COMMUNICATION COMPLEXITY

AMIR YEHUDAYOFF

Communication appears in many places. People communicate, computers communicate, the components of a computer communicate, prices can be thought of as a means of communication, etc.

These notes were written for "Complexity as a kaleidoscope" held at CIRM, 2025.

1. Introduction

We start with the simplest case of two parties communicating (introduced by Yao). There are two players Alice and Bob. Alice knows some information x and Bob knows y. They communicate to achieve a common goal. In our setting, the goal is to compute some function $f(x,y)$. They communicate in turns and in bits. For example, Alice starts by saying 01, then Bob replies by 111, then Alice and so forth, until they reach a state when they both know $f(x,y)$ and the goal is achieved. For example, Alice can send an encoding of x to Bob, and then Bob sends the answer $f(x,y)$ to Alice. The parties are not assumed to be computationally bounded, because we mostly care about communication.

Let us introduce the model. The inputs are $x, y \in \{0,1\}^n$. The target is to compute $f : \{0,1\}^n \times \{0,1\}^n \to Z$. The communication is represented by a protocol π. This is a rooted binary tree. Each internal node is labelled by either A or B, indicating "who is talking now". Each node v is also labelled by $g_v : \{0,1\}^n \to \{0,1\}$. The leaves are labelled by elements of Z.

The communication is determined by the protocol tree. If the current node v is A then Alice computes $g_v(x)$ and sends it to Bob and they go to the "$g_v(x)$ child of v". And similarly for Bob. When they reach a leaf labelled z they "output z". The protocol computes f if the output of the protocol on input x, y is $f(x,y)$.

Protocols have costs and functions have complexities. The depth of the tree $|\pi|$ is the cost of the protocol. The deterministic communication complexity $D(f)$ of a function is the minimum cost of a protocol for f.

Example. $XOR(x,y) = x_1 \oplus x_2 \oplus \cdots \oplus x_n \oplus y_1 \oplus \cdots \oplus y_n$.

Example (equality)**.** $eq(x,y) = 1_{x=y}$.

Remark. *The trivial upper bound for boolean f is $D(f) \leq n+1$. We think of polylog(n) as "efficient".*

Remark. *Another way to think of the same model is as follows. The input x chooses for each node v owned by Alice one of the children of v. The input y does the same for nodes owned by Bob. Now, all inner nodes in the tree have "arrows down". The protocol is executed by following the unique root-leaf path.*

Randomness

Randomized complexity.

A randomized protocol is a distribution on deterministic protocols. The players first jointly sample a deterministic protocol π, and then they get x, y and run $\pi(x,y)$. The standard cost of the randomized protocol, which we also denote by $|\pi|$, is the maximum cost of a deterministic protocol in its support. The success criterion is that for all x, y,

$$\Pr[\pi(x,y) = f(x,y)] \geq \frac{2}{3}.$$

We get the randomized communication complexity $R(f)$ of f.

Remark. *We can amplify the success probability from $\frac{2}{3}$ to $1-\delta$ by repeating the protocol for $O(\log 1/\delta)$ times and taking a majority vote.*

Remark. *It also makes sense to measure $|\pi|$ as expected cost.*

Example (equality)**.** *The players jointly sample p, q independently and uniformly at random $\{0,1\}^n$. Alice compute the two bits $\oplus_i p_i x_i$ and $\oplus_i q_i x_i$ and send them to Bob. Bob replies by "equal" if the two analogous bits for y are the same, and by "not equal" otherwise. The cost is three and the correctness holds:*

— *If $x = y$ then $\Pr[\pi(x,y) = 1] = 1$.*

— *If $x \neq y$ then $\Pr[\oplus_i p_i x_i = \oplus_i p_i y_i] = \frac{1}{2}$ and so $\Pr[\pi(x,y) = 1] \leq \frac{1}{4}$.*

Distributional complexity.

Randomness can also be thought of in terms of average-case complexity. The input can be random. Denote by μ a distribution on $\{0,1\}^n \times \{0,1\}^n$. In this setting, the success criterion is

$$\Pr[\pi(x,y) = f(x,y)] \geq \frac{2}{3}$$

where the distribution is both over π and over (x,y). In this setting, the optimal protocols can be assumed to be deterministic (because we can fix the randomness of π above). Consequently, we denote the complexity of f over μ by $D_\mu(f)$.

A very important and general result relates randomized complexity and distributional complexity. It follows from von Neumann's minimax theorem (from game theory) and is often referred to as Yao's minimax principle in CS context.

Theorem. *For every f,*

$$R(f) = \max_\mu D_\mu(f).$$

Remark. *Variants of this theorem hold in many other computational models.*

Proof sketch. The inequality

$$R(f) \geq \max_\mu D_\mu(f)$$

holds because every randomized protocol is a distributional protocol. The other inequality

$$R(f) \leq \max_\mu D_\mu(f)$$

requires the minimax theorem. The idea is that there are two players; the "input" player chooses (x,y) and the "protocol" player chooses a deterministic protocol π. The protocol player wins the game if $\pi(x,y) = f(x,y)$. The minimax theorem says that there is a distribution μ_* on the choices of the input player which is "optimal". For this μ_*, we have

$$D_{\mu_*}(f) \geq R(f).$$

□

Remark. *The minimax theorem tells us that "the only way" to prove lower bounds on $R(f)$ is to find "hard distributions". Typically, an important part of a lower bound proof is finding "hard distributions". The meaning of "hard distribution" depends on the context (above μ_* is hard). The minimax theorem often tells us that they exists, and that we "just" need to find them.*

The power of randomness

Remark. *There are many computational resources of interest: randomness, non-determinism ($\exists$ quantifier), co-non-determinism ($\forall$ quantifier), quantum, and more. Their strength leads to many fundamental problems in computational complexity theory (like the P versus NP question).*

In two-party communication complexity, randomness is extremely powerful.

Theorem.

$$D(eq) = n + 1$$

and

$$R(eq) \leq 3.$$

We have already seen the upper bound on R, so it remains to prove the lower bound.

Remark. *The analogous statement for P and BPP is believed to be false.*

Lower bounds on $D(\cdot)$

Proving lower bound in computational complexity theory is, generally speaking, difficult. Communication complexity provides a clean enough model so that we can actually prove optimal lower bounds. This is one of the key reasons to use this perspective also for other "more complicated" models of computation.

There are several mechanisms for proving lower bound on $D(\cdot)$. There are "combinatorial notions" like fooling pairs, there are "linear programming" methods, there are "information theoretic" methods, etc. We start with "linear algebra" methods.

To prove lower bounds for some computational model, we need to identify some "weakness" of the model, some structure that we can exploit.

Example. *Let us start by a well-known example of "computational complexity" from linear algebra. Computations have two components "basic blocks" and "operations with costs". The object we consider now are matrices M. The "building blocks" are matrices of the form*

$$M_{ij} = u_i v_j;$$

these are rank one matrices. The "operations" are sums and the cost is the number of summands. Every matrix now has a complexity; the minimum cost required to generated it. Every computation has a cost, and the complexity of an object is the minimum cost to generate it. The complexity we get is rank. The "structure" that allows to prove lower bounds on the rank (in fact to compute rank) is hidden in the various tools of linear algebra (like Gaussian elimination).

For deterministic complexity, the basic building blocks of protocols are rectangles.

Definition (rectangle)**.** *A rectangle ρ in $\{0,1\}^n \times \{0,1\}^n$ is a set of the form $\rho = \alpha \times \beta$ where $\alpha, \beta \subseteq \{0,1\}^n$.*

Remark. *In other words, a rectangle corresponds to the "simplest communication protocol"; a one-round simultaneous deterministic protocol in which Alice sends $1_{x \in \alpha}$ and Bob sends $1_{y \in \beta}$.*

Remark. *Every boolean f defines a boolean matrix*

$$M_f(x,y) = f(x,y).$$

A rectangle can be thought of as an indicator of a sub-matrix.

Example. *Draw an illustration of a protocol for a matrix.*

Lemma. *Let π be a protocol and v be a node in the protocol tree. Then, the set of x, y so that $\pi(x,y)$ passes through v is a rectangle.*

Proof. The proof is by induction (the details are left as easy an exercise). We shall explain by a picture. □

Exercise. *If f is boolean has a protocol π then there are rectangles $\rho_1, \ldots, \rho_L$ with $L \leq 2^{|\pi|}$ so that*

$$f = \sum_{i=1}^{L} \rho_i.$$

Details for reader. Let $v_1, \ldots, v_L$ be the leaves of the protocol tree that are labelled by 1. In particular,

$$L \leq 2^{|\pi|}.$$

Each leaf v_i defines a rectangle. Every (x,y) goes to a single leaf (which could be a zero leaf). Because the protocol is correct, the label of the leaf reached by (x,y) is $f(x,y)$. □

Definition. *Over a field $\mathbb{F}$, the matrix M_f has some rank which we denote by $rank_{\mathbb{F}}(f)$. We shall mostly work over $\mathbb{R}$ and then write $rank(f) = rank_{\mathbb{R}}(f)$.*

Exercise. *$rank_{\mathbb{R}}(f) = rank_{\mathbb{Q}}(f)$.*

Corollary (Mehlhorn and Schmidt)**.** *For all non-constant f and over all fields,*

$$D(f) > \log_2 rank(f).$$

Proof. Rectangles are rank one matrices (the inequality is strict, because one of the rectangles is a zero rectangle). □

Proof of theorem. It follows that $D(eq) = n + 1$ because the matrix is the identity matrix (with full rank). □

An application

Let us see how to reason about Turing machines (TM) using CC. A TM can have one of more tapes. It is known that the number of tapes “does not really matter”.

Example. *A two-tape TM that runs in time T can be simulated by a single-tape TM that runs in time $O(T^2)$.*

CC can help us to argue that this loss is required. Consider the language of all strings of the form $x0^n x$ where $x \in \{0,1\}^n$.

Claim. *There is a two-tape TM that decides this language in time $O(n)$.*

Sketch. Copy the input to the second tape in reversed form and check equality. □

Exercise. *Every single-tape TM for this language runs in time $\Omega(n^2)$.*

Sketch. A TM for the language must solve equality (on input $x0^ny$, decide $x \overset{?}{=} y$). We have seen that the deterministic CC of this function is $n+1$.

For $n < j < 2n$, we can think of Alice as holding the part of the computation to the left of j and Bob to the right of j. When the head passes through j, the players need to communicate $O(1)$ bits. If the head spends T_j times in position j, then the total communication is $O(T_j)$. For for each j,

$$T_j \geq \Omega(n).$$

Summing over all j's, the total running time is at least

$$\sum_j T_j \geq \Omega(n^2).$$

□

Summary

— We discussed three of the most basic models of two-party communication.
— Randomness is powerful.
— Linear algebra can help to understand computation.
— CC can help to understand TMs.

Remark. *There are more methods for proving lower bounds.*

Remark. *There are more computational resources of interest. Each brings new questions and new ideas.*

Remark. *There are also models for more than two parties. For three parties and more, there are actually two different extensions. The number-in-hand model; there are k inputs $x_1, \ldots, x_k$ and player i sees x_i. The number-on-the-forehead model; there are k inputs $x_1, \ldots, x_k$ and player i sees all inputs except x_i.*

2. Matrices

A two-party boolean function $f(x,y)$ correspond to boolean matrices $M = M_f$. The CC of f can be understood via properties of the M. For example, $D(f)$ is related to partitioning M to rectangles. Is there an analog statement for $R(f)$?

The two types of randomness

There are two types of randomness: public and private. Public randomness is more powerful than private randomness. Newman showed that private randomness is almost as powerful as public randomness.

Lemma. *If f has a public randomness protocol π then f has a private randomness protocol with cost $\leq |\pi| + O(\log n)$.*

Proof. Standard concentration bounds imply that there is a list of deterministic protocols $\pi_1, \ldots, \pi_T$ for $T = 10n$ so that for every (x,y) it holds that

$$\frac{1}{T}\sum_t 1_{\pi_t(x,y)\neq f(x,y)} \leq \frac{4}{10}.$$

The public randomness protocol is as follows. Alice privately chooses $t \in [T]$ uniformly at random, sends it to Bob and then they run π_{r_t}. □

A bit more details. Denote by r the public randomness of π, and by π_r the deterministic protocol obtained by fixing r. Sample $T = 10n$ i.i.d. copies of r. Standard concentration bounds (and the union bound) show that

$$\Pr\left[\exists x,y : \left|f(x,y) - \frac{1}{T}\sum_t \pi_{r_t}(x,y)\right| > \frac{1}{10}\right] \leq 2^{2n}2^{-3n} < 1.$$

□

Approximate rank

For private randomness (private-coin) protocols, there is an additional structure. We can relate $R(f)$ to approximate rank:

$$rank_\varepsilon(M) = \min\{rank(A) : \|A - M\|_\infty \leq \varepsilon\}.$$

Exercise (Krause). *If π is a private-coin protocol for π then $|\pi| \geq \log rank_{1/3}(f)$.*

Remark. *There are two way to think of the way an algorithm A uses randomness. One option is that a random string r is chosen and then algorithm becomes deterministic A_r. Another option is that the randomness is chosen as the algorithm progresses, independently of previous choices. To solve the exercise, the latter view is important.*

We saw that there is no "significant difference" between private-coin and public-coin protocols. But there was an additive $O(\log n)$ cost. We now show that this cost is unavoidable for the "equality" function.

Theorem (Alon). *If I is the $N \times N$ identity matrix then $rank_{1/3}(I) \geq \Omega(\log N)$.*

Exercise. *The theorem is sharp.*

Remark. *The theorem and the exercise show that private-coin randomized complexity of "equality" is at least $\Omega(\log n)$. In addition, the exercise is false for public-randomness protocols because $R(eq) \leq 3$.*

Remark. *The theorem is useful in many other settings, like coding theory and pseudorandomness.*

The proof of the theorem is spectral and is based on two steps.

Claim (small off-diagonal, high rank)**.** *If M is an $N \times N$ matrix so that $M_{ii} = 1$ and*

$$\sum_{i \neq j} M_{ij}^2 \leq N$$

then

$$rank(M) \geq \frac{N}{4}.$$

Remark. *The claim is a "noisy version" of the the extreme and trivial case*

$$\sum_{i \neq j} M_{ij}^2 \leq 0.$$

Remark. *Recall the following linear algebra. If M is a symmetric real $N \times N$ matrix then it has eigenvalues $\lambda_1, \ldots, \lambda_N$ and*

$$trace(M) = \sum_i \lambda_i.$$

Proof of claim. The matrix $A := (M + M^T)/2$ is symmetric, satisfies the same property and

$$rank(A) \leq 2rank(M).$$

Denote by λ_i the eigenvectors of A.

$$\begin{aligned} N^2 &= (trace(A))^2 \\ &= \Big(\sum_i \lambda_i\Big)^2 \\ &\leq rank(A) \cdot \sum_i \lambda_i^2 \\ &= rank(A) \cdot trace(A^2) \\ &= rank(A) \cdot \Big(\sum_{i,j} A_{ij}^2\Big) \\ &\leq rank(A) \cdot (N + N). \end{aligned}$$

□

Remark. *Before the second step, recall the following well-known counting problem. The number of ways to choose k items with repetition out of a domain of r items is*

$$\binom{k+r-1}{r-1} = \binom{k+r-1}{k}.$$

Remark. *Sometimes, it sufficed to bound $\binom{r+k}{k} \leq (r+k)^k$. This simple estimate is too crude here.*

Exercise (rank of point-wise products)**.** *For a matrix M of rank r and even $k > 0$,*

$$rank(M_{ij}^k) \leq \binom{r+k}{k}.$$

Hint.

$$(M_{ij})^k = \Big(\sum_{t=1}^{r}(v_t)_i(u_t)_j\Big)^k.$$

□

Proof of theorem. Let M be a matrix of rank r so that $M_{ii} \geq 2/3$ and $|M_{ij}| \leq 1/3$ for $i \neq j$. Without changing the rank, we can "normalize" M to have $M_{ii} = 1$ and $|M_{ij}| \leq 1/2$ for $i \neq j$. For even k we have that for $i \neq j$,

$$0 \leq M_{ij}^k \leq (1/2)^k.$$

For $k = \lceil \log_2 N \rceil$, we have

$$\sum_{i \neq j} M_{ij}^k \leq N$$

so that

$$\frac{N}{4} \leq rank(M_{ij}^k) \leq \binom{r+k}{k}.$$

This implies that $r \geq \Omega(\log N)$. □

Summary

— Saw an analog of the log-rank lower bound for randomized communication.
— Separated private- and public-coin protocols.
— A general lower bound for approximate rank of identity matrices.

Log ranks

Remark. *Saks and Lovaśz asked whether this lower bound is always tight. Is there* $C > 0$ *so that* $D(f) \leq C \log_2^C D(f)$*? This is the famous log-rank conjecture. This conjecture is at the heart of our mis-understanding of boolean matrices.*

Remark. *The approximate log-rank conjecture is false [Chattopadhyay-Mande-Sherif].*

Other ranks

Again, a boolean two-party function is a boolean matrix. We discussed two types of linear algebra complexity measures for boolean matrices: rank and approximate rank. There are many other notions of ranks that differ from each other in what the "basic blocks" are and what "A=B" means.

Sign rank. One example is sign rank;

$$sign - rank(M) = \min\{rank(A) : sign(A) = M\}.$$

Here we can assume that $A_{ij} \neq 0$ for all i, j. This notion of rank is related to kernel methods in machine learning and to unbounded error communication complexity. Forester (and works that followed) built a beautiful mechanism for lower bounding sign rank, which in turn lead to lower bounds in other computational models (like $TH \circ MAJ$ circuits).

Non negative. Another way to get even stronger lower bound is noting that the lower bounds on $D(f)$ and $R(f)$ we proved above actually rely on write $M = \sum_i M_i$ with all M_i's of rank one and non-negative. This leads to non-negative rank and non-negative approximate rank, which are always at least as large as rank and approximate rank. Non-negative rank is also known to be related to extension complexity of polytopes which is motivated by problems in optimization (Yannakakis and many works that followed).

Lovaśz solved the log-rank conjecture in this setting:

$$D(f) \leq O(\log^2 rank^+(f))$$

where $rank_+$ is non negative rank. The non-negative approximate log-rank conjecture is also false [Chattopadhyay-Mande-Sherif]. But the two sided version is still open

$$R(f) \leq polylog\ rank^+_{1/3}(M_f) \cdot rank^+_{1/3}(M_{1-f})?$$

Norms. The ranks so far are integer quantities, and as such should be expected to be "hard to understand". We can also define "smoother" complexity measures. There are many notions of vector norms and of matrix norms, each suitable for a specific objective.

Let us mention the following norm (it has some other names as well). Here M is a sign-matrix:

$$\|M\|_\nu = \inf\Big\{\sum_z |c_z| : M = \sum_z c_z M_z\Big\}$$

where M_z is a rank one matrix so that $\|M_z\|_\infty \leq 1$. It immediately follows that

$$D(f) \geq \log_2 \|M_f\|_\nu.$$

Sometimes this leads to a strong lower bound. But sometimes it does not.

Exercise. *If I is an identity matrix then $\|I\|_\nu = 1$.*

Remark. *An important and useful idea for understanding norms is duality. Every norm $\|x\|$ has a dual norm*

$$\|y\|_* = \max\{\langle y, x\rangle : \|x\| \leq 1\}.$$

One way to think about this is "x is a test function" and we shall discuss this in more detail tomorrow.

Abstract. We far we relaxed the notion of rank by "redefining $M = A$". There is a different general mechanism for relaxations. Matrices are themselves vectors. So, now work over $\mathbb{R}^N$. Let $S \subset \mathbb{R}^N$ be a set of "simple vectors". For rank, e.g., they are the rank one matrices. For a given $v \in \mathbb{R}^N$, define the complexity of v as

$$complexity(v) = \min\Big\{r \in \mathbb{N} : \exists s_1, \ldots, s_r \in S : v = \sum_i s_i\Big\}.$$

For matrix rank, Gaussian elimination gives us an algorithm for computing rank. But in other settings, there are no such efficient algorithms. One example is the definition of blocky rank, where the rank one matrices are "blowups of permutation matrices" introduced by [Hambardzumyan, Hatami, Hatami].

Discrepancy

Discrepancy is a measure of pseudorandomness (it is "dual to norms"). The smaller the discrepancy, the more random-like the object is.

The objets we care about are functions $f : [N] \to \{\pm 1\}$. (For discrepancy, we replace $\{0,1\}$ by $\{\pm 1\}$.) Let T be a (finite) collection of maps $t : [N] \to [0,1]$, which we think of as "tests". The discrepancy of f for tests T is

$$disc_T(f) = \max_{t \in T} \Big| \frac{1}{N} \sum_z f(z)t(z) \Big|.$$

Example. *If $f \in \{\pm 1\}^n$ is uniformly at random and T the indicators of $[1], [2], \ldots, [n]$ then*

$$\mathbb{E}\, dist_T(f) \approx \sqrt{n}.$$

In other words, this is expected value of the maximum of the absolute value of a simple random walk on $\mathbb{Z}$ up to time n.

In communication complexity, the test functions R are combinatorial rectangles:

$$disc(f) = disc_R(f) = \max_\rho 2^{-2n} \Big| \sum_{x,y} f(x,y)\rho(x,y) \Big|,$$

where ρ is a combinatorial rectangle.

Exercise. *$disc(f) \geq \Omega(2^{-n/2})$ for every f.*

Exercise. *$R(f) \geq \log_2 \Omega\Big(\frac{1}{disc(f)}\Big)$.*

Remark. *Sometimes we should measure discrepancy with respect to a non uniform distribution. For a distribution μ,*

$$disc_\mu(f) = \max_\rho \Big| \sum_{x,y} \mu(x,y) M_{x,y} \rho(x,y) \Big|.$$

Inner product

One of the main examples of functions with low discrepancy is the inner product function. The inner product function

$$f(x,y) = (-1)^{\langle x,y \rangle} = (-1)^{\sum_i x_i y_i}$$

is important in many settings. A central property which makes it useful is the following lemma.

Lemma. *The inner product function has discrepancy $2^{-n/2}$.*

Proof. Let $M = M_f$. First, we claim that it is an Hadamard matrix:[1]

$$MM^T = 2^n I.$$

[1] The main thing to verify is that the columns are orthogonal, which is true because for $x \neq x'$,

$$\sum_y (-1)^{\langle x,y \rangle}(-1)^{\langle x',y \rangle} = \sum_y (-1)^{\langle x+x',y \rangle},$$

where $x + x'$ is addition modulo two. Without loss of generality $x_1 \neq x_1'$, so that the y's can be partitioned to pairs $(0,y'), (1,y')$ for y' with $n-1$ bits. For each such pair, the sum is zero.

Now, for a rectangle ρ we have $\rho(x,y) = a(x)b(y)$ so that (by Cauchy-Schwarz)

$$\begin{aligned}\Big(\sum_{x,y} \frac{1}{2^{2n}} M_{x,y} a(x) b(y)\Big)^2 &= \frac{1}{2^{4n}} \Big(\sum_y b(y) \sum_x M_{x,y} a(x)\Big)^2 \\ &\leq \frac{1}{2^{4n}} \Big(\sum_y b(y)^2\Big)\Big(\sum_y \Big(\sum_x M_{x,y} a(x)\Big)^2\Big) \\ &\leq \frac{1}{2^{4n}} 2^n \Big(2^{2n} + \sum_{x \neq x'} a(x) a(x') \sum_y M_{x,y} M_{x',y}\Big) \\ &= \frac{1}{2^n}.\end{aligned}$$

□

Greater-than

Another central example of a matrix is the greater-than matrix $G_{ij} = 1_{j \geq i}$ where i, j are integers in $0, \ldots, 2^n - 1$. It also appears in analysis, compression of communication, differential privacy, etc.

Exercise (binary search)**.** *$R(geq) \leq O(\log(n) \cdot \log\log(n))$.*

Exercise (up-down binary search)**.** *$R(geq) \leq O(\log n)$.*

Theorem (Viola)**.** *$R(geq) \geq \Omega(\log n)$.*

Viola's argument is based on information theory. Another way to prove lower bounds is using discrepancy.

Remark. *For greater-than, the "hard distribution" is not the uniform distribution because for uniform input the value of geq is determined by the first few bits. The matrix also contains a "huge constant submatrix" so how can its discrepancy be small?*

Definition (the "hard distribution)**.** *Let μ_* be the distribution on (x,y) defined as: choose $i \in [n]$ uniformly at random, choose (x,y) uniformly at random conditioned on $x_{<i} = y_{<i}$.*

Lemma (Braverman-Weinstein)**.** *$disc_{\mu_*}(geq) \leq O(\frac{1}{\sqrt{n}})$.*

Exercise. *$disc_{\mu_*}(geq) \geq O(\frac{1}{\sqrt{n}})$.*

Lemma (Srinivasan-Y)**.** *There is μ so that $disc_\mu(geq) \leq O(\frac{1}{n})$.*

Remark. *So, there is an "even harder" distribution. Somewhat surprisingly, the distribution is not so intuitive, and the proof uses ideas of Bennet from the study of Schur multipliers. The distribution is constructed via the Hilbert matrix*

$$H_{ij} = \frac{1}{N + \frac{1}{2} - i - j}.$$

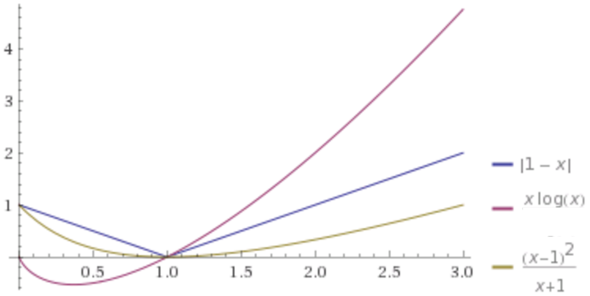
3. Information

When parties communicate, they learn information. A mathematical way to capture this is using probability theory. Our knowledge of the world is a probability distribution p on the universe $[N]$. The number $p(x)$ is the chance that we think that x is "correct". If p is uniform, we "know nothing" and when p is a single point then we "know everything".

Assume that Alice's current state of the knowledge is a distribution p. She now gets a message from Bob. Her state of knowledge changes to a distribution q. How much did Alice learn?

There are many way to measure this, we shall use the language of f-divergences [Csiszar, Morimoto, Ali, Silvey]. There is a convex $f : \mathbb{R}_+ \to \mathbb{R}$ so that $f(1) = 0$.

Example.



Definition. *For two distributions* p, q*,*

$$D_f(p,q) := \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right)$$

with the obvious values at zero or infinity.

Remark. *This is a measure of the "distance" or "divergence" between* p, q*. It allows to measure "how much did Alice's opinion change?".*

Example.

$$|1-x| \qquad \Longrightarrow \qquad \|p-q\|_1 := \sum_x |p(x)-q(x)| = 2\max_E p(E) - q(E).$$

$$x \log x \qquad \Longrightarrow \qquad D_{KL}(p||q) := \sum_x p(x) \log \frac{p(x)}{q(x)}.$$

$$\frac{(1-x)^2}{1+x} \qquad \Longrightarrow \qquad \Delta(p,q) = \sum_x \frac{(p(x)-q(x))^2}{p(x)+q(x)}.$$

Each f leads to unique properties of D_f, but there are some properties that are always true.

Lemma.

positivity: $D_f(p,q) \geq 0$.
convexity: $D_f(p,q)$ *is convex in* (p,q).
data processing: $D_f(p_X, q_X) \geq D_f(p_{g(X)}, q_{g(X)})$ *for all functions* g.[2]

[2] $p_{g(X)}$ is sometimes called the push-forward of p by g.

Remark. *Duality allows to related some quantity q to a dual quantity q^*. This appears in many areas of math. For D_f, we can also apply duality and give D_f "operational meaning". This meaning is similar in spirit to pseudorandomness and discrepancy. For example,*

$$\|p-q\|_1 = \max_{g : \|g\|_\infty \leq 1} \sum_x g(x)(p(x)-q(x))$$

and

$$\Delta(p,q) = \max_{g : \sum_x (p(x)+q(x))g^2(x) \leq 1} \Big(\sum_x (p(x)-q(x))g(x) \Big)^2 .$$

An example

The prior

Assume Alice's opinion X takes values in $\{0,1\}^n$ and has high "entropy", it is not so far away from the uniform distribution:

$$D_{KL}(p_X||u_n) \leq \delta n$$

where u_n is uniform over $\{0,1\}^n$. The first step we shall take is explaining why Alice does not know much on a typical entry X_i. We shall study this scenario using different f-divergences. The different divergences have different properties. Each f is suitable for a specific step in the argument.

KL-divergence.

Exercise (chain rule)**.**

$$D_{KL}(p_{X,Y}||q_{X,Y}) = D_{KL}(p_X||q_X) + \mathop{\mathbb{E}}_x D_{KL}(p_{Y|x}||q_{Y|x}).$$

Exercise (sub-additivity)**.** *If $q_{X,Y} = q_X \times q_Y$ then*

$$\mathop{\mathbb{E}}_x D_{KL}(p_{Y|x}||q_{Y|x}) \geq D_{KL}(p_Y||q_Y).$$

Details.

$$\begin{aligned}
\mathop{\mathbb{E}}_x D_{KL}(p_{Y|x}||q_{Y|x}) &= \mathop{\mathbb{E}}_x D_{KL}(p_{Y|x}||q_Y) \\
&= \sum_{x,y} p(x,y) \Big(\log \frac{p(y|x)}{p(y)} + \log \frac{p(y)}{q(y)} \Big) \\
&= \mathop{\mathbb{E}}_x D_{KL}(p_{Y|x}||p_Y) + D_{KL}(p_Y||q_Y) \\
&\geq D_{KL}(p_Y||q_Y).
\end{aligned}$$

□

Claim (subadditivity)**.** *If I is a uniform coordinate then X_I is close to uniform:*

$$\mathop{\mathbb{E}}_I D_{KL}(p_{X_i}||u_1) \leq \delta.$$

Proof. Because the uniform distribution is a product measure,

$$\begin{aligned}\mathop{\mathbb{E}}_{I} D_{KL}(p_{X_i}||u_1) &= \frac{1}{n}\sum_i D_{KL}(p_{X_i}||u_1)\\ &\le \frac{1}{n} D_{KL}(p_X||u_n)\\ &\le \frac{\delta n}{n}.\end{aligned}$$

□

Statistical distance.

Remark. *What does $D_{KL}(p_{X_i}||u_1) \le \delta$ mean? It reads "the i'th coordinate of X is close to uniform". But the notion of distance is given by D_{KL}. We would like to have a simpler to grasp measure.*

Lemma (Pinsker)**.** $\|p - q\|_1 \le \sqrt{2D(p||q)}$

Corollary. $\mathbb{E}_I \|p_{X_i} - u_1\|_1 \le \sqrt{2\delta}$.

Remark. *What does $\|p_{X_i} - u_1\|_1 \le \sqrt{2\delta}$ mean? We know that X_i is a bit, and the above means that the chance that X_i takes the value* 0 *or* 1 *is* $\frac{1}{2} \pm \sqrt{\delta}$*. This is already quite informative, but we lost a "square root".*

Remark. *This loss is often quite expensive but sometimes it is sharp. This loss is often extremely important in communication complexity but also in other settings, like in the parallel repetition theory. The fact that it is required for parallel repetition is related to "spherical cubes" [Feige, Kindler, O'Donnell], [Raz], [Kindler, Rao, O'Donnell, Wigderson].*

The posterior

Now, assume that Bob communicated some information about I to Alice. Denote by J the new distribution on $[n]$ Alice has in mind, and assume that Bob did not communicate a lot of information:

$$D_{KL}(p_J||p_I) \le \varepsilon.$$

Is it still true that Alice does not know much about X_j? By moving to statistical distance, the answer is yes:

$$\mathop{\mathbb{E}}_{J} \|p_{X_j} - u_1\|_1 \le \|p_J - p_I\|_1 + \mathop{\mathbb{E}}_{I} \|p_{X_i} - u_1\|_1 \le \sqrt{2\varepsilon} + \sqrt{2\delta};$$

the first inequality is natural and is left as an exercise. Now, we have two square root losses.

Triangular discrimination.

We can avoid quantitative losses by using the "correct" way to keep track of the flow of information. It turns out that triangular discrimination is more suitable here.

Lemma (Topsøe)**.** $\Delta(p||q) \le 2D_{KL}(p||q)$

Corollary. $\mathbb{E}_I \Delta(p_{X_i}||u_1) \le 2\delta$.

Now, recall that J is "after Alice learned something about I" and decompose:

$$\mathop{\mathbb{E}}_{J} \Delta(p_{X_j}||u_1) = \mathop{\mathbb{E}}_{I} \Delta(p_{X_j}||u_1) + \Big(\mathop{\mathbb{E}}_{J} \Delta(p_{X_j}||u_1) - \mathop{\mathbb{E}}_{I} \Delta(p_{X_j}||u_1)\Big).$$

We know that the left term is small. It remains to bound the cost of the "distribution shift", to show that the right term is small as well. Bounding the cost of a distribution shift is important in many settings.

Lemma (Y). $\Big|\mathbb{E}_J \Delta(p_{X_j}||u_1) - \mathbb{E}_I \Delta(p_{X_j}||u_1)\Big| \leq 2\varepsilon + 10\delta.$

The details are left as a (technical) exercise. □

Remark. *Triangular discrimination is a useful way to control "distribution shift". It was used in several settings (mostly implicitly), such as analyzing random walks on group and the Liouville property (bounded harmonic functions are constant).*

Pointer chasing

The pointer chasing problem was introduced by Papadimitriou and Sipser to study the importance of the number of rounds. In pointer chasing, there are two disjoint sets A, B each of size n, Alice gets $x : A \to B$ and Bob gets $y : B \to A$. There is a sequence of pointers z_0 is some fixed element of A and

$$z_1 = x(z_0),\ z_2 = y(z_1),\ \ldots$$

Remark. *Draw as edges in directed bipartite graph.*

The goal is to compute say if z_k is "even" or "odd". With a k-round protocol in which Alice speaks first, the communication complexity is $O(k \log n)$.

Remark. *There are other upper bounds. A $(k-1)$-round deterministic protocol of cost $O_k(n \log^{(k-1)} n)$ [Damm, Jukna, Sgall]. A $(k-1)$-round randomized protocol of cost $O((k + n/k)\log n)$ [Nisan, Wigderson].*

Remark. *We are mostly interested in lower bounds:*

D	$\Omega(\frac{n}{k^2})$	*[Duris, Galil, Schnitger]*
$k = 2$	$\Omega(n)$	*[Papadimitriou, Sipser]*
D	$\Omega(n - k\log n)$	*[Nisan, Wigderson]*
D_u	$\Omega(\frac{n}{k^2} - k\log n)$	*[Nisan, Wigderson]*
R	$\Omega_k(n \log^{(k-1)} n)$	*[Ponzio, Radhakrishnan, Venkatesh]*
R	$\Omega(\frac{n}{k})$	*[Klauck, NW]*

Remark. *The bounds above show the importance of interaction in some communication complexity problems. This has applications for monotone circuit complexity, distributed computing, and streaming algorithms.*

Remark. *The k^2 in the $\approx \frac{n}{k^2}$ lower bound on D_u over the uniform distribution is, roughly speaking, caused by the square root loss in Pinsker's inequality. Using the ideas discussed above, we can prove an $\Omega(\frac{n}{k} - k\log n)$ lower bound, which is essentially sharp.*

Remark. *The main lemma in the proof states the following. For a round $t < k$, denote by R_t the random variable*

$$R_t = (M_1, \ldots, M_t, Z_1, \ldots, Z_{k-1}).$$

Then,

$$\mathop{\mathbb{E}}_{r_t} \Lambda(p_{Z_t|r_t}||p_{Z_t}) \leq 6t\frac{|\pi| + k\log n}{n}.$$

Here, $M_1, M_2, \ldots$ are the messages in the protocol, and Λ is an asymmetric version of triangular discrimination:

$$\Lambda(p||q) = \sum_{x:p(x)>q(x)} \frac{(p(x)-q(x))^2}{p(x)+q(x)}.$$

Remark. *The lemma shows that if $|\pi|$ is small, then the distribution of Z_t did not change much so the player learned nothing on it.*

Remark. *The proof of the lemma is by induction on t. It is "computation free" in the sense that it just uses the properties we talked about as black boxes (and an additional property of communication protocols).*

Information theory

The particular case of D_{KL} is at the heart of information theory, which is extremely useful in many settings including communication complexity. A long sequence of works developed information theoretic tools in communication complexity ([Razborov], [Chakrabarti, Shi, Wirth, Yao], [Barak, Braverman, Chen, Rao], [Jain, Pereszlényi, P. Yao], [Braverman, Rao, Weinstein, Y], etc.). These tools allow to prove lower bounds as well as direct-sum-type results for randomized CC (and also for deterministic CC). There are still fundamental open questions in this area.