26102.

‡ Information Systems Laboratory, Electrical Engineering Department, Stanford University, Stanford, California 94305. The work of this author was partially supported by the Defense Advanced Research Projects Agency under contract MDA-0680 and by the U.S. Air Force under contract F49620-79C-0058.

In this paper, we examine the communication complexity of the Hamming distance function according to the three models. The Hamming distance between $\mathbf{x}, \mathbf{y} \in \{0, 1\}^n$ is defined as

$$d_H^n(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} 1(x_i \neq y_i),$$

where $1(\cdot)$ is the indicator function. When the lengths of $\mathbf{x}$ and $\mathbf{y}$ are not explicitly specified, the notation $d_H(\mathbf{x}, \mathbf{y})$ is used for their Hamming distance. In § 2, we consider the deterministic model. Section 2.1 contains the formal definition of a deterministic protocol and upper and lower bounds to $C(f)$ for an arbitrary function f are shown in § 2.2 (Theorems 2.1 and 2.2). In § 2.3, we prove a lower bound on $C(d_H^n)$ that is at most one bit less than the upper bound (Theorem 2.3). As a by-product of this result, we solve an independently interesting two-family extremal combinatorial problem (Theorem 2.4 and Corollary 2.3). Section 3 is concerned with the ε-error model. We first formally define an ε-error protocol. The communication complexity (under all three models) of the Hamming distance function is then related to those of the inner product function and the parity of the inner product function (Lemma 3.2). By proving a lower bound on the ε-error communication complexity of the latter (Lemma 3.3), we prove an $\Omega(n)$ lower bound for $C_\varepsilon(d_H^n)$ (Theorem 3.1) and $\bar{C}_\varepsilon(d_H^n)$ (Theorem 3.2). In § 4, we combine the result of § 3 and a Theorem of Yao [Yao1] to show that $D_\varepsilon(d_H^n) = \Omega(n)$ (Theorem 4.1), thus solving one of the open problems posed in [Yao3].

2. Deterministic model.

2.1. Formal definition. To formally define a deterministic protocol for f, we need the following definitions:

DEFINITION 2.1 [Yao2]. A *monochromatic rectangle* (*m-rect*) is a product set $\mathcal{U} \times \mathcal{V}$ where $\mathcal{U} \subseteq \mathcal{X}$, $\mathcal{V} \subseteq \mathcal{Y}$ such that $f(\mathbf{a}, \mathbf{b})$ is constant for all $\mathbf{u} \in \mathcal{U}$ and $\mathbf{v} \in \mathcal{V}$. An $m(f)$-partition is a partition of $\mathcal{X} \times \mathcal{Y}$ into m-rect's and $k(f)$ is defined as the minimum number of m-rect's over all $m(f)$-partitions of $\mathcal{X} \times \mathcal{Y}$.

DEFINITION 2.2. Denote $\mathcal{X}$ and $\mathcal{Y}$ the *row-projection* and *column-projection* of the product set $\mathcal{X} \times \mathcal{Y}$. The pair of product sets $(\mathcal{X}' \times \mathcal{Y}', \mathcal{X}'' \times \mathcal{Y}'')$ is called a *row-partition* of $\mathcal{X} \times \mathcal{Y}$ if $\mathcal{X}' = \mathcal{X}'' = \mathcal{X}$ and $\mathcal{Y}', \mathcal{Y}''$ partition $\mathcal{Y}$. A *Column-partition* is defined similarly. A *decomposition tree* (d-tree) for $\mathcal{X} \times \mathcal{Y}$ is a binary tree whose nodes are product sets $\subseteq \mathcal{X} \times \mathcal{Y}$. Each internal node is the *disjoint* union of its children. The root of the tree is $\mathcal{X} \times \mathcal{Y}$ and the leaves are m-rect's of f. It is clear that since the tree is binary each node is either row- or column-partitioned by its children.

Given a d-tree for f, we label the children of each node "0" and "1" and associate with it a protocol P as follows: At each step of the communication A and B consider one node in the tree (the first node being the root). If the node is column-partitioned by its children, A transmits the label of the child whose row-projection contains $\mathbf{x}$. If the node is row-partitioned, B transmits the label of the node whose column-projection contains $\mathbf{y}$. Next, A and B move to the node whose label was transmitted and repeat the process. The communication terminates when they arrive at a leaf and obtain the value of $f(\mathbf{x}, \mathbf{y})$.

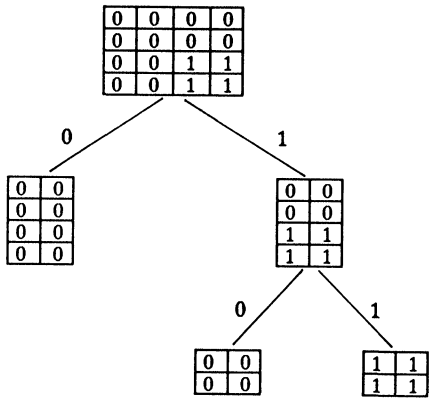
An easy induction (on the number of bits communicated) shows that at each step:

(i) A and B consider the same node of the tree.

(ii) This node contains $(\mathbf{x}, \mathbf{y})$.

(iii) If the node is internal then exactly one of its children contains $(\mathbf{x}, \mathbf{y})$.

An example of a d-tree for a function is shown in Fig. 2.1. Suppose in the d-tree for protocol P, the length of the (unique) path joining the root and the leaf containing



† $C_P = 2$ bits for this protocol.

FIG. 2.1. *An example of a protocol.*

$(\mathbf{x}, \mathbf{y})$ is $C_P(\mathbf{x}, \mathbf{y})$. The complexity of the communication protocol P is defined as

$$C_P(f) \triangleq \max_{(\mathbf{x},\mathbf{y}) \in \mathscr{X} \times \mathscr{Y}} C_P(\mathbf{x}, \mathbf{y}),$$

and the communication complexity of f is defined as

$$C(f) \triangleq \min_P C_P(f).$$

Remark. In our model, the transfer of information occurs in both directions, but *not* simultaneously. Since we are only concerned with the number of bits exchanged, it is straightforward to show that the lower bounds proved in this paper are not affected by this restriction.

2.2. General bounds for $C(f)$. A simple upper bound for $C(f)$ can be achieved by the following algorithm: Using $\lceil \log \|\mathscr{X}\| \rceil$[1] bits, A communicates $\mathbf{x}$ to B, who computes $f(\mathbf{x}, \mathbf{y})$. Another $\lceil \log \|\mathscr{Z}\| \rceil$ bits are then sufficient for B to inform A of the result. We therefore have the following theorems.

THEOREM 2.1. $C(f) \leqq \lceil \log \|\mathscr{X}\| \rceil + \lceil \log \|\mathscr{Z}\| \rceil$.

There are two general techniques for proving lower bounds for $C(f)$ for an arbitrary function f. The first one [Yao2] is based on $m(f)$-partitions of $\mathscr{X} \times \mathscr{Y}$. The other lower bound [MS] is obtained from the rank of the function table of f, which is being considered as an $\|\mathscr{X}\| \times \|\mathscr{Y}\|$ matrix. A statement of the first lower bound, according to our model, is the following:

THEOREM 2.2. $C(f) \geqq \lceil \log (k(f)) \rceil$.

Proof. The theorem follows from two properties of the d-tree corresponding to the protocol P:

(i) All the leaves of the d-tree are m-rect's (otherwise the result of the communication is not always correct).

(ii) The product set corresponding to the node of the d-tree is either row- or column-partitioned at each step of the protocol.

[1] All logarithms in this paper, unless otherwise specified, are of base 2.

By the definition of $k(f)$, every d-tree of f must have at least $k(f)$ leaves. Hence, the height of a d-tree is at least $\lceil \log (k(f)) \rceil$ and the theorem is proved. □

2.3. Upper and lower bounds for $C(d_H^n)$. In this section, we derive bounds for $C(d_H^n)$. The upper bound follows immediately as a corollary to Theorem 2.1:

$$C(d_H^n) \leqq n + \lceil \log (n+1) \rceil.$$

We next give a lower bound for $C(d_H^n)$, matching the upper bound up to smaller order terms, by a simple argument.

LEMMA 2.1. $C(d_H^n) \geqq n+1$.

Proof. For any $\mathbf{x} \in \{0,1\}^n$, $d_H^n(\mathbf{x}, \mathbf{x}) = 0$ and $d_H^n(\mathbf{x}, \bar{\mathbf{x}}) = n$, where $\bar{\mathbf{x}}$ is the complement of $\mathbf{x}$. Hence there are exactly 2^n m-rect's of Hamming Distance 0 and n respectively in any $m(d_H^n)$-decomposition of $\{0,1\}^n \times \{0,1\}^n$. Subsequently, $k(d_H^n) \geqq 2^{n+1}$ and by Theorem 2.2, $C_2(d_H^n) \geqq n+1$. □

This lower bound differs from the upper bound by $\lceil \log (n+1) \rceil - 1$ bits. A better lower bound, differing from the upper bound by no more than 1 bit, is stated in the following theorem.

THEOREM 2.3. $C_2(d_H^n) \geqq n + \lceil \log (n+1-\sqrt{n}) \rceil$.

Proof. We lower bound $k(d_H^n)$ by upper bounding the sizes of all the m-rect's in the function table. For $0 \leqq \delta \leqq n$, define

$$S_\delta^n \triangleq \{\mathcal{U} \times \mathcal{V} \subseteq \{0,1\}^n \times \{0,1\}^n : d_H^n(\mathbf{u}, \mathbf{v}) = \delta \text{ for all } \mathbf{u} \in \mathcal{U} \text{ and } \mathbf{v} \in \mathcal{V}\},$$

$$M(n, \delta) \triangleq \max \{\|\mathcal{U}\| \cdot \|\mathcal{V}\| : \mathcal{U} \times \mathcal{V} \in S_\delta^n\}.$$

In Lemma 2.2, we establish the fact $M(n, \delta) = M(n, n-\delta)$ by showing that for every m-rect $\in S_\delta^n$, there exists an m-rect $\in S_{n-\delta}^n$ with equal size. This reduces the task of upper bounding $M(n, \delta)$ to the range $0 \leqq \delta \leqq \lfloor n/2 \rfloor$. We then prove in Theorem 2.4 the crucial result that for $n = 2, 3, 4 \cdots$; $\delta = 0, 1, \cdots, \lfloor n/2 \rfloor$,

$$\frac{M(n, \delta)}{M(n-2, \delta-1)} \leqq \max \left[4, \frac{n(n-1)}{\delta(n-\delta)} \right].$$

As corollaries to Theorem 2.4, we show that

$$M(n, \delta) \leqq \begin{cases} \binom{n}{\delta} & \text{for all } n \text{ and } \delta < n/2 - \sqrt{n/4}, \\ \prod_{j=0}^{\delta-1} \dfrac{(n-2j)^2}{(\delta-j)(n-\delta-j)} & \text{for } n \geqq 4 \text{ and } \lfloor n/2 - \sqrt{n/4} \rfloor \leqq \delta \leqq \lfloor n/2 \rfloor. \end{cases} \tag{2.1}$$

Now denote

$$N(n, \delta) \triangleq \|\{(\mathbf{x}, \mathbf{y}) \in \{0,1\}^n \times \{0,1\}^n : d_H^n(\mathbf{x}, \mathbf{y}) = \delta\}\| \quad \text{for } 0 \leqq \delta \leqq n$$

$$= 2^n \cdot \binom{n}{\delta}.$$

It is clear that

$$k(d_H^n) \geqq \sum_{\delta=0}^{n} \frac{N(n, \delta)}{M(n, \delta)}$$

$$= 2 \cdot \sum_{\delta=0}^{\lceil n/2 - \sqrt{n/4} \rceil - 1} \frac{N(n, \delta)}{M(n, \delta)} + \sum_{\delta = \lceil n/2 - \sqrt{n/4} \rceil}^{\lceil n/2 + \sqrt{n/4} \rceil} \frac{N(n, \delta)}{M(n, \delta)}$$

$$\triangleq S_1 + S_2$$

where S_1 and S_2 are respectively the values of the first and second summations. From the first inequality of (2.1)

$$S_1 \geqq 2 \cdot \sum_{\delta=0}^{\lceil n/2-\sqrt{n/4}\rceil-1} \frac{2^n \cdot \binom{n}{\delta}}{\binom{n}{\delta}} = 2n \cdot \lceil n-\sqrt{n}\rceil.$$

On the other hand, as we show in Appendix 1,

$$S_2 \geqq 2^n \quad \text{for } n \geqq 4,$$

therefore $k(d_H^n) \geqq 2^n \lceil n+1-\sqrt{n}\rceil$ for $n \geqq 4$. For $1 \leqq n \leqq 3$, $k(d_H^n)$ can be verified from the Hamming Distance Function tables to be 4, 10 and 32 respectively, which still satisfy the lower bound above. Since $C_2(d_H^n) \geqq \lceil \log (k(d_H^n))\rceil$, we obtain

$$C_2(d_H^n) \geqq n + \lceil \log (n+1-\sqrt{n})\rceil. \qquad \square$$

We are ready to state and prove Lemma 2.2 and Theorem 2.4.

LEMMA 2.2. *$M(n, \delta) = M(n, n-\delta)$ for $\delta = 0, 1, \cdots, n$.*

Proof. Since for any $\mathbf{x}$ and $\mathbf{y} \in \{0,1\}^n$,

$$d_H^n(\mathbf{x}, \mathbf{y}) = \delta \Rightarrow \begin{cases} d_H^n(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = \delta, \\ d_H^n(\bar{\mathbf{x}}, \mathbf{y}) = n-\delta, \\ d_H^n(\mathbf{x}, \bar{\mathbf{y}}) = n-\delta, \end{cases}$$

the following are equivalent:

1) $\mathcal{U} \times \mathcal{V} \in S_\delta^n$,
2) $\bar{\mathcal{U}} \times \bar{\mathcal{V}} \in S_\delta^n$,
3) $\bar{\mathcal{U}} \times \mathcal{V} \in S_{n-\delta}^n$,
4) $\mathcal{U} \times \bar{\mathcal{V}} \in S_{n-\delta}^n$.

Since $\|\mathcal{U}\| = \|\bar{\mathcal{U}}\|$ and $\|\mathcal{V}\| = \|\bar{\mathcal{V}}\|$, the lemma is proved. $\square$

This lemma shows that the analysis of $M(n, \delta)$ can be reduced to the range $0 \leqq \delta \leqq \lfloor n/2\rfloor$. The basis for upper bounding $M(n, \delta)$ for $\delta = 1, 2, \cdots, \lceil n/2\rceil$ is provided by the following theorem.

THEOREM 2.4. *For $n = 3, 4 \cdots$, $\delta = 1, 2, \cdots, \lfloor n/2\rfloor$,*

$$\frac{M(n, \delta)}{M(n-2, \delta-1)} \leqq \max\left(4, \frac{n(n-1)}{\delta(n-\delta)}\right).$$

Proof. We first introduce some notation: For a set $C \subseteq \{0,1\}^n$ and $\varepsilon \in \{0,1\}$,

i) The ith bit of $c \in C$ is denoted by c_i.

ii) $C_\varepsilon^t \triangleq \{(c_1, \cdots, c_n) \in C: c_t = \varepsilon\} \subseteq \{0,1\}^n$.

iii) $C_\varepsilon^{*t} \triangleq \{(c_1, \cdots, c_{t-1}, c_{t+1}, \cdots, c_n): (c_1, \cdots, c_{t-1}, \varepsilon, c_{t+1}, \cdots, c_n) \in C\} \subseteq \{0,1\}^{n-1}$.

iv) Analogously, for two components s, t we define $C_{\varepsilon\eta}^{st} \subseteq \{0,1\}^n$ and $C_{\varepsilon\eta}^{*st} \subseteq \{0,1\}^{n-2}$.

v) For $a \in \{0,1\}^i$ and $\mathbf{b} \in \{0,1\}^j$, $\mathbf{ab} \in \{0,1\}^{i+j}$ represents the concatenation of $\mathbf{a}$ and $\mathbf{b}$.

Consider $A \times B \in S_\delta^n$. Construct $\mathcal{U} \subseteq \{0,1\}^{n(n-1)/2}$ from A by the following procedure:

i) Let $\Gamma \triangleq \{(i,j): i = 1, 2, \cdots, n-1; j = i+1, \cdots, n\}$, i.e. the set of pairs of distinct indices between 1 and n. Clearly $\nu \triangleq \|\Gamma\| = n(n-1)/2$. Order the pairs lexicographically and denote the kth pair as (k_i, k_j).

ii) Map every $\mathbf{a} \in A$ to a $\mathbf{u} \in \mathcal{U}$ such that for $k = 1, \cdots, \nu$,

$$u_k = \begin{cases} 1 & \text{if } (a_{k_i}, a_{k_j}) = (0,1) \text{ or } (1,0), \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, $\mathcal{V} \subseteq \{0,1\}^{n(n-1)/2}$ is constructed from B.

FACT 1. $\forall \mathbf{u} \in \mathcal{U}$ *and* $\mathbf{v} \in \mathcal{V}$: $d_H^n(\mathbf{u}, \mathbf{v}) = \delta(n-\delta)$.

Proof (*of Fact* 1). There is a one-to-one correspondence between each pair $(\mathbf{u}, \mathbf{v}) \in \mathcal{U} \times \mathcal{V}$ and $(\mathbf{a}, \mathbf{b}) \in A \times B$. $\mathbf{u}$ and $\mathbf{v}$ differ in a certain bit position iff there is exactly one component different in the corresponding pairs of bits from $\mathbf{a}$ and $\mathbf{b}$. The number of such pairs is $\delta(n-\delta)$. □

FACT 2. *Let*

$$\lambda_i \triangleq \frac{\|\{\mathbf{u} \in \mathcal{U}: u_i = 1\}\|}{\|\mathcal{U}\|} \quad \textit{and} \quad \mu_i \triangleq \frac{\|\{\mathbf{v} \in \mathcal{V}: v_i = 1\}\|}{\|\mathcal{V}\|}.$$

$\exists 1 \leqq k \leqq n(n-1)/2$ *such that either*:

$$\lambda_k \cdot \overline{\mu_k} \geqq \frac{\delta(n-\delta)}{n(n-1)}, \tag{2.2}$$

or

$$\overline{\lambda_k} \cdot \mu_k \geqq \frac{\delta(n-\delta)}{n(n-1)}. \tag{2.3}$$

Proof (*of Fact* 2). Define $\sigma_k \triangleq \|\{(\mathbf{u}, \mathbf{v}) \in \mathcal{U} \times \mathcal{V}: u_k \neq v_k\}\|$. Clearly, $\sigma_k = \|\mathcal{U}\| \cdot \|\mathcal{V}\| \cdot (\lambda_k \bar{\mu}_k + \bar{\lambda}_k \mu_k)$. On the other hand,

$$\begin{aligned} \sum_{k=1}^{n(n-1)/2} \sigma_k &= \sum_{\mathbf{u} \in \mathcal{U}} \sum_{\mathbf{v} \in \mathcal{V}} \|\{(u_k, v_k): u_k \neq v_k\}\| \\ &= \sum_{\mathbf{u} \in \mathcal{U}} \sum_{\mathbf{v} \in \mathcal{V}} d_H^n(\mathbf{u}, \mathbf{v}) \\ &= \|\mathcal{U}\| \cdot \|\mathcal{V}\| \cdot \delta(n-\delta). \end{aligned}$$

By the pigeon hole principle, there must exist an index k such that

$$\sigma_k \geqq \frac{2\delta(n-\delta)}{n(n-1)} \cdot \|\mathcal{U}\| \cdot \|\mathcal{V}\|$$

and one of the two terms making up σ_k must be no less than half of this value. □

Without loss of generality, assume (2.2) is satisfied by the specific value of k and that $k_i = 1$ and $k_j = 2$. To construct an m-rect $\in S_{\delta-1}^{n-2}$, we consider the following cases.

Case 1. $A_{10}^{12} \cup A_{01}^{12} = A$ **and** $B_{00}^{12} \cup B_{11}^{12} = B$: If $\mathbf{w} \in A_{01}^{*12}$ (resp. B_{00}^{*12}), then either $\mathbf{w} \in A_{10}^{*12}$ (resp. B_{11}^{*12}), or $10\mathbf{w}$ (resp. $11\mathbf{w}$) can be appended to A (resp. B), and this only increases $\|A\| \cdot \|B\|$. We can therefore assume $A_{10}^{*12} = A_{01}^{*12}$ and $B_{00}^{*12} = B_{11}^{*12}$. Define $C \triangleq A_{01}^{*12}$ or A_{10}^{*12} and $D \triangleq B_{00}^{*12}$ or B_{11}^{*12}. Clearly, $C \times D \in S_{\delta-1}^{n-2}$ and we have shown that

$$\begin{aligned} \|A\| \cdot \|B\| &= 4 \cdot \|C\| \cdot \|D\| \\ &\leqq \max\left(4, \frac{n(n-1)}{\delta(n-\delta)}\right) \cdot M(n-2, \delta-1). \end{aligned}$$

Case 2a. $A_{10}^{12} \cup A_{01}^{12} = A$ **and** $B_{00}^{12} \cup B_{11}^{12} \subset B$: If $\mathbf{w} \in B_{00}^{*12}$, then either $\mathbf{w} \in B_{11}^{*12}$, or $11\mathbf{w}$ can be appended to B which only increases $\|A\| \cdot \|B\|$. We can therefore assume that $B_{00}^{*12} = B_{11}^{*12}$. On the other hand, if $\mathbf{z} \in A_{01}^{*12}$ then $\mathbf{z} \notin A_{10}^{*12}$, i.e. $A_{01}^{*12} \cap A_{10}^{*12} = \varnothing$. Consider the following two cases:

a) $\delta(n-\delta)/n(n-1) \geqq 4$: Recall from the proof of Fact 2 that we have $\lambda_k \bar{\mu}_k + \bar{\lambda}_k \mu_k \geqq 2\delta(n-\delta)/n(n-1)$. $A_{10}^{12} \cup A_{01}^{12} = A$ implies that $\lambda_k = 1$ and therefore $\|A_{10}^{*12} \cup A_{01}^{*12}\| \cdot \|B_{00}^{*12}\| \geqq \delta(n-\delta)/n(n-1) \cdot \|A\| \cdot \|B\|$. Define $C \triangleq A_{10}^{*12} \cup A_{01}^{*12}$ and $D \triangleq B_{00}^{*12}$. Clearly, $C \times D \in S_{\delta-1}^{n-2}$ and we have shown that

$$\|A\| \cdot \|B\| = \frac{n(n-1)}{\delta(n-\delta)} \cdot \|C\| \cdot \|D\|$$

$$\leqq \max\left(4, \frac{n(n-1)}{\delta(n-\delta)}\right) \cdot M(n-2, \delta-1).$$

b) $\delta(n-\delta)/n(n-1) < 4$: We can construct an m-rect $P \times Q \in S_\delta^n$ from $A \times B$ such that $\|P \times Q\| > \|A \times B\|$. Specifically,

$$P \triangleq A \cup (\{(0,1)\} \times A_{10}^{*12}) \cup (\{(1,0)\} \times A_{01}^{*12}),$$

$$Q \triangleq B_{00}^{12} \cup B_{11}^{12}.$$

Note that

$$\|P\| \cdot \|Q\| \geqq 2 \cdot \|A\| \cdot \|B\| \cdot \frac{2\delta(n-\delta)}{n(n-1)} > \|A\| \cdot \|B\|.$$

Next, define $C \triangleq P_{01}^{*12}$ or P_{10}^{*12} and $D \triangleq Q_{00}^{*12}$ or Q_{11}^{*12}. Clearly, $C \times D \in S_{\delta-1}^{n-2}$ and that

$$\|A\| \cdot \|B\| \leqq \|P\| \cdot \|Q\| = 4 \cdot \|C\| \cdot \|D\| \leqq \max\left(4, \frac{n(n-1)}{\delta(n-\delta)}\right) \cdot M(n-2, \delta-1).$$

Case 2b. $A_{10}^{12} \cup A_{01}^{12} \subset A$ **and** $B_{00}^{12} \cup B_{11}^{12} = B$: The argument used in Case 2a is symmetric between A and B. Therefore we obtain the same upper bound on $\|A\| \cdot \|B\|$.

Case 3. $A_{01}^{12} \cup A_{10}^{12} \subset A$ **and** $B_{00}^{12} \cup B_{11}^{12} \subset B$: We observe that $A_{01}^{*12} \cap A_{10}^{*12} = \varnothing$ and $B_{00}^{*12} \cap B_{11}^{*12} = \varnothing$. Consider the following two cases.

a) $\delta(n-\delta)/n(n-1) \geqq 4$: Define $C \triangleq A_{01}^{*12} \cup A_{10}^{*12}$ and $D \triangleq B_{00}^{*12} \cup B_{11}^{*12}$. It is clear that $C \times D \in S_{\delta-1}^{n-2}$ and

$$\|A\| \cdot \|B\| = \frac{n(n-1)}{\delta(n-\delta)} \|C\| \cdot \|D\|$$

$$\leqq \max\left(4, \frac{n(n-1)}{\delta(n-\delta)}\right) \cdot M(n-2, \delta-1).$$

b) $\delta(n-\delta)/n(n-1) < 4$: Define

$$P \triangleq A_{01}^{12} \cup A_{10}^{12} \cup (\{(0,1)\} \times A_{10}^{*12}) \cup (\{(1,0)\} \times A_{01}^{*12}),$$

$$Q \triangleq B_{00}^{12} \cup B_{11}^{12} \cup (\{(0,0)\} \times B_{11}^{*12}) \cup (\{(1,1)\} \times B_{00}^{*12}).$$

Now $\|P\| \cdot \|Q\| \geqq 4 \cdot \|A\| \cdot \|B\| \cdot \delta(n-\delta)/n(n-1) > \|A\| \cdot \|B\|$. Define $C \triangleq P_{01}^{*12}$ or P_{10}^{*12} and $D \triangleq Q_{00}^{*12}$ or Q_{11}^{*12}. Clearly, $C \times D \in S_{\delta-1}^{n-2}$.

$$\|A\| \cdot \|B\| \leqq \|P\| \cdot \|Q\|$$

$$= 4 \cdot \|C\| \cdot \|D\|$$

$$\leqq \max\left(4, \frac{n(n-1)}{\delta(n-\delta)}\right) \cdot M(n-2, \delta-1).$$

Hence in all the four cases, we have succeeded in proving that

$$\frac{M(n,\delta)}{M(n-2,\delta-1)} \leqq \max\left(4, \frac{n(n-1)}{\delta(n-\delta)}\right). \qquad \square$$

The proofs of the following three corollaries are given in Appendix 2.

COROLLARY 2.1. *For* $\delta < \lceil n/2-\sqrt{n/4}\rceil$ *and* $\delta > \lfloor n/2+\sqrt{n/4}\rfloor$, $M(n,\delta) = \binom{n}{\delta}$.

COROLLARY 2.2. *For* $n/2-\sqrt{n/4} \leqq \delta \leqq n/2+\sqrt{n/4}$,

$$M(n,\delta) \leqq \prod_{j=0}^{\delta'-1} \frac{(n-2j)^2}{(\delta'-j)(n-\delta'-j)},$$

where

$$\delta' = \begin{cases} \delta & \text{for } \delta \leqq \lfloor n/2\rfloor, \\ \lfloor n/2\rfloor - \delta & \text{otherwise.} \end{cases}$$

COROLLARY 2.3. *For* $n = 1, 2, \cdots$, $\max_{0\leqq\delta\leqq n} M(n,\delta) = 2^n$ *and the maximum is achieved by* $\delta = \lfloor n/2\rfloor$ *and* $\lceil n/2\rceil$.

The last corollary, being a special case of Theorem 2.4, was previously derived using a less general argument and reported in [AEP].

3. The ε-error model.

3.1. Definitions and general lower bound. In the ε-error model, there is still a one-to-one correspondence between a protocol and a binary tree, which we call an ε-tree. An ε-tree is nearly identical to a d-tree defined in § 2.1, except that since errors are allowed by an ε-error protocol, the leaves of an ε-tree are no longer necessarily m-rect's. Each leaf is now a product set $A \times B \subseteq \mathcal{X} \times \mathcal{Y}$ such that most of its elements yield the same function value. The following definitions parallel those in § 2.1.

DEFINITION 3.1. Given a function $f: \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$, a *q-monochromatic rectangle* (abbreviated as q-rect) with error ε is a pair $(\mathcal{U} \times \mathcal{V}, \mathbf{z})$, where $\mathcal{U} \times \mathcal{V} \in \mathcal{X} \times \mathcal{Y}$ and $\mathbf{z} \in \mathcal{Z}$, such that $f(\mathbf{u}, \mathbf{v}) = \mathbf{z}$ for at least $(1-\varepsilon) \cdot \|\mathcal{U}\| \cdot \|\mathcal{V}\|$ pairs $(\mathbf{u}, \mathbf{v}) \in \mathcal{U} \times \mathcal{V}$. We denote the size of the largest q-rect with error ε by $M_\varepsilon(f)$. An $m_\varepsilon(f)$-partition is a partition of $\mathcal{X} \times \mathcal{Y}$ into q-rect's S_i with error ε_i, where $i = 1, \cdots, m_\varepsilon(f)$, such that

$$\sum_{i=1}^{m_\varepsilon(f)} \varepsilon_i \|S_i\| \leqq \varepsilon \|\mathcal{X}\| \cdot \|\mathcal{Y}\|.$$

We define $k_\varepsilon(f)$ as the minimum of $m_\varepsilon(f)$ over all $m_\varepsilon(f)$-partitions of $\mathcal{X} \times \mathcal{Y}$.

With these definitions, it is straightforward to pinpoint the differences between a d-tree and an ε-tree. In contrast to a d-tree which has m-rect's as its leaves, the leaves of an ε-tree are q-rect's. In addition, suppose there are k leaves in the tree, where the jth leaf has weight (i.e. the number of elements in it) ω_j and has error ε_j, then the following condition (which we shall refer to as the "ε-error requirement") must be satisfied:

$$\sum_{i=1}^{k} \varepsilon_i \omega_i \leqq \varepsilon \|\mathcal{X}\| \cdot \|\mathcal{Y}\|.$$

Let P be a protocol satisfying the ε-error requirement and $C_P(\mathbf{x}, \mathbf{y})$ be the depth of the leaf in the ε-tree representation of P to which $(\mathbf{x}, \mathbf{y})$ belongs. The ε-error communication complexity of f is defined as

$$C_\varepsilon(f) \triangleq \min_P \max_{(\mathbf{x},\mathbf{y})\in\mathcal{X}\times\mathcal{Y}} C_P(\mathbf{x}, \mathbf{y}).$$

With a uniform density on $\mathscr{X} \times \mathscr{Y}$, the average case ε-error communication complexity of f can also be defined:

$$\bar{C}_\varepsilon(f) \triangleq \frac{1}{\|\mathscr{X}\| \cdot \|\mathscr{Y}\|} \min_P \sum_{(\mathbf{x},\mathbf{y}) \in \mathscr{X} \times \mathscr{Y}} C_P(\mathbf{x}, \mathbf{y}).$$

The following lemma provides a lower bound for $C_\varepsilon(f)$ in terms of $M_\varepsilon(f)$.

LEMMA 3.1. $C_\varepsilon(f) \geqq \log \|\mathscr{X}\| + \log \|\mathscr{Y}\| - \log M_{2\varepsilon}(f) - 1$.

Proof. The proof of the first inequality is similar to that of Lemma 2.1. To prove the second inequality, consider the $m_\varepsilon(f)$-partition which achieves $m_\varepsilon(f) = k_\varepsilon(f)$. For convenience in notation, we shall abbreviate $k_\varepsilon(f)$ by k. Let S_i, $i = 1, \cdots, k$ be the q-rect's constructed. The ε-error requirement stipulates that

$$\sum_{i=1}^{k} \varepsilon_i \|S_i\| \leqq \varepsilon \|\mathscr{X}\| \|\mathscr{Y}\|,$$

which can be written as

$$\sum_{i:\, \varepsilon_i \leqq 2\varepsilon} \varepsilon_i \|S_i\| + \sum_{i:\, \varepsilon_i > 2\varepsilon} \varepsilon_i \|S_i\| \leqq \varepsilon \|\mathscr{X}\| \|\mathscr{Y}\|.$$

Clearly,

$$\sum_{i:\, \varepsilon_i \leqq 2\varepsilon} \|S_i\| \geqq (\|\mathscr{X}\| \cdot \|\mathscr{Y}\|)/2.$$

Suppose the number of q-rect's involved in the above sum is k'; we have

$$\frac{1}{k'} \sum_{i:\, \varepsilon_i \leqq 2\varepsilon} \|S_i\| \geqq (\|\mathscr{X}\| \cdot \|\mathscr{Y}\|)/(2k').$$

The left-hand side of the above equation is the average size of k' q-rect's. There must exist one q-rect S_i whose size is at least as large as the average, i.e.

$$\|S_i\| \geqq (\|\mathscr{X}\| \cdot \|\mathscr{Y}\|)/(2k').$$

Since $\|S_i\| \leqq M_{2\varepsilon}(f)$ and $k' \leqq k$, we have

$$M_{2\varepsilon}(f) \geqq (\|\mathscr{X}\| \cdot \|\mathscr{Y}\|)/(2k),$$

which gives the second inequality. □

There is a similar result for the average case complexity.

LEMMA 3.2. $\bar{C}_\varepsilon(f) \geqq (\log \|\mathscr{X}\| + \log \|\mathscr{Y}\| - \log M_{2\varepsilon}(f) - 1)/2$.

Proof. Let P be the protocol achieving $\bar{C}_\varepsilon(f)$. Consider the ε-tree representation of P. As there is no ambiguity, we also call this ε-tree P. Consider those leaves of this tree with no more than 2ε error. Without loss of generality, let them be the first m leaves of the tree and denote their weights by w_j, $1 \leqq j \leqq m$. We must have

$$W \triangleq \sum_{i=1}^{m} w_i \geqq (\|\mathscr{X}\| \cdot \|\mathscr{Y}\|)/2,$$

for if otherwise, the remaining leaves already violates the ε-error requirement. Clearly

$$\bar{C}_\varepsilon(f) = \bar{C}_P \geqq \frac{\sum_{j=1}^{m} l_j w_j}{\|\mathscr{X}\| \cdot \|\mathscr{Y}\|}.$$

By the entropy bound for the external path length of a binary tree,

$$\sum_{j=1}^{m} l_j w_j \geqq \sum_{j=1}^{m} w_j \cdot \log\left(\frac{W}{w_j}\right).$$

Since $w_j \leqq M_{2\varepsilon}(f)$ for all j, the left-hand side of the above inequality is at least

$$\sum_{j=1}^{m} \left(w_j \log \left(\frac{W}{M_{2\varepsilon}(f)} \right) \right) = W \log \left(\frac{W}{M_{2\varepsilon}(f)} \right).$$

Subsequently

$$\begin{aligned} \bar{C}_\varepsilon(f) &\geqq \frac{W}{\|\mathscr{X}\| \cdot \|\mathscr{Y}\|} \cdot \log \left(\frac{W}{M_{2\varepsilon}(f)} \right) \\ &\geqq (\log \|\mathscr{X}\| + \log \|\mathscr{Y}\| - \log M_{2\varepsilon}(f) - 1)/2. \end{aligned}$$

□

3.2. Lower bounds for $C_\varepsilon(d_H^n)$ and $\bar{C}_\varepsilon(d_H^n)$. Letting $\mathbf{x}, \mathbf{y} \in \{0, 1\}^n$, their *inner product* is defined as

$$f_I^n(\mathbf{x}, \mathbf{y}) \triangleq \sum_{i=1}^{n} x_i y_i$$

and the parity of $f_I^n(\mathbf{x}, \mathbf{y})$ is denoted by $f_{IP}^n(\mathbf{x}, \mathbf{y})$. We first relate $C_\varepsilon(d_H^n)$, $C_\varepsilon(f_I^n)$ and $C_\varepsilon(f_{IP}^n)$.

LEMMA 3.3. *Given* $0 \leqq \varepsilon < 1$, $C_\varepsilon(d_H^n) + 2\lceil \log (n+1) \rceil \geqq C_\varepsilon(f_I^n) \geqq C_\varepsilon(f_{IP}^n)$.

Proof. The second inequality is easily proved by noting that knowing the inner product, the parity of the inner product can always be computed. However when an erroneous value of $f_I^n(\mathbf{x}, \mathbf{y})$ is used to compute $f_{IP}^n(\mathbf{x}, \mathbf{y})$, the latter is not necessarily in error. Hence in order to compute $f_{IP}^n(\mathbf{x}, \mathbf{y})$, with error no more than ε, one can always first compute $f_I^n(\mathbf{x}, \mathbf{y})$ with the same designated error. To prove the first inequality, consider the different values that the pair (x_i, y_i) can take. Let

$$t_{1,1} \triangleq \sum_{i=1}^{n} 1(x_i = 1, y_i = 1).$$

Similarly, $t_{1,0}$, $t_{0,1}$ and $t_{0,0}$ are defined. We have the following relations:

(i) $d_H^n(\mathbf{x}, \mathbf{y}) = t_{0,1} + t_{1,0}$.

(ii) $\text{wt}(\mathbf{x}) = t_{1,1} + t_{1,0}$, where $\text{wt}(\mathbf{x})$ is the number of ones in $\mathbf{x}$.

(iii) $\text{wt}(\mathbf{y}) = t_{1,1} + t_{0,1}$.

(iv) $f_I^n(\mathbf{x}, \mathbf{y}) = t_{1,1}$.

It is easy to see that

$$(\text{wt}(\mathbf{x}) + \text{wt}(\mathbf{y}) - d_H^n(\mathbf{x}, \mathbf{y})) = 2f_I^n(\mathbf{x}, \mathbf{y}).$$

Hence knowing the weights of both $\mathbf{x}$ and $\mathbf{y}$, there is a one-to-one correspondence between the Hamming distance and the inner product. Since the weight of one sequence can be communicated to the other person in $\lceil \log (n+1) \rceil$ bits, we have the first inequality. □

Clearly, the argument also holds for the average case complexities. Thus

COROLLARY 3.1. $\bar{C}_\varepsilon(d_H^n) + 2\lceil \log (n+1) \rceil \geqq \bar{C}_\varepsilon(f_I^n) \geqq \bar{C}_\varepsilon(f_{IP}^n)$.

Finally, restricting Lemma 3.3 to the case $\varepsilon = 0$, we have the following relationship among the deterministic communication complexities of the three functions.

COROLLARY 3.2. $C(d_H^n) + 2\lceil \log (n+1) \rceil \geqq C(f_I^n) \geqq C(f_{IP}^n)$.

We next prove an upper bound for $M_\varepsilon(f_{IP}^n)$.

LEMMA 3.4. *For* $0 \leqq \varepsilon \leqq \frac{1}{8}$, $M_\varepsilon(f_{IP}^n) \leqq (1 + c\varepsilon) \cdot 2^n$, *where c is a constant dependent only on* ε.

Proof. Define $A(n)$, the *function table* for f_{IP}^n as a $2^n \times 2^n$ matrix, whose (i, j)th component is $f_{IP}^n(\mathbf{b}(i), \mathbf{b}(j))$ (where $\mathbf{b}(k)$ is the binary representation of $0 \leqq k \leqq 2^n - 1$). Consider $\{\mathbf{r}_i, i = 1, \cdots, 2^n\}$, the rows of $A(n)$ as a set of binary 2^n sequences. We have the following:

FACTS: (1) $\text{wt}(\mathbf{r}_1) = 0$ *and* $\text{wt}(\mathbf{r}_i) = 2^{n-1}$ *for* $1 < i \leqq 2^n$. (2) $d_H(\mathbf{r}_i, \mathbf{r}_j) = 2^{n-1}$ *for all* $i \neq j$.

We prove the facts by induction. The case $n = 1$ is easily settled by inspection. Suppose that the claim is true for n, using $A(n+1)$ (Fig. 3.1), we shall show that it also holds for $n+1$.

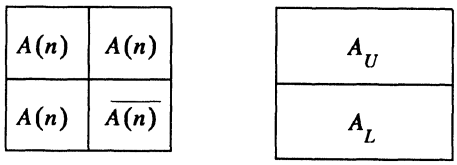


FIG. 3.1. *Inner product function table for sequence length* $n+1$.

To prove Fact 1, note that each row in $A(n+1)$ is a concatenation of two 2^n-sequences, each having weight 2^{n-1}. To prove Fact 2, we denote by A_U and A_L the upper and lower halves of $A(n+1)$ respectively, as shown in Fig. 3.1b. Consider any two sequences $\mathbf{r}_1$ and $\mathbf{r}_2$ in $A(n+1)$. If both of them are in A_U or A_L, then the Hamming distance between each half sequence is 2^{n-1}. If one of them is in A_U and the other in A_L, then there are two cases.

(i) $\mathbf{r}_i = \mathbf{aa}$, i.e. the concatenation of two copies of $\mathbf{a}$, which is a row in $A(n)$; and $\mathbf{r}_j = \mathbf{a}\bar{\mathbf{a}}$: The Hamming distance between the first half sequences is $d_H(\mathbf{a}, \mathbf{a}) = 0$ and that of the second half is $d_H(\mathbf{a}, \bar{\mathbf{a}}) = 2^n$.

(ii) $\mathbf{r}_i = \mathbf{aa}$, and $\mathbf{r}_j = \mathbf{b}\bar{\mathbf{b}}$, where $\mathbf{a}$ and $\mathbf{b}$ are rows in $A(n)$: The Hamming distance between the first half sequences is $d_H(\mathbf{a}, \mathbf{b}) = 2^{n-1}$ and that of the second half is $d_H(\mathbf{a}, \bar{\mathbf{b}}) = 2^{n-1}$.

In either case, the total Hamming distance is 2^n. □

For convenience of notation, denote 2^n by N. We are ready to prove that $M_\varepsilon(f_{IP}^n) \leqq (1+c\varepsilon) \cdot N$ for $0 \leqq \varepsilon \leqq \frac{1}{8}$. First consider q-rect's giving function value 0. Suppose there exists such a q-rect $P = A \times B$ of size $(L+1) \times M$, such that $(L+1)M > (1+c\varepsilon) \cdot N$. Construct the product set $Q = A' \times B$, of size $L \times M$ from P by the following procedure: If there is a row of all zeros in P, remove it, otherwise remove an arbitrary row. Consider $R = A' \times \{0, 1\}^n$ (i.e. the rows of the function table of which Q is a part). We define α_i, $i = 1, \cdots, N$ as the proportion of ones in the ith column of R. From Fact 1,

$$L \cdot \sum_{i=1}^{N} \alpha_i = LN/2. \tag{3.1}$$

From the restriction on the amount of impurities in Q (which is the first M columns of R),

$$0 \leqq L \cdot \sum_{i=1}^{M} \alpha_i \leqq \varepsilon ML. \tag{3.2}$$

Combining with (3.1),

$$N/2 \geqq \sum_{i=M+1}^{N} \alpha_i \geqq N/2 - \varepsilon M. \tag{3.3}$$

Next, we compute the Hamming distance of all combinations of two rows of R, first counting vertically and then horizontally. From Fact 2,

$$L^2 \cdot \sum_{i=1}^{N} \alpha_i \bar{\alpha}_i = \frac{L(L-1)}{2} \cdot \frac{N}{2} \quad \text{where } \bar{\alpha}_i = 1 - \alpha_i.$$

Separating the sum into two parts,

$$L^2 \cdot \sum_{i=M+1}^{N} \alpha_i \bar{\alpha}_i = \frac{L(L-1)}{2} \cdot \frac{N}{2} - L^2 \sum_{i=1}^{M} \alpha_i \bar{\alpha}_i$$

$$\geqq \frac{L(L-1)}{2} \cdot \frac{N}{2} - L^2 \sum_{i=1}^{M} \alpha_i.$$

Substituting the right-hand inequality of (3.2) and dividing both sides by $N-M$,

$$\frac{1}{N-M} \sum_{i=M+1}^{N} \alpha_i \bar{\alpha}_i \geqq \frac{1}{N-M}\left[\frac{(L-1)}{2L} \cdot \frac{N}{2} - \varepsilon M\right]. \tag{3.4}$$

Substituting the left-hand inequality of (3.3),

$$\frac{1}{N-M} \sum_{i=M+1}^{N} \alpha_i^2 \leqq \frac{(L+1)}{4L(N-M)} \cdot \left[N + \frac{4\varepsilon LM}{L+1}\right]. \tag{3.5}$$

It is easy to see that

$$\frac{1}{N-M} \sum_{i=M+1}^{N} \alpha_i^2 \leqq \left(\frac{1}{N-M} \sum_{i=M+i}^{N} \alpha_i\right)^2.$$

Applying this to the left-hand inequality of (3.3),

$$\frac{1}{N-M} \sum_{i=M+1}^{N} \alpha_i^2 \geqq \frac{N^2}{4(N-M)^2}\left[1 - \frac{2\varepsilon M}{N}\right]^2. \tag{3.6}$$

Combining (3.5) and (3.6), we have

$$\left(1+\frac{1}{L}\right)\left(1-\frac{M}{N}\right) \geqq \frac{\left[1-\frac{2\varepsilon M}{N}\right]^2}{\left[1+\frac{4(L-1)\varepsilon M}{(L+1)N}\right]}. \tag{3.7}$$

The assertion we have made is that $(L+1)M > (1+c\varepsilon)N$. Substituting this into (3.7) and invoking the fact that $\varepsilon \leqq \frac{1}{8}$, we show in Appendix 3 that there exists a constant $c = c(\varepsilon)$ such that (3.7) is a contradiction. Hence $(L+1) \leqq (1+c\varepsilon)N$ as claimed.

To prove that the size of the largest q-rect for function value 1 also satisfies the same upper bound, note that Facts 1 and 2 still hold if we replace them by the corresponding statement after taking componentwise complements. This completes the proof of the lemma. □

Applying Lemmas 3.1 and 3.3 to this result, the main theorem follows readily.

THEOREM 3.1. *For* $0 < \varepsilon \leqq \frac{1}{16}$,

$$C_\varepsilon(d_H^n) + 2\lceil \log(n+1) \rceil \geqq C_\varepsilon(f_I^n) \geqq C_\varepsilon(f_{IP}^n) \geqq n - \log(1+c\varepsilon) - 1,$$

where c is a constant dependent only on ε.

Similarly, for the average case complexity

THEOREM 3.2. *For* $0 < \varepsilon \leqq \frac{1}{16}$,

$$\bar{C}_\varepsilon(d_h^n) + 2\lceil \log(n+1) \rceil \geqq \bar{C}_\varepsilon(f_I^n) \geqq \bar{C}_\varepsilon(f_{IP}^n) \geqq (n - \log(1+c\varepsilon) - 1)/2,$$

where c is a constant dependent only on ε.

4. The ε-randomized model. The ε-randomized protocol was introduced by Yao and a definition can be found in [Yao2]. The ε-randomized communication complexity of computing the Hamming distance, $D_\varepsilon(d_H^n)$, was investigated in [Yao3], where it was proved that $D_\varepsilon(d_H^n)$ grows faster than $\Omega(\log n)$. In the following Theorem, we use the results derived in § 3 to show that $D_\varepsilon(d_H^n) = \Omega(n)$, thus resolving an open problem posed by Yao in [Yao3]. The proof uses the following.

LEMMA 4.1 [Yao1]. *For any function f and* $0 \leqq \varepsilon < \frac{1}{2}$,

$$D_\varepsilon(f) \geqq (\bar{C}_{2\varepsilon}(f))/2.$$

THEOREM 4.1. *For* $0 \leqq \varepsilon < \frac{1}{2}$,

$$D_\varepsilon(d_H^n) = \Omega(n).$$

Proof. For $0 \leqq \varepsilon < \frac{1}{32}$, the theorem follows readily from Theorem 3.1 and Lemma 4.1. For $\frac{1}{32} \leqq \varepsilon < \frac{1}{2}$, given a randomized protocol with complexity $D_\varepsilon(f)$ and error probability $\varepsilon = \frac{1}{2} - \delta$, we can construct one with error probability less than $\frac{1}{64}$ as follows: Given the pair of values $(\mathbf{x}, \mathbf{y})$, repeat the protocol $2m-1$ times, such that $m(1-4\delta^2)^m < \frac{1}{64}$, and take the majority of the outcome as $f(\mathbf{x}, \mathbf{y})$. It is easy to show that the resulting error probability is no more than

$$\sum_{k \leqq m} \binom{2m-1}{k} (1-\varepsilon)^{2m-1-k} \leqq \frac{1}{64}.$$

Clearly, m is a function of ε only. Hence, there exists a constant $c = c(\varepsilon)$ such that

$$c \cdot D_\varepsilon(d_H^n) \geqq D_{1/64}(d_H^n),$$

and the theorem follows from the lower bound on the left-hand expression. □

5. Concluding comments. The upper and lower bounds for $C(d_H^n)$ can be compared by examining $\lceil \log(n+1-\sqrt{n}) \rceil$ and $\lceil \log(n+1) \rceil$. One finds that for all n, the two terms never differ by more than 1. (Actually, except for those n which satisfy $n+1 > 2^m$ and $n+1-\sqrt{n} \leqq 2^m$ for some integer m, they are identical.) Hence, our bounds are tight to within one bit. This difference is probably due to a combination of the facts that we are only considering m-rect's of maximal size for each δ, and that the optimal $m(f)$-partition is simply not achievable. It does not seem likely that there exists an algorithm whose complexity is lower than the obvious upper bound.

In Corollary 2.1, we showed that $M(n, \delta) = \binom{n}{\delta}$ for $\delta < \lceil n/2 - \sqrt{n/4} \rceil$ and $\delta > \lfloor n/2 + \sqrt{n/4} \rfloor$. We also showed in Corollary 2.3 that $M(n, \lfloor n/2 \rfloor) = M(n, \lceil n/2 \rceil) = 2^n$. However, it is not known whether the $M(n, \delta)$ upper bounds for $n/2 - \sqrt{n/4} \leqq \delta \leqq n/2 + \sqrt{n/4}$ are achievable. We believe that they are not. The interesting question then is whether one can prove tighter upper bounds for them.

Appendix 1. We prove in this Appendix the lower bound on

$$S_2 = \sum_{\delta = \lceil n/2 - \sqrt{n/4} \rceil}^{\lceil n/2 + \sqrt{n/4} \rceil} \frac{N(n, \delta)}{M(n, \delta)}$$

defined in the proof of Theorem 2.3. There are two cases.

Case 1. $n=2m$ for some m: By Corollary 3, for each δ in the range of S_2,

$$\frac{N(2m,\delta)}{M(2m,\delta)} \geqq \prod_{i=0}^{\delta-1} \frac{2m-2i-1}{2m-2i}.$$

For $\delta = m$, we have

$$\frac{N(2m,m)}{M(2m,m)} \geqq \binom{2m}{m} \geqq \frac{2^{2m}}{\sqrt{4m}}.$$

For any values of δ in the range, we have

$$\frac{N(2m,\delta)}{M(2m,\delta)} \geqq \frac{2^{2m}}{\sqrt{4m}} \cdot \prod_{i=1}^{m-\delta} \frac{2i}{2i-1}.$$

Note that each of these terms $\geqq 2(2^{2m}/\sqrt{4m})$ and there are $\lceil\sqrt{2m}\rceil$ of them. Hence

$$S_2 \geqq (2 \cdot \lceil\sqrt{2m}\rceil + 1) \cdot \frac{2^{2m}}{\sqrt{4m}}$$

$$\geqq 2^{2m}.$$

Case 2. $n=2m-1$ for some m: There are two middle terms.

$$\frac{N(2m-1,m)}{M(2m-1,m)} = \frac{N(2m-1,m-1)}{M(2m-1,m-1)} \geqq \binom{2m-1}{m} \geqq \frac{2^{2m-1}}{\sqrt{4m}}$$

and for any value of δ in the range, we have

$$\frac{N(2m-1,\delta)}{M(2m-1,\delta)} \geqq \frac{2^{2m-1}}{\sqrt{4m}} \cdot \prod_{i=1}^{m-\delta} \frac{2i+1}{2i}.$$

Note that each of these terms is $\leqq \frac{3}{2} \cdot 2^{2m-1}/\sqrt{4m}$ and there are $\lceil\sqrt{2m-1}\rceil - 2$ of them. Therefore

$$S_2 \geqq \left(\frac{3}{2} \cdot (\lceil\sqrt{2m-1}\rceil - 2) + 2\right) \cdot \frac{2^{2m-1}}{\sqrt{4m}} \geqq 2^{2m-1}.$$

Hence in both cases, the assertion $S_2 \geqq 2^n$ is true. □

Appendix 2. In this Appendix, we give the proofs for Corollaries 2.1, 2.2 and 2.3.

COROLLARY 2.1. *For* $\delta < \lceil n/2 - \sqrt{n/4}\rceil$ *or* $\delta > \lfloor n/2 + \sqrt{n/4}\rfloor$, $M(n,\delta) = \binom{n}{\delta}$.

Proof. By Lemma 2.2 and the fact that $\binom{n}{\delta} = \binom{n}{n-\delta}$, we only have to consider the range $\delta < \lceil n/2 - \sqrt{n/4}\rceil$. For any δ, define $A \triangleq \{0\}$ and $B \triangleq \{\mathbf{x}: d(\mathbf{x},0) = \delta\}$. It is clear that $A \times B \in S_\delta^n$ and that $\|A \times B\| = \binom{n}{\delta}$. Therefore one side of the equality is proved. To prove the other side, just note that the equation

$$4 = \frac{n(n-1)}{x(n-x)}$$

has positive root $x = n/2 - \sqrt{n/4}$. Hence, for $\delta < \lceil n/2 - \sqrt{n/4}\rceil$,

$$\max\left(4, \frac{n(n-1)}{\delta(n-\delta)}\right) = \frac{n(n-1)}{\delta(n-\delta)}.$$

Moreover, this still holds if we replace n by $n-2j$ and δ by $\delta-j$, for all $j<\delta$. Subsequently, by Lemma 2.2,

$$M(n,\delta)\leqq\prod_{j=0}^{\delta-1}\frac{(n-2j)((n-2j)-1)}{(\delta-j)((n-2j)-(\delta-j))}\cdot M((n-2\delta),0)=\binom{n}{\delta}. \qquad \square$$

COROLLARY 2.2. *For* $n/2-\sqrt{n/4}\leqq\delta\leqq n/2+\sqrt{n/4}$,

$$M(n,\delta)\leqq\prod_{j=0}^{\delta'-1}\frac{(n-2j)^2}{(\delta'-j)(n-\delta'-j)},$$

where

$$\delta'=\begin{cases}\delta & \text{for } \delta\leqq\lfloor n/2\rfloor,\\ \lfloor n/2\rfloor-\delta & \text{otherwise.}\end{cases}$$

Proof. First note that for $n\geqq 4$, the following holds for all $0\leqq\delta\leqq\lfloor n/2\rfloor$:

$$\frac{n^2}{\delta(n-\delta)}\geqq\max\left(4,\frac{n(n-1)}{\delta(n-\delta)}\right)$$

and the relation is definitely true for the range of δ in this corollary. Hence $M(n,\delta)/M(n-2,\delta-1)\geqq n^2/\delta(n-\delta)$ and it is clear that this still holds true when we replace n by $n-2j$ and δ by $\delta-j$, for $j\leqq\delta$. Apply Theorem 2.4 recursively δ times and since $M(n-2\delta,0)=1$

$$M(n,\delta)\leqq\prod_{j=0}^{\delta-1}\frac{(n-2j)^2}{(\delta-j)((n-2j)-(\delta-j))}=\prod_{j=0}^{\delta-1}\frac{(n-2j)^2}{(\delta-j)(n-\delta-j)},$$

which completes the proof of the corollary. $\square$

COROLLARY 2.3. *For* $n=1,2,\cdots,\max_{0\leqq\delta\leqq n}M(n,\delta)=2^n$ *and the maximum is achieved by* $\delta=\lfloor n/2\rfloor$ *and* $\lceil n/2\rceil$.

Proof. We first show that $M(n,\lfloor n/2\rfloor)=2^n$. By Lemma 2.2, this also establishes $M(n,\lceil n/2\rceil)=2^n$. The crucial observation is that for $\delta=\lfloor n/2\rfloor$, $n(n-1)/\delta(n-\delta)\leqq 4$. Hence $M(n,\lfloor n/2\rfloor)/M(n-2,\lfloor n/2\rfloor-1)\leqq 4$. Moreover, this relation is still true if we replace n by $n-2j$ and δ by $\delta-j$, for $j\leqq\delta$. For even n, apply Theorem 2.4 recursively $n/2-1$ times and since $M(2,1)=2$, we obtain $M(n,n/2)\leqq 2^n$. For odd n, apply Theorem 2.4 recursively $(n+1)/2$ times and since $M(1,0)=1$, we obtain $M(n,\lfloor n/2\rfloor)\leqq 2^n$. On the other hand, for even n, define $A\triangleq\{01,10\}^{n/2}$ and $B\triangleq\{00,11\}^{n/2}$. Clearly $A\times B\in S^n_{n/2}$ and $\|A\times B\|=2^n$. For odd n, define $C\triangleq A\times\{0\}$ and $C\triangleq B\times\{0\}$ and $C\times D\in S^n_{\lfloor n/2\rfloor}$. Therefore $M(n,\lfloor n/2\rfloor)\geqq 2^n$.

Now, suppose there exist $A\times B\in S^n_\delta$ such that $\|A\times B\|>2^n$. Consider the following two cases:

a) $n=2m$ for some m. By Lemma 2.2, $\bar{A}\times B\in S^n_{n-\delta}$. Define $C\triangleq A\times\bar{A}$ and $D\triangleq B\times B$. Clearly, $C\times D\in S^{4m}_{2\delta}$ and $\|C\times D\|>2^{4m}$ which is a contradiction to Corollary 2.3.

b) $n=2m+1$ for some m. Again $C\times D\in S^{4m}_{2\delta}$. However, since n is odd, $A\cap\bar{A}=\varnothing$ and $B\cap\bar{B}=\varnothing$ (cf. Lemma 2.2). Hence $C\cap\bar{C}=\varnothing$ and $D\cap\bar{D}=\varnothing$. Define $P\triangleq C\cup\bar{C}$ and $Q\triangleq D\cup\bar{D}$. Therefore $P\times Q\in S^{4m+2}_{2\delta}$ and $\|P\times Q\|>2^{4m+2}$ which is a contradiction to Corollary 2.3. $\square$

Appendix 3. We prove in this Appendix that

$$\left(1+\frac{1}{L}\right)\left(1-\frac{M}{N}\right)\geqq\frac{(1-2\varepsilon M/N)^2}{1+4(L-1)\varepsilon M/((L+1)M)} \tag{A3.1}$$

is a contradiction if $(L+1)M=\Lambda>(1+c\varepsilon)N$ for a constant $c=c(\varepsilon)$ and $\varepsilon>\frac{1}{8}$. The

left-hand side of (A3.1) is

$$\left(1+\frac{1}{L}\right)\left(1-\frac{M}{N}\right)=\left(1+\frac{1}{L}\right)\left(1-\frac{\Lambda}{N(L+1)}\right)$$
$$=1+\frac{1}{L}-\frac{\Lambda}{N(L+1)}-\frac{\Lambda}{NL(L+1)}.$$

The right-hand side of (A3.1) is larger than

$$\left(1-\frac{4\varepsilon}{N}\right)\cdot\left(1-\frac{4(L-1)\varepsilon M}{(L+1)N}\right)\geqq 1-\frac{8\varepsilon M}{N}\left(1+\frac{L-1}{L+1}\right)\geqq 1-\frac{8\varepsilon M}{N}.$$

For (A3.1) to be a contradiction, we want

$$1-\frac{8\varepsilon M}{N}>1+\frac{1}{L}-\frac{\Lambda}{N(L+1)}-\frac{\Lambda}{NL(L+1)},$$

which simplifies to

$$\frac{L\Lambda}{L+1}\left(1+\frac{1}{L}-8\varepsilon\right)>N. \tag{A3.2}$$

Since

$$1-8\varepsilon<\frac{L}{L+1}\left(1+\frac{1}{L}-8\varepsilon\right),$$

(A3.2) is true if

$$\Lambda(1-8\varepsilon)>N$$

which is equivalent to $\Lambda>(1+c\varepsilon)N$ for a constant $c=c(\varepsilon)$. Note that the condition $\varepsilon<\frac{1}{8}$ is certainly required for the above to hold. □

Acknowledgment. We are grateful to Alon Orlitsky for his contribution in the formalization of deterministic protocols.

REFERENCES

[AEP] R. AHLSWEDE, A. EL GAMAL AND K. F. PANG, *A two family extremal problem in Hamming space*, Discrete Math., 49 (1984), pp. 1-5.

[MS] K. MEHLHORN AND E. M. SCHMIDT, *Las Vegas is better then determinism in* VLSI *and distributed computing*, Proc. 14th Annual ACM Symposium on Theory of Computing, April 1982, pp. 330-337.

[Yao1] A. C. YAO, *Probabilistic computations: towards a unified measures of complexity*, Proc. 18th Symposium on Foundations of Computer Science, Oct. 1977, pp. 222-227.

[Yao2] ———, *Some complexity questions related to distributive computing*, Proc. 11th Annual ACM Symposium on Theory of Computing, May 1979, pp. 209-213.

[Yao3] ———, *Lower bounds by probabilistic arguments*, Proc. 25th Symposium on Foundations of Computer Science, November 1983, pp. 420-428.