Communication Complexity under Product and Nonproduct Distributions

Alexander A. Sherstov

The University of Texas at Austin
Department of Computer Sciences
Austin, TX 78712 USA
sherstov@cs.utexas.edu

May 14, 2007

Abstract

We completely solve an open problem of Kushilevitz and Nisan (1997) in communication complexity. Let $\mathrm{R}_\epsilon^{\mathrm{pub}}(f)$ and $\mathrm{D}_\epsilon^\mu(f)$ denote the randomized and μ-distributional communication complexities of f, respectively (ϵ a small constant). Yao's well-known Minimax Principle states that

$$\mathrm{R}_\epsilon^{\mathrm{pub}}(f) = \max_\mu \left\{ \mathrm{D}_\epsilon^\mu(f) \right\}.$$

Kushilevitz and Nisan (1997) ask whether this equality is approximately preserved if the maximization is taken over product distributions only, rather than all distributions μ:

$$\mathrm{R}_\epsilon^{\mathrm{pub}}(f) \overset{?}{=} \left(\max_{\mu \text{ product}} \left\{ \mathrm{D}_\epsilon^\mu(f) \right\} \right)^{O(1)}.$$

We refute this hypothesis in the strongest possible terms. Namely, we show the existence of a function $f : \{-1,+1\}^n \times \{-1,+1\}^n \to \{-1,+1\}$ for which $\max_{\mu \text{ product}} \left\{ \mathrm{D}_\epsilon^\mu(f) \right\} = O(1)$ but $\mathrm{R}_{1/3}^{\mathrm{pub}}(f) = \Omega(n)$. Furthermore, f has discrepancy $O(2^{-n(1/2-\epsilon)})$, which is almost the smallest possible. In particular, f is a hardest function for every major model of communication. Yet, the distributional method restricted to product distributions can certify at best an $\Omega(1)$ communication lower bound for f.

Our result also gives an essentially optimal separation, $\Omega(1)$ vs. $O(2^{-n(1/2-\epsilon)})$, between discrepancy under product and nonproduct distributions, improving on the author's recent result (Sherstov 2007). Finally, we give an essentially optimal separation, $O(1)$ vs. $\Omega(N^{1-\epsilon})$, between between the statistical-query complexity and sign rank of an $N \times N$ sign matrix. This settles a open question recently posed by the author (Sherstov 2007) and completes the taxonomy of the main complexity measures of sign matrices.

1 Introduction

Among the primary models of communication complexity is the *randomized model* [10, Chapter 3]. Let X and Y be finite sets. Two parties, Alice and Bob, have access to disjoint parts $x \in X$ and $y \in Y$ of the input to a fixed function $f : X \times Y \to \{-1, +1\}$ and must therefore communicate to evaluate $f(x, y)$. They can use an unlimited number of shared random bits. On every input, the players must compute the correct value with probability at least $2/3$. The *cost* of a protocol is number of bits exchanged in the worst case. The *randomized complexity* $\mathrm{R}^{\mathrm{pub}}_{1/3}(f)$ of a function f is the cost of the best protocol for f.

A closely related notion is that of distributional complexity. Let μ be a probability distribution on $X \times Y$. The μ-distributional communication complexity of f, denoted $\mathrm{D}^{\mu}_{1/3}(f)$, is the cost of the optimal deterministic protocol for f with error at most $1/3$ with respect to μ. Using the Minimax Theorem for zero-sum games, Yao [16] gave a simple proof that

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

where the constant $1/3$ can be replaced by any other. Yao's 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 $\mathrm{D}^{\mu}_{1/3}(f)$ 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. While *product* distributions $\mu(x, y) = \mu_X(x)\mu_Y(y)$ are easier to analyze, they do not always yield the optimal lower bounds. A standard example of this phenomenon is the set disjointness function DISJ on n-bit strings: every product distribution μ has $\mathrm{D}^{\mu}_{1/3}(\mathrm{DISJ}) = O(\sqrt{n} \log n)$ (see [10]), although $\mathrm{R}^{\mathrm{pub}}(\mathrm{DISJ}) = \Theta(n)$ (see [6, 14]). Let

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

The above considerations motivated Kushilevitz and Nisan (1997) to pose the following problem:

Research Problem (Kushilevitz and Nisan [10, p. 37])**.** Can restricting the distribution μ to be a product distribution affect the resulting lower bound on $\mathrm{R}^{\mathrm{pub}}(f)$ by more than a polynomial factor? Formally, is $\mathrm{R}^{\mathrm{pub}}_{1/3}(f) = (\mathrm{D}^{\times}_{1/3}(f))^{O(1)}$?

Since its formulation, this problem has seen little progress. Kremer, Nisan, and Ron [9] studied its restriction to *one-way* protocols and obtained a separation of $O(1)$ vs. $\Omega(n)$ for the "greater than" function GT. Unfortunately, a function can have vastly different communication complexity in the one-way and usual (two-way) randomized models. Such is the case of GT, whose two-way randomized complexity is a mere $O(\log n)$.

Another step toward solving the Kushilevitz-Nisan question has been recently taken by the author [15]. 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 indeed essential to the discrepancy method, a common 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 $f : \{-1,+1\}^n \times \{-1,+1\}^n \to \{-1,+1\}$ with $\mathrm{D}^{\times}_{1/3}(f) = O(1)$ and $\mathrm{R}^{\mathrm{pub}}_{1/3}(f) = \Omega(n)$. In fact, the prove the following more delicate result:

Theorem 1.1. *Let $\epsilon > 0$ be an arbitrary constant. Then there exists a function $f : \{-1,+1\}^n \times \{-1,+1\}^n \to \{-1,+1\}$ with all of the following properties:*

$$\mathrm{D}^{\times}_{\epsilon}(f) = O(1),$$

$$\mathrm{R}^{\mathrm{pub}}_{1/3}(f) = \Omega(n),$$

$$\mathrm{disc}^{\times}(f) = \Omega(1),$$

$$\mathrm{disc}(f) = O(2^{-n(\frac{1}{2}-\epsilon)}).$$

The notation $\mathrm{disc}^{\times}(f)$ stands for the smallest discrepancy of f under a product distribution, by analogy with $\mathrm{D}^{\times}(f)$.

A key aspect of Theorem 1.1 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 2.7 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 exponentially small advantage on every input (say, $2^{-n/4}$). Finally, it is clear from our proof (see Remark 3.4) that f has complexity $\Omega(n)$ in the unbounded-error model [13], which has an even weaker success criterion.

To summarize the previous paragraph, f has the highest communication complexity in every major model. Yet, the distributional method restricted to product distributions can certify at best an $\Omega(1)$ lower bound. In this sense, we refute the hypothesis of the Kushilevitz-Nisan problem in the strongest possible terms.

Finally, Theorem 1.1 also improves on our previously obtained [15] exponential separation for discrepancy. 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 1.1 amplifies this gap to what is essentially optimal, although the function is no longer explicit.

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 algebraic 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 [1–5, 11, 12, 15].

Fundamental complexity measures of 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 [7] and turns out be be intimately linked with discrepancy.)
- $\mathrm{dc}(A)$, 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. [1]:

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

Linial and Shraibman [11] showed that $\mathrm{mc}(A)$ and $1/\mathrm{disc}(A)$ are always within a factor of 8. The author [15] has recently extended these two results to the following picture:

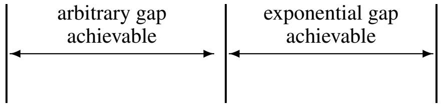
$$\frac{1}{\mathrm{disc}^\times(A)} \quad =_{\mathsf{poly}} \quad \mathrm{sq}(A) \quad \leqslant_{\mathsf{poly}} \quad \overbrace{\mathrm{dc}(A) \quad \leqslant_{\mathsf{poly}} \quad \mathrm{mc}(A)}^{\text{exponential gap achievable}} \quad \approx \quad \frac{1}{\mathrm{disc}(A)}$$

The symbols $\leqslant_{\mathsf{poly}}$ and $=_{\mathsf{poly}}$ in the above diagram have their intuitive meaning; we give precise statements in Section 2.2. The only missing piece from this diagram is the gap between sq(A) and dc(A), which is left as an open problem in [15]. We solve this problem, showing that the gap between sq(A) and dc(A) can be *arbitrary*:

Theorem 1.2 (SQ dimension vs. dimension complexity)**.** *Let $\epsilon > 0$ be an arbitrary constant. Then there exists a matrix family $A \in \{-1,+1\}^{N\times N}$ with*

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

It is easy to show (see Section 2.2) that $\mathrm{dc}(A) \leqslant \min\{M,N\}$ for every $A \in \{-1,+1\}^{M\times N}$. In this light, Theorem 1.2 gives essentially the best gap that can exist by definition. This completes our taxonomy to the following overall picture:



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

Our Techniques. A key feature of our approach is to view a sign matrix both as a communication problem and as a set of Boolean functions to learn (namely, the matrix rows). This perspective invites the use of tools from communication complexity and learning theory. One of the ingredients in our proofs is a simulation due to Kremer, Nisan, and Ron [9] that links the one-way communication complexity of a matrix to its *VC dimension*. We also recall a combinatorial fact due to Ben-David et al. [1] about matrices with low VC dimension. To combine these two results, we use the above taxonomy of complexity measures.

2 Preliminaries

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

2.1 Communication Complexity

We consider Boolean functions $f : X\times Y \to \{-1,+1\}$. Typically $X = Y = \{-1,+1\}^n$, but we also allow X and Y to be arbitrary sets, possibly of unequal cardinality. We identify a function f with its *communication matrix* $A = [f(x,y)]_{y,x} \in$

$\{-1,+1\}^{|Y|\times|X|}$. 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 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

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

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

Proposition 2.1 (Kushilevitz and Nisan [10, pp. 36–38])**.** *For every Boolean function $f(x,y)$, every distribution μ, and every $\gamma > 0$,*

$$\mathrm{R}^{\mathrm{pub}}_{1/2-\gamma/2}(f) \;\geqslant\; \mathrm{D}^{\mu}_{1/2-\gamma/2}(f) \;\geqslant\; \log_2 \frac{\gamma}{\mathrm{disc}_\mu(f)}.$$

A definitive resource for further details is the book of Kushilevitz and Nisan [10].

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

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

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 $\frac{1}{\sqrt{M}}\mathbf{a}_1, \ldots, \frac{1}{\sqrt{M}}\mathbf{a}_N$. 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 2.2. *Let $A \in \{-1,+1\}^{M\times N}$. Then*

$$\begin{aligned} 1 &\leqslant \mathrm{dc}(A) \leqslant \min\{M, N\}, \\ 1 &\leqslant \mathrm{mc}(A) \leqslant \min\{\sqrt{M},\ \sqrt{N}\}. \end{aligned}$$

We say that a matrix $R \in \mathbb{R}^{M\times N}$ *sign-represents* a matrix $A \in \{-1,+1\}^{M\times N}$, denoted $A = \mathrm{sign}(R)$, if $A_{ij}R_{ij} > 0$ for all i, j. Observe that the dimension complexity of a sign matrix is the minimum rank of any real matrix that sign-represents it.

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

$$f_1,\ f_2,\ \ldots,\ f_d \in \mathscr{C}$$

and a probability distribution μ on X such that

$$\left| \mathop{\mathbf{E}}_{x\sim\mu} [f_i(x)f_j(x)] \right| \leqslant \frac{1}{d} \qquad \text{for all } i \neq j. \tag{2.1}$$

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 $\{1, 2, \ldots, N\} \to \{-1,+1\}$. It is a simple exercise to show that any functions $f_1, f_2, \ldots, f_d$ that satisfy (2.1) must be linearly independent, and thus

$$\mathrm{sq}(A) \leqslant \mathrm{rank}(A) \qquad \text{for all } A. \tag{2.2}$$

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 [7]. But, as we shall see shortly, it naturally fits in our taxonomy of complexity measures of sign matrices.

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 $\mathrm{disc}(A)$. They are related in an elegant way, as follows:

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

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

Theorem 2.3 (Sherstov [15, Thm. 7.1])**.** *Let A be a sign matrix. Then*

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

Theorem 2.4 (Sherstov [15, Thm. 3.2])**.** *Let A be a sign matrix. Then*

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

Theorem 2.5 (Ben-David, Eiron, and Simon [1])**.** *Let $A \in \{-1,+1\}^{M\times N}$. Then*

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

Theorem 2.6 (Linial and Shraibman [11])**.** *Let A be a sign matrix. Then*

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

The following observation is immediate from Proposition 2.2 and Theorem 2.6:

Proposition 2.7. *Let $A \in \{-1,+1\}^{M\times N}$. Then*

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

2.3 Learning Theory

Let X be a finite set, such as $X = \{-1,+1\}^n$. 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 functions of $\mathscr{C}$, columns indexed by inputs $x \in X$, and entries given by $A(f,x) = f(x)$. In other words, A's rows are precisely the functions of $\mathscr{C}$. 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:

$$\Delta_\mu(f,g) \stackrel{\mathrm{def}}{=} \mathop{\mathbf{Pr}}_{x\sim\mu}[f(x) \neq g(x)].$$

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 $x^1, \ldots, x^m \sim \mu$ and their labels $f(x^1), \ldots, f(x^m)$, and with probability at least $1 - \delta$ produces a hypothesis h with $\Delta_\mu(h, f) \leqslant \epsilon$. The latter 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:

Theorem 2.8 (VC Theorem; see [8, 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*

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

examples.

Theorem 2.8 almost matches the information-theoretic lower bounds on the number of examples necessary. These lower bounds come in different flavors; for example, see [8, Thm. 3.5]. We will need the following specialized version, which we state with a proof for the reader's convenience.

Proposition 2.9 (Information-theoretic barrier)**.** *Let μ be a probability distribution and $\mathscr{C}$ be 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* $\log|\mathscr{C}| + \log(1 - \delta)$ *examples.*

Proof. Let L be a learner for $\mathscr{C}$ that uses m examples. View L as a deterministic function $L(x^1, y^1, \ldots, x^m, y^m, r)$ that takes training examples and a random string as input and outputs a hypothesis. With this notation, we have:

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

Reordering the expectation and probability operators yields

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

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

$$\mathop{\mathbf{Pr}}_{f \in \mathscr{C}}\left[\Delta_\mu(f,\ L(x^1, f(x^1), \ldots, x^m, f(x^m), r)) \leqslant \frac{\epsilon}{2}\right] \geqslant 1 - \delta. \qquad (2.3)$$

With $x^1, \ldots, x^m, r$ thus fixed, 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 (2.3) 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 theory, see the textbook by Kearns and Vazirani [8].

3 Communication Gap

In this section, we prove our main results 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, Nisan, and Ron [9, Thm. 3.2])**.** *Let A be a sign matrix, $\epsilon > 0$ an arbitrary constant. Then* $\mathrm{D}^{\times}_{\epsilon}(A) = O(\mathrm{vc}(A))$.

Proof (Kremer, Nisan, Ron [9]). 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 protocol for A. Alice and Bob use their public coin to pick points

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

independently at random, according to μ_X. Here m is a parameter we will fix 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 2.8), any row A'_y consistent with $m = O(\mathrm{vc}(A))$ labeled examples will, with probability $\epsilon/2$, 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 protocol for A with cost $O(\mathrm{vc}(A))$ and error at most ϵ over $\mu = \mu_X \times \mu_Y$. By a standard averaging argument, there must be a deterministic protocol with the same cost and error at most ϵ. □

Our next ingredient is a combinatorial fact about sign matrices.

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

The key property of $\mathscr{Z}(N, c)$ for our purposes is the following result:

Theorem 3.3 (Ben-David, Eiron, and Simon [1, 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 $\Omega(N^{1-\frac{2}{c}})$.*

We are now in a position to prove the main result of this section.

Theorem 1.1 (Restated from p. 2). *Let $\epsilon > 0$ be an arbitrary constant. Then there exists a function $f : \{-1, +1\}^n \times \{-1, +1\}^n \to \{-1, +1\}$ with all of the following properties:*

$$\mathrm{D}^{\times}_{\epsilon}(f) = O(1),$$

$$\mathrm{R}^{\mathrm{pub}}_{1/3}(f) = \Omega(n),$$

$$\mathrm{disc}^{\times}(f) = \Omega(1),$$

$$\mathrm{disc}(f) = O(2^{-n(\frac{1}{2}-\epsilon)}).$$

Proof. Let $c = 2\lceil 1/\epsilon \rceil$. Theorem 3.3 ensures the existence of $A \in \mathscr{Z}(2^n, c)$ with $\mathrm{dc}(A) = \Omega(2^{n(1-\epsilon)})$. Then

$$\mathrm{disc}(A) \overset{\text{Thm. 2.6}}{\leqslant} \frac{8}{\mathrm{mc}(A)} \overset{\text{Thm. 2.5}}{\leqslant} O\left(\sqrt{\frac{n}{\mathrm{dc}(A)}}\right) = O\left(2^{-n(\frac{1}{2}-\epsilon)}\right). \tag{3.1}$$

By Proposition 2.1, we immediately conclude that

$$\mathrm{R}^{\mathrm{pub}}_{1/3}(A) = \Omega(n). \tag{3.2}$$

On the other hand, it is clear that every matrix in $\mathscr{Z}(2^n, c)$ has VC dimension at most $2c$. Theorem 3.1 now implies that

$$\mathrm{D}^{\times}_{\epsilon}(A) = O(1). \tag{3.3}$$

In light of (3.3), Proposition 2.1 shows that

$$\mathrm{disc}^{\times}(A) = \Omega(1). \tag{3.4}$$

The theorem follows from (3.1)–(3.4). □

Remark 3.4. It is clear from the proof that the function f in question satisfies $\mathrm{dc}(f) \geqslant 2^{\Omega(n)}$. This is equivalent to saying that f has communication complexity $\Omega(n)$ in the unbounded error model of Paturi and Simon [13].

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*

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

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

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

for all $i \neq j$. In particular, $\Delta_\mu(f_i, f_j) \geqslant 1/4$. Thus, the information-theoretic barrier (Proposition 2.9) 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 2.8), the number of examples needed is at most

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

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

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

Theorem 1.2 (Restated from p. 4). *Let $\epsilon > 0$ be an arbitrary constant. Then there exists a matrix family $A \in \{-1,+1\}^{N\times N}$ with*

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

Proof. Let $c = 2\lceil 1/\epsilon \rceil$. By Theorem 3.3, there exists a matrix $A \in \mathscr{X}(N, c)$ with

$$\mathrm{dc}(A) = \Omega(N^{1-\epsilon}). \tag{4.1}$$

On the other hand, it is clear that every matrix in $\mathscr{X}(N, c)$ has VC dimension at most $2c$. Therefore, Lemma 4.1 shows that

$$\mathrm{sq}(A) \leqslant 2^{O(c)} = O(1). \tag{4.2}$$

The theorem follows from (4.1) and (4.2). □

5 Further Notes on the VC and SQ Dimensions

In Section 4, we obtained a separation between between the SQ dimension and dimension complexity. Instrumental to that result was the relationship between two learning-theoretic quantities, the VC and SQ dimensions. This section concludes with a closer look at them.

For a given sign matrix A, and let $\mathrm{vc}(A^{\mathsf{T}})$ denote the VC dimension of the columns of A when viewed as Boolean functions. Define $\mathrm{sq}(A^{\mathsf{T}})$ analogously. It is well-known that there are matrices A with an exponential gap between $\mathrm{vc}(A)$ and $\mathrm{vc}(A^{\mathsf{T}})$. For example, the $2^n \times n$ matrix IND whose rows are the distinct vectors in $\{-1,+1\}^n$ satisfies:

$$\mathrm{vc}(\mathrm{IND}) = n, \qquad \mathrm{vc}(\mathrm{IND}^{\mathsf{T}}) = \lfloor \log n \rfloor.$$

This gap is in fact the largest possible since

$$\mathrm{vc}(A^{\mathsf{T}}) \geqslant \lfloor \log\ \mathrm{vc}(A) \rfloor \qquad \text{for all } A.$$

The reason for this inequality is as follows. If $\mathrm{vc}(A) = d$, then A contains a submatrix of size $\lfloor \log d \rfloor \times 2^{\lfloor \log d \rfloor}$ whose columns are all the possible vectors in $\{-1,+1\}^{\lfloor \log d \rfloor}$.

By contrast, the gap between $\mathrm{sq}(A)$ and $\mathrm{sq}(A^\mathsf{T})$ is always at most polynomial. Specifically, the author has shown [15, Cor. 7.1.1] that

$$\left(\frac{1}{32}\,\mathrm{sq}(\mathscr{C})\right)^{1/4} < \mathrm{sq}(\mathscr{C}^\mathsf{T}) < 32\,\mathrm{sq}(\mathscr{C})^4 \qquad \text{for all } A. \tag{5.1}$$

This result follows from Theorem 2.3 and the fact that $\mathrm{disc}^\times(A) = \mathrm{disc}^\times(A^\mathsf{T})$.

Finally, we examine the relationship between the VC and SQ dimensions for the same matrix. The result were are about to state is an extension of Lemma 4.1 above.

Proposition 5.1. *Let A be a sign matrix. Then:*

$$\max\{\tfrac{1}{2}\,\mathrm{vc}(A), \mathrm{vc}(A^\mathsf{T})\} \;\leqslant\; \mathrm{sq}(A) \;\leqslant\; 2^{O(\min\{\mathrm{vc}(A),\mathrm{vc}(A^\mathsf{T})\})}.$$

Proof. It is clear from the definitions that

$$\mathrm{sq}(A) \geqslant \mathrm{vc}(A^\mathsf{T}) \qquad \text{and} \qquad \mathrm{sq}(A) \geqslant 2^{\lfloor \log\ \mathrm{vc}(A)\rfloor} \geqslant \frac{1}{2}\,\mathrm{vc}(A).$$

On the other hand, Lemma 4.1 shows that

$$\mathrm{sq}(A) \leqslant 2^{O(\mathrm{vc}(A))}.$$

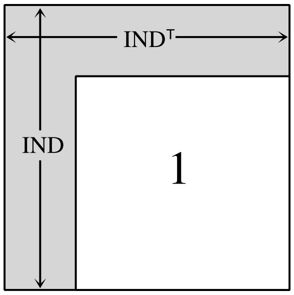
In view of (5.1), we also have

$$\mathrm{sq}(A) \leqslant 2^{O(\mathrm{vc}(A^\mathsf{T}))}.$$

These four inequalities complete the proof. □

We now show that Proposition 5.1 is best possible in that the quantity $\mathrm{sq}(A)$ really can range anywhere between the stated lower and upper bounds.

- Consider the following sign matrix A of size $2^n \times 2^n$:



The $2^n \times n$ submatrix IND is as defined earlier. It is clear that $\mathrm{vc}(A) = \mathrm{vc}(A^{\mathsf{T}}) = n$. By (2.2), we have $\mathrm{sq}(A) \leqslant \mathrm{rank}(A) \leqslant 2n$.

- Consider now the $2^n \times 2^n$ Hadamard matrix

$$A = [\mathrm{PARITY}(x_1 \wedge y_1, \ldots, x_n \wedge y_n)]_{x,y}.$$

  Here again one can show that $\mathrm{vc}(A) = \mathrm{vc}(A^{\mathsf{T}}) = n$. However, we now have $\mathrm{sq}(A) = 2^n$ since the rows of A are orthogonal.

To summarize, in both examples above we have $\mathrm{vc}(A) = \mathrm{vc}(A^{\mathsf{T}}) = n$, and thus Proposition 5.1 implies that

$$n \leqslant \mathrm{sq}(A) \leqslant 2^{O(n)}.$$

In the first example it turns out that $\mathrm{sq}(A) \leqslant 2n$, while in the second $\mathrm{sq}(A) = 2^n$. Hence, Proposition 5.1 cannot be strengthened in general.

References

[1] S. Ben-David, N. Eiron, and H. U. Simon. Limitations of learning via embeddings in Euclidean half spaces. *J. Mach. Learn. Res.*, 3:441–461, 2003.

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

[3] J. Forster, M. Krause, S. V. Lokam, R. Mubarakzjanov, N. Schmitt, and H.-U. Simon. Relations between communication complexity, linear arrangements, and computational complexity. In *FST TCS '01: Proceedings of the 21st Conference on Foundations of Software Technology and Theoretical Computer Science*, pages 171–182, London, UK, 2001. Springer-Verlag.

[4] J. Forster, N. Schmitt, H. U. Simon, and T. Suttorp. Estimating the optimal margins of embeddings in Euclidean half spaces. *Mach. Learn.*, 51(3):263–281, 2003.

[5] J. Forster and H. U. Simon. On the smallest possible dimension and the largest possible margin of linear arrangements representing given concept classes. *Theor. Comput. Sci.*, 350(1):40–48, 2006.

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

[7] M. Kearns. Efficient noise-tolerant learning from statistical queries. In *Proceedings of the 25th Annual ACM Symposium on Theory of Computing (STOC)*, pages 392–401, 1993.

[8] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, MA, USA, 1994.

[9] I. Kremer, N. Nisan, and D. Ron. On randomized one-round communication complexity. *Computational Complexity*, 8(1):21–49, 1999.

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

[11] N. Linial and A. Shraibman. Learning complexity vs. communication complexity. Manuscript at `http://www.cs.huji.ac.il/~nati/PAPERS/lcc.pdf`, December 2006.

[12] N. Linial and A. Shraibman. Lower bounds in communication complexity based on factorization norms. Manuscript at `http://www.cs.huji.ac.il/~nati/PAPERS/ccfn.pdf`, December 2006.

[13] R. Paturi and J. Simon. Probabilistic communication complexity. *J. Comput. Syst. Sci.*, 33(1):106–123, 1986.

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

[15] A. A. Sherstov. Halfspace matrices. In *Proc. of the 22nd Conference on Computational Complexity (CCC)*, 2007.

[16] A. C.-C. Yao. Lower bounds by probabilistic arguments. In *FOCS*, pages 420–428, 1983.