On the power of lower bound methods for one-way quantum communication complexity

Shengyu Zhang

The Chinese University of Hong Kong
syzhang@cse.cuhk.edu.hk

Abstract. This paper studies the three known lower bound methods for one-way quantum communication complexity, namely the Partition Tree method by Nayak, the Trace Distance method by Aaronson, and the two-way quantum communication complexity. It is shown that all these three methods can be exponentially weak for some total Boolean functions. In particular, for a large class of functions generated from Erdös-Rényi random graphs $G(N,p)$ with p in some range of $1/poly(N)$, the two-way quantum communication complexity gives linear lower bound while the other two methods only gives constant lower bounds. This denies the possibility of using any of these known quantum lower bounds for proving the fundamental conjecture that the classical and quantum one-way communication complexities are polynomially related. En route of the exploration, we also discovered that the power of Nayak's method is exactly equal to the *extended equivalence query complexity* in learning theory.

1 Introduction

Communication complexity studies the minimum amount of communication needed to compute a function of inputs distributed over two (or more) parties. Through more than three decades of studies since its invention by Yao [33], it has flourished into a research field with connections to numerous other computational settings such as circuit complexity, streaming algorithms, data structures, decision tree complexity, VLSI design, and so on. See [20] for a comprehensive introduction of the field, and [32] for exhibitions of connections to more areas such as algorithmic game theory, optimization, and pseudo-randomness.

Different models were proposed and studied in the basic two-party setting, where the two parties, usually called Alice and Bob, are given inputs $x \in \{0,1\}^n$ and $y \in \{0,1\}^m$, respectively, and they need to compute $f(x,y)$. In the *two-way* model, Alice and Bob are allowed to send messages back and forth; in the one-way model, only Alice sends a message to Bob. The protocol (with the parties) can be deterministic, randomized and quantum, with the latter two allowing a bounded error. The least amount of communication needed for the worst-case input is the communication complexity in the corresponding models. We denote by $\mathsf{D}(f)$, $\mathsf{R}(f)$ and $\mathsf{Q}(f)$ the deterministic, randomized and quantum communication complexities of function f in the two way model, and $\mathsf{D}^1(f)$, $\mathsf{R}^1(f)$ and $\mathsf{Q}^1(f)$ the corresponding complexities in the one-way model.

Though much weaker than the two-way model, the one-way model has also caused much attention for various reasons: First, the model is powerful enough to admit many efficient protocols, including both cases for specific functions (such as Equality) and cases for general functions (such as the one with cost in terms of the γ_2^∞-norm [22]). Second, the one-way communication complexity has close connections to some other areas such as space complexity of streaming algorithms [24]. Third, proving lower bounds of the one-way communication complexity for general functions turns out to be mathematically quite challenging.

Lower bound methods for communication complexity are of particular interest since in most, if not all, connections to other theoretical areas, communication complexity serves as a lower bound of the other complexity measures under concern. In the quantum scenarios, lower bounds on quantum communication complexity are interesting for another important reason. One of the most fundamental questions in the field is to pin down the largest gap between classical and quantum communication complexities for total Boolean functions. Despite its importance, however, the problem is notoriously hard and our knowledge is very limited: the largest known gap between $Q(f)$ and $R(f)$ for a total function is quadratic (achieved by, for example, Disjointness [17, 28, 7, 3, 14, 29, 30]), while the best upper bound of $R(f)$ in terms of $Q(f)$ is still exponential. The situation in the one-way model is more embarrassing: Despite a lot of efforts [1, 2, 16, 31], we have still not able to find any super-constant separation between $\mathsf{Q}^1(f)$ and $R^1(f)$, while the best upper bound of gap is exponential. Actually, it was highly nontrivial to find even relations [5] or partial functions [11, 18] with exponential gaps between $\mathsf{Q}^1(f)$ and $R^1(f)$. Based on this and various other facts such as Holevo's bound, it is reasonable to conjecture that $R^1(f)$ and $\mathsf{Q}^1(f)$ are polynomially related for all total Boolean functions f.

Two approaches were previously taken to understand the relation. One is trying to simulate a quantum protocol directly by a randomized one. Unfortunately, the cost of all classical simulations so far have parameters other than $\mathsf{Q}^1(f)$, and those parameters can be easily as large as n for some total functions; see the end of this section for more details of related work. The other natural approach is to firstly derive a general lower bound $\mathsf{Q}^1(f) = \Omega(L(f))$, and then prove a matching upper bound $R^1(f) = poly(L(f))$. Recently Jain, Klauck and Nayak proved that the one-way rectangle bound tightly characterizes $R^1(f)$ [15], which raised the hope of proving the polynomial relation by establishing a

matching quantum lower bound. Jain and Zhang tried along this way [16], but only succeeded partially[1]. This second approach was also taken by some other researchers before [31], but there were always subtle gaps to the goal.
Note that for this approach to succeed for a general total function f, the tightness of the quantum lower bound $L(f)$ is crucial: If it is not always polynomially tight, then any attempt on establishing a matching classical upper bound is doomed to fail. There are two methods particularly for the quantum one-way model. One is the *trace distance method* for general functions by Aaronson [1]. The other lower bound method is given by Nayak [25] for the Random Access Code (RAC) problem, based on a simple but elegant information theoretical argument; we will refer to this technique as the *partition tree* method (for the reason that will be clear from later discussions). Besides these two methods, the two-way quantum communication complexity $\mathsf{Q}(f)$ can also serve as a lower bound for $\mathsf{Q}^1(f)$ by definition. In this paper, we show that

Theorem 1. *None of the above three known lower bound methods for $\mathsf{Q}^1(f)$ is polynomially tight. Actually, they can all be exponentially weak.*

This theorem implies that any one of the known methods does not suffice to prove the conjecture that $\mathsf{R}^1(f) = poly(\mathsf{Q}^1(f))$. It can also be viewed as an partial explanation on why the conjecture, if true, is so hard to prove. These negative results on the tightness call for new lower bound methods for $\mathsf{Q}^1(f)$, and we hope that the exhibited weakness of the methods can guide us searching for new and more powerful ones, which is helping on proving this conjecture as well as other potential applications.

Next we discuss in more details about our studies of the various lower bound methods. First, unlike the trace distance method, the partition tree method does not have a well-defined formula for general functions. This paper starts from cleaning up the picture of the partition tree method, leading to a robust generalization to arbitrary total Boolean function. The new formula also enjoys a nicer form which makes analysis of its ultimate power much easier. As an unexpected connection, it turns out that the best lower bound achievable by this method is exactly equal to the *extended equivalence query complexity* in computational learning theory.
We then analyze the power of the three lower bound techniques. Various relations between these techniques are studied, among which the advantage of $\mathsf{Q}(f)$ over the other two methods is particularly interesting. Presumably $\mathsf{Q}(f)$ should not be a good lower bound for $\mathsf{Q}^1(f)$ which in general can be much larger; for example, for Index function $\mathsf{Q}(f) \leq \log_2 n$ but $\mathsf{Q}^1(f) = 1$. However, it turns out that for most functions induced by a random graph $G(N, p)$ for a large range of $p = 1/poly(N)$, both the partition tree method and the trace distance methods can only give a constant lower bound for $\mathsf{Q}^1(f)$, while we can show that $\mathsf{Q}(f) = \Omega(n)$ by the generalized discrepancy method [22].

More related work On the relation of $\mathsf{R}^1(f)$ and $\mathsf{Q}^1(f)$, if parameters other than $\mathsf{Q}^1(f)$ are allowed, then nontrivial classical upper bounds are known: Nayak's $\Omega(n)$ lower bounds on RAC, Klauck's observation that Nayak's result actually shows $\mathsf{Q}^1(f) \geq VC(f)$ (the VC-dimension of f) [19], and Sauer's lemma that $D^1(f) = O(mVC(f))$ together imply the upper bound $D^1(f) = O(m\mathsf{Q}^1(f))$ for all total functions f. Aaronson later generalized this to partial functions $D^1(f) = O(m\mathsf{Q}^1(f)\log \mathsf{Q}^1(f))$ [1] and $\mathsf{R}^1(f) = O(m\mathsf{Q}^1(f))$ [2]. Jain and Zhang [16] improved the last bound to $\mathsf{R}^1(f) = O((I_\mu(X;Y)+1)VC(f))$ for total functions where $I_\mu(X;Y)$ is the mutual information of the correlated inputs (X, Y) under a hard distribution μ.

[1] Technically, we proved that the distributional quantum one-way communication complexity is lowered bounded by the distributional rectangle bound for all product distributions.

There are quite a few results on separations of classical and quantum communication complexities for total functions in the so-called SMP model [6] and for partial functions or relations in various other models [7, 27, 12, 10].

2 Preliminaries

Suppose Alice's input set is $\mathcal{X}$ with $N = 2^n$ and Bob's input set is $\mathcal{Y}$ with size $M = 2^m$. We usually use x to denote Alice's input and y to denote Bob's input. The set of inputs $\{(x, y) : f(x, y) = b\}$ is called b-inputs.
For a graph $G = (V, E)$, the function $f_G : V \times V \to \{0, 1\}$ is defined as $f_G(x, y) = 1$ iff $(x, y) \in E$. We assume that $f(x, x) = 0$. For a vertex $v \in V$, its neighbor set is denoted by $N(v)$. An N-node random graph in the Erdös-Rényi model $G(N, p)$ is obtained by connecting each pair of vertices independently with probability p. For a graph G, its adjacency matrix is A_G. For a matrix A, let $\sigma_1(A)$, ..., $\sigma_r(A)$ be the singular values of A in the decreasing order, where $r = \mathtt{rank}(A)$.

For general background of quantum computing, we refer to the textbook [26]. The following Holevo's bound is a fundamental result about the accessible information of a quantum state.

Theorem 2 (Holevo, [13]). *Suppose* Alice *prepares a state ρ_X where $X = 1, ..., n$ with probability $p_1, ..., p_n$, and* Bob *performs a POVM measurement $\{E_1, ..., E_n\}$ on the state with outcome Y. Then*

$$I(X; Y) \leq S\left(\sum_x p_x \rho_x\right) - \sum_x p_x S(\rho_x) \tag{1}$$

The following Fano's inequality relating the "error" of two random variables to their mutual information can be found in textbooks such as [9].

Lemma 1 (Fano's inequality). *Let X and Y be two random variable taking values in S. Let $\epsilon \stackrel{\text{def}}{=} \mathbf{Pr}[X \neq Y]$, then*

$$H(\epsilon) + \epsilon \log(|S| - 1) \geq H(X|Y). \tag{2}$$

The trace distance method was introduced by Aaronson [1] as a general lower bound for $\mathsf{Q}^1(f)$.

Theorem 3 (Aaronson, [1]). *Let $f : \{0, 1\}^n \times \{0, 1\}^m \to \{0, 1\}$ be a total Boolean function, and μ is a probability distribution on the 1-input set $\{(x, y) : f(x, y) = 1\}$. Let D_k be the distribution over $(\{0, 1\}^n)^k$ formed by first choosing $y \in \mu$ and then choosing k samples independently from the conditional distribution μ_y. Suppose that $\mathbf{Pr}_{x \leftarrow \mu,\ y \leftarrow \mu}[f(x, y) = 0] = \Omega(1)$, then*

$$\mathsf{Q}^1(f) = \Omega(\log \frac{1}{\|D_2 - D_1^2\|}). \tag{3}$$

Here "$x \leftarrow \mu,\ y \leftarrow \mu$" is to draw x and y independently from the two marginal distributions of μ.

Definition 1. *The* trace distance bound *for $\mathsf{Q}^1(f)$ is $\mathsf{TD}(f) = \max_\mu \log_2 \frac{1}{\|D_2 - D_1^2\|}$ where the maximum is taken over all probability distributions μ on the 1-inputs.*

Linial and Shraibman introduced the following lower bound for $\mathsf{Q}(f)$ based on the factorization norm. For a matrix A, define $\gamma_2(A) = \min_{A=BC} \|B\|_{2\to\infty} \|C\|_{1\to 2}$ where for vector norms $\|\cdot\|_X$ and $\|\cdot\|_Y$, the operator norm $\|A\|_{X\to Y} \stackrel{\text{def}}{=} \max_{\|x\|_X = 1} \|Ax\|_Y$. For a sign matrix A and $\alpha \geq 1$, let $\gamma_2^\alpha(A) = \min_{B: 1 \leq a_{ij} b_{ij} \leq \alpha} \gamma_2(B)$.

Theorem 4 (Linial and Shraibman, [22]). $Q_\epsilon(f) \geq \log_2 \gamma_2^{1/(1-2\epsilon)}(f) - O_\epsilon(1)$.

The bound is also known as the *generalized discrepancy method*. The bound actually holds even for $Q^*(f)$, the quantum communication complexity with entanglement shared by Alice and Bob, is lower bounded by the above quantity. Here we are mainly concerned with the case without entanglement because the no-separation conjecture becomes trivial (due to quantum teleportation) if we compare $Q^{1,*}(f)$ and $R^{1,*}(f)$.

Definition 2. *The* ϵ-factorization norm bound *for* $Q_\epsilon(f)$ *is* $\mathsf{FN}_\epsilon(f) = \log_2 \gamma_2^{1/(1-2\epsilon)}(f)$, *and the* factorization norm bound *for* $Q^*(f)$ *is* $\mathsf{FN}(f) = \mathsf{FN}_{1/3}(f)$.

3 The partition tree method

The partition tree bound is defined as follows. Consider a binary *partition tree* $\mathcal{T}$ *of* $\mathcal{X}$, where each node $v = v_1...v_i$ (i is the depth of v) is associated with an input y_v of Bob. Let X be a random variable according to the distribution p over $\mathcal{X}$. This tree induced a subset $\mathcal{X}_v \subseteq \mathcal{X}$ for each node v in the following inductive way: the root corresponds to $\mathcal{X}$, and suppose the set $\mathcal{X}_v$ is defined then the two subsets $\mathcal{X}_{v0}$ and $\mathcal{X}_{v1}$ for its two children $v0$ and $v1$ is defined by $\mathcal{X}_{vb} = \{x \in \mathcal{X}_v : f(x, y_v) = b\}$. Define a sequence of random variables $V_1, ..., V_{depth(\mathcal{T})}$ by $\mathbf{Pr}[V_{i+1} = b|V_1...V_i] = p(\mathcal{X}_{V_1...V_ib})/p(\mathcal{X}_{V_1...V_i})$. For a node $v = v_1...v_i$, define $p(v) \stackrel{\text{def}}{=} p(\mathcal{X}_v)$ and $p_v(b) \stackrel{\text{def}}{=} p(\mathcal{X}_{vb}|\mathcal{X}_v)$ for $b \in \{0, 1\}$ and $p_v(\min) \stackrel{\text{def}}{=} \min\{p_v(0), p_v(1)\}$.

Definition 3. *The* partition tree bound *for* $\mathsf{Q}^1(f)$ *is* $\mathsf{PT}(f) = \max_{\mathcal{T},p} \sum_{v\in\mathcal{T}} p(v)p_v(\min)$.

It is not quite immediate to generalize Nayak's argument (for RAC) to this formula as a lower bound of $\mathsf{Q}^1(f)$. Please see Appendix for detailed explanations.
To study $\mathsf{PT}(f)$, first observe that if one can find a balanced binary subtree of height h, then $\mathsf{PT}(f) \geq h(1 - H(\epsilon))$ since one can put half-half probabilities on both branches of each node of the subtree. (Note that this is at least $VC(f)$ but could be much larger than it, as shown in the Greater Than function in Appendix B.) The following theorem says that this is actually also the best lower bound the partition tree method can provide.

Theorem 5. *There exists a subset* $S \subseteq \mathcal{X}$ *and a partition tree* $\mathcal{T}^*$ *for* f *on* $(S, \mathcal{Y})$ *s.t.*

$$\mathsf{PT}(f) = \text{ the length of the shortest path of } \mathcal{T}^*. \tag{4}$$

Proof. Fix a maximizer $(\mathcal{T}, p)$ in the definition of $\mathsf{PT}(f)$. Observe that the distribution p conditioned on any subtree is also the optimal for the subtree, thus the overall best p can be computed inductively. Suppose the best p has been assigned to both left and right subtrees, and the resulting lower bounds are $l(T_0)$ and $l(T_1)$ respectively. Then the best assignment of p at the root is easy to compute as follows. Suppose it gives p_b to subtree b, then the overall lower bound is $\max_{(p_0,p_1):\ p_0+p_1=1} \min\{p_0, p_1\} + p_0 l(T_0) + p_1 l(T_1)$. By case analysis (whether $p_0 \leq p_1$, and then whether $l(T_1) - l(T_0) \geq \pm 1$), it is easy to get that the overall lower bound T is

$$l(T) = \begin{cases} l(T_0) & l(T_1) - l(T_0) \leq -1 \quad (\text{maximizer}: \ p_1 = 0) \\ \frac{1}{2} + \frac{1}{2}(l(T_0) + l(T_1)) & -1 \leq l(T_1) - l(T_0) \leq 1 \quad (\text{maximizer}: \ p_0 = p_1 = 1/2) \\ l(T_1) & l(T_1) - l(T_0) \geq 1 \quad (\text{maximizer}: \ p_0 = 0) \end{cases} \tag{5}$$

This implies that with loss of generality we can apply the partition tree method only on a sub-function, obtained by deleting some rows. And we can always use half-half probabilities

for the two branches. In this way, the second case in the above formula for $l(T)$ is always taken in the tree (for internal nodes). Now we start at the root and do the following argument along a shortest path P from the root to a leaf. Suppose the length of the path is d; namely there are d edges on the path. (Without loss of generality, assume that the leaf is in T_0.) We have

$$l(T) = (l(T_0) + l(T_1) + 1)/2 \leq l(T_0) + 1 \tag{6}$$

because $|l(T_0) - l(T_1)| \leq 1$. Continue the argument until we reach the leaf, we get $l(T) \leq l(v) + d = d$, as desired.

Note that the standard decision tree complexity is to minimize the the length of the longest path, but here the best PT bound is to maximize the length of the shortest path.
It turns out to have an interesting connection to the *extended equivalence query complexity* in learning theory, which we will define using the language of communication complexity as follows. Alice has an input x and Bob wants to exactly learn x by making queries to Alice, who then responses with an answer. Different query models were studied in learning theory.

1. *membership query*: Bob's query is a column index y, and Alice's response is $f(x, y)$;
2. *equivalence query*: Bob's query is a string $a \in \{0,1\}^M$ as a guess of x. If $a = x$, then Alice tells Bob so and the game is over. Otherwise, Alice not only tells Bob that his guess is wrong, but also provides a column y which $f(x, y) \neq a_y$.
   If Bob is restricted to use strings $a \in \{0,1\}^M$ appearing as rows in the matrix f as queries, then this is called the *equivalence query*; if Bob is allowed to use any string $a \in \{0,1\}^M$ as an query, it is called the *extended equivalence query*.

The minimal number of a particular type of queries Bob needs to make for the worst-case input x is called the query complexity of that type. Denote by $MQ(f)$, $EQ(f)$ and $XEQ(f)$ the membership query complexity, the equivalence query complexity and the extended equivalence query complexity of the function f, respectively. The following theorem gives a characterization of $XEQ(f)$ by relating it to membership query computation.

Theorem 6 (Littlestone, [23]).

$$XEQ(f) = \max_{\mathcal{T}} \min_{x} d(x, \mathcal{T}), \tag{7}$$

where $\mathcal{T}$ is a membership query computation tree and $d(x, \mathcal{T})$ is the depth of x in $\mathcal{T}$.

A membership query computation tree is a decision tree with membership queries in the natural way; see the survey [4] for a formal definition (as well as an extensive review of different types of queries in learning theory). Its important relation to our work is that the membership query computation tree is exactly our partition tree, and thus the above theorem and Theorem 5 combined give the following full characterization of PT.

Theorem 7. $\mathsf{PT}(f) = XEQ(f)$.

4 Comparisons between the powers

In this section we will study the power of the lower bound methods, the PT bound part of which uses the limitation result established in the previous section. We will prove Theorem 1 by a circle of comparison results in the order of PT >> Q >> TD >> PT. The first separation is easily exhibited by Index function. Next we will show that though as a lower bound method merely for the two-way complexity, the factorization norm method can be

much stronger than the other two methods for the one-way complexity. In fact, for almost all functions f in some range (the precise meaning of which will be clear shortly) the factorization norm gives a linear lower bound for $\mathsf{Q}(f)$ while the other two cannot even prove a super constant lower bound for $\mathsf{Q}^1(f)$. The advantage of FN over TD is given next, and that of FN over PT is given in Section 4.3.

4.1 On the advantage of the factorization norm method over the trace distance method

In this section we will show that for a random Erdös-Rényi graph $G(N,p)$ for some range of p, its adjacency matrix as a function f has $\mathsf{FN}(f) = \Omega(n)$ but $\mathsf{TD}(f) = O(1)$.
Here we consider random graph $G(N,p)$ since the corresponding limitations for TD and PT are easier to show. Let A be the adjacency matrix of $G = (V,E)$ and denote by d_i the degree of vertex i. Let $P = [p_{ij}]$ with $p_{ij} = 1/d_i$, $D = diag(d_1, ..., d_N)$, and the normalized Laplacian is $\mathcal{L} = I - D^{-1/2}AD^{-1/2}$. Consider $\bar{\mathcal{L}} \stackrel{\text{def}}{=} I - \mathcal{L}$: Since it is symmetric, it has a spectral decomposition $\bar{\mathcal{L}} = \sum_{i=1}^{N} \lambda_i |\eta_i\rangle\langle\eta_i|$. A standard fact is that $\lambda_1 = 1$ and $|\eta_1\rangle = (\sqrt{d_1/(2|E|)}, ..., \sqrt{d_N/(2|E|)})$. It is known, for example from a general result in [8], that

Lemma 2 (Chung, Lu, Vu, [8]). *For $p = \omega(\frac{\log^4 N}{N})$, it holds with probability $1 - o(1)$ that*

$$\max_{i=2,\ldots,n} |\lambda_i| \le O(1/\sqrt{pN}) \tag{8}$$

This implies that the factorization norm method gives a good lower bound for most such random graphs.

Theorem 8. *If $\omega(\log^4 N/N) \le p \le 1 - \Omega(1)$, then an N-node random graph $G(N,p)$ has*

$$\mathsf{FN}(f_G) - O(1) \ge \frac{1}{2}\log_2(pN) - O(1) \tag{9}$$

with probability $1 - o(1)$.

Proof. Let $S = 2A - J$ be the sign matrix of A, and let $\bar{\mathcal{L}}_{-1} = \bar{\mathcal{L}} - \lambda_1|\eta_1\rangle\langle\eta_1|$. It is known that $\gamma_2^*(A) \le N\sigma_1(A)$ for every $N \times N$ real matrix A (see [21]), thus

$$\gamma_2^*(\bar{\mathcal{L}}_{-1}) \le N\sigma_1(\bar{\mathcal{L}}_{-1}) = N\sigma_2(\bar{\mathcal{L}}) \le O(N/\sqrt{pN}) \tag{10}$$

with probability $1 - o(1)$. On the other hand, we have

$$\gamma_2^\infty(S) \ge \frac{\langle S, \bar{\mathcal{L}}_{-1}\rangle}{\gamma_2^*(\bar{\mathcal{L}}_{-1})} \ge \Omega\Big(\frac{\sqrt{pN}}{N}\Big)\Big\langle 2A - J, D^{-1/2}AD^{-1/2} - |\eta_1\rangle\langle\eta_1|\Big\rangle \tag{11}$$

It is not hard to verify that the (i,j)-entry of $D^{-1/2}AD^{-1/2}$ is $1/\sqrt{d_i d_j}$ if $(i,j) \in E$ and 0 otherwise. Thus the above inner product is equal to

$$\sum_{(i,j)\in E}\left(\frac{1}{\sqrt{d_i d_j}} - \frac{\sqrt{d_i d_j}}{|E|}\right) + \frac{(\sum_i \sqrt{d_i})^2}{2|E|} \tag{12}$$

By the assumption, $p < 1 - c$ for some small constant c. By Chernoff bound, one can see that $\mathbf{Pr}[\exists d_i \notin ((1-\delta)pN, (1+\delta)pN)] \le Ne^{-\delta^2 pN/2} = o(1)$ for the small constant $\delta = c/4$ and $p = \omega(\log N/N)$. Putting this concentration bound into the above quantity, we have

$$\gamma_2^\infty(S) \ge \frac{\sqrt{pN}}{N}\Big(\frac{N(1-\delta)pN}{(1+\delta)pN} - 2(1+\delta)pN + \frac{N^2(1-\delta)pN}{N(1+\delta)pN}\Big) \tag{13}$$

$$= \frac{\sqrt{pN}}{N}2N\Big(\frac{1-\delta}{1+\delta} - p(1+\delta)\Big) = \Omega(\sqrt{pN}) \tag{14}$$

for $p < 1 - c$ and $\delta = c/4$. Thus

$$\mathsf{FN}(S) = \log_2 \gamma_2^{1/(1-2\epsilon)}(S) \geq \log_2 \gamma_2^{\infty}(S) \geq \frac{1}{2}\log(pN) - O(1). \tag{15}$$

Next we show that the trace distance method can only give a constant lower bound for random graph functions.

Theorem 9. *For $p = o(N^{-6/7})$, a random graph $G(N,p)$ has $\mathsf{TD}(f_G) = O(1)$ with probability $1 - o(1)$.*

Here we are not aiming to maximize the range of p, though we believe that the result still holds for larger p. The main goal is to show the existence of a range $p = 1/poly(N)$.
We will first show in the following Lemma 4 that with probability $1-o(1)$, a random graph $G(N,p)$ has some good properties. The proof uses Lemma 3 which is on the relation of the number of edges and that of vertices with some connection requirement. After these, we will show that for graphs with those properties, the TD bound is very low.

Lemma 3. *For any constant $\delta > 0$, there are constants c and d s.t. for all distinct vertices $V_x = \{x_1, ..., x_c\}$ and $V_z = \{z_1, ..., z_d\}$, if*

1. *any x_i and z_k share at least one common neighbor, and*
2. *there is no vertex y connecting to all x_i's and z_k's.*

then there exists $V_y = \{y_1, ..., y_e\}$, s.t. any pair (x_i, z_k) of vertices are connected via y_j for some $j \in [e]$, and

$$\frac{g}{c+d+e} \geq \frac{4}{3} - \delta \tag{16}$$

where g is the number of edges between $V_x \cup V_z$ and V_y.

Proof. For each (x_i, z_k), there is at least one y connecting them. Collect all these y's to form the set V_y. (For pairs (x_i, z_k) with more than one connectors y, we pick an arbitrary one.) Thus each y has degree at least 2, and therefore

$$g/(c+d+e) \geq 2e/(c+d+e). \tag{17}$$

Now we will give another lower bound

$$g/(c+d+e) \geq 1 + (d-2)/(e+6). \tag{18}$$

Combining the two inequalities gives the desired result.

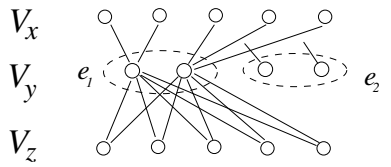

Fig. 1. Illustration for the proof of Lemma 3

To show the second bound, fix a setting with $g/(e+6)$ minimized. See Figure 1 (where every $N(z_k) \cap V_y$ is the same set simply for convenience of illustration). The way we chose V_y guarantees that $N(V_y)$ contains the whole V_x. Pick a subset $S \subseteq V_y$ with the minimum

size *s.t.* the $N(S) \supseteq V_x$. By definition, the number of edges from S to V_x is at least $|V_x| = c$. Define $e_1 = |S|$ and $e_2 = e - |S|$; Condition 2 implies $e_1 \geq 2$. Note that for each z_k, its neighbor set in V_y, *i.e.* $N(z_k) \cap V_y$, also connects to all V_x, thus the number of edges from z_k to V_y is $|N(x_i) \cap V_y| \geq e_1$. Also note that each node in $V_y - S$ has at least one edge to V_x. Thus the total number of edges in this small graph $V_x \cup V_y \cup V_z$ is at least

$$de_1 + c + e_2 = (d-1)e_1 + c + e \geq 2(d-1) + c + e = 1 + (d-2)/(e+6), \qquad (19)$$

as desired.

Lemma 4. *For $p = o(N^{-6/7})$, a random graph $G = G(N, p)$ has all the following properties with probability $1 - o(1)$.*

1. *For any vertex x with (at least) three neighbors y_1, y_2, y_3, at least one of the two pairs (y_1, y_2) and (y_2, y_3) only has x as their common neighbor.*
2. *There are universal constants c and d s.t. for any c vertices $x_1, ..., x_c$ that do not share a common neighbor, there are at most $d-1$ vertices $z_1, ..., z_{d-1}$ which have distance exactly 2 to all x_i's.*
3. *The graph does not contain a $K_{3,2}$, the $(3, 2)$-complete bipartite graph, as a subgraph.*

Proof. We will show that each condition is satisfied with probability $1 - o(1)$ for a random graph $G(N, p)$ with the setting of p as in the stated range, thus the probability that all are satisfied is also $1 - o(1)$.

1. Consider the event that the statement is false, *i.e.* both pairs have a common neighbor other than x. If these two other common neighbors are actually the same one, there are 5 vertices and 6 edges needed, so this event happens with probability at most $N^5 p^6$; if these two common neighbors are not the same, then there are 6 vertices and 7 edges and thus the probability of this happening is at most $N^6 p^7$. For the probability p chosen in the Lemma, both probabilities are $o(1)$.
2. By Lemma 3, we know that the intersection of the cover sets of c vertices is of size d with probability at most $N^{c+d+e} p^g = N^{-\Omega(1)} = o(1)$.
3. A random graph contains a $K_{3,2}$ with probability at most $N^5 p^6 = N^{-\Omega(1)} = o(1)$.

Lemma 5. *Suppose there is a distribution μ on 1-inputs with $\mathbf{Pr}_{x \leftarrow \mu, y \leftarrow \mu}[f(x, y) = 0] = \Omega(1)$ satisfied. If there is a submatrix A s.t. $\mu(A) = 1 - o(1)$, and A as a function has $Q^{1,pub}(A) = q$, then $\|D_2 - D_1^2\|_1 = 2^{\Omega(-q)}$. In particular, $\|D_2 - D_1^2\|_1 = \Omega(1)$ for the following two special cases*

1. *there is a subset $S \subseteq \mathcal{X}$ s.t. $|S| = O(1)$ and $\mu(S) = 1 - o(1)$,*
2. *there is a submatrix A s.t. each column is monochromatic except for at most $O(1)$ entries, and $\mu(A) = 1 - o(1)$.*

Proof. The only possibly nontrivial statement is the last sentence, which we need to show that $Q^{1,pub}(f) = O(1)$. But this is easy because, for example, by first flipping 0's and 1's in some columns we can assume that each column has at most a constant number of 1's. Then the function becomes the OR of a constant number of functions each of which is essentially an Equality function with some relabeling of rows and columns.

Now we are ready to prove the theorem.

Proof. (of Theorem 9) We take the graphs with all good properties in Lemma 4. It is enough to show that any distribution μ on the edge set E with the following condition satisfied

$$\mathbf{Pr}_{x \leftarrow \mu,\ y \leftarrow \mu}[f(x, y) = 0] = \Omega(1), \qquad (20)$$

has that $\|D_2 - D_1^2\|_1 = \Omega(1)$. Assuming that it is not true, *i.e.* $\|D_2 - D_1^2\|_1 = o(1)$, we will first show that this assumption forces μ to put most mass on just one star-shape cluster of vertices, then show that in this case, it is also unavoidable to have $\|D_2 - D_1^2\|_1 = \Omega(1)$ finally.

For two vertices x and x', we say x *covers* x', denoted by $x \sim x'$, if they share a common neighbor y. Otherwise we write $x \nsim x'$. For a vertex itself, we assume $x \sim x$ as long as x has a non-zero degree. Define the set $Cover(x) = \{x' : x \sim x'\}$. By definition,

$$\|D_2 - D_1^2\|_1 = \sum_{x,x'} \left| \sum_y \mu(y)\mu(x|y)\mu(x'|y) - \mu(x)\mu(x') \right| \tag{21}$$

$$= \sum_{x,x':\ x \sim x'} \left| \sum_y \mu(y)\mu(x|y)\mu(x'|y) - \mu(x)\mu(x') \right| + \sum_{x,x':\ x \nsim x'} \mu(x)\mu(x') \tag{22}$$

Thus

$$\sum_x \mu(x) \mathbf{Pr}_{x' \leftarrow \mu}[x \nsim x'] = \sum_{x,x':\ x \nsim x'} \mu(x)\mu(x') \leq \|D_2 - D_1^2\|_1 = o(1) \tag{23}$$

This means that an average x covers most other vertices x' (weighted under μ). In particular, there exists one x_0 *s.t.* $\mathbf{Pr}_{x' \leftarrow \mu}[x' \nsim x_0] = o(1)$. Suppose its neighbor set is $N(x_0) = \{y_1, ..., y_t\}$, and define *clusters* $S_i = N(y_i) - \{x_0\}$, and put $S = \cup_i S_i$. Also note that $Cover(x_0)$ is nothing but $S \cup \{x_0\}$, thus $\mu(S \cup \{x_0\}) = 1 - o(1)$. Note that by Lemma 5, we can assume that $\mu(x_0) = 1 - \Omega(1)$ (otherwise the desired conclusion has already been proved). Denote by $\mu_{\neg x_0}$ the distribution $\mu(x)$ conditioned on $x \neq x_0$.

Claim 1 *At least one of the following two statements is true:*

1. *There are two disjoint subsets G_1 and G_2 of S s.t. $\mu_{\neg x_0}(G_b) = \Omega(1)$ for both $b = 1, 2$, and any two vertices $x_1 \in G_1$ and $x_2 \in G_2$ belong to two different clusters.*
2. *There is a single cluster S_i with $\mu_{\neg x_0}(S_i) = 1 - o(1)$.*

Proof. It is not hard to verify that the first property in Lemma 4 actually guarantees that any cluster S_i intersects at most one other cluster. We say two clusters form an *overlapping pair* (of clusters) if they overlap. A *block* is either an overlapping pair or an isolated cluster (*i.e.* one that does not overlap with any other cluster). Find a block B with the largest mass under $\mu_{\neg x_0}$. Three cases are discussed in order.

If $\mu_{\neg x_0}(B) = o(1)$, then all blocks have $o(1)$ mass (under $\mu_{\neg x_0}$). Since all blocks collectively contain $\mu_{\neg x_0}(S) \geq \mu(S) = \mu(S \cup \{x_0\}) - \mu(x_0) = 1 - o(1) - (1 - \Omega(1)) = \Omega(1)$, we can easily partition S into two disjoint groups $G_1 \uplus G_2$ where $G_b = \cup_{i \in T_b} S_i$ ($b = 1, 2$, $T_1 \uplus T_2 = [t]$) *s.t.* $\mu_{\neg x_0}(G_1) = \mu_{\neg x_0}(S)/2 - o(1) = \Omega(1)$ and $\mu_{\neg x_0}(G_2) = \mu_{\neg x_0}(S)/2 + o(1) = \Omega(1)$.

If $\mu_{\neg x_0}(B) \in [\Omega(1), 1 - \Omega(1)]$, then let $G_1 = B$ and G_2 contain the rest blocks.

Finally, for the case $\mu_{\neg x_0}(B) = 1 - o(1)$, if B is a single cluster, then the second statement of the claim is satisfied. If B is an overlapping pair, say (S_1, S_2), then either it holds that one of them has $\mu_{\neg x_0}(S_b) = 1 - o(1)$, in which case the second statement holds, or $\mu_{\neg x_0}(S_1 - S_2) = \Omega(1)$ and so is $\mu_{\neg x_0}(S_2 - S_1)$, in which case putting $G_1 = S_1 - S_2$ and $G_2 = S_2 - S_1$ makes the first statement to hold.

We continue the proof of Theorem 9. If the second statement of the above claim is true, it means that there is a single $y_i \in N(x_0)$ *s.t.* $\mu(N(y_i)) = 1 - o(1)$. (Note that $N(y_i)$ includes a cluster and x_0 itself.) By the third property of Lemma 4, each vertex y other than y_i only connects to at most two vertices in $N(y_i)$ (to avoid a $(3, 2)$-complete bipartite graph). Thus the submatrix on $N(y_i) \times \mathcal{Y}$ has $1 - o(1)$ μ-mass but has all 1's in column y_i and at most two 1's in all other columns. By Lemma 5, we see that $\|D_2 - D_1^2\|_1 = \Omega(1)$.

Therefore, we can assume that the first statement of the claim is the case, so

$$\mathbf{E}_{x\leftarrow\mu_{\neg x_0}}\left[\mathbf{Pr}_{x'\leftarrow\mu}[x'\nsim x]\right] \geq \sum_{b=1,2}\mu_{\neg x_0}(G_b)\mathbf{E}_{x\leftarrow\mu_{\neg x_0}}\left[\mathbf{Pr}_{x'\leftarrow\mu}[x'\nsim x] \mid x\in G_b\right] \quad (24)$$

On the other hand, we have

$$\mathbf{E}_{x\leftarrow\mu_{\neg x_0}}\left[\mathbf{Pr}_{x'\leftarrow\mu}[x'\nsim x]\right] = \frac{\sum_{x\neq x_0}\mu(x)\mathbf{Pr}_{x'\leftarrow\mu}[x'\nsim x]}{1-\mu(x_0)} = \frac{o(1)}{\Omega(1)} = o(1), \quad (25)$$

Since both $\mu_{\neg x_0}(G_b) = \Omega(1)$, we have $\mathbf{E}_{x\leftarrow\mu_{\neg x_0}}\left[\mathbf{Pr}_{x'\leftarrow\mu}[x'\nsim x] \mid x\in G_b\right] = o(1)$ for both $b=1,2$. Therefore, we can find two points $x_b\in G_b$ both with $\mathbf{Pr}_{x'\leftarrow\mu}[x'\nsim x_b] = o(1)$. This means that for both $b=1,2$, most of mass of μ is put on $Cover(x_b)$. Combined with the same fact for x_0, we see that actually $\mu(\cap_{i=0,1,2}Cover(x_i)) = 1-o(1)$. But note that both x_1 and x_2 are covered by x_0 since they are chosen from S, and they are not in the same cluster as guaranteed by the first statement of the above claim. Consequently, x_0, x_1, x_2 do not share a common neighbor.

Now define set $T=\{x_0,x_1,x_2\}$. As long as $|T|$ is constant, we can assume by Lemma 5 that $\mu(T) = 1-\Omega(1)$. Then similar to Eq (25), it follows that $\mathbf{E}_{x\leftarrow\mu}[\mathbf{Pr}_{x'\leftarrow\mu}[x'\nsim x]|x\notin T] = o(1)$. Thus there exists another point x in $S-T$ *s.t.* $\mathbf{Pr}_{x'\leftarrow\mu}[x'\nsim x] = o(1)$. Add this point to T and continue this process until $|T|=c$. Each point $x\in T$ has the property that $\mu(Cover(x)) = 1-o(1)$, and consequently $\mu(\cap_{x\in T}Cover(x)) = 1-o(1)$ by noting that $|T|=c$ is a constant. Also recall that the vertices in T do not share a common neighbor since actually even x_0,x_1,x_2 do not. By the second property of Lemma 4, the intersection of their cover sets has only constant size, and thus using Lemma 5 we get $\|D_2 - D_1^2\|_1 = \Omega(1)$. This completes the proof.

4.2 On the advantage of the Trace Distance method over the partition tree method

We observed that the partition tree method can be much better than the factorization norm method, and have shown that the factorization norm method can be much better than the trace distance method. To finish the circle, we now show that the trace distance method can be much better than the partition tree method. Different than Theorem 10, this time we can give an explicit function to show the separation.

The Coset function $Coset(G)$ is defined as follows. For a fixed group G, Alice is given a coset x as her input and Bob is given an element $y\in G$ as his input; the question is whether $y\in x$. Aaronson [1] studied the function for the group $\mathbb{Z}_p^2$ (where p is a prime number) and proved that $\mathsf{Q}^1(Coset(\mathbb{Z}_p^2)) = \Theta(\log p)$; that is, Alice asymptotically needs to send the whole input to Bob. Here we show that the partition tree method can only give a very small constant lower bound for this function.

Proposition 1. $\mathsf{PT}(Coset(\mathbb{Z}_p^2)) = 2$.

Proof. We denote Alice's input by (a,b) and Bob's input by (x,y). The function is to see whether $y = ax+b$. The basic reason is that the tree for any sub-function can never have a path with 3 1-branches, thus the shortest path is of length at most 2. Actually, for any Bob's input (x,y), we want to argue that there is at most one Alice's input (a,b) after two 1-branches. Indeed, suppose the two 1-branches happens at two pivoting columns (x_1,y_1) and (x_2,y_2). Then any (a,b) has to satisfy that

$$ax_1+b = y_1, \qquad ax_2+b = y_2 \quad (26)$$

If $x_1 = x_2$, then by the above equalities it holds that $y_1 = y_2$, which means that the two pivoting columns are exactly the same. But this is impossible in a partition tree since repeatedly picking the same column does not give any new partition. Thus $x_1 \neq x_2$, implying that the above system of equations has one unique solution for (a, b).

Comment We can also consider the general dimension case $Coset(Z_p^d)$. Aaronson's proof can be easily adapted to show a lower bound of $\mathsf{Q}^1(f) = \Omega(d \log p)$, and it is not hard to see that $\mathsf{PT}(f) = \Omega(d)$. Therefore the above separation between $\mathsf{PT}(f)$ and $\mathsf{Q}^1(f)$ deteriorates when the dimension goes higher.

4.3 Other discussions of the power comparisons

The main goal of this paper is to study the ultimate power of the known lower bound methods for $\mathsf{Q}^1(f)$, and in particular their tightness because of the no-separation conjecture reason mentioned in Section 1. Though it is not our goal to thoroughly study all the six relations between the three methods, it is good to know for more insights. This section so far showed three of them as a circle, leaving the three other relations to discuss. First, it turns out that PT is also weak for random graph functions.

Theorem 10. *For any $\alpha = \Omega(1)$, if $p = N^{-\alpha}$, then an N-node random graph $G(N, p)$ has $\mathsf{PT}(f_G) = O(1)$ with probability $1 - o(1)$.*

Proof. Suppose there exists a partition tree of height h. Then there are $2^{h+1} - 1$ internal nodes which are distinct and labeled by column indexes y, and 2^{h+1} leaves in the tree which are distinct and labeled by row indexes x. We can use a string $s(y)$ to encode y, with the length of $s(y)$ equal to the depth of y; similarly for $(h+1)$-bit strings $s(x)$. By the definition of partition tree, we know that there are exactly 2^h edges from the nodes y in each layer of the tree to the leaves x. For any fixed labeled x's and y's, this happens in $G(N, p)$ with probability $p^{h2^h}(1-p)^{h2^h}$. Thus

$$\mathbf{Pr}[\exists 2^{h+1} \text{ labeled } x\text{'s and } (2^{h+1}-1) \text{ labeled } y\text{'s to form a partition tree}] \tag{27}$$

$$\leq \binom{N}{2^{h+1}}^2 (p(1-p))^{h2^h} \leq N^{2^{h+2}} p^{h2^h} = (N^4 p^h)^{2^h} \tag{28}$$

If $p \leq N^{-\alpha}$ for some constant α, then letting $h = 4/\alpha$ makes the above bound no more than $N^{-2^h} = o(1)$. In other words, with probability $1 - o(1)$, there is no partition tree of depth $h = 4/\alpha = O(1)$.

For PT over TD, we believe that actually $\mathsf{TD}(\mathsf{Index}) = O(\log n)$, though we can only show it for the *symmetric* distribution μ, *i.e.* $\mu(x, y) = \mu(x', y')$ if $|x| = |x'|$.

Theorem 11. *For any distribution p on $\{0, 1, ..., n\}$, let $\mu_p(x, y) = p(|x|)$ for all (x, y) with $x_y = 1$. Then the trace distance bound under μ_p for the Index function is only $O(\log n)$.*

Proof. We assume that n is a prime and will prove that for any p over $\{0, ..., n\}$, the induced μ has $\|D_1^2 - D_2\| = 1/poly(n)$ and thus the TD bound under μ is $\Omega(\log n)$. If n is not a prime, we find a prime $n' \in [n, 2n]$ and apply the result on n' by noting that any distribution p over $\{0, ..., n\}$ is also a distribution over $\{0, ..., n'\}$ (with $p_{n+1} = ... = p_{n'} = 0$).
Let us first examine the condition of $\mathbf{Pr}_{x \leftarrow \mu, y \leftarrow \mu}[f(x, y) = 0] = \Omega(1)$.

$$\sum_{(x,y):x_y=0} \mu(x)\mu(y) = \sum_k p_k \frac{n-k}{n} = 1 - \frac{\sum_k k p_k}{n} \tag{29}$$

where the first inequality is because of the symmetry over columns. So by the condition, we have $\sum_k kp_k = cn$ for some constant $c < 1$.
In what follows we may use x to also indicate the set $\{i : x_i = 1\}$.

$$\|D_1^2 - D_2\| = \sum_{x,x'} \left| \sum_{i \in x \cap x'} \mu(i) \mathbf{Pr}[x|i] \mathbf{Pr}[x'|i] - \mu(x)\mu(x') \right| \tag{30}$$

$$= \sum_{x,x'} \left| \sum_{i \in x \cap x'} \frac{1}{n} \frac{q_{|x|}}{1/n} \frac{q_{|x'|}}{1/n} - \sum_{i \in x} q_{|x|} \sum_{i \in x'} q_{|x'|} \right| \tag{31}$$

$$= \sum_{x,x'} q_{|x|} q_{|x'|} \big| n|x \cap x'| - |x||x'| \big| \tag{32}$$

$$= \sum_{kl} \frac{p_k p_l}{kl} \mathbf{E}_{|x|=k,|x'|=l} \big[\big| n|x \cap x'| - kl \big| \big] \tag{33}$$

where in the expectation x is uniformly at random over all strings with weight k and similarly for x'. Now the expectation is at least $\mathbf{Pr}_{|x|=k,|x'|=l} \big[n|x \cap x'| \neq kl \big]$. Since we assume that n is a prime, the event $n|x \cap x'| = kl$ happens only if $(k, l) \in \{(1, n), (n, 1), (n, n)\}$. Thus

$$\|D_1^2 - D_2\| \geq \sum_{kl} \frac{p_k p_l}{kl} (1 - \mathbf{Pr}_{|x|=k,|x'|=l}[n|x \cap x'| = kl]) \tag{34}$$

$$\geq \frac{1}{n^2}(1 - (p_1 p_n + p_n p_1 + p_n^2)) = \frac{1}{n^2}(1 - (p_1 + p_n)^2 + p_1^2) \tag{35}$$

If $p_1 + p_n = 1 - o(1)$, then the condition $\sum_k kp_k \leq cn$ implies $p_n \leq c$, and thus

$$p_1 \geq 1 - o(1) - c \qquad \text{and} \qquad 1 - (p_1 + p_n)^2 + p_1^2 \geq p_1^2 = \Omega(1) \tag{36}$$

If $p_1 + p_n$ is less than a constant $d < 1$, then

$$1 - (p_1 + p_n)^2 + p_1^2 \geq 1 - d^2 = \Omega(1) \tag{37}$$

as well. Thus in any case, we have $\|D_1^2 - D_2\| = \Omega(1/n^2)$ and thus the TD bound at this distribution is $O(\log n)$.

For general distributions, we can first use symmetrization to get a symmetric distribution. Then $\mathsf{TD}(\mathsf{Index}) = O(\log n)$ as long as the function $\|D_2 - D_1^2\|$ is convex over all distributions μ; unfortunately we do not know whether the convexity is true.
Finally, for TD over FN, we do not know much about the advantage of yet.

5 Concluding remarks and open questions

The tightness results in this paper call for new lower bound methods for $\mathsf{Q}^1(f)$. With the light shed by comparisons in Section 4, one (vague) approach is trying to somehow combine the advantages of the methods to get a more powerful one.
The factorization norm method appears pretty strong for lower bounding the two-way quantum communication complexity. Can we modify it to obtain a "useful" lower bound for $\mathsf{Q}^1(f)$, in the sense of either showing strong lower bounds for specific questions, or establishing connections to other measures for general functions?

Acknowledgment We would like to thank Rahul Jain for many valuable discussions during the collaboration of paper [16], and Yi-Kai Liu for pointing out the reference [4].

References

1. Scott Aaronson. Limitations of quantum advice and one-way communication. *Theory of Computing*, 1:1–28, 2005.
2. Scott Aaronson. The learnability of quantum states. *Proceedings of the Royal Society A*, 463:2088, 2007.
3. Scott Aaronson and Andris Ambainis. Quantum search of spatial regions. *Theory of Computing*, 1:47–79, 2005.
4. Dana Angluin. Queries revisited. *Theoretical Computer Science*, 313(2):175–194, 2004.
5. Ziv Bar-Yossef, T. S. Jayram, and Iordanis Kerenidis. Exponential separation of quantum and classical one-way communication complexity. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC)*, pages 128–137, 2004.
6. Harry Buhrman, Richard Cleve, John Watrous, and Ronald de Wolf. Quantum fingerprinting. *Physical Review Letters*, 87(16), 2001.
7. Harry Buhrman, Richard Cleve, and Avi Wigderson. Quantum vs. classical communication and computation. In *Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing (STOC)*, pages 63–68, 1998.
8. Fan Chung, Linyuan Lu, and Van Vu. The spectra of random graphs with given expected degrees. *Internet Mathematics*, 1(3):257–275, 2004.
9. Thomas Cover and Joy Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, NY, USA, 2 edition, 2006.
10. Dmitry Gavinsky. Classical interaction cannot replace a quantum message. In *Proceedings of the Fortieth Annual ACM Symposium on the Theory of Computing (STOC)*, pages 95–102, 2008.
11. Dmitry Gavinsky, Julia Kempe, Iordanis Kerenidis, Ran Raz, and Ronald de Wolf. Exponential separation of quantum and classical one-way communication complexity. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing (STOC)*, pages 516–525, 2007.
12. Dmitry Gavinsky and Pavel Pudlák. Exponential separation of quantum and classical non-interactive multi-party communication complexity. In *Proceedings of the 23rd Annual IEEE Conference on Computational Complexity*, pages 332–339, 2008.
13. A.S. Holevo. Some estimates of the information transmitted by quantum communication channels. *Problemy Peredachi Informatsii*, 9:311, 1973. English translation in Problems of Information Transmission 9, pp. 177-183, 1973.
14. Peter Høyer, Michele Mosca, and Ronald de Wolf. Quantum search on bounded-error inputs. In *Proceedings of the 30th International Colloquium on Automata, Languages and Programming (ICALP)*, pages 291–299, 2003.
15. Rahul Jain, Hartmut Klauck, and Ashwin Nayak. Direct product theorems for classical communication complexity via subdistribution bounds. In *Proceedings of the Fortieth Annual ACM Symposium on the Theory of Computing (STOC)*, pages 599–608, 2008.
16. Rahul Jain and Shengyu Zhang. New bounds on classical and quantum one-way communication complexity. *Theoretical Computer Science*, 410(26):2463–2477, 2009.
17. Bala Kalyanasundaram and Georg Schintger. The probabilistic communication complexity of set intersection. *SIAM Journal on Discrete Mathematics*, 5(4):545–557, 1992.
18. Bo'az Klartag and Oded Regev. Quantum one-way communication can be exponentially stronger than classical communication. In *Proceedings of the 44th Annual ACM Symposium on the Theory of Computing (STOC)*, 2011. To appear.
19. Hartmut Klauck. Lower bounds for quantum communication complexity. *SIAM Journal on Computing*, 37(1):20–46, 2007.
20. Eyal Kushilevitz and Noam Nisan. *Communication Complexity*. Cambridge University Press, Cambridge, UK, 1997.

21. Nati Linial, Shahar Mendelson, Gideon Schechtman, and Adi Shraibman. Complexity measures of sign matrices. *Combinatorica*, 27:439–463, 2007.
22. Nati Linial and Adi Shraibman. Lower bounds in communication complexity based on factorization norms. In *Proceedings of the Thirty-ninth Annual ACM symposium on Theory of Computing (STOC)*, pages 699–708, 2007.
23. Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988.
24. S. M. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1(2), 2005.
25. Ashwin Nayak. Optimal lower bounds for quantum automata and random access codes. In *Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 124–133, 1999.
26. Michael Nielsen and Isaac Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge, UK, 2000.
27. Ran Raz. Exponential separation of quantum and classical communication complexity. In *Proceedings of the Thirty-First Annual ACM Symposium on the Theory of Computing (STOC)*, pages 358–367, 1999.
28. Alexander Razborov. On the distributional complexity of disjointness. *Theoretical Computer Science*, 106:385–390, 1992.
29. Alexander Razborov. Quantum communication complexity of symmetric predicates. *Izvestiya: Mathematics*, 67(1):145–159, 2003.
30. Alexander Sherstov. The pattern matrix method for lower bounds on quantum communication. In *Proceedings of the 40th Annual ACM Symposium on the Theory of Computing (STOC)*, pages 85–94, 2008.
31. Yaoyun Shi. Personal communication.
32. Avi Wigderson. Depth through breadth, or why should we attend talks in other areas? In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC)*, page 579, 2004. Slides available at http://www.math.ias.edu/~avi/TALKS/STOC04.ppt.
33. Andrew Chi-Chih Yao. Some complexity questions related to distributive computing. In *Proceedings of the Eleventh Annual ACM Symposium on Theory of Computing (STOC)*, pages 209–213, 1979.

A Clarifying the picture of the partition tree method

At a high level, the method works in the following way. One first defines a random variable X according to a distribution p over Alice's input set $\mathcal{X}$, and then corresponding to different inputs y, Bob can perform different measurements to distinguish between two possibilities of Alice's random input. By Holevo's bound and the correctness of the protocol, each measurement provides some mutual information for X: When $\epsilon < p \stackrel{\text{def}}{=} \min\{p_0, p_1\}$, an ϵ-error protocol can distinguish two mixed states $p_0\rho_0$ and $p_1\rho_1$, where ρ_b corresponds to some inputs x's with $f(x, y) = b$. By Fano's inequality, this distinguisher gives a $H(p)-H(\epsilon)$ mutual information as the lower bound we gain at this step. This argument works perfectly well for the RAC problem, where it always has $p_0 = p_1 = 1/2$. When we try to apply the method to general functions, an immediate question is: What if $\epsilon \geq p$? We can first use amplification to drop the error probability to some ϵ^* with a loss of factor of $\Theta(\log 1/\epsilon^*)$, and give up those nodes with $p < \epsilon$. But it is not clear at all what ϵ^* should be used and what the resulting bound is for a general function. In this paper, we observe that the necessity of $\epsilon < p$ is actually a conceptual pitfall — we can obtain mutual information of an amount of $\Theta(p)$ for arbitrary $\epsilon < 1/2$ and p. Using this observation, we give a unified lower bound formula which is better than the old one as above. The nice form of the formula also makes analysis of its limitation easy.

A.1 Review of the method

The partition tree method has its origin from Nayak's lower bound for RAC [25], or equivalently, the Index function defined as follows. Suppose $x \in \{0,1\}^n$, $y \in \{0,1\}^{\lceil \log_2 n \rceil}$, then $\mathsf{Index}(x,y) = x_y$, that is, the y-th bit of x. For simplicity, assume n is a power of 2. Nayak proved that $\mathsf{Q}^1_\epsilon(\mathsf{Index}) \geq (1 - H(\epsilon))n$, as briefly reviewed below. Suppose on input x, Alice sends the mixed state ρ_x. Consider $\rho = 2^{-n}\sum_x \rho_x$, and let $\rho_{b_1 \ldots b_l} = 2^{-n+l}\sum_{x:x_1=b_1,\ \ldots,\ x_l=b_l} \rho_x$, then

$$S(\rho) = S\Big(\frac{1}{2}\rho_0 + \frac{1}{2}\rho_1\Big) \geq I(b_1, T) + \frac{1}{2}S(\rho_0) + \frac{1}{2}S(\rho_1) \geq (1 - H(\epsilon)) + \frac{1}{2}S(\rho_0) + \frac{1}{2}S(\rho_1) \quad (38)$$

Here T is the output of the protocol when Bob's input is $y = 1$ and Alice's input is chosen uniformly at random. Use this output as an ϵ-error distinguisher for $(\frac{1}{2}\rho_0, \frac{1}{2}\rho_1)$ and apply Holevo's bound to get the first inequality. The second inequality is due to Fano's inequality. Keep applying this argument for each $\rho_{b_1 \ldots b_i}$ by using $y = i$ at level i all the way to the end of $i = n$ and we will get the lower bound $n(1 - H(\epsilon))$.

It is not hard to observe that in the above argument, though for each fixed i, Bob uses the same input $y = i$ for different states $\rho_{b_1 \ldots b_i}$, this is not necessary in general. Indeed we can define a binary *partition tree* $\mathcal{T}$ of $\mathcal{X}$, where each node $v = v_1 \ldots v_i$ (i is the depth of v) is associated with an input y_v of Bob. Let X be a random variable according to the distribution p over $\mathcal{X}$. This tree induced a subset $\mathcal{X}_v \subseteq \mathcal{X}$ for each node v in the following way: the root corresponds to $\mathcal{X}$, and suppose the set $\mathcal{X}_v$ is defined then the two subsets $\mathcal{X}_{v0}$ and $\mathcal{X}_{v1}$ for its two children $v0$ and $v1$ is defined by $\mathcal{X}_{vb} = \{x \in \mathcal{X}_v : f(x, y_v) = b\}$. Define a sequence of random variables $V_1, \ldots, V_{depth(\mathcal{T})}$ by $\mathbf{Pr}[V_{i+1} = b | V_1 \ldots V_i] = p(\mathcal{X}_{V_1 \ldots V_i b})/p(\mathcal{X}_{V_1 \ldots V_i})$. Then Nayak's argument gives a lower bound of

$$\sum_i I(V_{i+1}; \tilde{f}(x, y_{V_1 \ldots V_i}) | V_1 \ldots V_i). \quad (39)$$

where $\tilde{f}(x,y)$ is the output of the protocol on input (x,y). We recommend readers to read Appendix B to see a lower bound of $\Omega(n)$ for the Greater Than (GT) function as a very illustrative example to show the tree structure (as oppose to the line structure in the Index function case).

For a node $v = v_1 \ldots v_i$, define $p(v) \stackrel{\text{def}}{=} p(\mathcal{X}_v)$ and $p_v(b) \stackrel{\text{def}}{=} p(\mathcal{X}_{vb} | \mathcal{X}_v)$ for $b \in \{0,1\}$ and $p_v(\min) \stackrel{\text{def}}{=} \min\{p_v(0), p_v(1)\}$. Then if $p_v(\min) \geq \epsilon$, we have $I(V_{i+1}; \tilde{f}(x, y_{v_1 \ldots v_i}) | V_1 = v_1, \ldots, V_i = v_i) \geq H(p_v(\min)) - H(\epsilon)$ by Fano's inequality. Now the question is: what if $p_v(\min) < \epsilon$? Two immediate approaches for this issue are: First, give up those vertex v with $p_v(\min) < \epsilon$ and hope the gained mutual information on the rest nodes are large enough to give a good lower bound; second, use error reduction to drop the error probability to some ϵ^* *s.t.* $\epsilon^* < p_v(\min)$ by repeating the protocol $\Theta(\log 1/\epsilon^*)$ times, which causes a $\Theta(\log 1/\epsilon^*)$ factor of loss. Combining these two approaches, one gets a lower bound of

$$\mathsf{Q}^1(f) = \Omega\left(\max_{\mathcal{T},\ p,\ \epsilon^*} \frac{\sum_{v \in \mathcal{T}} p(v)[H(p_v(\min)) - H(\epsilon^*)]^+}{\log 1/\epsilon^*}\right) \quad (40)$$

where the function a^+ means a if $a \geq 0$ and 0 otherwise.

A.2 An improved bound

As we have seen, a key issue is $p_v(\min)$ versus ϵ. In general, suppose we have $(p_0\rho_0, p_1\rho_1)$ where ρ_b's are two mixed states; let $p = \min\{p_0, p_1\}$. If our distinguisher has error probability greater than p, can it give any useful information about b? The old argument by Fano's

inequality does not seem to apply now and it is attempting to give up those nodes. However, this turns out to be a conceptual pitfall, as shown by the following key observation that even if the error ϵ is a constant and the current partition has small min probability p, we can still gain mutual information of $\Omega(p)$.

Proposition 2. *Let $\rho = \sum_{T=0,1} p_T \rho_T$ and denote $p = \min\{p_0, p_1\}$. Suppose we have an measurement on ρ to output T' with property $\mathbf{Pr}[T \neq T'] \leq \epsilon$. Then regardless of the relation of p and ϵ, it holds that*

$$S(\rho) - (p_1 S(\rho_1) + p_0 S(\rho_0)) \geq 2(1 - H(\epsilon))p. \tag{41}$$

Proof. Without loss of generality, assume that $p = p_1 \leq 1/2$. The key is to allocate a small part of ρ_0 and to let the distinguisher only deal with ρ_1 and this small part of ρ_0.

$$\begin{aligned}
S(\rho) &= S(p_1\rho_1 + p_0\rho_0) && (42)\\
&= S(p_1\rho_1 + p_1\rho_0 + (1-2p_1)\rho_0) && (43)\\
&\geq 2p_1 S\left(\frac{1}{2}\rho_1 + \frac{1}{2}\rho_0\right) + (1-2p_1)S(\rho_0)) && (44)\\
&\qquad \textit{// We give up distinguishing } p_1(\rho_0+\rho_1) \textit{ and } (1-2p_1)\rho_0\\
&\geq 2p_1\left(\frac{1}{2}S(\rho_1) + \frac{1}{2}S(\rho_0) + 1 - H(\epsilon)\right) + (1-2p_1)S(\rho_0) && (45)\\
&= p_1 S(\rho_1) + p_0 S(\rho_0) + 2p_1(1 - H(\epsilon)) && (46)
\end{aligned}$$

When $p_1 = 1/2$, the bound coincides with the usual $1 - H(\epsilon)$ one as in Nayak's original proof for the Index function. Using the above observation, we get the following lower bound.

Theorem 12.

$$Q^1_\epsilon(f) \geq 2(1 - H(\epsilon)) \max_{\mathcal{T},\ p} \sum_{v \in \mathcal{T}} p(v) p_v(\min) \tag{47}$$

and in particular,

$$\mathsf{Q}^1(f) \geq \Omega\left(\max_{\mathcal{T},\ p} \sum_{v \in \mathcal{T}} p(v) p_v(\min)\right) \tag{48}$$

Note that this is always better than the previous one in Eq. (40). Actually, no matter what ϵ^* is picked in Eq. (40), for those v whose $p_v(\min) \leq \epsilon^*$, the new bound has a gain of $\Omega(p_v(\min))$ but the old bound has nothing. For those v whose $p_v(\min) > \epsilon^*$, we have

$$\frac{H(p_v(\min)) - H(\epsilon^*)}{\log 1/\epsilon^*} < \frac{H(p_v(\min))}{\log 1/\epsilon^*} = O(\frac{p_v(\min)\log(1/p_v(\min))}{\log 1/\epsilon^*}) < O(p_v(\min)). \tag{49}$$

Thus the gain in the new bound is also better. Actually the old bound amounts to first fixing a threshold ϵ^* but the new bound can adaptively gain $\Omega(p_v(\min))$, which is much larger than $\frac{H(p_v(\min)) - H(\epsilon^*)}{\log 1/\epsilon^*}$ if $p_v(\min) >> \epsilon^*$.
Therefore this form can serve as a unified form of PT, and we will show a full characterization of its power.

B The partition tree method on Greater Than function

In this section we will illustrate the partition tree method using the Greater Than function for an example. Recall that $GT(x, y) = 1$ iff $x \geq y$. Let the distribution p over $\mathcal{X}$ be the uniform one. Bob first uses input 10...0 to distinguish between $(\frac{1}{2}\rho_0, \frac{1}{2}\rho_1)$ where $\rho_{b_1} =$

$\sum_{x:x_1=b_1} \rho_x$. Then for ρ_{b_1}, Bob uses input $b_1 10...0$ as his input to distinguish $(\frac{1}{2}\rho_{b_1 0}, \frac{1}{2}\rho_{b_1 1})$ where $\rho_{b_1 b_2} = \sum_{x:x_1=b_1,x_2=b_2} \rho_x$. Note that different than Nayak's proof for the Index function, now for different b_1, Bob's inputs $b_1 10...0$ are different. Continue this process, at step i, Bob uses input $b_1...b_{i-1}10...0$ as his input to distinguish $(\frac{1}{2}\rho_{b_1...b_{i-1}0}, \frac{1}{2}\rho_{b_1...b_{i-1}1})$ where $\rho_{b_1...b_i} = \sum_{x:x_1=b_1,...,x_i=b_i} \rho_x$. Note that the two states to be distinguished are always of half-half probabilities because p is uniform. Similarly as in the Index function case, for each level i, each node of depth i gives a $(1 - H(\epsilon))$ mutual information, thus finally after n steps the total amount of mutual information gained as our lower bound is $n(1 - H(\epsilon))$. As in the Index function case, the partition tree for GT is also a complete binary tree of depth n. The difference is that in GT function, for each level i, the nodes of depth i use different y's as Bob's input to extract the mutual information.