COMMUNICATION COMPLEXITY UNDER PRODUCT AND NONPRODUCT DISTRIBUTIONS

Alexander A. Sherstov

Abstract. We solve an open problem in communication complexity posed by Kushilevitz and Nisan (1997). Let $R_\epsilon(f)$ and $D^\mu_\epsilon(f)$ denote the randomized and μ-distributional communication complexities of f, respectively (ϵ a small constant). Yao's well-known minimax principle states that $R_\epsilon(f) = \max_\mu \{D^\mu_\epsilon(f)\}$. Kushilevitz and Nisan (1997) ask whether this equality is approximately preserved if the maximum is taken over product distributions only, rather than all distributions μ. We give a strong negative answer to this question. Specifically, we prove the existence of a function $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ for which $\max_{\mu \text{ product}} \{D^\mu_\epsilon(f)\} = \Theta(1)$ but $R_\epsilon(f) = \Theta(n)$. We also obtain an exponential separation between the statistical query dimension and sign-rank, solving a problem previously posed by the author (2007).

Keywords. Randomized and distributional communication complexity, product and nonproduct distributions, Yao's Minimax Principle

Subject classification. 03D15, 68Q17.

1. Introduction

Among the primary models of communication complexity is the *randomized model*. Let $f\colon X \times Y \to \{-1,+1\}$ be a given function, where X and Y are finite sets. Alice receives an input $x \in X$, Bob receives $y \in Y$, and their objective is to compute $f(x,y)$ with minimal communication. To this end, Alice and Bob share an unlimited supply of random bits. Their protocol is said to *compute* f if on every input (x,y), the output is correct with probability at least $2/3$. The *cost* of a protocol is the worst-case number of bits exchanged on any input. The *randomized communication complexity* of f, denoted $R_{1/3}(f)$, is the least cost of a protocol that computes f.

Randomized communication complexity is intimately related to distribu-

tional complexity. Let μ be a probability distribution on $X \times Y$. The *μ-distributional communication complexity* of f, denoted $D^{\mu}_{1/3}(f)$, is the least cost of a deterministic protocol for f with error probability at most $1/3$ with respect to μ. Using the Minimax Theorem for zero-sum games, Yao (1983) gave a simple proof that

$$R_{1/3}(f) = \max_{\mu} \left\{ D^{\mu}_{1/3}(f) \right\},$$

where the constant $1/3$ can be replaced by any other (see also Kushilevitz & Nisan 1997, Thm. 3.20). This equation has been the basis for essentially all lower bounds on randomized communication complexity: one defines a probability distribution μ on $X \times Y$ and argues that the cost $D^{\mu}_{1/3}(f)$ of the best deterministic protocol with error at most $1/3$ over μ must be high.

The main design question, then, is what distribution μ to consider. A *product distribution* μ on $X \times Y$ is a distribution that can be expressed as $\mu(x, y) = \mu_X(x)\,\mu_Y(y)$, where μ_X and μ_Y are distributions over X and Y, respectively. Product distributions are particularly attractive because they are easier to analyze. Unfortunately, they do not always lead to optimal lower bounds. A standard example is the DISJOINTNESS function on n-bit strings, whose randomized complexity is $\Theta(n)$ (Kalyanasundaram & Schnitger 1992; Razborov 1992) and whose distributional complexity is $O(\sqrt{n} \log n)$ under every product distribution (Babai *et al.* 1986, Sec. 8). Define

$$D^{\times}_{1/3}(f) = \max_{\mu \text{ product}} \left\{ D^{\mu}_{1/3}(f) \right\}.$$

These considerations motivate the following problem:

PROBLEM 1.1 (Kushilevitz & Nisan 1997, p. 37). *Can restricting the distribution μ to be a product distribution affect the resulting lower bound on $R_{1/3}(f)$ by more than a polynomial factor? Formally, is $R_{1/3}(f) = (D^{\times}_{1/3}(f))^{O(1)}$?*

Since its formulation, this problem has seen little progress. Kremer, Nisan, and Ron (1999) studied its restriction to *one-way* randomized protocols and obtained a separation of $O(1)$ versus $\Omega(n)$ for the function GREATER-THAN. Unfortunately, a function can have vastly different communication complexity in the one-way model and the usual, two-way model. Such is the case of GREATER-THAN, whose two-way randomized complexity is a mere $O(\log n)$.

Another step toward solving the Kushilevitz-Nisan question was recently taken by the author (Sherstov 2008). Namely, we gave an exponential separation between the *discrepancy* under product and nonproduct distributions, for

an explicit function. In particular, we showed that the use of nonproduct distributions is essential to the *discrepancy method,* a major technique for communication lower bounds.

This paper solves the Kushilevitz-Nisan problem completely and in its original form: we prove the existence of a function f with $D^{\times}_{1/3}(f) = O(1)$ and $R_{1/3}(f) = \Omega(n)$.

THEOREM 1.2 (Main Theorem). *Let $\epsilon > 0$ be an arbitrary constant. Then there exists a function $f\colon \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ with all of the following properties:*

$$\begin{aligned} D^{\times}_{\epsilon}(f) &= O(1), \\ R_{1/3}(f) &= \Omega(n), \\ \mathrm{disc}^{\times}(f) &= \Omega(1), \\ \mathrm{disc}(f) &= O(2^{-n(\frac{1}{2}-\epsilon)}). \end{aligned}$$

(See Section 3 for a more detailed statement, with all dependencies on ϵ spelled out.) The symbols $\mathrm{disc}(f)$ and $\mathrm{disc}^{\times}(f)$ above stand for the smallest discrepancy of f under arbitrary distributions and product distributions, respectively. We review the discrepancy of functions and sign matrices thoroughly in Section 2. For now, it suffices to know that low discrepancy implies high communication complexity in most models.

A key aspect of Theorem 1.2 is that the function f in question has exponentially small discrepancy. Indeed, its discrepancy essentially meets the $\Omega(2^{-n/2})$ lower bound for *any* function on n bit strings (see Proposition 2.8 below). As a result, f has communication complexity $\Omega(n)$ not only in the randomized model, but also in the nondeterministic and various quantum models. Furthermore, the communication complexity of f remains $\Omega(n)$ even if one simply seeks a randomized/quantum protocol with advantage $2^{-n/4}$ over random guessing, on every input. Finally, it is clear from our proof (see Remark 3.10) that f has complexity $\Omega(n)$ in the *unbounded-error model* due to Paturi and Simon (1986), which has an even weaker success criterion.

To summarize the previous paragraph, f has the highest communication complexity in all standard models. Yet, the distributional method restricted to product distributions can certify at best an $\Omega(1)$ lower bound.

Finally, Theorem 1.2 improves on our previously obtained exponential separation for discrepancy (Sherstov 2008). In that earlier work, we constructed

an explicit matrix $A \in \{-1,+1\}^{2^n \times 2^{n^2}}$ with $\mathrm{disc}^{\times}(A) = \Omega(1/n^4)$ and $\mathrm{disc}(A) = O(\sqrt{n}/2^{n/4})$. Theorem 1.2 amplifies that gap to what is essentially optimal, although the function is no longer explicit.

Complexity measures of sign matrices. We now consider a different contribution of this work, which pertains to the complexity measures of sign matrices. This comparatively new area studies matrices with ± 1 entries from a complexity-theoretic point of view, focusing on their analytic rather than combinatorial structure. The study of sign matrices has strong ties to classical complexity theory, computational learning, and functional analysis, and has drawn considerable interest (Ben-David *et al.* 2003; Forster 2002; Forster *et al.* 2001, 2003; Forster & Simon 2006; Linial *et al.* 2007; Linial & Shraibman 2008; Sherstov 2008).

Fundamental complexity measures of a sign matrix A are:

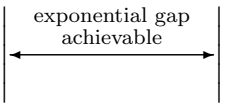
- $\mathrm{disc}^{\times}(A)$, the smallest discrepancy of A under a product distribution;
- $\mathrm{sq}(A)$, the statistical-query (SQ) dimension of A viewed as a concept class. This quantity arises in Kearns' *statistical query* model of learning (1993) and turns out to be intimately related to discrepancy (Sherstov 2008).
- $\mathrm{dc}(A)$, the dimension complexity of A, also known as sign-rank;
- $\mathrm{mc}(A)$, the margin complexity of A;
- $\mathrm{disc}(A)$, the smallest discrepancy of A under an arbitrary distribution.

Precise definitions of these quantities appear in Section 2. Among the early findings is the following inequality due to Ben-David *et al.* (2003):

$$\mathrm{dc}(A) \leqslant O(\mathrm{mc}(A)^2 \log(M+N)) \qquad \text{for every } A \in \{-1,+1\}^{M \times N}.$$

Linial and Shraibman (2008) showed that $\mathrm{mc}(A)$ and $1/\mathrm{disc}(A)$ are always within a factor of 8 of each other. The author extended these two results to the following picture (Sherstov 2008):

$$\frac{1}{\mathrm{disc}^{\times}(A)} \;=_{\mathrm{poly}}\; \mathrm{sq}(A) \;\leqslant_{\mathrm{poly}}\; \mathrm{dc}(A) \;\leqslant_{\mathrm{poly}}\; \mathrm{mc}(A) \;\approx\; \frac{1}{\mathrm{disc}(A)}$$

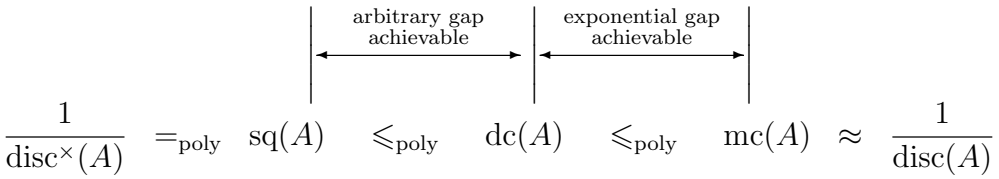


The symbols $\leqslant_{\text{poly}}$ and $=_{\text{poly}}$ in the above diagram have their intuitive meaning; we give precise statements in Section 2.2. The only piece missing from this diagram is the gap between $\text{sq}(A)$ and $\text{dc}(A)$, which we left as an open problem (Sherstov 2008, §7). We solve this problem here, showing that this gap can be *arbitrary*:

THEOREM 1.3 (SQ dimension vs. dimension complexity). *Let $\epsilon > 0$ be an arbitrary constant. Then there exists a matrix $A \in \{-1,+1\}^{N \times N}$ with*

$$\begin{aligned} \text{sq}(A) &= O(1), \\ \text{dc}(A) &= \Omega(N^{1-\epsilon}). \end{aligned}$$

It is easy to show (see Section 2.2) that $\text{dc}(A) \leqslant \min\{M,N\}$ for every $A \in \{-1,+1\}^{M \times N}$. In this light, Theorem 1.3 gives essentially the best gap that can exist by definition. This completes our taxonomy to the following overall picture:



Our techniques. An important ingredient in our proof is a simulation due to Kremer *et al.* (1999) that relates the one-way communication complexity of a sign matrix to its Vapnik-Chervonenkis dimension. Another ingredient is an analytic fact due to Ben-David *et al.* (2003) about matrices with low Vapnik-Chervonenkis dimension, which in turn relies on fundamental results in combinatorics by Alon *et al.* (1985) and Bollobás (1978). To combine these ingredients, we use the above taxonomy of complexity measures.

2. Preliminaries

This section surveys facts regarding communication complexity, sign matrices, and learning theory that figure in our proofs.

2.1. Communication complexity. We consider Boolean functions of the form $f\colon X \times Y \to \{-1,+1\}$, where -1 and $+1$ correspond to "true" and "false," respectively. Typically $X = Y = \{0,1\}^n$, but we also allow X and Y to be arbitrary finite sets, possibly of unequal cardinality. We identify f with its *communication matrix* $A = [f(x,y)]_{y\in Y,\, x\in X}$. In particular, we use the terms "communication complexity of f" and "communication complexity of A" interchangeably (and likewise for other complexity measures, such as discrepancy). The two communication models of interest to us are the public-coin randomized model and the deterministic model, both reviewed in Section 1.

For a fixed distribution μ over $X \times Y$, the *discrepancy* of f is defined as

$$\operatorname{disc}_\mu(f) = \max_{\substack{X'\subseteq X,\\ Y'\subseteq Y}} \left| \sum_{(x,y)\in X'\times Y'} \mu(x,y) f(x,y) \right|.$$

We define $\operatorname{disc}(f) = \min_\mu\{\operatorname{disc}_\mu(f)\}$. We let $\operatorname{disc}^\times(f)$ denote the minimum discrepancy of f under *product* distributions. The *discrepancy method* is a powerful technique that gives lower bounds on the randomized and distributional complexity in terms of the discrepancy:

Proposition 2.1 (see Kushilevitz & Nisan 1997, pp. 36–38). *For every Boolean function $f(x,y)$, every distribution μ, and every $\gamma > 0$,*

$$R_{1/2-\gamma/2}(f) \geqslant D^\mu_{1/2-\gamma/2}(f) \geqslant \log_2 \frac{\gamma}{\operatorname{disc}_\mu(f)}.$$

As an illustration, consider the well-studied function inner product modulo 2, defined on $\{0,1\}^n \times \{0,1\}^n$ by

$$\textsc{ip}(x,y) = (-1)^{\sum_{i=1}^n x_i y_i}.$$

The following fact is well-known; see Kushilevitz & Nisan (1997, §3.5) as well as earlier work (Babai *et al.* 1986; Chor & Goldreich 1988).

Proposition 2.2. *The function* ip *has discrepancy at most $2^{-n/2}$ with respect to the uniform distribution on $\{0,1\}^n \times \{0,1\}^n$. In particular, $D^\times_{1/3}(\textsc{ip}) = \Theta(n)$ and $R_{1/3}(\textsc{ip}) = \Theta(n)$.*

A definitive resource for further details is the book of Kushilevitz and Nisan (1997).

2.2. Sign matrices. We frequently use "generic-entry" notation to specify a matrix succinctly: we write $A = [F(i,j)]_{i,j}$ to mean that the (i,j)th entry of A is given by the expression $F(i,j)$. A *(Euclidean) embedding* of a matrix $A \in \{-1,+1\}^{M\times N}$ is a collection of vectors $\mathbf{u}_1,\dots,\mathbf{u}_M \in \mathbb{R}^k$ and $\mathbf{v}_1,\dots,\mathbf{v}_N \in \mathbb{R}^k$ (for some k) such that $\langle \mathbf{u}_i, \mathbf{v}_j\rangle \cdot A_{ij} > 0$ for all i,j. The integer k is the *dimension* of the embedding. The quantity

$$\gamma = \min_{i,j} \frac{|\langle \mathbf{u}_i, \mathbf{v}_j\rangle|}{\|\mathbf{u}_i\|_2 \cdot \|\mathbf{v}_j\|_2}$$

is the *margin* of the embedding. The *dimension complexity* $\mathrm{dc}(A)$ is the smallest dimension of an embedding of A. The *margin complexity* $\mathrm{mc}(A)$ is the minimum $1/\gamma$ over all embeddings of A.

Let $\mathbf{e}_i$ denote the vector with 1 in the ith component and zeroes elsewhere. The following is a trivial embedding of a sign matrix $A = [\,\mathbf{a}_1 \,|\, \dots \,|\, \mathbf{a}_N\,] \in \{-1,+1\}^{M\times N}$: label the rows by vectors $\mathbf{e}_1,\dots,\mathbf{e}_M \in \mathbb{R}^M$ and the columns by vectors $\mathbf{a}_1,\dots,\mathbf{a}_N$. It is easy to see that this embedding has dimension M and margin $1/\sqrt{M}$. By interchanging the roles of the rows and columns, we obtain the following well-known fact:

PROPOSITION 2.3. *Let $A \in \{-1,+1\}^{M\times N}$. Then*

$$1 \leqslant \mathrm{dc}(A) \leqslant \min\{M,N\},$$
$$1 \leqslant \mathrm{mc}(A) \leqslant \min\left\{\sqrt{M},\sqrt{N}\right\}.$$

Let X be a finite set. For a family $\mathscr{C}$ of functions $X \to \{-1,+1\}$, define its *statistical query (SQ) dimension* $\mathrm{sq}(\mathscr{C})$ to be the largest integer d for which there are functions $f_1, f_2, \dots, f_d \in \mathscr{C}$ and a probability distribution μ on X such that

$$\left|\mathop{\mathbf{E}}_{x\sim\mu}\left[f_i(x)f_j(x)\right]\right| \leqslant \frac{1}{d} \qquad \text{for all } i \neq j.$$

For a sign matrix $A \in \{-1,+1\}^{M\times N}$, we define $\mathrm{sq}(A)$ to be the SQ dimension of the rows of A viewed as functions $\{1,2,\dots,N\} \to \{-1,+1\}$. The SQ dimension is an important quantity in learning theory. It was originally defined as a complexity measure in Kearns' *statistical query* model of learning (1993).

At this point, we have introduced five complexity measures of a sign matrix: $\mathrm{disc}^{\times}(A)$, $\mathrm{sq}(A)$, $\mathrm{dc}(A)$, $\mathrm{mc}(A)$, and $\mathrm{disc}(A)$. The relationships among them can

be stated concisely as follows:

$$\frac{1}{\operatorname{disc}^{\times}(A)} \;=_{\text{poly}}\; \operatorname{sq}(A) \;\leqslant_{\text{poly}}\; \operatorname{dc}(A) \;\leqslant_{\text{poly}}\; \operatorname{mc}(A) \;\approx\; \frac{1}{\operatorname{disc}(A)}$$

This diagram summarizes work by different authors at different times. We now traverse it left to right, giving precise quantitative statements.

THEOREM 2.4 (Sherstov 2008, Thm. 7.1). *Let A be a sign matrix. Then*

$$\sqrt{\frac{1}{2}\operatorname{sq}(A)} < \frac{1}{\operatorname{disc}^{\times}(A)} < 8\operatorname{sq}(A)^2.$$

THEOREM 2.5 (Sherstov 2008, Thm. 3.5). *Let A be a sign matrix. Then*

$$\operatorname{sq}(A) < 2\operatorname{dc}(A)^2.$$

THEOREM 2.6 (Ben-David *et al.* 2003). *Let $A \in \{-1,+1\}^{M\times N}$. Then*

$$\operatorname{dc}(A) \leqslant O(\operatorname{mc}(A)^2 \log(M+N)).$$

THEOREM 2.7 (Linial & Shraibman 2008). *Let A be a sign matrix. Then*

$$\frac{1}{8}\operatorname{mc}(A) \leqslant \frac{1}{\operatorname{disc}(A)} \leqslant 8\operatorname{mc}(A).$$

The following observation is immediate from Proposition 2.3 and Theorem 2.7:

PROPOSITION 2.8. *Let $A \in \{-1,+1\}^{M\times N}$. Then*

$$\operatorname{disc}(A) \geqslant \frac{1}{8\min\left\{\sqrt{M},\sqrt{N}\right\}}.$$

2.3. Learning theory. Let X be a finite set, such as $X = \{0,1\}^n$. A *concept class* $\mathscr{C}$ is any set of functions $X \to \{-1,+1\}$. We identify $\mathscr{C}$ with the sign matrix A whose rows are indexed by the functions of $\mathscr{C}$, columns indexed by the inputs $x \in X$, and entries given by $A(f,x) = f(x)$. In what follows, we use $\mathscr{C}$ and its corresponding sign matrix interchangeably.

Let μ be a probability distribution over X. Then the following is a natural notion of distance between functions $f, g\colon X \to \{-1,+1\}$:

$$\Delta_\mu(f,g) = \mathop{\mathbf{P}}_{x\sim\mu}\Big[f(x) \neq g(x)\Big].$$

A concept class $\mathscr{C}$ is *learnable* to accuracy ϵ and confidence δ under distribution μ from m examples if there is an algorithm L that, for every unknown $f \in \mathscr{C}$, takes as input i.i.d. examples $x^{(1)}, x^{(2)}, \ldots, x^{(m)} \sim \mu$ and their labels $f(x^{(1)}), f(x^{(2)}), \ldots, f(x^{(m)})$, and with probability at least $1-\delta$ produces a hypothesis $h\colon X \to \{-1,+1\}$ with $\Delta_\mu(h,f) \leqslant \epsilon$. The probability is over the random choice of examples and any internal randomization in L.

For a sign matrix A (and thus its corresponding concept class), define its *Vapnik-Chervonenkis (VC) dimension* $\mathrm{vc}(A)$ to be the largest d such that A features a $2^d \times d$ submatrix whose rows are the distinct elements of $\{-1,+1\}^d$. The VC dimension is a combinatorial quantity that exactly captures the learning complexity of a concept class. This is borne out by the following classical theorem (Vapnik & Chervonenkis 1971; Blumer *et al.* 1989).

Theorem 2.9 (VC Theorem; see Kearns & Vazirani 1994, Thm. 3.3). *Let $\mathscr{C}$ be a concept class and μ a distribution. Then $\mathscr{C}$ is learnable to accuracy ϵ and confidence δ under μ from*

$$m = O\left(\frac{1}{\epsilon}\log\frac{1}{\delta} + \frac{\mathrm{vc}(\mathscr{C})}{\epsilon}\log\frac{1}{\epsilon}\right)$$

examples. Moreover, any algorithm that outputs as a hypothesis some member of $\mathscr{C}$ consistent with the given m examples will successfully learn $\mathscr{C}$.

Theorem 2.9 almost matches the information-theoretic lower bounds on the number of examples necessary. These lower bounds come in different flavors; for example, see Kearns & Vazirani (1994, Thm. 3.5). We will need the following specialized version, which we state with a proof for the reader's convenience.

Proposition 2.10 (Information-theoretic barrier). *Let μ be a probability distribution and $\mathscr{C}$ a concept class such that $\Delta_\mu(f,f') > \epsilon$ for every two distinct $f, f' \in \mathscr{C}$. Then learning $\mathscr{C}$ to accuracy $\epsilon/2$ and confidence δ under μ requires $\log_2|\mathscr{C}| + \log_2(1-\delta)$ examples.*

PROOF. Let L be a learner for $\mathscr{C}$ that uses m examples, achieving accuracy $\epsilon/2$ and confidence δ. View L as a function $L(x^{(1)}, y_1, \ldots, x^{(m)}, y_m, r)$ that takes labeled training examples and a random string as input and outputs a hypothesis. With this notation, we have:

$$\mathop{\mathbf{E}}_{f \in \mathscr{C}} \left[\mathop{\mathbf{P}}_{x^{(1)},\ldots,x^{(m)},r} \left[\Delta_\mu \Big(f,\ L\Big(x^{(1)}, f(x^{(1)}), \ldots, x^{(m)}, f(x^{(m)}), r \Big) \Big) \leqslant \frac{\epsilon}{2} \right] \right] \geqslant 1 - \delta.$$

Reordering the expectation and probability operators yields

$$\mathop{\mathbf{E}}_{x^{(1)},\ldots,x^{(m)},r} \left[\mathop{\mathbf{P}}_{f \in \mathscr{C}} \left[\Delta_\mu \Big(f,\ L\Big(x^{(1)}, f(x^{(1)}), \ldots, x^{(m)}, f(x^{(m)}), r \Big) \Big) \leqslant \frac{\epsilon}{2} \right] \right] \geqslant 1 - \delta.$$

Thus, there is a fixed choice of $x^{(1)}, \ldots, x^{(m)}, r$ for which

$$\mathop{\mathbf{P}}_{f \in \mathscr{C}} \left[\Delta_\mu \Big(f,\ L\Big(x^{(1)}, f(x^{(1)}), \ldots, x^{(m)}, f(x^{(m)}), r \Big) \Big) \leqslant \frac{\epsilon}{2} \right] \geqslant 1 - \delta. \tag{2.11}$$

With $x^{(1)}, \ldots, x^{(m)}, r$ fixed in this way, algorithm L becomes a deterministic mapping from $\{-1,+1\}^m$ to the hypothesis space. In particular, L can output at most 2^m different hypotheses. Equation (2.11) says that L succeeds in producing an $\frac{\epsilon}{2}$-approximator for at least $(1-\delta)\,|\mathscr{C}|$ functions in $\mathscr{C}$. Since no hypothesis can be an $\frac{\epsilon}{2}$-approximator for two different functions in $\mathscr{C}$, we have $2^m \geqslant (1-\delta)\,|\mathscr{C}|$. □

For a thorough introduction to computational learning, see the textbook by Kearns and Vazirani (1994).

3. The communication gap

In this section, we prove our main result concerning communication under product vs. nonproduct distributions. We first recall an elegant simulation that relates the communication complexity of a sign matrix to its VC dimension.

THEOREM 3.1 (Kremer *et al.* 1999, Thm. 3.2). *Let A be a sign matrix. Let ϵ be given with $0 < \epsilon \leqslant 1/3$. Then*

$$D^{\times}_{\epsilon}(A) = O\left(\frac{1}{\epsilon}\, \mathrm{vc}(A) \log \frac{1}{\epsilon} \right).$$

Moreover, this communication cost can be achieved with a one-way protocol.

PROOF (Kremer *et al.* 1999). Let X and Y be the finite sets that index the columns and rows of A, respectively. Let $\mu = \mu_X \times \mu_Y$ be a given product distribution. Consider the following public-coin randomized one-way protocol for A. On input $(x, y) \in X \times Y$, Alice and Bob use their shared random bits to pick points

$$x^{(1)},\ x^{(2)},\ \ldots,\ x^{(m)} \in X$$

independently at random, according to μ_X. Here m is a parameter to be fixed later. Next, Bob sends Alice the values

$$A(y, x^{(1)}),\ A(y, x^{(2)}),\ \ldots,\ A(y, x^{(m)}).$$

At this point, Alice identifies any $y' \in Y$ with

$$\begin{aligned} A(y', x^{(1)}) &= A(y, x^{(1)}), \\ A(y', x^{(2)}) &= A(y, x^{(2)}), \\ &\vdots \\ A(y', x^{(m)}) &= A(y, x^{(m)}), \end{aligned}$$

and announces $A(y', x)$ as the output of the protocol.

In learning-theoretic terms, the protocol amounts to Alice learning the unknown row A_y of the matrix A from random labeled examples distributed according to μ_X. By the VC Theorem (Theorem 2.9), any row A'_y consistent with $m = O(\frac{1}{\epsilon} \operatorname{vc}(A) \log \frac{1}{\epsilon})$ labeled examples will, with probability $\geqslant 1 - \epsilon/2$, have $\Delta_{\mu_X}(A'_y, A_y) \leqslant \epsilon/2$. In particular, Alice's answer will be correct with probability at least $1 - \epsilon$ (with respect to μ_X and regardless of Bob's input y).

To summarize, we have obtained a public-coin randomized one-way protocol for A with cost m and error at most ϵ over $\mu = \mu_X \times \mu_Y$. By a standard averaging argument, there must be a one-way deterministic protocol with the same cost and error at most ϵ with respect to $\mu_X \times \mu_Y$. □

Our next ingredient is a combinatorial fact about sign matrices.

DEFINITION 3.2 (Zarankiewicz matrices). *Let $\mathscr{Z}(N, c)$ denote the family of $N \times N$ matrices with ± 1 entries that contain no submatrix of size $c \times c$ with all entries equal to* 1.

A classical result due to Bollobás (1978) states that $\mathscr{Z}(N, c)$ contains a considerable fraction of the matrices in $\{-1, +1\}^{N \times N}$. On the other hand, Alon *et al.* (1985) proved that all but a tiny fraction of the matrices in $\{-1, +1\}^{N \times N}$ have high dimension complexity. These two results were combined in the work of Ben-David *et al.* (2003) to the following effect:

THEOREM 3.3 (Ben-David *et al.* 2003, Thm. 12). *Let $c \geqslant 2$ be a fixed integer. Then all but a vanishing fraction of the matrices in $\mathscr{Z}(N, c)$ have dimension complexity $\Omega(N^{1-\frac{2}{c}})$.*

A glance at the proof of Ben-David *et al.* reveals the following somewhat more delicate result, which is what we will need in this paper.

THEOREM 3.4. *Let $\alpha > 0$ be a suitably small absolute constant. Let c be a given integer with*

$$2 \leqslant c \leqslant \alpha \sqrt{\frac{\log N}{\log \log N}}.$$

Then all but a vanishing fraction of the matrices in $\mathscr{Z}(N, c)$ have dimension complexity at least $\alpha N^{1-\frac{2}{c}}$.

We are now in a position to prove the main result of this section, which contains Theorem 1.2 from the Introduction as a special case.

THEOREM 3.5. *Let ϵ be given, $0 < \epsilon \leqslant \frac{1}{3}$. Then there exists a function $f\colon \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ with all of the following properties:*

$$\begin{aligned} D^{\times}_{\epsilon}(f) &\leqslant \Theta\left(\tfrac{1}{\epsilon^2} \log \tfrac{1}{\epsilon}\right), \\ R_{1/3}(f) &\geqslant \Theta(n), \\ \operatorname{disc}^{\times}(f) &\geqslant \epsilon^{\Theta(1/\epsilon^2)}, \\ \operatorname{disc}(f) &\leqslant O\left(2^{-n(\frac{1}{2}-\epsilon)}\right). \end{aligned}$$

PROOF. Let $\alpha > 0$ be the absolute constant from Theorem 3.4. If $\epsilon < \frac{4}{\alpha}\sqrt{\log n/n}$, the theorem holds trivially for the function INNER PRODUCT MODULO 2 (see Propositions 2.2 and 2.8).

In the contrary case, Theorem 3.4 is applicable with $c = 2\lceil 1/\epsilon \rceil$ and ensures the existence of $A \in \mathscr{Z}(2^n, c)$ with $\operatorname{dc}(A) \geqslant \alpha 2^{n(1-\epsilon)}$. Then

$$\begin{aligned} \operatorname{disc}(A) &\leqslant \frac{8}{\operatorname{mc}(A)} && \text{by Theorem 2.7} \\ &\leqslant O\left(\sqrt{\frac{n}{\operatorname{dc}(A)}}\right) && \text{by Theorem 2.6} \\ &\leqslant \Theta\left(2^{-n(\frac{1}{2}-\epsilon)}\right). && (3.6) \end{aligned}$$

By Proposition 2.1, we immediately conclude that

$$R_{1/3}(A) \geqslant \Theta(n). \tag{3.7}$$

Since every matrix in $\mathscr{Z}(2^n, c)$ has VC dimension at most $2c$, it follows from Theorem 3.1 that

$$D_\epsilon^\times(A) \leqslant \Theta\left(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon}\right). \tag{3.8}$$

In light of (3.8), Proposition 2.1 shows that

$$\operatorname{disc}^\times(A) \geqslant \epsilon^{\Theta(1/\epsilon^2)}. \tag{3.9}$$

The theorem follows from (3.6)–(3.9). □

Remark 3.10. As mentioned in the proof, the function f in question satisfies $\operatorname{dc}(f) \geqslant 2^{\Theta(n)}$. This is equivalent to saying that f has communication complexity $\Theta(n)$ in the unbounded-error model of Paturi and Simon (1986).

Remark 3.11. The bound $D_\epsilon^\times(f) \leqslant \left(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon}\right)$ in Theorem 3.5 can be achieved even by one-way protocols. This is because we bounded $D_\epsilon^\times(f)$ using Theorem 3.1, which gives a one-way communication protocol for the task.

4. SQ dimension and dimension complexity

The purpose of this section is to exhibit a large gap between the SQ dimension and dimension complexity of an $N \times N$ sign matrix. We start with a technical lemma.

Lemma 4.1 (VC and SQ dimensions). *Let $\mathscr{C}$ be a concept class. Then*

$$\operatorname{sq}(\mathscr{C}) \leqslant 2^{O(\operatorname{vc}(\mathscr{C}))}.$$

Proof. Let $\operatorname{sq}(\mathscr{C}) = d \geqslant 2$. Our goal is to show that $\operatorname{vc}(\mathscr{C}) = \Omega(\log d)$. By definition of the SQ dimension, there is a distribution μ and and functions $f_1, \ldots, f_d \in \mathscr{C}$ such that

$$\Delta_\mu(f_i, f_j) \geqslant \frac{1}{2} - \frac{1}{2d}$$

for all $i \neq j$. In particular, $\Delta_\mu(f_i, f_j) > 1/5$. Thus, the information-theoretic barrier (Proposition 2.10) shows that learning $\mathscr{C}$ to accuracy $1/10$ and confidence $1/2$ requires

$$m \geqslant \Omega(\log d)$$

examples. Yet by the VC Theorem (Theorem 2.9), the number of examples needed is at most

$$m \leqslant O(\mathrm{vc}(\mathscr{C})).$$

Comparing these lower and upper bounds on m yields the desired result. □

We are now prepared for the main result of this section:

THEOREM 1.3 (Restated from p. 5). *Let $\epsilon > 0$ be an arbitrary constant. Then there exists a matrix $A \in \{-1,+1\}^{N \times N}$ with*

$$\begin{aligned} \mathrm{sq}(A) &= O(1), \\ \mathrm{dc}(A) &= \Omega(N^{1-\epsilon}). \end{aligned}$$

PROOF. Let $c = 2\lceil 1/\epsilon \rceil$. By Theorem 3.4, there exists a matrix $A \in \mathscr{Z}(N, c)$ with $\mathrm{dc}(A) = \Omega(N^{1-\epsilon})$. Since every matrix in $\mathscr{Z}(N, c)$ has VC dimension at most $2c$, it follows from Lemma 4.1 that $\mathrm{sq}(A) \leqslant 2^{O(c)} = O(1)$. □

Acknowledgements

I would like to thank Anna Gál, Dmitry Gavinsky, Adam Klivans, and the anonymous reviewers for their helpful feedback on an earlier version of this manuscript. I learned about the Kushilevitz-Nisan problem from Anna Gál's communication complexity class at the University of Texas at Austin, which she taught in a most inspiring way. This research was supported by Adam Klivans' NSF CAREER Award and NSF Grant CCF-0728536.

References

NOGA ALON, PETER FRANKL & VOJTECH RÖDL (1985). Geometrical Realization of Set Systems and Probabilistic Communication Complexity. In *Proc. of the 26th Symposium on Foundations of Computer Science (FOCS)*, 277–280.

LÁSZLÓ BABAI, PETER FRANKL & JANOS SIMON (1986). Complexity classes in communication complexity theory. In *Proc. of the 27th Symposium on Foundations of Computer Science (FOCS)*, 337–347.

Shai Ben-David, Nadav Eiron & Hans Ulrich Simon (2003). Limitations of learning via embeddings in Euclidean half spaces. *J. Mach. Learn. Res.* **3**, 441–461.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler & Manfred K. Warmuth (1989). Learnability and the Vapnik-Chervonenkis dimension. *J. ACM* **36**(4), 929–965.

Béla Bollobás (1978). *Extremal Graph Theory.* Academic Press, New York.

Benny Chor & Oded Goldreich (1988). Unbiased Bits from Sources of Weak Randomness and Probabilistic Communication Complexity. *SIAM J. Comput.* **17**(2), 230–261.

Jürgen Forster (2002). A linear lower bound on the unbounded error probabilistic communication complexity. *J. Comput. Syst. Sci.* **65**(4), 612–625.

Jürgen Forster, Matthias Krause, Satyanarayana V. Lokam, Rustam Mubarakzjanov, Niels Schmitt & Hans-Ulrich Simon (2001). Relations Between Communication Complexity, Linear Arrangements, and Computational Complexity. In *Proc. of the 21st Conf. on Foundations of Software Technology and Theoretical Computer Science (FST TCS)*, 171–182.

Jürgen Forster, Niels Schmitt, Hans Ulrich Simon & Thorsten Suttorp (2003). Estimating the Optimal Margins of Embeddings in Euclidean Half Spaces. *Mach. Learn.* **51**(3), 263–281.

Jürgen Forster & Hans Ulrich Simon (2006). On the smallest possible dimension and the largest possible margin of linear arrangements representing given concept classes. *Theor. Comput. Sci.* **350**(1), 40–48.

Bala Kalyanasundaram & Georg Schnitger (1992). The Probabilistic Communication Complexity of Set Intersection. *SIAM J. Discrete Math.* **5**(4), 545–557.

Michael Kearns (1993). Efficient noise-tolerant learning from statistical queries. In *Proc. of the 25th Symposium on Theory of Computing (STOC)*, 392–401.

Michael J. Kearns & Umesh V. Vazirani (1994). *An Introduction to Computational Learning Theory.* MIT Press, Cambridge.

Ilan Kremer, Noam Nisan & Dana Ron (1999). On Randomized One-Round Communication Complexity. *Computational Complexity* **8**(1), 21–49.

Eyal Kushilevitz & Noam Nisan (1997). *Communication complexity.* Cambridge University Press, New York.

Nathan Linial & Adi Shraibman (2008). Learning complexity vs. communication complexity. In *Proc. of the 23rd Conf. on Computational Complexity (CCC)*, 53–63.

Nati Linial, Shahar Mendelson, Gideon Schechtman & Adi Shraibman (2007). Complexity Measures of Sign Matrices. *Combinatorica* **27**(4), 439–463.

Ramamohan Paturi & Janos Simon (1986). Probabilistic communication complexity. *J. Comput. Syst. Sci.* **33**(1), 106–123.

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

Alexander A. Sherstov (2008). Halfspace Matrices. *Comput. Complex.* **17**(2), 149–178. Preliminary version in 22nd CCC, 2007.

Vladimir Naumovich Vapnik & Alexey Yakovlevich Chervonenkis (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications* **16**(2), 264–280.

Andrew Chi-Chih Yao (1983). Lower Bounds by Probabilistic Arguments. In *Proc. of the 24th Symposium on Foundations of Computer Science (FOCS)*, 420–428.

Manuscript received August 24, 2008

Alexander A. Sherstov
Department of Computer Sciences
The University of Texas at Austin
1 University Station C0500
Austin, TX 78712-0233 USA
sherstov@cs.utexas.edu
http://www.cs.utexas.edu/~sherstov