Electronic Colloquium on Computational Complexity, Report No. 206 (2018)

Equality Alone Does not Simulate Randomness

Arkadev Chattopadhyay[*] Shachar Lovett[†] Marc Vinyals[‡]

December 3, 2018

Abstract

The canonical problem that gives an exponential separation between deterministic and randomized communication complexity in the classical two-party communication model is 'Equality'. In this work, we show that even allowing access to an 'Equality' oracle, deterministic protocols remain exponentially weaker than randomized ones. More precisely, we exhibit a total function on n bits with randomized one-sided communication complexity $O(\log n)$, but such that every deterministic protocol with access to 'Equality' oracle needs $\Omega(n/\log n)$ cost to compute it.

1 Introduction

A deterministic communication protocol in Yao's two-party model is a strategy for a collaborative game between two parties, Alice and Bob, each of whom receives an input and whose task is to compute a function while communicating as little as possible.

It has been known since the origins of communication complexity that randomized protocols, where the parties are given access to a source of randomness and are allowed to make errors with small probability, are strictly more powerful than deterministic protocols. The classic example is the Equality function over n-bit strings, which has a randomized protocol with $\mathrm{O}(\log n)$ bits of communication, while every deterministic protocol requires at least $n+1$ bits [Yao79].

An efficient protocol for Equality is obtained by using a fingerprinting technique: use the randomness source to obtain a fingerprint of the strings to be compared of length $\mathrm{O}(\log n)$, exchange the fingerprints, and answer whether the fingerprints are equal.

A few more examples of functions where randomness is helpful are the 'Greater-Than' function [Nis93], the sparse set disjointness problem [HW07], and the Hamming distance problem with a small threshold [Yao03]. In all cases the fingerprinting technique is enough to efficiently solve the problems. Is fingerprinting all there is to randomized protocols?

To state this question in a formal way we consider a model of communication where the parties are given access to an oracle that solves the Equality problem and are charged a cost of one bit each time the parties call the oracle. The set of functions that can be computed by some protocol in this model of cost polylog n bits, is called P^{EQ}. The set of functions that have randomized protocols of cost polylog n is called BPP. We overload notation to use P^{EQ} and BPP to refer to both the class of functions and the corresponding communication models respectively. The question then is whether every function that has a randomized protocol with c bits of communication, also has a P^{EQ} protocol with poly$(c, \log n)$ bits of communication and oracle calls. In other words, is $\mathsf{P}^{\mathrm{EQ}} = \mathsf{BPP}$?

[*] School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai. arkadev.c@tifr.res.in. Part of the work done while visiting the Simons Institute for the Theory of Computing.

[†] Department of Computer Science and Engineering, University of California, San Diego. slovett@ucsd.edu. Supported by NSF award 1614023.

[‡] School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai. marc.vinyals@tifr.res.in. Supported by the Prof. R Narasimhan Foundation.

ISSN 1433-8092

The P^{EQ} model was first considered in [BFS86]. The knowledge about it until our work (for total functions, see discussion below) can be summarized as follows:

$$\mathsf{P} \subsetneq \mathsf{P}^{\mathrm{EQ}} \subseteq \mathsf{BPP}.$$

P^{EQ} is also strictly weaker than the P^{NP} model, since EQ calls can be simulated with an NP oracle but P^{EQ} cannot efficiently solve the coNP-complete set-disjointness problem. It also is worth mentioning that giving access to an Equality oracle is equivalent to giving access to a Greater-Than oracle up to a logarithmic factor, and that the latter model of communication was introduced by Krajíček [Kra98] as real communication.

Partial functions. There are many examples in the literature of *partial functions* that separate P^{EQ} from BPP. One such example is the gap Hamming distance problem with a large gap. Concretely, the problem is to distinguish between pairs of input strings whose Hamming distance is less than a 1/3-fraction and more than a 2/3-fraction. This can be solved with a randomized protocol with $O(1)$ bits, that samples a position in the strings uniformly at random and answers whether the strings are the same at that position. On the other hand, this problem has cost $\Omega(n)$ in the P^{NP} model [PSS14], and hence in the P^{EQ} model too.

A different example follows from the simulation theorem of [BEGJ00], made explicit in [dRNV16], to lift a (partial) function that exhibits an exponential gap between deterministic and randomized query complexity, say promised majority. To be more precise, we consider the majority function of n bits with the promise that the fraction of zeros is either less than $1/3$ or more than $2/3$, which can be computed with a randomized decision tree by querying the input at a constant number of randomly sampled points, but requires linearly many queries to be solved by a deterministic decision tree. If we compose this function with the indexing gadget with pointers of size $\mathrm{O}(\log n)$ then we have a randomized protocol of cost $\mathrm{O}(\log n)$ that evaluates a constant number of instances of the gadget, while the simulation theorem tells us that it requires real communication $\Omega(n \log n)$.

Total functions. The question about separation between P^{EQ} and BPP for *total functions* requires a different approach. If one uses the same approach as before, namely, lifting theorems, then a quadratic separation follows for example from the pointer chasing function [ABB+17] composed with indexing. However, this is where the lifting from query complexity approach seems to end, since deterministic and randomized query complexity are known to be polynomially related for total functions [Nis91]. Our main result is a non-lifted total function, which exhibits exponential separation between P^{EQ} and randomized communication.

Definition 1.1. The integer inner product problem $\mathrm{IIP}_{m,t}(x,y)$ is defined as follows. The inputs are integer vectors $x, y \in [-M, M]^t$ where $M = 2^m$. The output is 1 if $\langle x, y\rangle = 0$, where the inner product is computed over the integers.

Note that the input size of $\mathrm{IIP}_{m,t}$ is $n = (m+1)t$. We consider it for large m and $t = O(1)$.

Theorem 1.2 (Main theorem, informal). *For any $t \geq 6$, the total function* IIP *on n bits can be computed with* $\mathrm{O}(\log n)$ *bits of randomized communication but requires* $\Omega(n/\log n)$ *cost to be solved by* P^{EQ} *protocols.*

2 Preliminaries

We assume knowledge with standard definitions in communication complexity, such as in [KN97]. The only somewhat non-standard definition we need is that of protocols with access to an oracle.

If A is a communication problem, then the parties involved in a P^A protocol communicate via an oracle for A. This is, every message is a pair of inputs for the function A, and the output $A(x,y)$ is visible to both parties. We assume that A is nontrivial in the sense that it can simulate sending one-bit messages from each party to the other one. The cost of such a protocol is the number of bits the oracle outputs, and $\mathsf{P}^A(f)$ is the minimum over all protocols. In particular, P^{EQ} is a protocol with oracle access to the equality oracle, and P^{GT} is a protocol with oracle access to the greather-than oracle, both of which are nontrivial.

3 A Lower Bound Technique for P^{EQ}

The goal of this section is to develop a lower bound technique for P^{EQ}. It will be convenient to consider instead the model of P^{GT}, where the players have oracle access to a greater-than oracle. Note that as an EQ oracle can be simulated by two calls to a GT oracle, the latter model is stronger.

Given a matrix M we denote by $|M|$ the number of elements in M. We say that M is *monotone* if $M_{i_1,j_1} \le M_{i_2,j_2}$ for all pairs of entries such that $i_1 \le i_2$ and $j_1 \le j_2$.

Lemma 3.1. *A monotone matrix M can be partitioned into four rectangles R_1, R_2, R_3, R_4, such that R_1, R_2 are monochromatic and $|R_1| + |R_2| \ge |R_3| + |R_4|$.*

Proof. Let a and b be the dimensions of the matrix M and assume without loss of generality that $a \ge b$. Let a_1 be the maximal number such that $M_{a_1,b_1} = 0$, with $b_1 = \lceil a_1 b/a \rfloor$. Then the rectangle $R_1 = [1,a_1] \times [1,b_1]$ is 0-monochromatic, while the rectangle $R_2 = [a_1+1,a] \times [b_1+1,b]$ is 1-monochromatic. We define $R_3 = [1,a_1] \times [b_1+1,b]$ and $R_4 = [a_1+1,a] \times [1,b_1]$. To complete the proof let $a_2 = a - a_1$ and $b_2 = b - b_2$, and observe that if $a_1 > a_2$ then $b_1 \ge b_2$, while if $a_1 < a_2$ then $b_1 \le b_2$. Therefore by the rearrangement inequality

$$|R_1| + |R_2| = a_1 b_1 + a_2 b_2 \ge a_1 b_2 + a_2 b_1 = |R_3| + |R_4| \ .$$

□

If $\mathcal{R}$ is a set of rectangles $R_i = A_i \times B_i$, we denote the *perimeter* of $\mathcal{R}$ by $p(\mathcal{R}) = \sum_{R_i \in \mathcal{R}} |A_i| + |B_i|$.

Lemma 3.2. *Assume that f is an n-bit function which has a P^{GT} protocol with cost c. Then there exists a partition $\mathcal{R}$ of f into monochromatic rectangles with perimeter $p(\mathcal{R}) \le 2^{n+1}(2n)^c$.*

Proof. Let $p(M)$ be the minimum perimeter over all partitions of a matrix M into monochromatic rectangles. Let $p(a,b)$ be the maximum of $p(M)$ over all monotone matrices M of size $a \times b$. We prove by induction over a and b that $p(a,b) \le (a+b)\log(ab)$. Consider a monotone matrix M of size $a \times b$. Apply Lemma 3.1 to partition the matrix into 4 rectangles, two monochromatic rectangles of size $a_1 \times b_1$ and $a_2 \times b_2$, and two other rectangles of size $a_2 \times b_1$ and $a_1 \times b_2$, with $a_1 + a_2 = a$, $b_1 + b_2 = b$, and $a_1 b_2 + a_2 b_1 \le ab/2$. We then apply the induction hypothesis to each non-monochromatic rectangle, while noting that for the monochromatic rectangles, their perimeter is the sum of their dimensions:

$$\begin{aligned}
p(M) &\le (a_1+b_1) + (a_2+b_2) + p(a_1,b_2) + p(a_2,b_1) \\
&\le (a_1+b_1) + (a_2+b_2) + (a_1+b_2)\log(a_2 b_1) + (a_2+b_1)\log(a_1 b_2) \\
&\le (a+b) + (a+b)\log(ab/2) \\
&= (a+b)\log(ab) \ .
\end{aligned}$$

Next, assume that we have a matrix M with a partition $\mathcal{R}$ into monochromatic rectangles $R_i = A_i \times B_i$ with perimeter $p(\mathcal{R})$. Assume that we obtain a matrix M' by instantiating a GT oracle call in each rectangle. We will show that there exists a partition $\mathcal{R}'$ of M' into monochromatic rectangles such that

$$p(\mathcal{R}') \leq 2n \cdot p(\mathcal{R}).$$

For each rectangle R_i in M', first sort the rows in increasing order and the columns in decreasing order so that R_i becomes monotone. Then by the previous bound for monotone matrices:

$$p(\mathcal{R}') \leq \sum_{R_i \in \mathcal{R}} p(|A_i|, |B_i|) \leq \sum_{R_i \in \mathcal{R}} (|A_i| + |B_i|) \log |R_i| \leq 2n \sum_{R_i \in \mathcal{R}} |A_i| + |B_i| = 2n \cdot p(\mathcal{R}) \ .$$

To conclude the proof, let M be a matrix obtained by c iterative calls to the GT oracle. Let $M_0, \ldots, M_c$ denote the intermediate matrices, where M_i is the matrix obtained after the first i calls. Then M_0 is a monochromatic matrix of size $2^n \times 2^n$ and $M = M_c$. Thus $p(M_0) = 2^{n+1}$ and $p(M_i) \leq 2n \cdot p(M_{i-1})$ for $i = 1, \ldots, c$. We conclude that $p(M) \leq 2^{n+1} \cdot (2n)^c$ as claimed. □

The next lemma gives an easy to verify condition under which Lemma 3.2 can be applied.

Lemma 3.3. *Let f be an n-bit function with a corresponding $2^n \times 2^n$ communication matrix M. Assume that:*

1. *The number of entries i, j with $M_{i,j} = 1$ is $\alpha 2^{2n}$.*

2. *For any 1-monochromatic rectangle R in M it holds that $|R| \leq \beta 2^{2n}$.*

Then the communication complexity of f in P^{EQ} is $\Omega\left(\frac{\log(\alpha/\sqrt{\beta})}{\log n}\right)$.

Proof. Let $\mathcal{R}$ be a partition of $f^{-1}(1)$ into rectangles $R_i = A_i \times B_i$ which minimizes $\sum |A_i| + |B_i|$. Observe that

$$\sum_{R_i \in \mathcal{R}} |A_i| + |B_i| \geq 2 \sum_{R_i \in \mathcal{R}} \sqrt{|A_i||B_i|} \ . \tag{1}$$

Let $x_i = |A_i||B_i|/2^{2n}$ denote the relative area of each rectangle R_i. Then the following minimization problem lower bounds the right hand side of (1):

$$2^{n+1} \cdot \min_{\sum_i x_i = \alpha, 0 \leq x_i \leq \beta} \sum_i \sqrt{x_i} \ .$$

The minimum of a concave function over a convex polytope is attained at a vertex, in this case any point with $\lfloor \alpha/\beta \rfloor$ coordinates equal to β, one coordinate equal to $\alpha - \lfloor \alpha/\beta \rfloor \beta$, and the rest equal to 0. Hence

$$\sum_{R_i \in \mathcal{R}} |A_i| + |B_i| \geq 2^{n+1} \lfloor \alpha/\beta \rfloor \sqrt{\beta} \ .$$

If f has a P^{EQ} protocol with cost c, then it has a P^{GT} protocol with cost $2c$, and hence by Lemma 3.2

$$\sum_{R_i \in \mathcal{R}} |A_i| + |B_i| \leq 2^{n+1} (2n)^{2c} \ .$$

Rearranging these gives $c \geq \Omega(\log(\alpha/\sqrt{\beta})/\log n)$ as claimed. □

4 Separation

We demonstrate the separation by considering the inner product function over the integers. We recall the definition from the introduction.

Definition 4.1. The integer inner product problem $\mathrm{IIP}_{m,t}(x,y)$ is defined as follows. The inputs are integer vectors $x, y \in [-M,M]^t$ where $M = 2^m$. The output is 1 if $\langle x,y\rangle = 0$, where the inner product is computed over the integers.

Note that the input size of $\mathrm{IIP}_{m,t}$ is $n = (m+1)t$. We consider it for large m and $t = O(1)$.

Lemma 4.2. *There is a* coRP *protocol for* $\mathrm{IIP}_{m,t}$ *of cost* $\mathrm{O}(t \log m)$.

Proof. Consider the following protocol: sample a uniformly random prime q among the first $4m + 2\log t$ primes, compute $\langle x,y\rangle \pmod q$ by having Alice send t integers $x_i \pmod q$ to Bob, and accept if and only if $\langle x,y\rangle = 0 \pmod q$. The protocol uses $\mathrm{O}(t\log q) = \mathrm{O}(t \log m)$ bits of communication.

The protocol is always correct on 1-inputs. To see that it is correct on 0-inputs with probability at least $1/2$ we observe that the probability of failure is the probability of picking a prime q that divides $\langle x,y\rangle$. Since the number $\langle x,y\rangle$ is bounded by tM^2 in absolute value, it is divisible by at most $\log(tM^2) = 2m + \log t$ primes, and since we have $4m + 2\log t$ primes to choose from, the probability of failure is at most $1/2$. □

Lemma 4.3. *If t is even then* $\Pr_{x,y}[\mathrm{IIP}_{m,t}(x,y) = 1] = \Omega(1/tM^2)$.

Proof. Write $x = (x', -x'')$ and $y = (y', y'')$ where $x', y', x'', y'' \in [-M,M]^{t/2}$, so that $\langle x,y\rangle = \langle x',y'\rangle - \langle x'',y''\rangle$. The distribution of $\langle x',y'\rangle$ and $\langle x'',y''\rangle$ are i.i.d and take at most $\mathrm{O}(tM^2)$ possible values. So the collision probability is $\Omega(1/tM^2)$. □

Lemma 4.4. *For any rectangle* $R \subseteq \mathrm{IIP}^{-1}_{m,t}(1)$ *we have* $|R| \le (4M)^t$.

Proof. Let $A, B \subset [-M,M]^t$ such that $\langle x,y\rangle = 0$ for all $x \in A, y \in B$. Let p be a prime between $2M+1$ and $4M$, and consider the problem modulo p. Note that we can injectively identify A, B with subsets of $\mathbb{F}_p^t$. Let V, W denote the linear subspaces of $\mathbb{F}_p^t$ spanned by A, B, respectively. Then $V \perp W$ and hence $|V||W| \le p^t$. This implies that $|A||B| \le p^t \le (4M)^t$.

□

Lemma 4.5. *Any* P^{EQ} *protocol for* $\mathrm{IIP}_{m,t}$ *with even* $t \ge 6$ *has cost* $\Omega(n/\log n)$.

Proof. Apply Lemma 3.3 with $\alpha = \Omega(1/tM^2)$ as given by Lemma 4.3, and $\beta = (4M)^t/(2M+1)^{2t} \le 1/M^t$ as given by Lemma 4.4. We obtain

$$\mathsf{P}^{\mathrm{EQ}}(\mathrm{IIP}_{m,t}) = \Omega\left(\frac{\log(\alpha/\sqrt{\beta})}{\log n}\right) = \Omega\left(\frac{\log(M^{t/2-2}/t)}{\log n}\right) = \Omega(tm/\log n) = \Omega(n/\log n) \ .$$

□

A related example. We give a similar separation by the inner product function over polynomials. Let $\mathbb{F}_2[z]$ denote the ring of univariate polynomials over $\mathbb{F}_2$.

Definition 4.6. The polynomial inner product problem $\mathrm{PIP}_{m,t}(x,y)$ is defined as follows. The inputs x, y are t-tuples of polynomials in $\mathbb{F}_2[z]$, each of degree at most m. The output is 1 if $\langle x,y\rangle = 0$, where the inner product is computed over $\mathbb{F}_2[z]$.

Note that also here the input size is $n = (m+1)t$. Again we consider large m and $t = O(1)$.

Lemma 4.7. *There is a* coRP *protocol for* $\mathrm{PIP}_{m,t}$ *of cost* $\mathrm{O}(t\log m)$.

Proof. Consider the following protocol. Alice and Bob interpret their polynomials as polynomials in $\mathbb{F}_q[z]$ with $q = 2^{\lceil \log m\rceil + 2}$. They sample a uniformly random point $z \in \mathbb{F}_q$ and compute $\langle x, y\rangle(z)$ by having Alice send the result of evaluating each of her polynomials at z. The protocol uses $\mathrm{O}(t\log q) = \mathrm{O}(t\log m)$ bits of communication.

The protocol is always correct on 1-inputs. To see that it is correct on 0-inputs with probability at least $1/2$ we observe that the probability of failure is the probability of picking a root of $\langle x, y\rangle$. Since the number of roots is at most $2m$ and we have $q \geq 4m$ points in $\mathbb{F}_q$ to choose from, the probability of failure is at most $1/2$. □

Lemma 4.8. *Any* P^{EQ} *protocol for* $\mathrm{IIP}_{m,t}$ *with even* $t \geq 6$ *has cost* $\Omega(n/\log n)$.

The proof is analogous to that of Lemma 4.5. We can use Lemma 4.3 unchanged, and we adapt Lemma 4.4 by considering the inner product function over $\mathbb{F}_q$ with $q = 2^m$.

5 Concluding Remarks

This work belongs to the general area of understanding the power of randomness in communication complexity. We use this opportunity to remind the readers of a fascinating open problem, posed explicitly by Göös, Pitassi and Watson [GPW18], which is whether $\mathsf{BPP} \subset \mathsf{P}^{\mathsf{NP}}$ for *total functions*. It is known that this containment is not true for partial functions. Göös et al. suggested, as a first step, separating the class of total functions in BPP from an interesting subclass of P^{NP}. In this work, we took this step by providing the first (exponential) separation between BPP and P^{EQ}, the latter being one of the most natural subclasses of P^{NP} . However, the original problem of separating BPP from P^{NP} remains open.

To state this in combinatorial terms, a function f has a P^{NP} protocol of cost c if the following holds. There exists a list of 2^c rectangles R_i and values $z_i \in \{0, 1\}$, such that $f(x, y) = z_i$ for the *first* rectangle R_i in the list for which $(x, y) \in R_i$ (We may assume that the last rectangle contains all possible inputs, to make this model well defined). In particular, if $\mathsf{BPP} \subset \mathsf{P}^{\mathsf{NP}}$ then there must exist a monochromatic rectangle in f of density $2^{-O(c)}$. Understanding this question seems to be pivotal towards understanding the relation between BPP and P^{NP}.

Problem 1. *Let f be an n-bit total boolean function with a randomized protocol of cost c. Is it true that f must contain a monochromatic rectangle R of size $|R| \geq 2^{-O(c)}2^{2n}$?*

Acknowledgements

We are grateful to Bruno Loff, Jaikumar Radhakrishnan and especially Suhail Sherif for helpful discussions in the early stages of this work. We also thank the Simons Institute for the Theory of Computing at Berkeley where some of this work took place.

References

[ABB+17] Andris Ambainis, Kaspars Balodis, Aleksandrs Belovs, Troy Lee, Miklos Santha, and Juris Smotrovs. Separations in query complexity based on pointer functions. *Journal of the ACM*, 64(5):32:1–32:24, 2017. Preliminary version in *STOC '16*.

[BEGJ00] María Luisa Bonet, Juan Luis Esteban, Nicola Galesi, and Jan Johannsen. On the relative complexity of resolution refinements and cutting planes proof systems. *SIAM Journal on Computing*, 30(5):1462–1484, 2000. Preliminary version in *FOCS '98*.

[BFS86] László Babai, Péter Frankl, and János Simon. Complexity classes in communication complexity theory. In *Proceedings of the 27th Annual Symposium on Foundations of Computer Science (FOCS '86)*, pages 337–347, October 1986.

[dRNV16] Susanna F. de Rezende, Jakob Nordström, and Marc Vinyals. How limited interaction hinders real communication (and what it means for proof and circuit complexity). In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science (FOCS '16)*, pages 295–304, October 2016.

[GPW18] Mika Göös, Toniann Pitassi, and Thomas Watson. The landscape of communication complexity classes. *Computational Complexity*, 27(2):245–304, June 2018. Preliminary version in *ICALP '16*.

[HW07] Johan Håstad and Avi Wigderson. The randomized communication complexity of set disjointness. *Theory of Computing*, 3(1):211–219, 2007.

[KN97] Eyal Kushilevitz and Noam Nisan. *Communication Complexity*. Cambridge University Press, New York, NY, USA, 1997.

[Kra98] Jan Krajíček. Interpolation by a game. *Mathematical Logic Quarterly*, 44(4):450–458, 1998.

[Nis91] Noam Nisan. CREW PRAMs and decision trees. *SIAM Journal on Computing*, 20(6):999–1007, December 1991.

[Nis93] Noam Nisan. The communication complexity of threshold gates. In *Proceedings of Combinatorics, Paul Erdős is Eighty*, volume 1, pages 301–315, 1993.

[PSS14] Periklis A. Papakonstantinou, Dominik Scheder, and Hao Song. Overlays and limited memory communication. In *Proceedings of the 29th IEEE Conference on Computational Complexity (CCC '14)*, pages 298–308, June 2014.

[Yao79] Andrew Chi-Chih Yao. Some complexity questions related to distributive computing (preliminary report). In *Proceedings of the 11th Annual ACM Symposium on Theory of Computing (STOC '79)*, pages 209–213, April 1979.

[Yao03] Andrew Chi-Chih Yao. On the power of quantum fingerprinting. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing (STOC '03)*, pages 77–81, June 2003.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il